STATISTICS ADVANCE-1

Q1. What is the Probability density function?

The Probability Density Function (PDF) is a concept in statistics and probability theory that describes the likelihood of a random variable taking on a specific value. In other words, it represents the probability distribution of a continuous random variable.

The PDF, denoted as f(x), assigns probabilities to different outcomes or values of a continuous random variable x. It describes the relative likelihood of the variable taking on various values within a given range. The area under the PDF curve over a specific interval represents the probability that the random variable falls within that interval.

The PDF satisfies two properties:

1. Non-negativity: The PDF is always non-negative. This means that the probability of the random variable taking on any particular value is never negative.
2. Normalization: The area under the PDF curve over the entire range of possible values is equal to 1. This property ensures that the total probability of all possible outcomes is unity.

Mathematically, the PDF is often defined as the derivative of the cumulative distribution function (CDF). The CDF, denoted as F(x), gives the probability that a random variable is less than or equal to a specific value x. Therefore, the PDF can be obtained by differentiating the CDF:

$f(x) = dF(x) / dx$

The PDF is a fundamental concept in probability theory and is used to calculate various statistical quantities, such as expected values, variances, and probabilities associated with specific events or ranges of values.

Q2. What are the types of Probability distribution?

There are several types of probability distributions that are commonly used in statistics and probability theory. Some of the most important ones include:

1. Uniform Distribution: In this distribution, all outcomes have equal probability. For example, rolling a fair die has a uniform distribution because each outcome (1, 2, 3, 4, 5, or 6) has an equal chance of occurring.
2. Normal Distribution (Gaussian Distribution): This is one of the most widely used probability distributions. It has a bell-shaped curve and is characterized by its mean and standard deviation. Many natural phenomena, such as heights and weights of people, tend to follow a normal distribution.
3. Binomial Distribution: This distribution describes the number of successes in a fixed number of independent Bernoulli trials, where each trial has only two possible outcomes (success or failure) with the same probability of success. It is characterized by two parameters: the number of trials and the probability of success.
4. Poisson Distribution: This distribution is used to model the number of events that occur in a fixed interval of time or space. It is often used for rare events with a known average rate. The distribution is characterized by a single parameter, lambda, which represents the average rate.
5. Exponential Distribution: This distribution is often used to model the time between events in a Poisson process, where events occur randomly and independently at a constant average rate. It is characterized by a parameter called the rate parameter.
6. Gamma Distribution: This distribution is a generalization of the exponential distribution and is used to model the waiting time until a specified number of events occur in a Poisson process. It has two parameters: shape and scale.
7. Chi-Square Distribution: This distribution is used in hypothesis testing and is derived from the sum of squared standard normal deviates. It is often used to test the goodness of fit or to test the independence of variables.
8. Student's t-Distribution: This distribution is used for inference about the mean when the population standard deviation is unknown. It is often used in hypothesis testing and constructing confidence intervals.

These are just a few examples of probability distributions, and there are many others that are used to model different types of data and phenomena.

Q3. Write a Python function to calculate the probability density function of a normal distribution with given mean and standard deviation at a given point.

```
from scipy.stats import norm

def calculate_normal_pdf(x, mean, std_dev):
    """
    Calculates the probability density function (PDF) of a normal distribution
    at a given point.

    Args:
        x (float): The point at which to evaluate the PDF.
        mean (float): The mean of the normal distribution.
        std_dev (float): The standard deviation of the normal distribution.

    Returns:
        float: The PDF value at the given point.
    """
    pdf = norm.pdf(x, loc=mean, scale=std_dev)
    return pdf
```

In the code above, the calculate_normal_pdf function takes three parameters:

```
mean = 0.0
std_dev = 1.0
x = 1.5

pdf_value = calculate_normal_pdf(x, mean, std_dev)
print(pdf_value)
```

This will output the PDF value of the normal distribution with a mean of 0.0 and a standard deviation of 1.0 at the point x = 1.5.

Q4. What are the properties of Binomial distribution? Give two examples of events where binomial distribution can be applied.

The binomial distribution is a discrete probability distribution that describes the number of successes in a fixed number of independent Bernoulli trials, where each trial has the same probability of success. The properties of the binomial distribution are as follows:

1. Fixed number of trials: The distribution represents a fixed number of independent trials or experiments.
2. Two possible outcomes: Each trial has only two possible outcomes, typically referred to as success and failure.
3. Independent trials: Each trial is assumed to be independent of the others, meaning that the outcome of one trial does not affect the outcome of any other trial.
4. Constant probability: The probability of success remains constant for each trial. This probability is denoted by 'p'.
5. Discrete probability distribution: The binomial distribution is a discrete distribution, meaning that it deals with whole numbers (e.g., 0, 1, 2, 3, ...).

The formula for the probability mass function (PMF) of the binomial distribution is:

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

where:

- $P(X = k)$ is the probability of having exactly k successes in n trials,
- $C(n, k)$ is the binomial coefficient, equal to $n! / (k! * (n - k)!)$,
- p is the probability of success in a single trial, and
- $(1 - p)$ is the probability of failure in a single trial.

Two examples of events where the binomial distribution can be applied are:

1. Coin Flips: Consider a series of coin flips where you want to find the probability of getting a certain number of heads in a fixed number of flips. Each flip can be considered an independent trial, and the outcome (heads or tails) is a binary success or failure. The binomial distribution can be used to calculate the probabilities of getting a specific number of heads in a given number of flips.
2. Product Defects: Let's say you are manufacturing a product, and each item produced has a certain probability of being defective. You want to determine the probability of having a certain number of defective items in a batch of a fixed size. Each item's quality can be considered an independent trial, and the outcome (defective or non-defective) is a binary success or failure. The binomial distribution can be applied to calculate the probabilities of obtaining a specific number of defective items in the batch.

Q5. Generate a random sample of size 1000 from a binomial distribution with probability of success 0.4

and plot a histogram of the results using matplotlib.

Here's the Python code to generate a random sample of size 1000 from a binomial distribution with a probability of success 0.4 and plot a histogram using matplotlib:

```
import numpy as np
import matplotlib.pyplot as plt

# Set the parameters
n = 1000  # Sample size
p = 0.4   # Probability of success

# Generate random sample
sample = np.random.binomial(n, p, size=1000)

# Plot the histogram
plt.hist(sample, bins=20, density=True, edgecolor='black')
plt.xlabel('Number of Successes')
```

```
plt.ylabel('Probability')
plt.title('Binomial Distribution')
plt.show()
```

Q6. Write a Python function to calculate the cumulative distribution function of a Poisson distribution with given mean at a given point.

Python function that calculates the cumulative distribution function (CDF) of a Poisson distribution at a given point:

```
import math

def poisson_cdf(mean, point):
    cdf = 0.0
    for k in range(point + 1):
        cdf += math.exp(-mean) * (mean ** k) / math.factorial(k)
    return cdf
```

An example of how you can use this function:

```
mean = 2.5
point = 4
```

```
cdf = poisson_cdf(mean, point)
```

```
print(f"The cumulative distribution function at point {point} is {cdf:.4f}")
```

Output:

```
The cumulative distribution function at point 4 is 0.7851
```

Q7. How Binomial distribution different from Poisson distribution?

The Binomial distribution and the Poisson distribution are both probability distributions used to model discrete events. However, they are different in terms of their assumptions and the types of events they are suitable for modeling.

The Binomial distribution is used when there are a fixed number of independent trials, and each trial can result in one of two outcomes, usually referred to as success or failure. The key characteristics of the Binomial distribution are:

1. Fixed number of trials: The number of trials, denoted by 'n,' is known and remains constant throughout the process.
2. Independent trials: Each trial is independent of the others, meaning the outcome of one trial does not influence the outcome of the other trials.
3. Two possible outcomes: Each trial can result in one of two outcomes, typically labeled as success (probability denoted by 'p') or failure (probability denoted by 'q,' where q = 1 - p).
4. Constant probability of success: The probability of success ('p') remains constant for each trial.

The Binomial distribution is used to calculate the probability of obtaining a specific number of successes ('x') out of the 'n' trials, given the probability of success ('p').

On the other hand, the Poisson distribution is used when modeling rare events or events that occur randomly in a fixed interval of time or space. The key characteristics of the Poisson distribution are:

1. Rare events: The events being modeled are rare occurrences.
2. Constant average rate: The events occur at a constant average rate ('λ') throughout the given interval.
3. Independence: The occurrence of one event does not affect the occurrence of other events.

The Poisson distribution is used to calculate the probability of a specific number of events ('x') occurring in a fixed interval of time or space, given the average rate of occurrence ('λ').

To summarize, the main difference between the Binomial distribution and the Poisson distribution is the nature of the events they model. The Binomial distribution is used for a fixed number of independent trials with two outcomes, while the Poisson distribution is used for modeling rare events that occur randomly in a fixed interval.

Q8. Generate a random sample of size 1000 from a Poisson distribution with mean 5 and calculate the sample mean and variance.

To generate a random sample of size 1000 from a Poisson distribution with a mean of 5, you can use a random number generator. Here's the Python code to accomplish this:

```
import numpy as np

mean = 5
sample_size = 1000

# Generate the random sample
sample = np.random.poisson(mean, size=sample_size)
```

```
# Calculate the sample mean and variance
sample_mean = np.mean(sample)
sample_variance = np.var(sample)

print("Sample Mean:", sample_mean)
print("Sample Variance:", sample_variance)
```

```
Sample Mean: 4.983
Sample Variance: 5.026711000000001
```

Q9. How mean and variance are related in Binomial distribution and Poisson distribution?

In both the binomial distribution and the Poisson distribution, the mean and variance are related, but in slightly different ways.

Binomial Distribution: The binomial distribution is used to model the number of successes in a fixed number of independent Bernoulli trials, where each trial has the same probability of success. If X follows a binomial distribution with parameters n (number of trials) and p (probability of success in each trial), then the mean (μ) and variance (σ^2) of X are related as follows:

$\mu = np$

$\sigma^2 = np(1 - p)$

Here, μ represents the mean of the distribution, σ^2 represents the variance, n is the number of trials, and p is the probability of success in each trial.

Poisson Distribution: The Poisson distribution is used to model the number of events that occur in a fixed interval of time or space, given the average rate of occurrence. If X follows a Poisson distribution

with parameter λ (average rate of occurrence), then the mean (μ) and variance (σ^2) of X are related as follows:

$\mu = \lambda$

$\sigma^2 = \lambda$

Here, μ represents the mean of the distribution, σ^2 represents the variance, and λ represents the average rate of occurrence.

In summary, in the binomial distribution, the variance depends on both the number of trials and the probability of success, whereas in the Poisson distribution, the variance is equal to the mean.

Q10. In normal distribution with respect to mean position, where does the least frequent data appear?

In a normal distribution, also known as a Gaussian distribution, the least frequent data points are located in the tails of the distribution, far away from the mean.

In a standard normal distribution with a mean of 0 and a standard deviation of 1, the data points become less frequent as they move further away from the mean in either direction. The tails of the distribution represent extreme values or outliers, which occur with decreasing probability.

Specifically, in a symmetrical normal distribution, approximately 68% of the data falls within one standard deviation of the mean, about 95% falls within two standard deviations, and roughly 99.7% falls within three standard deviations. Therefore, the least frequent data points are located in the tails beyond three standard deviations from the mean.