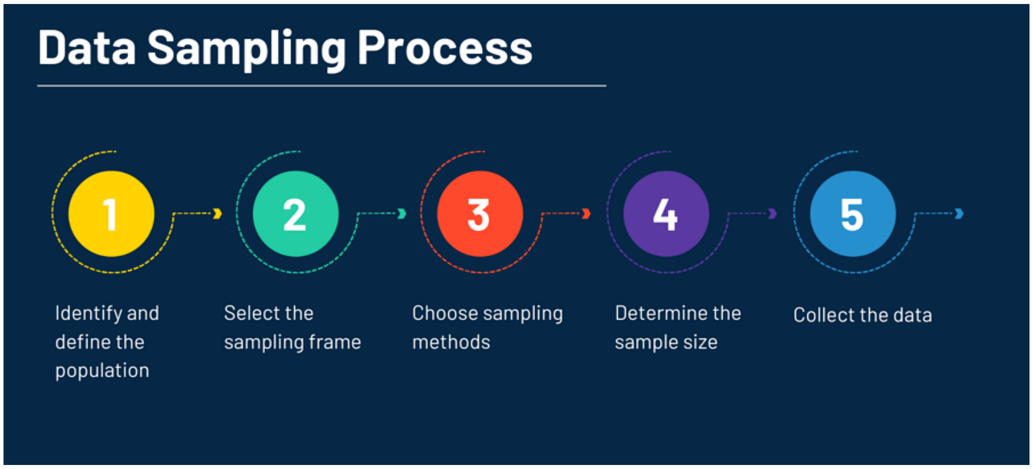
INSAID

STATISTICS INTERVIEW QUESTIONS

1. What is Sampling?

Data sampling is a common statistics technique that’s used to analyze patterns and trends in a subset of data that’s representative of a larger data set being examined. Sampling is used to determine how much data to collect and how often it should be collected.

The overall process of data sampling is a statistical analysis method that helps to draw conclusions about populations from samples.



Data Sampling Process

1. Identify and define the population
2. Select the sampling frame
3. Choose sampling methods
4. Determine the sample size
5. Collect the data

It’s useful when the data set that needs to be examined is too large to be analyzed as a whole. An example of this is big data analytics, which looks at raw, massive sets of data in an attempt to uncover trends.

In these cases, identifying and analyzing a representative sample of data is more efficient, as well as cost-effective, than trying to survey the entirety of data or population. In addition to being low-cost, analyzing a sample of data takes less time than trying to analyze the entire population of data.

2. Explain all the sampling methods.

These methods are broken down into two main categories: probability sampling and non-probability sampling.

Probability Sampling - In the category of probability sampling, every aspect of the population has an equal chance of being selected to be studied and analyzed. These methods typically provide the best chance of creating a sample that's as representative as possible.

i) Simple Random Sampling -
Each individual is chosen by chance, and each member of the population or group has an equal chance of being selected.

ii) Systematic sampling -
In this method, the first individual is selected randomly, while others are selected using a "fixed sampling interval". Therefore, a sample is created by setting an interval that derives data from the larger population.

iii) Stratified sampling -
Stratified sampling is a method where elements of the population are divided into small subgroups, called stratas, based on their similarities or a common factor. Samples are then randomly collected from each subgroup.

iv) Cluster sampling -
The method of clustering divides the entire population, or large data set, into clusters, or sections, based on a defining factor. Then the clusters are randomly selected to be put in the sample and then analyzed.

v) Multistage sampling -
Multistage sampling is a more complicated form of cluster sampling. Essentially, this method works by dividing the larger population into many clusters. The second-stage clusters are then broken down further based on a secondary factor. Then, those clusters are sampled and analyzed.

Non-probability sampling - The data sampling methods in the non-probability category have elements that don't have an equal chance of being selected to be included in the sample, meaning they don't rely on randomization. These techniques rely on the ability of the data scientist, data analyst, or whoever is doing the selecting, to choose the elements for a sample.

i) Convenience sampling -
In convenience sampling, sometimes called accidental or availability sampling, the data is collected from an easily accessible and available group. Essentially, individuals are selected based on their availability and willingness to be a part of the sample.

ii) Quota sampling -
When the quota method is used in data sampling, items are chosen based on predetermined characteristics. The researcher doing the data sampling ensures equal representation within the sample for all subgroups within the data set or population.

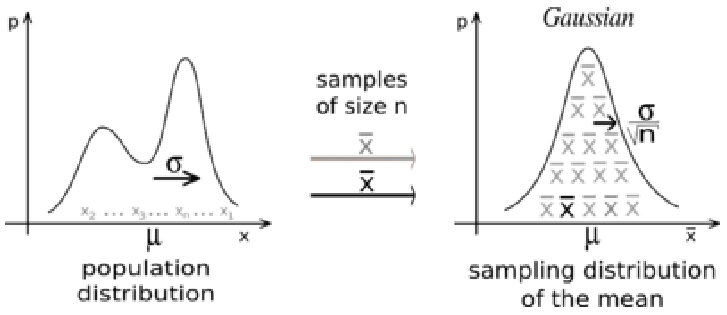
iii) Judgment sampling -
Judgment sampling, which is also known as selective sampling, is based on the assessment of experts in the field when choosing who to ask to be included in the sample.

iv) Snowball sampling -
The snowball sampling, sometimes also called referral sampling or chain referral sampling, is used when the population is completely rare and unknown. This is typically done by selecting one, or a small group, of individuals based on the specific criteria. Then, the person(s) selected are then used to find more individuals to be analyzed.

3. What is Central Limit Theorem?

it states that if we sample from a population using a sufficiently large sample size, the mean of the samples (also known as the sample population) will be normally distributed (assuming true random sampling). What’s especially important is that this will be true regardless of the distribution of the original population.

The central limit theorem is important because it is used in hypothesis testing and also to calculate confidence intervals.



Central Limit Theorem explained visually

4. What is the difference between Type 1 and Type 2 error?

If the result of the test corresponds with reality, then a correct decision has been made (e.g., the person is healthy and is tested as healthy, or the person is not healthy and is tested as not healthy). However, if the result of the test does not correspond with reality, then two types of error are distinguished: type I error and type II error.

A tabular relationship between truthfulness/falseness of the null hypothesis and outcomes of the test can be seen in the table below:

	Null Hypothesis is true	Null Hypothesis is false
Reject null hypothesis	**Type I Error** **False Positive**	Correct Outcome True Positive
Fail to reject null hypothesis	Correct Outcome True Negative	**Type II Error** **False Negative**

5. What do the terms p-value, coefficient, and r-squared value mean? What is the significance of each of these components?

The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In

other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable.
Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant. This statistical control that regression provides is important because it isolates the role of one variable from all of the others in the model.

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model

6. What is the Binomial Probability Formula?

The binomial distribution consists of the probabilities of each of the possible numbers of successes in N trials for independent events that each have a probability of π (the Greek letter pi) of occurring.

7. How do you assess the statistical significance of an insight?

You would perform hypothesis testing to determine statistical significance. First, you would state the null hypothesis and alternative hypothesis. Second, you would calculate the p-value, the probability of obtaining the observed results of a test assuming that the null hypothesis is true. Last, you would set the level of the significance (alpha) and if the p-value is less than the alpha, you would reject the null — in other words, the result is statistically significant.

8. What is the statistical power?

‘Statistical power’ refers to the power of a binary hypothesis, which is the probability that the test rejects the null hypothesis given that the alternative hypothesis is true.

$$Power = P(reject\ Null\ |alternative\ is\ true)$$

9. Explain selection bias (with regard to a dataset, not variable selection). Why is it important?

Selection bias is the phenomenon of selecting individuals, groups or data for analysis in such a way that proper randomization is not achieved, ultimately resulting in a sample that is not representative of the population.
Understanding and identifying selection bias is important because it can significantly skew results and provide false insights about a particular population group.

10. Is mean imputation of missing data acceptable practice? Why or why not?

Mean imputation is the practice of replacing null values in a data set with the mean of the data.
Mean imputation is generally bad practice because it doesn't take into account feature correlation. For example, imagine we have a table showing age and fitness score and imagine that an eighty-year-old has a missing fitness score. If we took the average fitness score from an age range of 15 to 80, then the eighty-year-old will appear to have a much higher fitness score that he actually should.
Second, mean imputation reduces the variance of the data and increases bias in our data. This leads to a less accurate model and a narrower confidence interval due to a smaller variance.

11. Define quality assurance, six sigma.

Quality assurance: an activity or set of activities focused on maintaining a desired level of quality by minimizing mistakes and defects.

Six sigma: a specific type of quality assurance methodology composed of a set of techniques and tools for process improvement. A six sigma process is one in which 99.99966% of all outcomes are free of defects.

12. Give examples of data that does not have a Gaussian distribution, nor log-normal.

Any type of categorical data won’t have a gaussian distribution or lognormal distribution. Exponential distributions — eg. the amount of time that a car battery lasts or the amount of time until an earthquake occurs.

13. Give examples of data that does not have a Gaussian distribution, nor log-normal.

Any type of categorical data won’t have a gaussian distribution or lognormal distribution. Exponential distributions — eg. the amount of time that a car battery lasts or the amount of time until an earthquake occurs.

14. 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 the mean > median > mode.

15. What is the difference between descriptive and inferential statistics?

Descriptive statistics: Descriptive statistics is used to summarize from 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.