

Statistics Advance-1

January 17, 2024

0.1 Q1. What is the Probability density function?

A Probability Density Function (PDF) is a mathematical function that describes the relative likelihood of a continuous random variable taking on a given value. The PDF represents the distribution of probabilities for all possible values of the random variable, and the area under the curve of the PDF between any two values of the variable gives the probability of the variable taking a value within that range.

The PDF can be thought of as a continuous version of a probability distribution. Unlike a discrete probability distribution, which assigns a probability to each possible value of a discrete random variable, the PDF assigns probabilities to ranges of possible values of a continuous random variable.

Mathematically, the PDF is defined as the derivative of the cumulative distribution function (CDF) of the random variable. The PDF is non-negative and integrates to one over the entire range of the random variable.

0.2 Q2. What are the types of Probability distribution?

There are many different types of probability distributions, but they can be broadly classified into two categories: discrete probability distributions and continuous probability distributions.

Discrete Probability Distributions: Discrete probability distributions are used to describe the probability of occurrence of discrete random variables. Examples of discrete probability distributions include the Bernoulli distribution, binomial distribution, Poisson distribution, and geometric distribution.

Continuous Probability Distributions: Continuous probability distributions are used to describe the probability of occurrence of continuous random variables. Examples of continuous probability distributions include the normal distribution, uniform distribution, exponential distribution, and beta distribution.

The choice of probability distribution depends on the nature of the data and the specific problem at hand. The appropriate distribution should be chosen to accurately model the data and to make inferences and predictions about the population of interest.

0.3 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.

```
[2]: import math

def normal_pdf(x, mu, sigma):
    """
    Calculates the probability density function of a normal distribution with
    mean mu and standard deviation sigma at the point x.
    """
    exponent = -0.5 * ((x - mu) / sigma) ** 2
    coefficient = 1 / (sigma * math.sqrt(2 * math.pi))
    return coefficient * math.exp(exponent)

pdf_value = normal_pdf(x=2, mu=0, sigma=1)
print(pdf_value)
```

0.05399096651318806

0.4 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 probability of a fixed 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:

Fixed number of trials: The number of trials is fixed and denoted by n .

Independent trials: Each trial is independent of the others.

Two possible outcomes: Each trial has only two possible outcomes, success or failure.

Constant probability of success: The probability of success, denoted by p , is constant for all trials.

Discrete values: The number of successes, denoted by k , is a discrete variable that takes values from 0 to n .

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

Tossing a Coin: When tossing a fair coin, there are two possible outcomes - heads or tails - with a fixed probability of 0.5 for each outcome. If we toss the coin n times and want to know the probability of getting k heads, we can model this as a binomial distribution with n trials, a probability of success of 0.5, and the number of successes k .

Manufacturing Defects: A company manufactures electronic components that have a defect rate of 2%. If we randomly sample n components from the production line and want to know the probability that k of them are defective, we can model this as a binomial distribution with n trials, a probability of success of 0.02 (the probability of a component being defective), and the number of successes k .

In both of these examples, the binomial distribution allows us to model the probability of a fixed number of successes in a fixed number of independent trials, where each trial has the same probability of success.

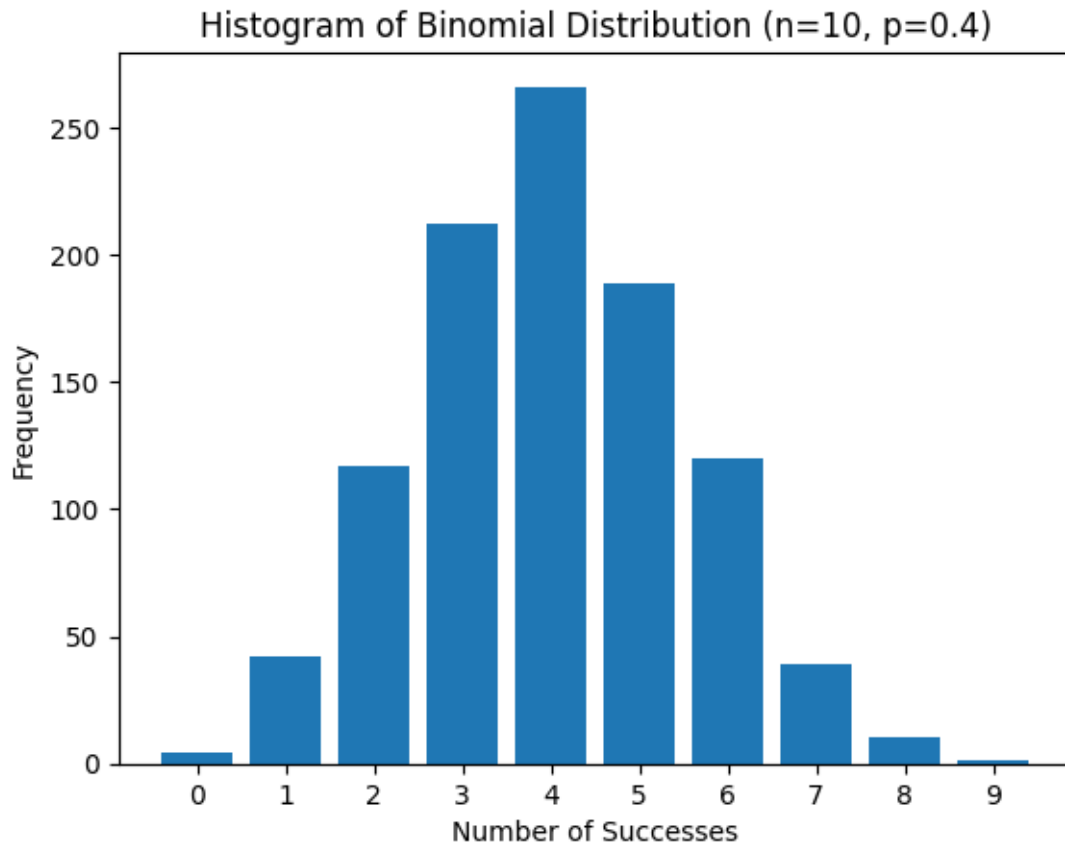
0.5 Q5. Generate a random sample of size 1000 from a binomial distribution with probability of success 0.4 and plot a histogram of the results using matplotlib.

```
[3]: import numpy as np
import matplotlib.pyplot as plt

# Set the parameters of the binomial distribution
n = 10
p = 0.4

# Generate a random sample of size 1000 from the binomial distribution
sample = np.random.binomial(n, p, size=1000)

# Plot a histogram of the results
plt.hist(sample, bins=range(n+1), align='left', rwidth=0.8)
plt.xticks(range(n))
plt.xlabel('Number of Successes')
plt.ylabel('Frequency')
plt.title('Histogram of Binomial Distribution (n={}, p={})'.format(n, p))
plt.show()
```



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

```
[4]: import math

def poisson_cdf(mu, k):
    cdf = 0.0
    for i in range(k+1):
        cdf += math.exp(-mu) * (mu ** i) / math.factorial(i)
    return cdf
mu = 2.5
k = 3
cdf = poisson_cdf(mu, k)
print("The cumulative distribution function of a Poisson distribution with mean_
↵{} at point {} is {:.4f}".format(mu, k, cdf))
```

The cumulative distribution function of a Poisson distribution with mean 2.5 at point 3 is 0.7576

0.7 Q7. How Binomial distribution different from Poisson distribution?

Binomial distribution and Poisson distribution are both discrete probability distributions that describe the number of events that occur in a given number of trials. However, there are some key differences between the two distributions.

Parameters: The Binomial distribution is characterized by two parameters: the number of trials n and the probability of success p . In contrast, the Poisson distribution is characterized by one parameter: the average rate at which events occur.

Application: The Binomial distribution is used to model the number of successes in a fixed number of independent trials, where the probability of success is constant across all trials. The Poisson distribution, on the other hand, is used to model the number of rare events that occur over a fixed interval of time or space.

Shape: The shape of the Binomial distribution is symmetric when $p = 0.5$, and becomes increasingly skewed as p approaches 0 or 1. In contrast, the Poisson distribution is typically skewed to the right, and becomes more symmetric as the rate parameter increases.

Limitations: The Binomial distribution is appropriate when the number of trials n is large and the probability of success p is small, but the mean np is not too small. The Poisson distribution is appropriate when the rate parameter is small, but the product of the rate and the observation time or space is moderate to large.

In summary, the main differences between the Binomial distribution and the Poisson distribution are the number of parameters, the application, the shape, and the limitations of the distributions.

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

```
[5]: import numpy as np

# Set the random seed for reproducibility
np.random.seed(123)

# Generate a random sample of size 1000 from a Poisson distribution with mean 5
sample = np.random.poisson(5, size=1000)

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

print("Sample mean: {:.4f}".format(mean))
print("Sample variance: {:.4f}".format(variance))
```

Sample mean: 4.9190

Sample variance: 5.0304

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

In both the Binomial and Poisson distributions, the mean and variance are related in a similar way.

For the Binomial distribution with parameters n and p , the mean is given by $\mu = np$ and the variance is given by $\sigma^2 = np(1-p)$. Thus, the variance is a function of both the number of trials n and the probability of success p . As p gets closer to 0 or 1, the variance decreases, reflecting the fact that the distribution becomes increasingly concentrated around the mean.

For the Poisson distribution with parameter λ , the mean is also given by $\mu = \lambda$ and the variance is also given by $\sigma^2 = \lambda$. Thus, the variance is a function of the rate parameter λ . As λ increases, the variance also increases, reflecting the fact that the distribution becomes more spread out around the mean.

In both distributions, when the mean is large, the variance is also large, indicating that the distribution is more spread out. Conversely, when the mean is small, the variance is also small, indicating that the distribution is more tightly clustered around the mean.

In summary, for both the Binomial and Poisson distributions, the variance is related to the mean in a way that reflects the degree of spread or concentration of the distribution.

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

In a normal distribution, the least frequent data appears at the tails of the distribution, farthest from the mean. This is because the normal distribution is symmetric and bell-shaped, with the majority of the data clustered around the mean. The tails of the distribution represent the extreme values that are less likely to occur, since they are located far away from the mean.

Specifically, in a normal distribution with mean μ and standard deviation σ , the least frequent data appears at the values that are 3 σ or more away from the mean in either direction. This is because about 99.7% of the data falls within 3 σ of the mean, leaving only 0.3% of the data to be spread out across the tails of the distribution.

Therefore, the least frequent data in a normal distribution appears at the extreme values that are located at the tails of the distribution, farthest from the mean.