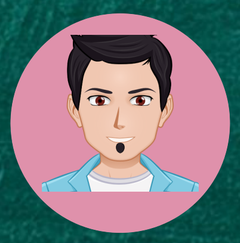
AI News

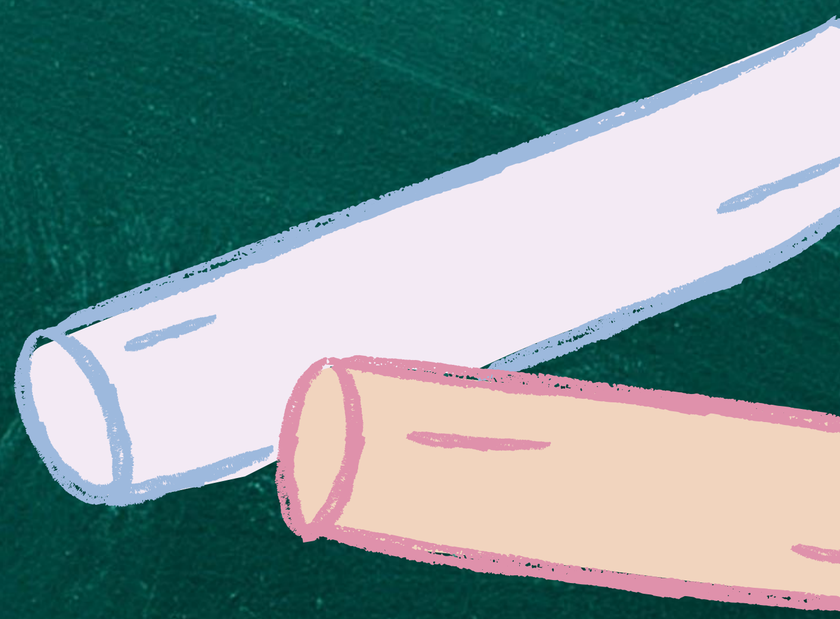
INTERVIEW

QUESTIONS YOU MUST PREPARE BEFORE A DATA SCIENCE INTERVIEW

(LET'S CRACK IT)



SHAPE
@shape.ai



QUESTION 1: WHAT IS THE CENTRAL LIMIT THEOREM AND WHY IS IT IMPORTANT?

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.

QUESTION 2: WHAT IS SAMPLING? HOW MANY SAMPLING METHODS DO YOU KNOW?

Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined.

There are many different methods for drawing samples from data; the ideal one depends on the data set and situation.

Sampling can be based on probability, an approach that uses random numbers that correspond to points in the data set to ensure that there is no correlation between points chosen for the sample.

Further variations in probability sampling include:
Simple random sampling, Stratified sampling, Cluster sampling, Multistage sampling, Systematic sampling

Non-probability data sampling method includes: Convenience sampling, Consecutive sampling, Purposive or judgmental sampling, Quota sampling

QUESTION 3: WHAT IS DIFFERENCE BETWEEN TYPE1 AND TYPE 2 ERROR?

A type 1 error (False Positive Error) occurs when the null hypothesis is true but is rejected. A type I error, or false positive, is asserting something as true when it is actually false. A type 2 error (False Negative) occurs when the null hypothesis is false but erroneously fails to be rejected. A type II error, or false negative, is where a test result indicates that a condition failed, while it actually was successful.

QUESTION 4: WHAT IS SELECTION BIAS?

Selection ('or' sampling) bias occurs when the sample data that is gathered and prepared for modeling has characteristics that are not representative of the true, future population of cases the model will see. That is, active selection bias occurs when a subset of the data is symmetrical (i.e. non randomly) excluded from the analysis.

QUESTION 5: WHAT IS AN EXAMPLE OF DATASET WITH A NON-GAUSSIAN DISTRIBUTION?

Binomial: multiple toss of a coin
Bin(n, p): the binomial distribution consists of the probability of each of the possible numbers of successes on n trials for independent events that each have a probability of p of occurring.

Bernoulli: Bin(1,p) = Be(p)
Poisson: Pois(λ)

DETAILED ANSWERS OF THE QUESTIONS MENTIONED EARLIER:

1. https://spin.atomicobject.com/2015/02/12/central-limit-theorem-intro/
2. https://searchbusinessanalytics.techtarget.com/definition/data-sampling
3. https://www.datasciencecentral.com/profiles/blogs/understanding-type-i-and-type-ii-errors
4. https://www.elderresearch.com/blog/selection-bias-in-analytics
5. https://www.quora.com/Most-machine-learning-datasets-are-in-Gaussian-distribution-Where-can-we-find-the-dataset-which-follows-Bernoulli-Poisson-gamma-beta-etc-distribution

AI

SHAPE
@shape.ai

COMMENT
LET US KNOW WHAT DO YOU WANT TO SEE NEXT
(LET'S CRACK IT)