Statistics Advance Assignment - 2

January 17, 2024

0.1 Q1: What are the Probability Mass Function (PMF) and Probability Density Function (PDF)? Explain with an example.

Probability Mass Function (PMF) and Probability Density Function (PDF) are both ways of describing the probability distribution of a random variable.

A Probability Mass Function (PMF) is a function that gives the probability that a discrete random variable is exactly equal to a certain value. In other words, the PMF gives the probability mass at each possible value of the random variable. The PMF is often denoted as P(X=x), where X is the random variable and x is a possible value that X can take.

For example, consider a coin that is flipped three times, where each flip has a probability of 0.5 of resulting in heads. The number of heads that come up can be modeled by a discrete random variable X with possible values 0, 1, 2, or 3. The PMF of X is:

P(X=0) = 0.125 P(X=1) = 0.375 P(X=2) = 0.375 P(X=3) = 0.125

The PMF indicates that the probability of getting 0 heads is 0.125, the probability of getting 1 head is 0.375, and so on. The sum of all the probabilities is 1, since one of these outcomes must occur.

A Probability Density Function (PDF), on the other hand, is a function that describes the relative likelihood that a continuous random variable takes on a certain value. In other words, the PDF gives the density of probability per unit of the random variable. The area under the curve of the PDF over a certain interval gives the probability of the random variable falling within that interval. The PDF is often denoted as f(X), where X is the random variable.

For example, consider a continuous random variable X that follows a normal distribution with mean 0 and standard deviation 1. The PDF of X is:

f(X) = (1 / sqrt(2)) * exp(-X^2/2)

This PDF describes the shape of the normal distribution, which is a bell-shaped curve centered at the mean of 0. The density of probability is highest at the mean and decreases symmetrically as X moves away from the mean in either direction. The area under the curve of the PDF between two values of X gives the probability that X falls within that range. For example, the probability that X falls between -1 and 1 can be calculated by integrating the PDF between -1 and 1:

P(-1 X 1) = [-1,1] f(X) dX = 0.6827

This gives a probability of 0.6827, which represents the proportion of the total area under the PDF that falls between -1 and 1.

0.2 Q2: What is Cumulative Density Function (CDF)? Explain with an example. Why CDF is used?

The Cumulative Density Function (CDF) is a function that describes the probability that a random variable X is less than or equal to a certain value x. In other words, the CDF of X gives the cumulative probability of X taking on a value less than or equal to x. The CDF is often denoted as F(x), where X is the random variable and x is a value that X can take.

For example, consider a continuous random variable X that follows a uniform distribution on the interval [0,1]. The CDF of X is:

F(x) = 0, if x $<$ 0 F(x) = x, if 0 x $<$ 1 F(x) = 1, if x 1

This CDF describes the cumulative probability of X taking on a value less than or equal to x. For example, the probability that X is less than or equal to 0.5 can be calculated by evaluating the CDF at x=0.5:

F(0.5) = 0.5

This gives a probability of 0.5, which represents the proportion of the total probability mass of X that falls below 0.5.

The CDF is used because it provides a complete description of the probability distribution of a random variable, including both discrete and continuous distributions. It can be used to calculate probabilities for any interval of values, not just specific values like the PMF or PDF. Additionally, the CDF is a cumulative measure of probability, which can be useful in certain applications where we are interested in the probability of a certain outcome occurring after a certain number of trials or events.

In summary, the CDF is a function that describes the cumulative probability of a random variable taking on a value less than or equal to a certain value. It is a useful tool for describing and analyzing probability distributions, and provides a complete description of the distribution that can be used to calculate probabilities for any interval of values.

0.3 Q3: What are some examples of situations where the normal distribution might be used as a model?

0.4 Explain how the parameters of the normal distribution relate to the shape of the distribution.

The normal distribution is one of the most commonly used probability distributions in statistics and is often used as a model for many natural phenomena. Some examples of situations where the normal distribution might be used as a model include:

Height: The distribution of human height tends to follow a normal distribution, with a mean height of around 5'7" for men and 5'3" for women.

Test Scores: The distribution of test scores in a large population tends to follow a normal distribution, with most people scoring near the mean and fewer people scoring very high or very low.

IQ Scores: IQ scores also tend to follow a normal distribution, with the mean score set at 100 and a standard deviation of 15.

Measurement Errors: The distribution of errors in measurement tends to follow a normal distribution.

The normal distribution is a bell-shaped probability distribution with the following properties:

The mean ( ) is the center of the distribution and determines the location of the peak.

The standard deviation ( ) determines the spread or width of the distribution. Larger standard deviations result in a wider distribution.

The shape of the distribution is symmetrical about the mean.

The total area under the curve is equal to 1.

The parameters of the normal distribution, namely the mean and the standard deviation, determine the shape of the distribution. If the mean increases, the distribution shifts to the right, and if the mean decreases, the distribution shifts to the left. If the standard deviation increases, the distribution becomes more spread out or wider, and if the standard deviation decreases, the distribution becomes more narrow. The mean and standard deviation also determine the height and shape of the bell curve. For example, a larger mean would result in a taller and narrower curve, while a smaller mean would result in a shorter and wider curve.

1 Q4: Explain the importance of Normal Distribution. Give a few real-life examples of Normal Distribution.

The normal distribution is one of the most important probability distributions in statistics and has a wide range of applications in various fields. Some of the key reasons for the importance of the normal distribution are:

It is a very common distribution: Many natural phenomena tend to follow a normal distribution, making it a useful model for statistical analysis.

It has well-defined properties: The normal distribution is well-studied and well-understood, which makes it easy to work with mathematically and to interpret statistically.

It is flexible: The normal distribution can be scaled and shifted to fit a wide range of data, making it a versatile tool for statistical analysis.

Some real-life examples of the normal distribution include:

Heights: The distribution of human heights tends to follow a normal distribution, with most people falling near the mean height and fewer people falling at the extremes of the distribution.

Test scores: The distribution of test scores in a large population tends to follow a normal distribution, with most people scoring near the mean and fewer people scoring very high or very low.

Body weight: The distribution of body weight in a population tends to follow a normal distribution, with most people having a weight near the mean and fewer people having a weight at the extremes of the distribution.

IQ scores: IQ scores also tend to follow a normal distribution, with most people having a score near the mean and fewer people having scores at the extremes of the distribution.

Errors in measurement: The distribution of errors in measurement tends to follow a normal distribution, with most errors being small and fewer errors being very large.

Overall, the normal distribution is an important tool for statistical analysis and has a wide range of applications in various fields, including finance, physics, psychology, and more.

1.1 Q5: What is Bernaulli Distribution? Give an Example. What is the difference between Bernoulli Distribution and Binomial Distribution?

Bernoulli distribution is a discrete probability distribution that models the outcome of a single experiment that can have only two possible outcomes - success or failure. It is named after Swiss mathematician Jacob Bernoulli, who developed the concept of probability theory.

The Bernoulli distribution has a single parameter, p, which represents the probability of success in a single trial. The probability mass function (PMF) of the Bernoulli distribution is given by:

P(X=x) = p^x * (1-p)^(1-x) for x = 0 or 1

where X is the random variable representing the outcome of the experiment, and p is the probability of success.

An example of the Bernoulli distribution is the coin toss experiment, where the outcome can be either heads or tails. Assuming a fair coin, the probability of getting heads (success) is 0.5, and the probability of getting tails (failure) is also 0.5.

The key difference between Bernoulli and Binomial distribution is that Bernoulli distribution models the outcome of a single trial, whereas Binomial distribution models the outcome of multiple independent trials with the same probability of success. In other words, Bernoulli distribution is a special case of Binomial distribution, where the number of trials is 1.

The Binomial distribution has two parameters - n (the number of trials) and p (the probability of success in each trial). The probability mass function (PMF) of the Binomial distribution is given by:

P(X=k) = C(n,k) * p^k * (1-p)^(n-k)

where X is the random variable representing the number of successes in n independent trials, k is the number of successes, and C(n,k) is the binomial coefficient.

An example of the Binomial distribution is the number of heads obtained in 10 coin tosses, where the probability of getting heads (success) in each trial is 0.5. In this case, n=10 and p=0.5.

1.2 Q6. Consider a dataset with a mean of 50 and a standard deviation of 10. If we assume that the dataset is normally distributed, what is the probability that a randomly selected observation will be greater than 60? Use the appropriate formula and show your calculations.

To find the probability that a randomly selected observation from a normally distributed dataset with a mean of 50 and standard deviation of 10 will be greater than 60, we can use the standard normal distribution and standardize the value of 60 using the z-score formula:

z = (x -) /

where z is the standardized score, x is the observation, is the mean, and is the standard deviation.

In this case, we have:

z = (60 - 50) / 10 = 1

We can then use a standard normal distribution table or calculator to find the probability of a z-score greater than 1.

Using a standard normal distribution table, we find that the probability of a z-score greater than 1 is approximately 0.1587.

Therefore, the probability that a randomly selected observation from the normally distributed dataset with a mean of 50 and standard deviation of 10 will be greater than 60 is approximately 0.1587 or 15.87%.

1.3 Q7: Explain uniform Distribution with an example.

Uniform distribution is a continuous probability distribution where all values within a specific range have an equal chance of occurring. In other words, it is a probability distribution where the likelihood of any given value occurring is the same as the likelihood of any other value occurring within the same range.

For example, let's say we have a fair six-sided die. When we roll the die, the outcome can be any number between 1 and 6, and each outcome has an equal probability of occurring. Therefore, the distribution of outcomes is uniform.

We can represent the uniform distribution mathematically as:

f(x) = 1 / (b - a)

where f(x) is the probability density function, a is the minimum value in the range, and b is the maximum value in the range.

For example, if we consider a uniform distribution between 0 and 10, the probability density function would be:

f(x) = 1 / (10 - 0) = 0.1

This means that the probability of getting any value between 0 and 10 is the same and equal to 0.1.

1.4 Q8: What is the z score? State the importance of the z score.

A z-score, also known as a standard score, is a statistical measure that represents the number of standard deviations a data point is from the mean of its population. It is calculated by subtracting the mean of the population from the data point and dividing the result by the standard deviation of the population.

The formula for calculating z-score is:

z = (x -) /

where x is the data point, is the population mean, and is the population standard deviation.

The importance of the z-score is that it allows us to standardize and compare different sets of data that may have different scales and units. By converting data points to z-scores, we can compare them to a standard normal distribution with a mean of 0 and a standard deviation of 1. This allows us to make meaningful comparisons and draw conclusions about the relative standing of different data points within their respective populations.

For example, if we have two sets of data with different means and standard deviations, we can standardize the data using z-scores and compare them on a common scale. This is useful in fields such as finance, where we might want to compare the performance of different stocks, or in education, where we might want to compare the performance of students on different tests.

1.5 Q9: What is Central Limit Theorem? State the significance of the Central Limit Theorem.

The Central Limit Theorem (CLT) is a fundamental concept in probability theory that states that, under certain conditions, the sample means from a large number of independent and identically distributed (i.i.d) random variables will be approximately normally distributed, regardless of the underlying distribution of the population.

In other words, the CLT states that the sum or average of a large number of independent and identically distributed random variables will converge to a normal distribution as the sample size increases, regardless of the shape of the original distribution. This is true even if the original population is not normally distributed.

The significance of the Central Limit Theorem is that it allows us to make inferences about a population using a sample, without having to know the underlying distribution of the population. This is because the CLT ensures that the sample mean will be approximately normally distributed, and we can use this fact to construct confidence intervals and perform hypothesis testing.

For example, suppose we want to estimate the mean height of all students in a university, but we cannot measure the height of every student. Instead, we take a random sample of 100 students and calculate the sample mean. By the Central Limit Theorem, we know that the distribution of the sample mean will be approximately normal, regardless of the distribution of heights in the population. This allows us to make inferences about the population mean height based on our sample mean and standard error.

1.6 Q10: State the assumptions of the Central Limit Theorem.

The Central Limit Theorem (CLT) relies on the following assumptions:

Independence: The samples or observations are independent of each other.

Sample size: The sample size is sufficiently large. While there is no hard and fast rule for the minimum sample size, a commonly used rule of thumb is that the sample size should be at least 30.

Identically distributed: The samples or observations are drawn from the same population, and therefore have the same mean and standard deviation.

Finite variance: The population from which the samples are drawn must have a finite variance.

If these assumptions are met, then the sample means will converge to a normal distribution, regardless of the underlying distribution of the population. It is important to note that violating

these assumptions can lead to inaccurate conclusions and the CLT may not hold.