Data Science interview questions

1. Probability and Statistics

1.1 What is a random variable?

In probability and statistics, random variables are used to quantify outcomes of a random occurrence, and therefore, can take on many values. Random variables are required to be measurable and are typically real numbers. Random variables can be discrete or continuous. For example, the letter X may be designated to represent the sum of the resulting numbers after three dice are rolled. In this case, X could be 3 (1 + 1+ 1), 18 (6 + 6 + 6), or somewhere between 3 and 18, since the highest number of a die is 6 and the lowest number is 1.

1.2 What is the main condition for a function to be a probability mass function?

A **probability mass function (PMF)**— also called a *frequency function*— gives you probabilities for **discrete random variables**. “Random variables” are variables from experiments like dice rolls, choosing a number out of a hat, or getting a high score on a test. The “discrete” part means that there’s a set number of outcomes. For example, you can only roll a 1, 2, 3, 4, 5, or 6 on a die.

Its counterpart is the probability density function, which gives probabilities for **continuous random variables**.

1.3 What is conditional probability?

The **conditional probability** of an event B is the probability that the event will occur given the knowledge that an event A has already occurred. This probability is written P(B|A), notation for the probability of B given A. In the case where events A and B are independent (where event A has no effect on the probability of event B), the conditional probability of event B given event A is simply the probability of event B, that is P(B).
If events A and B are not independent, then the probability of the intersection of A and B (the probability that both events occur) is defined by

P(A and B) = P(A)P(B|A).

From this definition, the conditional probability P(B|A) is easily obtained by dividing by *P(A)*:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

1.4 What is a Bernoulli distribution?

A Bernoulli distribution is a **discrete distribution** with only two possible values for the random variable. The distribution has only two possible outcomes and a single trial which is called a Bernoulli trial. The two possible outcomes in Bernoulli distribution are labeled by n=0 and n=1 in which n=1 (success) occurs with probability **p** and n=0 (failure) occurs with probability **1-p,** and since it is a probability value so 0<=p<=1.

The probability mass function (PMF) of a Bernoulli distribution is defined as:

If an experiment has only two possible outcomes, "success" and "failure," and if p is the probability of success, then-

$$P(n) = p^n (1-p)^{1-n}.$$

Another common way to write this is-

$$P(n) = \begin{cases} 1 - p & \text{for } n = 0 \\ p & \text{for } n = 1 \end{cases}$$

Note: Success here refers to an outcome that we want to keep track of. For example, in the dice rolling example, a double six in both dice would be a success, anything else rolled would be failure.

1.5 What isa normal distribution?

The normal distribution is the most important probability distribution in statistics because it fits many natural phenomena. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution. It is also known as the Gaussian distribution and the bell curve.

The normal distribution is a probability function that describes how the values of a variable are distributed. It is a symmetric distribution where most of the observations cluster around the central peak and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely.

1.6 What Is the Central Limit Theorem (CLT)?

In the study of probability theory, the central limit theorem (CLT) states that the distribution of sample approximates a normal distribution (also known as a “bell curve”) as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population distribution shape.

Said another way, CLT is a statistical theory stating that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population. Furthermore, all the samples will follow an approximate normal distribution pattern, with all variances being approximately equal to the variance of the population, divided by each sample's size.

1.7 What does P-value signify about the statistical data?

In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

1.8 What is the difference between skewed and uniformdistribution?

Uniform distribution refers to a condition when all the observations in a dataset are equally spread across the range of distribution. Skewed distribution refers to the condition when one side of the graph has more dataset in comparison to the other side.

1.9 What is the difference between covarianceand correlation?

Covariance

It is a metric which is used to measure the direction of relationship between two random variables and evaluates how two variables change together. Difference between variance and covariance lies in the fact that variance measures how one variable varies, however, covariance measures how two variables vary with respect to each other. In other words, variance reveals the covariance of variable with itself. Covariance can take up any value from (-) infinity to (+) infinity and gives the direction of relationship between two variables.

One important point to note here is that it only measures how two variables change together, not the dependency of one variable on another one.

Correlation:

Correlation is used to measure the strength of relationship between two variables and is the scaled measure of covariance. Correlation coefficient is a dimensionless metric and its value varies from (-1) to (+1). Here, (-1) refers to strong negative relationship between two variables while (+1) refers to strong positive relationship

Correlation and Covariance:

Covariance and correlation are related to each other in the sense that covariance determines the type of interaction between two variables while correlation determines the direction as well as strength of the relationship between two variables.

1.10 What is the differece between Variance and Covariance?

Variance

Variance is used in statistics to describe the spread between a data set from its mean value. It is calculated by finding the probability-weighted average of squared deviations from the expected value. So the larger the variance, the larger the distance between the numbers in the set and the mean. Conversely, a smaller variance means the numbers in the set are closer to the mean.

Covariance

A covariance refers to the measure of how two random variables will change when they are compared to each other. Covariance is calculated as expected value or average of the product of the differences of each random variable from their expected values.