Batch:H2_1 Roll No.: 16010122151

Experiment No. 3

Title : To implement probability based statistical modelling

Aim: To implement probability based statistical modelling such as Binomial Distribution, Poisson Distribution and Normal/Gaussian distribution.

Expected Outcome of Experiment:
CO1 : Develop an understanding of data science and business analytics.

Books/ Journals/ Websites referred:

1. Binomial distribution

The "binomial" in binomial distribution means two terms—the number of successes and the number of attempts. Each is useless without the other. Binomial distribution is a common discrete distribution used in statistics, as opposed to a continuous distribution, such as normal distribution. This is because binomial distribution only counts two states, typically represented as 1 (for a success) or 0 (for a failure), given a number of trials in the data. Binomial distribution thus represents the probability for x successes in n trials, given a success probability p for each trial.

The binomial distribution function is calculated as:

$P_{(x:n,p)} = {}^{n}C_{x}\, p^{x} (1 - p)^{n-x}$

Where:

- n is the number of trials (occurrences)
- x is the number of successful trials
- p is the probability of success in a single trial
- ${}^{n}C_{x}$ is the combination of n and x. A combination is the number of ways to choose a sample of x elements from a set of n distinct objects where order does not matter, and replacements are not allowed. Note that ${}_{n}C_{x}$ = n! / r! (n − r) !), where ! is factorial (so, 4! = 4 × 3 × 2 × 1).

Program:

```
# Setting the parameters for the binomial distribution

n_trials <- 10  # Number of trials

prob_success <- 0.3  # Probability of success



# Generate a random sample from a binomial distribution
```

```
random_sample <- rbinom(n = 1, size = n_trials, prob = prob_success)

cat("Random sample:", random_sample, "\n")


# Calculate the probability mass function (PMF) at specific values

values <- c(0, 1, 2, 3)

pmf_values <- dbinom(x = values, size = n_trials, prob = prob_success)

cat("PMF at", values, ":", pmf_values, "\n")


# Calculate the cumulative distribution function (CDF) at specific values

cdf_values <- pbinom(q = values, size = n_trials, prob = prob_success)

cat("CDF at", values, ":", cdf_values, "\n")


# Find quantiles given probabilities

quantiles <- qbinom(p = c(0.1, 0.5, 0.9), size = n_trials, prob = prob_success)

cat("Quantiles at  probabilities 0.1, 0.5, 0.9:", quantiles, "\n")
```

OUTPUT:

```
Random sample: 3

PMF at 0 1 2 3 : 0.02824752 0.1210608 0.2334744 0.2668279

CDF at 0 1 2 3 : 0.02824752 0.1493083 0.3827828 0.6496107

Quantiles at probabilities 0.1, 0.5, 0.9: 1 3 5
```

2. Poisson Distribution

In statistics, a Poisson distribution is a probability distribution that is used to show how many times an event is likely to occur over a specified period. In other words, it is a count distribution. Poisson distributions are often used to understand independent events that occur at a constant rate within a given interval of time. It was named after French mathematician Siméon Denis Poisson.

$$f(x) = \frac{\lambda^{\mathrm{x}}}{x!} e^{-\text{ഉ}}$$

Where:

- e is Euler's number (e = 2.71828...)
- x is the number of occurrences
- x! is the factorial of x
- λ is equal to the expected value (EV) of x when that is also equal to its variance

Program:

```
# Setting the parameter for the Poisson distribution

lambda <- 3  # Average number of events per unit of time or space

# Generate a random sample from a Poisson distribution

random_sample <- rpois(n = 10, lambda = lambda)

cat("Random sample:", random_sample, "\n")

# Calculate the probability mass function (PMF) at specific values

values <- c(0, 1, 2, 3)

pmf_values <- dpois(x = values, lambda = lambda)

cat("PMF at", values, ":", pmf_values, "\n")

# Calculate the cumulative distribution function (CDF) at specific values

cdf_values <- ppois(q = values, lambda = lambda)

cat("CDF at", values, ":", cdf_values, "\n")

# Find quantiles given probabilities

quantiles <- qpois(p = c(0.1, 0.5, 0.9), lambda = lambda)

cat("Quantiles at probabilities 0.1, 0.5, 0.9:", quantiles, "\n")
```

OUTPUT:

```
Random sample: 2 4 1 2 3 3 4 2 2 3

PMF at 0 1 2 3 : 0.04978707 0.1493612 0.2240418 0.2240418

CDF at 0 1 2 3 : 0.04978707 0.1991483 0.4231901 0.6472319

Quantiles at probabilities 0.1, 0.5, 0.9: 1 3 5
```

3. Normal Distribution

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.In graphical form, the normal distribution appears as a "bell curve". The standard normal distribution has two parameters: the mean and the standard deviation. In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.

The normal distribution follows the following formula. Note that only the values of the mean (μ) and standard deviation (σ) are necessary

Normal Distribution Formula.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where:

- *x* = value of the variable or data being examined and f(x) the probability function
- μ = the mean
- σ = the standard deviation

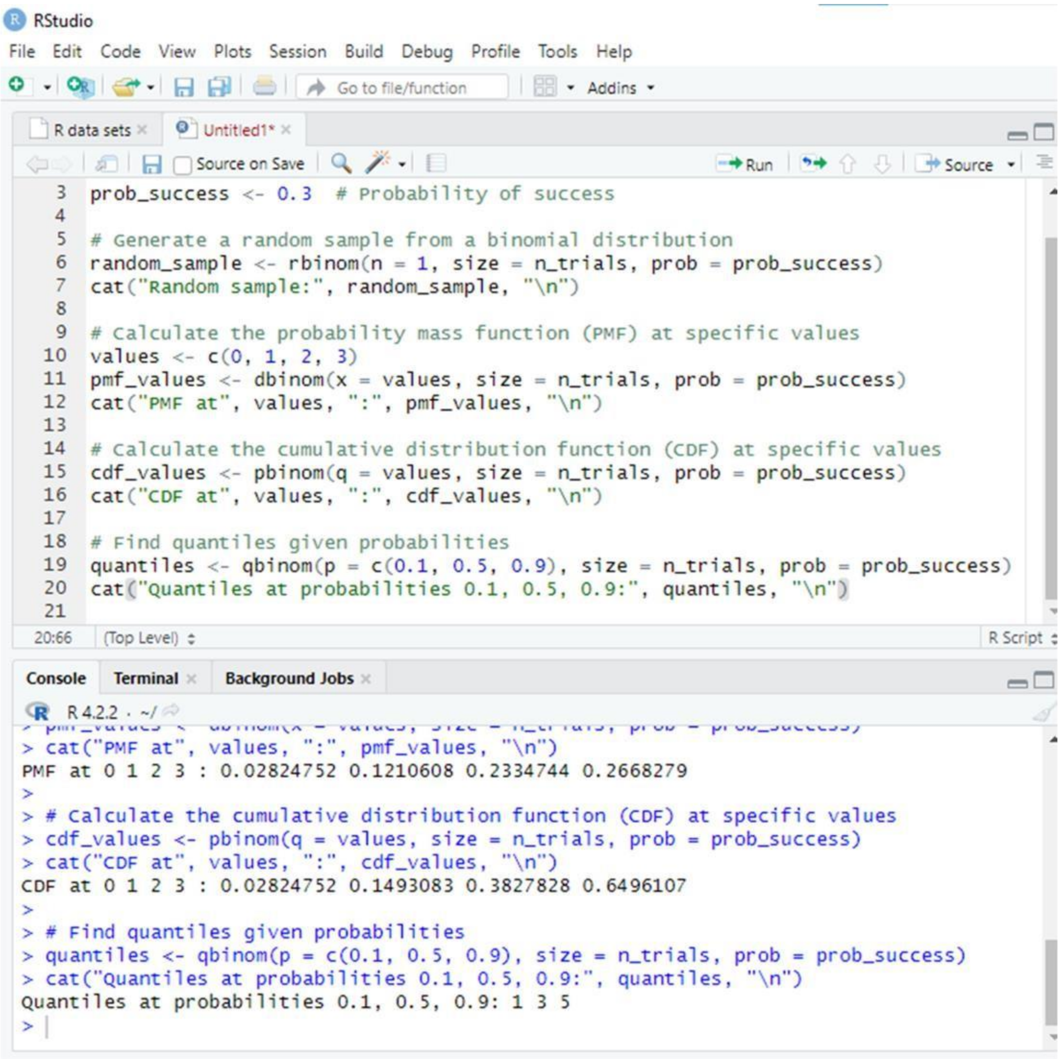
Program:

```
# Setting the parameters for the normal distribution

mean_value <- 0  # Mean of the distribution

sd_value <- 1     # Standard deviation of the distribution

# Generate a random sample from a normal distribution

random_sample <- rnorm(n = 10, mean = mean_value, sd = sd_value)

cat("Random sample:", random_sample, "\n")

# Calculate the probability density function (PDF) at specific values

values <- c(-2, -1, 0, 1, 2)

pdf_values <- dnorm(x = values, mean = mean_value, sd = sd_value)

cat("PDF at", values, ":", pdf_values, "\n")
```

```
# Calculate the cumulative distribution function (CDF) at specific values

cdf_values <- pnorm(q = values, mean = mean_value, sd = sd_value)

cat("CDF at", values, ":", cdf_values, "\n")

# Find quantiles given probabilities

quantiles <- qnorm(p = c(0.1, 0.5, 0.9), mean = mean_value, sd = sd_value)

cat("Quantiles at probabilities 0.1, 0.5, 0.9:", quantiles, "\n")
```

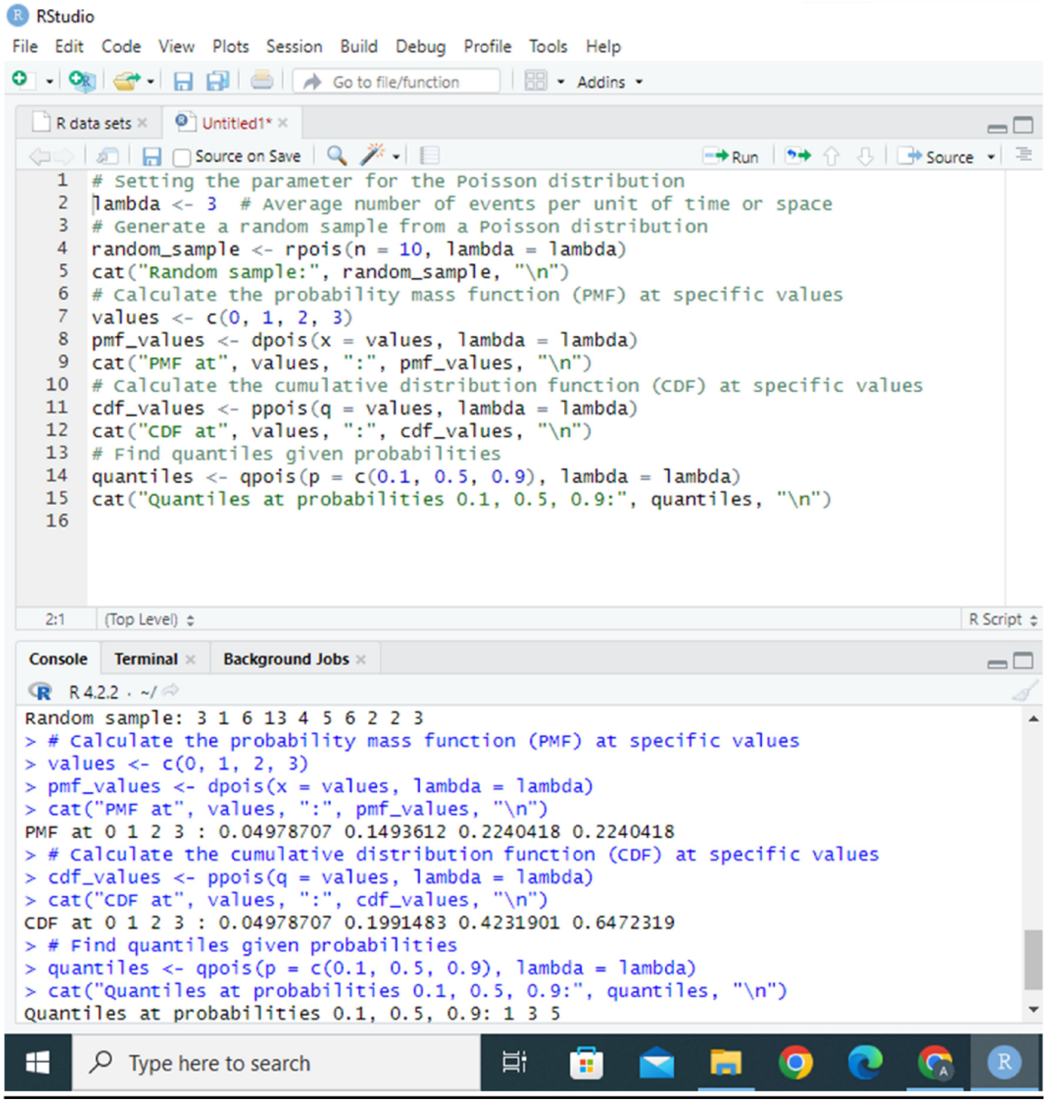
OUTPUT:

```
Random sample: -2.450496 0.3155664 0.469913 -0.656226 -0.6094917 -1.41421 -0.124466 -1.610715 -0.4915843 -0.3460785

PDF at -2 -1 0 1 2 : 0.05399097 0.2419707 0.3989423 0.2419707 0.05399097

CDF at -2 -1 0 1 2 : 0.02275013 0.1586553 0.5 0.8413447 0.9772499

Quantiles at probabilities 0.1, 0.5, 0.9: -1.281552 0 1.281552
```



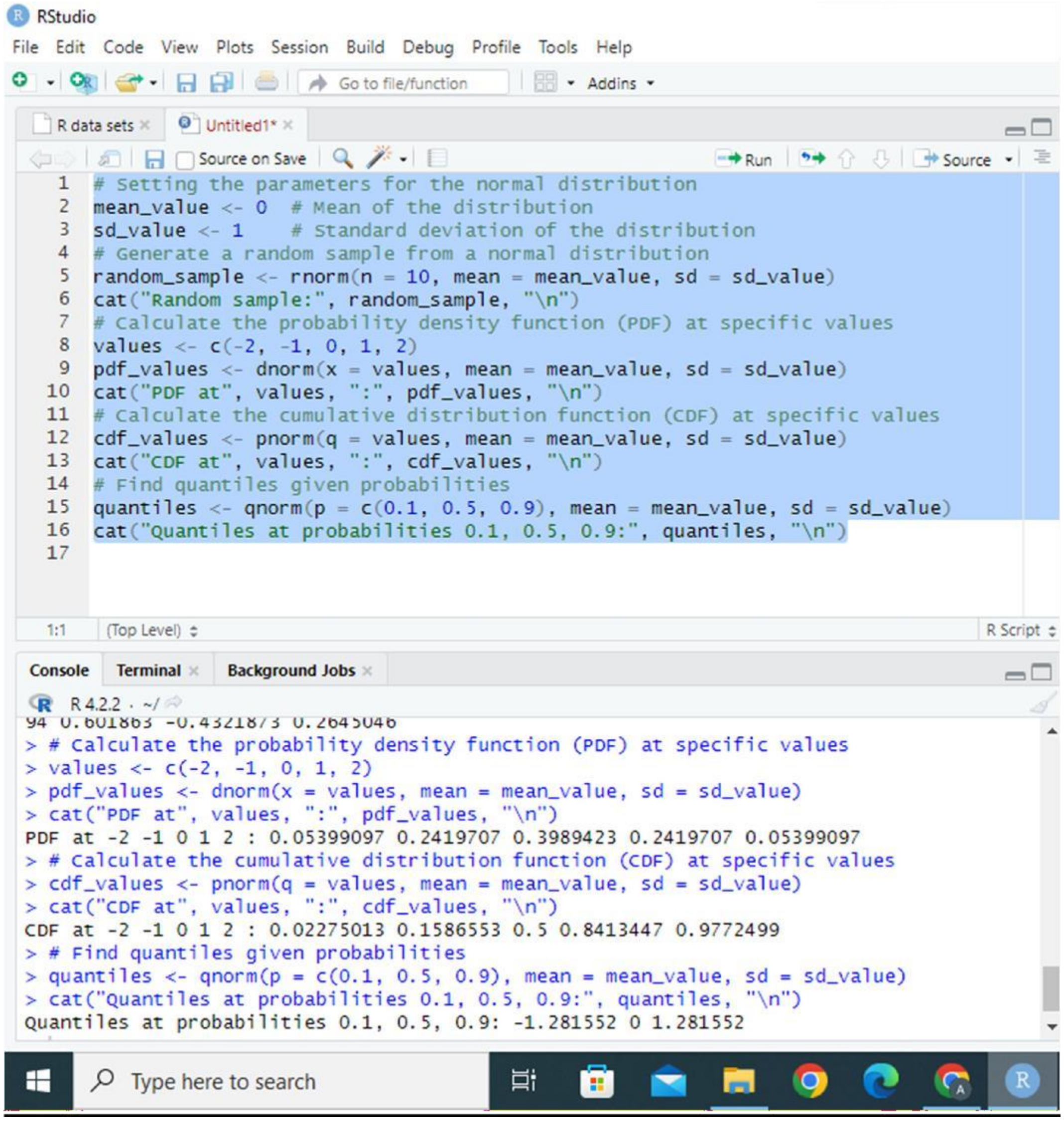
```
prob_success <- 0.3  # Probability of success

# Generate a random sample from a binomial distribution
random_sample <- rbinom(n = 1, size = n_trials, prob = prob_success)
cat("Random sample:", random_sample, "\n")

# Calculate the probability mass function (PMF) at specific values
values <- c(0, 1, 2, 3)
pmf_values <- dbinom(x = values, size = n_trials, prob = prob_success)
cat("PMF at", values, ":", pmf_values, "\n")

# Calculate the cumulative distribution function (CDF) at specific values
cdf_values <- pbinom(q = values, size = n_trials, prob = prob_success)
cat("CDF at", values, ":", cdf_values, "\n")

# Find quantiles given probabilities
quantiles <- qbinom(p = c(0.1, 0.5, 0.9), size = n_trials, prob = prob_success)
cat("Quantiles at probabilities 0.1, 0.5, 0.9:", quantiles, "\n")
```

```
> pmf_values <- dbinom(x = values, size = n_trials, prob = prob_success)
> cat("PMF at", values, ":", pmf_values, "\n")
PMF at 0 1 2 3 : 0.02824752 0.1210608 0.2334744 0.2668279
>
> # Calculate the cumulative distribution function (CDF) at specific values
> cdf_values <- pbinom(q = values, size = n_trials, prob = prob_success)
> cat("CDF at", values, ":", cdf_values, "\n")
CDF at 0 1 2 3 : 0.02824752 0.1493083 0.3827828 0.6496107
>
> # Find quantiles given probabilities
> quantiles <- qbinom(p = c(0.1, 0.5, 0.9), size = n_trials, prob = prob_success)
> cat("Quantiles at probabilities 0.1, 0.5, 0.9:", quantiles, "\n")
Quantiles at probabilities 0.1, 0.5, 0.9: 1 3 5
> 
```



```
# Setting the parameter for the Poisson distribution
lambda <- 3  # Average number of events per unit of time or space
# Generate a random sample from a Poisson distribution
random_sample <- rpois(n = 10, lambda = lambda)
cat("Random sample:", random_sample, "\n")
# Calculate the probability mass function (PMF) at specific values
values <- c(0, 1, 2, 3)
pmf_values <- dpois(x = values, lambda = lambda)
cat("PMF at", values, ":", pmf_values, "\n")
# Calculate the cumulative distribution function (CDF) at specific values
cdf_values <- ppois(q = values, lambda = lambda)
cat("CDF at", values, ":", cdf_values, "\n")
# Find quantiles given probabilities
quantiles <- qpois(p = c(0.1, 0.5, 0.9), lambda = lambda)
cat("Quantiles at probabilities 0.1, 0.5, 0.9:", quantiles, "\n")
```

```
Random sample: 3 1 6 13 4 5 6 2 2 3
> # Calculate the probability mass function (PMF) at specific values
> values <- c(0, 1, 2, 3)
> pmf_values <- dpois(x = values, lambda = lambda)
> cat("PMF at", values, ":", pmf_values, "\n")
PMF at 0 1 2 3 : 0.04978707 0.1493612 0.2240418 0.2240418
> # Calculate the cumulative distribution function (CDF) at specific values
> cdf_values <- ppois(q = values, lambda = lambda)
> cat("CDF at", values, ":", cdf_values, "\n")
CDF at 0 1 2 3 : 0.04978707 0.1991483 0.4231901 0.6472319
> # Find quantiles given probabilities
> quantiles <- qpois(p = c(0.1, 0.5, 0.9), lambda = lambda)
> cat("Quantiles at probabilities 0.1, 0.5, 0.9:", quantiles, "\n")
Quantiles at probabilities 0.1, 0.5, 0.9: 1 3 5
```



```r
# Setting the parameters for the normal distribution
mean_value <- 0  # Mean of the distribution
sd_value <- 1    # Standard deviation of the distribution
# Generate a random sample from a normal distribution
random_sample <- rnorm(n = 10, mean = mean_value, sd = sd_value)
cat("Random sample:", random_sample, "\n")
# Calculate the probability density function (PDF) at specific values
values <- c(-2, -1, 0, 1, 2)
pdf_values <- dnorm(x = values, mean = mean_value, sd = sd_value)
cat("PDF at", values, ":", pdf_values, "\n")
# Calculate the cumulative distribution function (CDF) at specific values
cdf_values <- pnorm(q = values, mean = mean_value, sd = sd_value)
cat("CDF at", values, ":", cdf_values, "\n")
# Find quantiles given probabilities
quantiles <- qnorm(p = c(0.1, 0.5, 0.9), mean = mean_value, sd = sd_value)
cat("Quantiles at probabilities 0.1, 0.5, 0.9:", quantiles, "\n")
```

```
94 0.601863 -0.4321873 0.2645046
> # Calculate the probability density function (PDF) at specific values
> values <- c(-2, -1, 0, 1, 2)
> pdf_values <- dnorm(x = values, mean = mean_value, sd = sd_value)
> cat("PDF at", values, ":", pdf_values, "\n")
PDF at -2 -1 0 1 2 : 0.05399097 0.2419707 0.3989423 0.2419707 0.05399097
> # Calculate the cumulative distribution function (CDF) at specific values
> cdf_values <- pnorm(q = values, mean = mean_value, sd = sd_value)
> cat("CDF at", values, ":", cdf_values, "\n")
CDF at -2 -1 0 1 2 : 0.02275013 0.1586553 0.5 0.8413447 0.9772499
> # Find quantiles given probabilities
> quantiles <- qnorm(p = c(0.1, 0.5, 0.9), mean = mean_value, sd = sd_value)
> cat("Quantiles at probabilities 0.1, 0.5, 0.9:", quantiles, "\n")
Quantiles at probabilities 0.1, 0.5, 0.9: -1.281552 0 1.281552
```

Post Lab questions

Q.1 You are managing a quality control process for a production line where each item produced can be classified as either defective or non-defective. The probability of producing a defective item is 0.05.

a. Define the binomial distribution and explain the key components involved.

b. How does the binomial distribution differ from other probability distributions?

c. Discuss the conditions that must be satisfied for a random variable to follow a binomial distribution.

a. **Binomial Distribution:** 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 only two possible outcomes: success (usually denoted as 1) or failure (usually denoted as 0). The key components involved are:

- **n**: The number of trials or experiments.
- **p**: The probability of success in each trial. In your case, p=0.05.
- **x**: The number of successes (defective items in your scenario) out of n trials.
- **P(X = x)**: The probability of observing x successes in n trials.

b. **Difference from Other Probability Distributions:** The binomial distribution differs from other probability distributions, such as the normal distribution or Poisson distribution, primarily in terms of the nature of the data it describes and the assumptions it makes. The binomial distribution deals with discrete data (successes/failures), while the normal distribution deals with continuous data. Additionally, the binomial distribution assumes a fixed number of trials with two possible outcomes, whereas other distributions may have different assumptions about the underlying process being modeled.

c. **Conditions for a Binomial Distribution:** For a random variable to follow a binomial distribution, the following conditions must be satisfied:

1. **Fixed Number of Trials (n)**: The number of trials or experiments must be fixed in advance.
2. **Independent Trials**: Each trial must be independent of the others. The outcome of one trial should not influence the outcome of another.
3. **Two Possible Outcomes**: Each trial must have only two possible outcomes: success or failure.
4. **Constant Probability of Success (p)**: The probability of success (denoted as p) must remain constant for each trial.

Q.2 Provide an example scenario from a real-world application where the binomial distribution and Poisson distribution is applicable. Explain why it fits the respective models.
Solution :-

1. **Binomial Distribution: Quality Control in Manufacturing** Scenario: A manufacturing company produces electronic components, and each component can either be defective or non-defective. The company's quality control process involves randomly selecting 50 components from each batch produced and testing them for defects.
   - **Applicability**: This scenario fits the binomial distribution because:
     - There is a fixed number of trials (50 components tested in each batch).
     - Each trial (testing a component) has only two possible outcomes: defective or non-defective.
     - The trials are independent, assuming that testing one component does not affect the testing of another.
     - The probability of a component being defective remains constant across trials, assuming consistent manufacturing processes.
   - **Example Calculation**: If the probability of a component being defective is p=0.05, we can use the binomial distribution to calculate the probability of finding a certain number of defective components in a batch of 50.

2. **Poisson Distribution: Traffic Accidents in a City** Scenario: A city's traffic department wants to model the number of traffic accidents that occur at a particular intersection during a given hour of the day. They collect data over several weeks and find that, on average, there are 2 accidents per hour at that intersection.
    - **Applicability**: This scenario fits the Poisson distribution because:
        - The number of accidents occurring in each hour can be modeled as a count of rare events (accidents).
        - The events (accidents) occur independently of each other within each time interval (hour).
        - The average rate of accidents (λ) is constant over time, assuming no significant changes in traffic patterns during the observation period.
    - **Example Calculation**: With an average rate of 2 accidents per hour ($=2\lambda=2$), the Poisson distribution can be used to calculate the probability of observing a specific number of accidents (e.g., 0, 1, 2) in a given hour at the intersection.

Q.3 The normal distribution is a fundamental concept in statistics and probability. Provide a comprehensive description of the normal distribution, covering the following aspects:

a. Define the normal distribution and explain its key characteristics.

b. Discuss the standard normal distribution and the role of the z-score in standardizing values.

c. Describe situations or phenomena in the real world where the normal distribution is commonly observed. Discuss why the normal distribution is a suitable model for these scenarios.

Solution :-

a. **Normal Distribution:** The normal distribution, also known as the Gaussian distribution or bell curve, is a continuous probability distribution that is symmetric about its mean. It is characterized by its probability density function (PDF) which is defined by two parameters: the mean (μ) and the standard deviation (σ). The PDF of the normal distribution is given by the formula: $2f(x)=2\pi\sigma 1e-2\sigma 2(x-\mu)2$ where:

- μ represents the mean, which determines the center of the distribution.
- σ represents the standard deviation, which measures the spread or dispersion of the distribution.
- e is the base of the natural logarithm.
- π is the mathematical constant pi.

Key Characteristics:

- Symmetry: The normal distribution is symmetric about its mean, with the tails extending indefinitely in both directions.
- Unimodality: It has a single peak at the mean, making it unimodal.
- Empirical Rule: Approximately 68% of the data falls within one standard deviation of the mean, 95% within two standard deviations, and 99.7% within three standard deviations.

- Asymptotic: The tails of the normal distribution approach, but never touch, the x-axis as they extend infinitely in both directions.

b. **Standard Normal Distribution and Z-score:** The standard normal distribution is a special case of the normal distribution with a mean (μ) of 0 and a standard deviation (σ) of 1. It is denoted by Z and has the PDF:$2f(z)=2\pi 1e-2z2$

The Z-score, also known as the standard score, represents the number of standard deviations a data point is from the mean of a normal distribution. It is calculated as: $Z=\sigma x-\mu$ where:

- x is the individual data point.
- μ is the mean of the distribution.
- σ is the standard deviation of the distribution.

The standardization process involving Z-scores allows for the comparison of values from different normal distributions and facilitates various statistical analyses and hypothesis testing.

c. **Real-World Applications of Normal Distribution:** The normal distribution is observed in numerous real-world situations due to the central limit theorem and the prevalence of random variables influenced by multiple factors. Common examples include:

- **Height and Weight:** Human height and weight distributions often follow a normal distribution due to the combined effects of genetic and environmental factors.
- **Test Scores:** Scores on standardized tests, such as IQ tests or SAT exams, often approximate a normal distribution among a large population.
- **Measurement Errors:** Errors in measurements, such as experimental errors in scientific experiments or manufacturing errors in industrial processes, tend to follow a normal distribution.
- **Financial Markets:** Daily returns on financial assets, such as stocks or currencies, are often assumed to follow a normal distribution, although this assumption may be challenged in practice.

The normal distribution is suitable for these scenarios because:

- Many natural phenomena are influenced by multiple independent factors, leading to a bell-shaped distribution due to the central limit theorem.
- The normal distribution is mathematically tractable and well-studied, allowing for easy analysis and inference using statistical techniques.
- It serves as a useful approximation for many real-world datasets, especially when sample sizes are sufficiently large.

Conclusion:

By carrying out this experiment, we were able to show how probability-based statistical modeling techniques—such as the Poisson, Normal, and Binomial distributions—can be used in practice to analyze and interpret data from the real world.