**Batch: H2\_3 Roll No.: 16010122221**

**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:

### 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 nCx = 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

### 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) =

𝜆𝑥

# 𝑥!

𝑒−𝜆

### 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

### 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.

1 −1 𝑥−𝜇 2

### Where:

𝑓(𝑥) = 𝑒 2 (

# 𝜎√2𝜋

𝜎 )

* *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

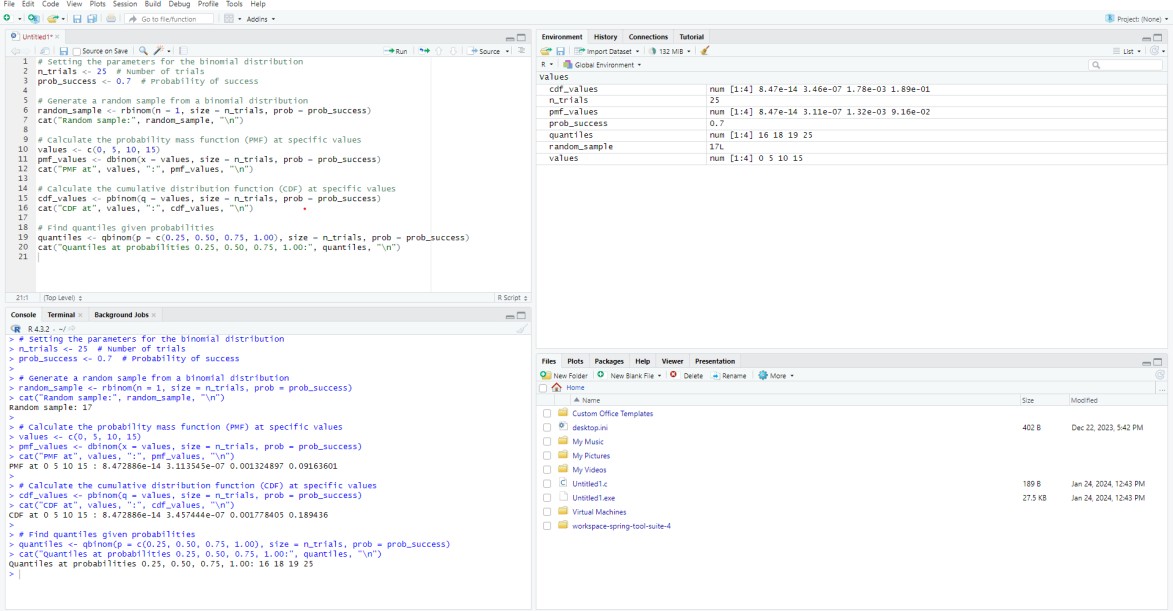
PDF at -2 -1 0 1 2 : 0.05399097 0.2419707 0.3989423 0.2419707 0.05399097

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

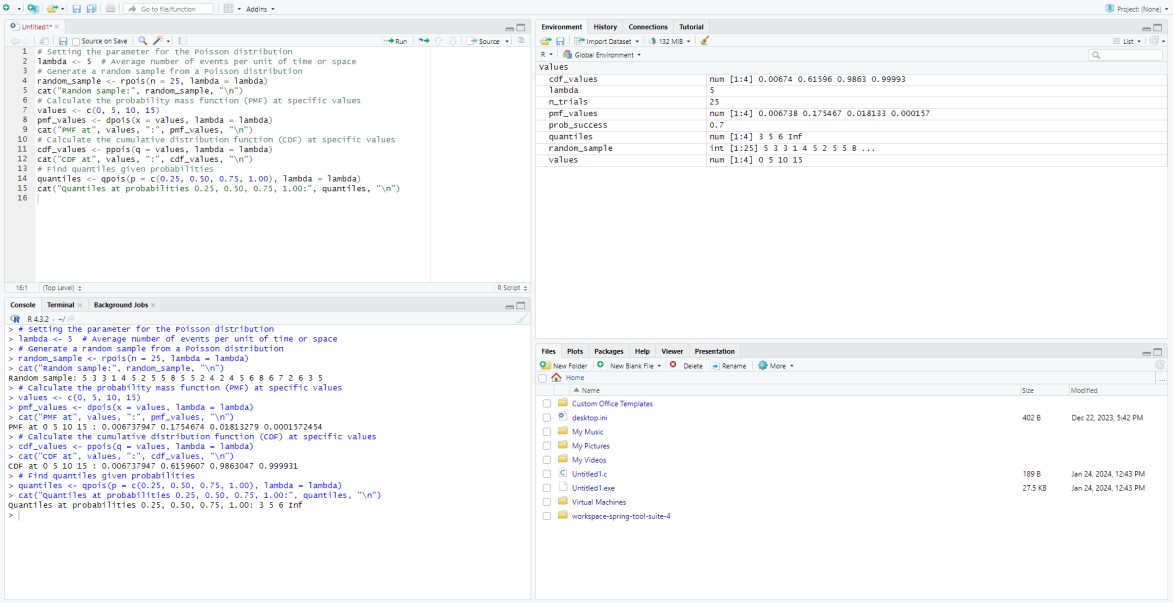
Quantiles at probabilities 0.1, 0.5, 0.9: -1.281552 0 1.281552

**EXECUTION:**

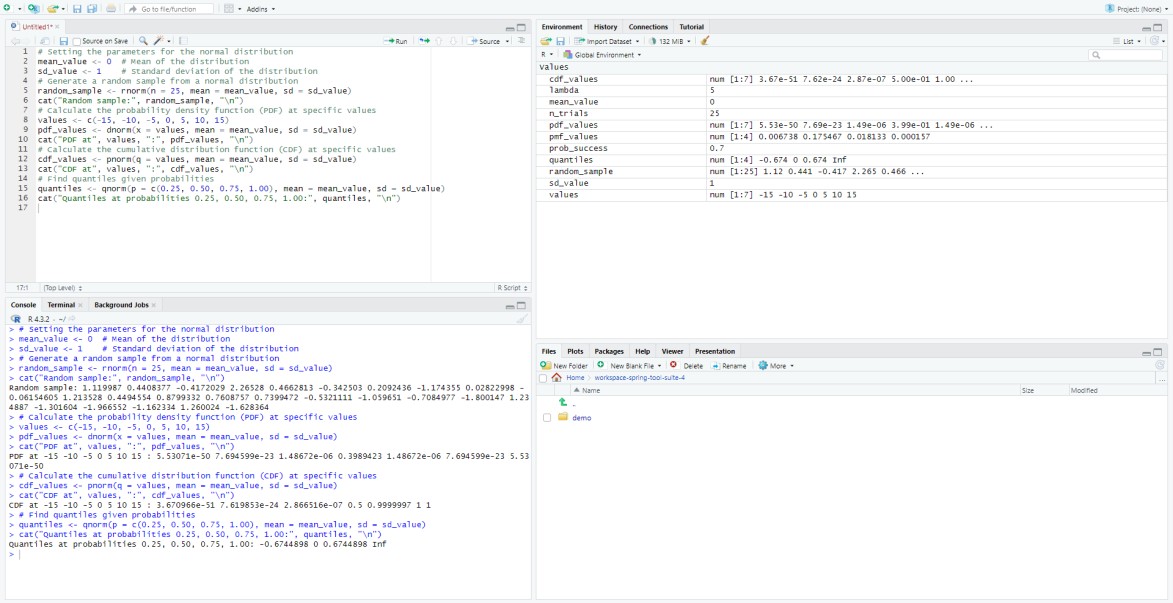
1. **Binomial Distribution:**



## Poisson Distribution:



1. **Normal Distribution:**



## Post Lab questions:

* 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.**

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

The binomial distribution models the number of successes in a fixed number of independent Bernoulli trials. Key components include the number of trials (n), probability of success (p), probability of failure (q), random variable (X), and the probability mass function (PMF) determining success counts.

## How does the binomial distribution differ from other probability distributions?

Distinct from continuous distributions, the binomial distribution deals with discrete outcomes in a fixed number of trials. Unlike normal or Poisson distributions, it specifically models binary outcomes, making it suitable for situations with repeated independent experiments yielding success or failure.

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

Conditions for a binomial distribution include a fixed number of trials, independence among trials, identical probability of success, and binary outcomes. These criteria ensure that the random variable representing the number of successes adheres to the binomial distribution, aiding quality control analyses in production processes.

## 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:**

## Binomial Distribution Example:

In quality control for manufacturing, the binomial distribution is applicable when inspecting a fixed number of items. If the probability of producing a defective item is 0.02, each inspection becomes an independent Bernoulli trial. With a fixed number of trials, such as 100 items, the binomial model predicts the probability of different counts of defective items, aiding in quality assessment.

## Poisson Distribution Example:

For call center arrival rates, the Poisson distribution is suitable when modeling the number of incoming calls per minute. If the average rate is 5 calls per minute, the Poisson model predicts the likelihood of specific call counts, assuming rare, independent events in continuous time.

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

* + 1. **Define the normal distribution and explain its key characteristics.**

The normal distribution is a symmetric bell-shaped probability distribution characterized by a mean (μ) and standard deviation (σ). It follows the empirical rule, with about 68%, 95%, and 99.7% of data falling within one, two, and three standard deviations from the mean, respectively.

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

The standard normal distribution has a mean (μ) of 0 and standard deviation (σ) of 1. The z-score standardizes values, representing the number of standard deviations a data point is from the mean. Z-scores facilitate comparison and analysis across different normal distributions.

## 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.

The normal distribution is commonly observed in phenomena such as human height, IQ scores, and measurement errors. Its suitability arises from the central limit theorem, stating that the sum or average of a large number of independent, identically distributed random variables approximates a normal distribution, making it a versatile model in various fields.

**Conclusion:** Implementing probability-based statistical models, including Binomial, Poisson, and Normal distributions, is vital in applied data science. These models empower data scientists to analyze real-world phenomena and make informed decisions.