**Batch: A - 2 Roll No.: 16014022050**

**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. <https://www.r-tutor.com/elementary-statistics/probability-distributions/binomial-distribution>
2. <https://www.r-tutor.com/elementary-statistics/probability-distributions/poisson-distribution>
3. <https://www.r-tutor.com/elementary-statistics/probability-distributions/normal-distribution>

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Theory:**

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 ) = nC x p x( 1 - p ) n - x*

**Where –**

* n is the number of trials (occurances)
* 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

1. **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.

**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

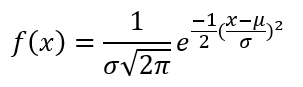
Quantiles at probabilities 0.1, 0.5, 0.9: 1 3 5

1. **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 is given by,

****

**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

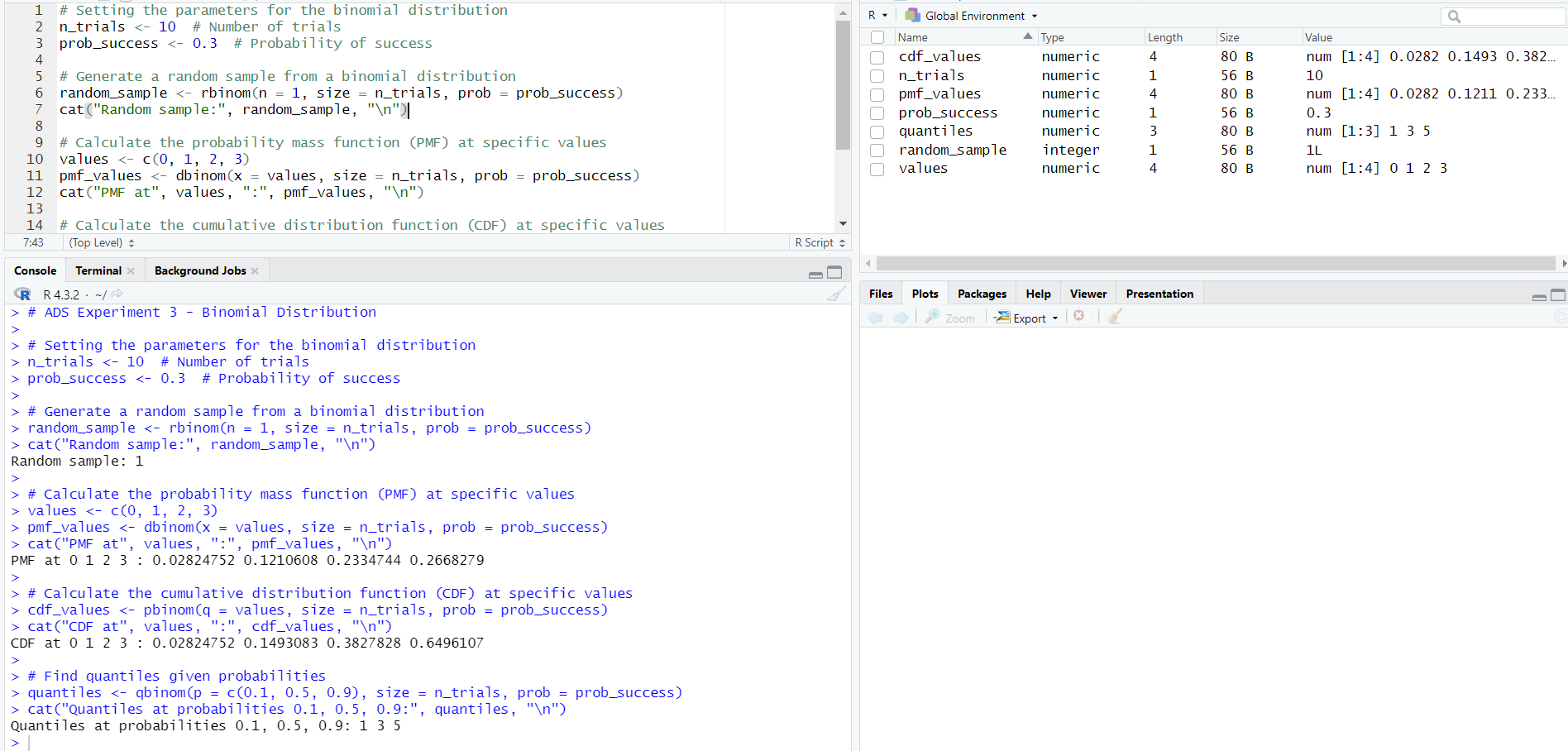
Quantiles at probabilities 0.1, 0.5, 0.9: -1.281552 0 1.281552

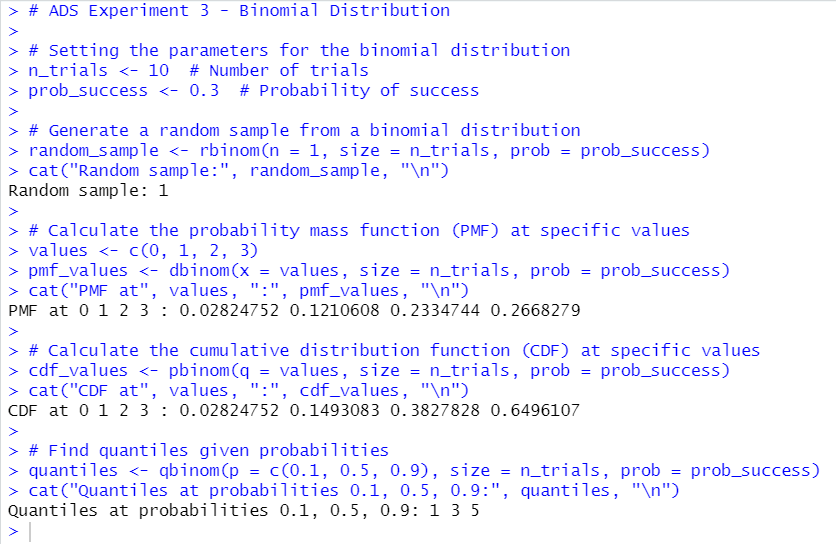
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

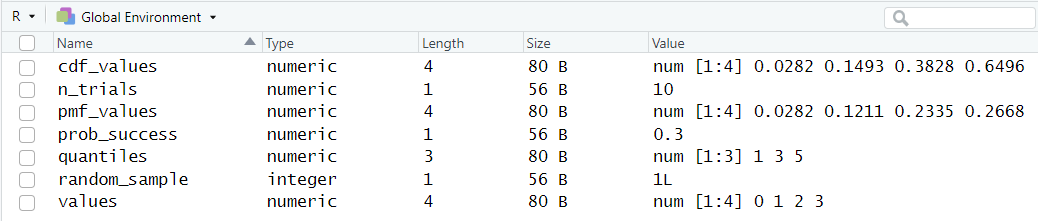
**Implementation:**

Students have to perform **ALL** the operations shown above and add their screenshots here.

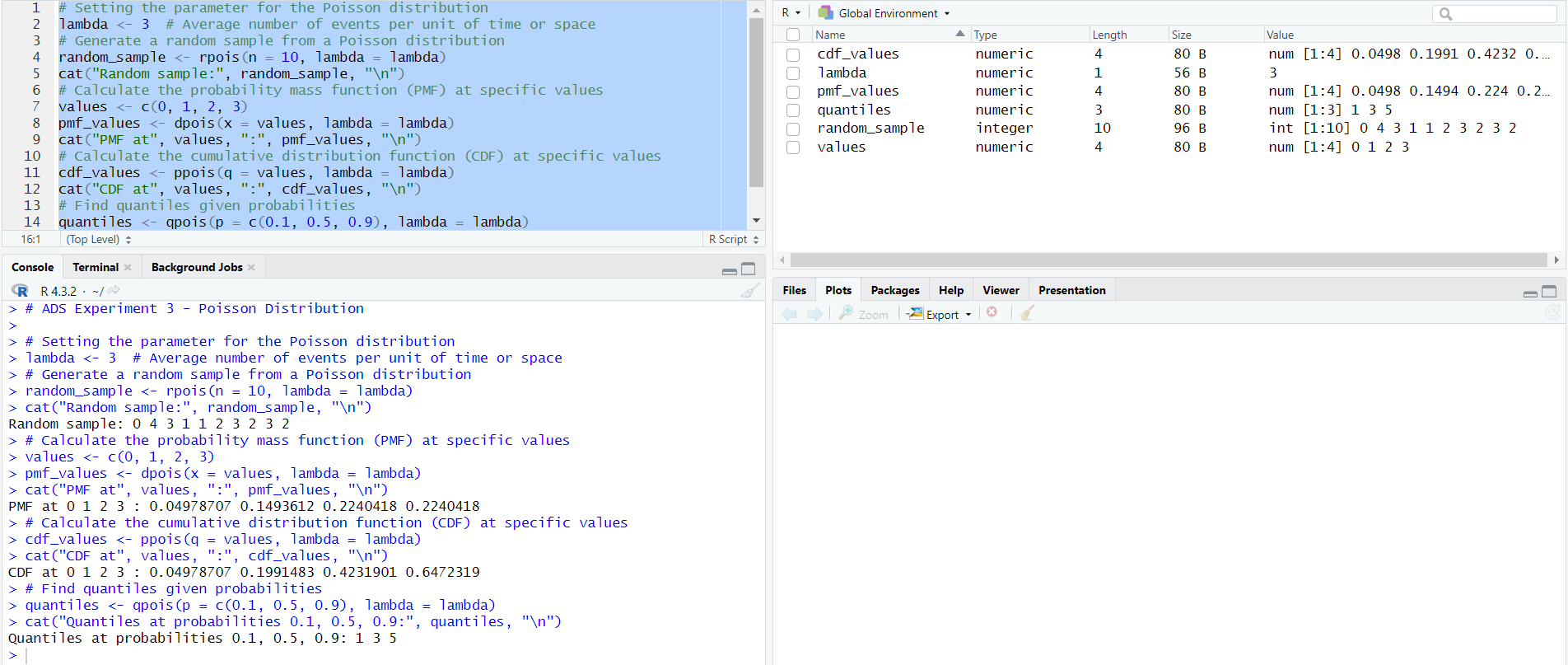
1. **Binomial Distribution:**

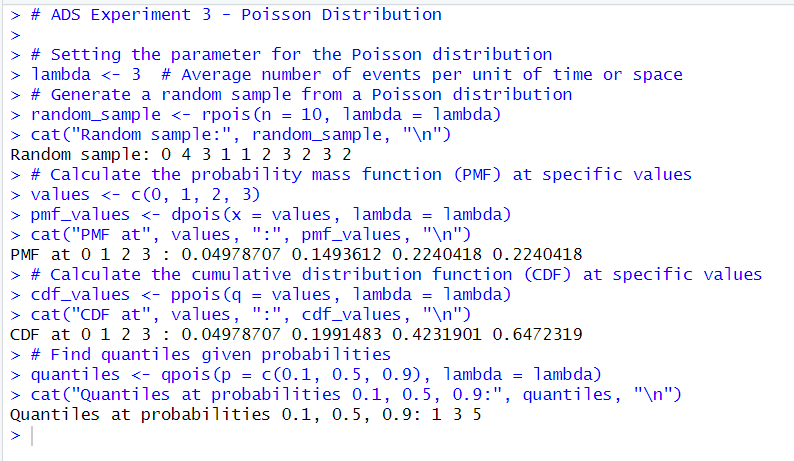
****

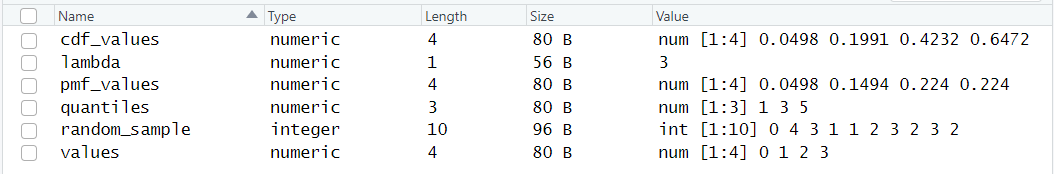
****

****

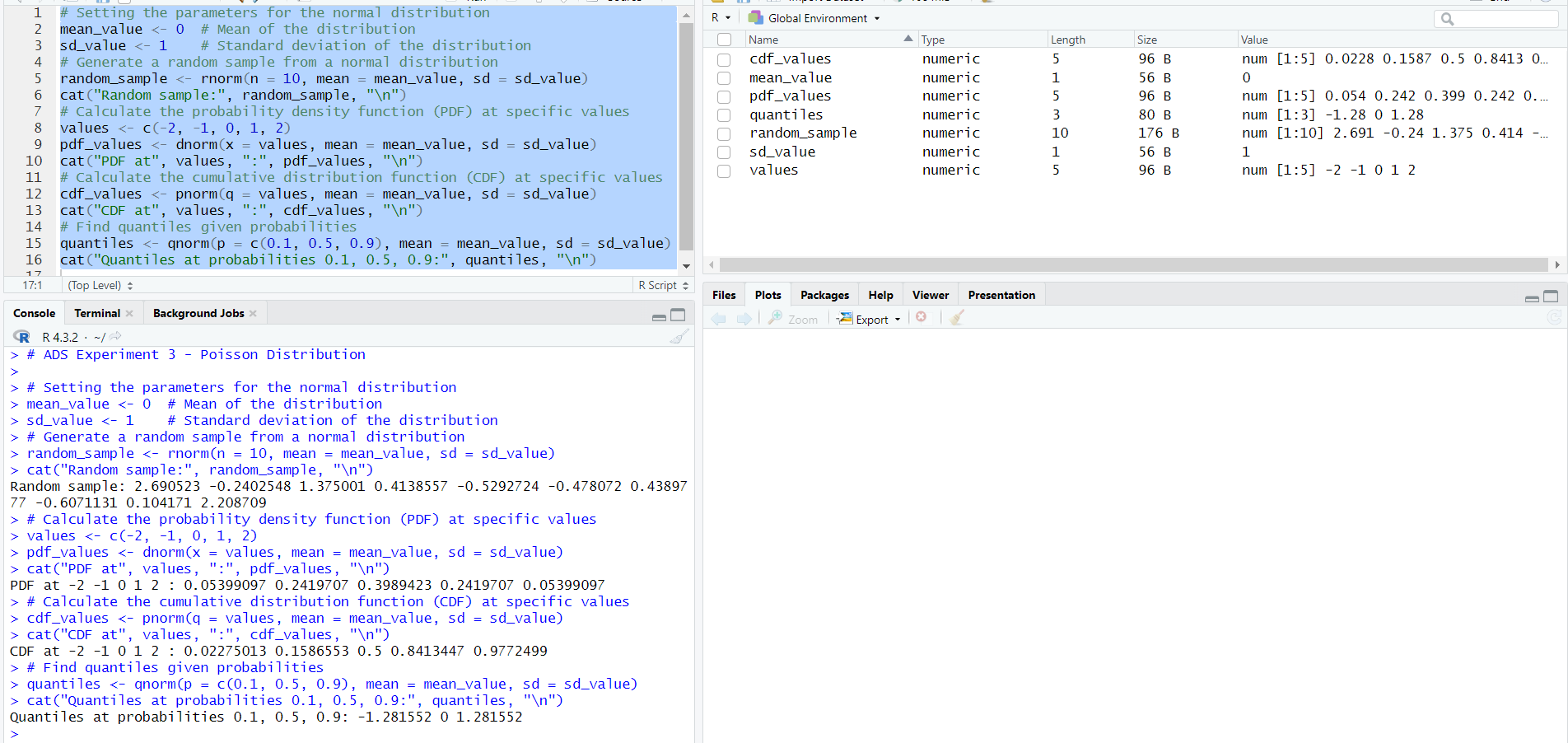
1. **Poisson Distribution:**

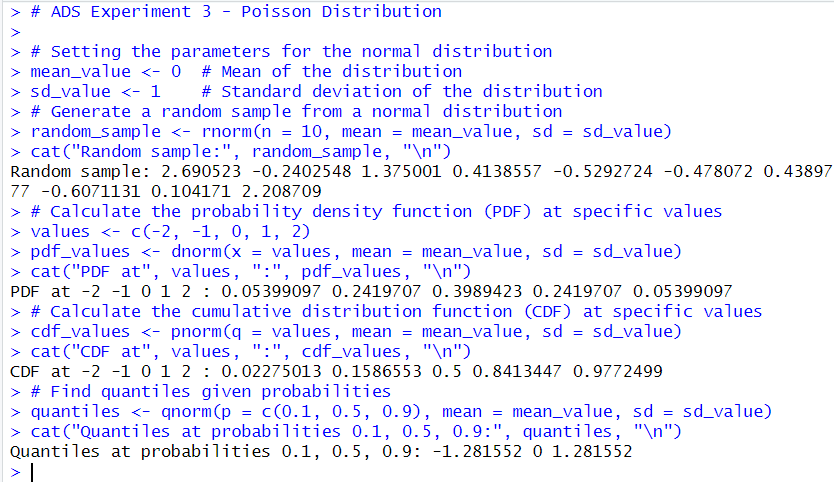
****

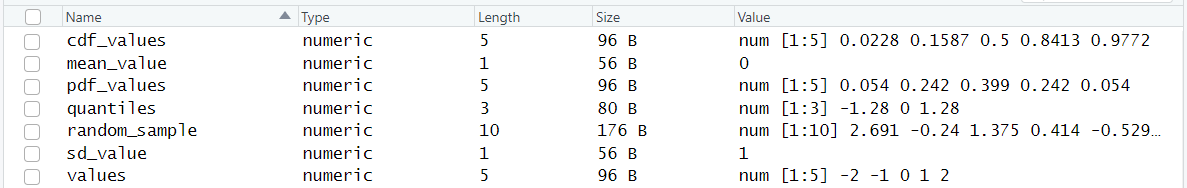
****

****

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.**
2. **Define the binomial distribution and explain the key components involved.**

The binomial distribution is a discrete probability distribution that models the number of successes (or failures) in a fixed number of independent and identical trials.

Key components:

* Number of Trials (n): The total number of independent trials or experiments.
* Probability of Success (p): The probability of success in a single trial.
* Probability of Failure (q): The complement of the probability of success, i.e., q = 1 - p.
* Random Variable (X): Represents the number of successes in n trials.
* Binomial Coefficient (n choose k): Denoted as C(n, k), represents the number of ways to choose k successes from n trials.

The probability mass function (PMF) of the binomial distribution is given by:

*P(X = k) = C(n, k) pk q(n - k)*

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

* Discrete Nature: The binomial distribution is discrete, dealing with a fixed number of discrete events or trials.
* Two Possible Outcomes: Each trial has only two possible outcomes - success or failure.
* Fixed Number of Trials: The number of trials is predetermined and remains constant.
* Independent Trials: The trials are independent of each other.

Unlike the normal distribution, which is continuous, or the Poisson distribution, which deals with the number of events in a fixed interval, the binomial distribution specifically focuses on the number of successes in a fixed number of trials.

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

* Fixed Number of Trials (n): The number of trials must be fixed in advance.
* Independent Trials: Each trial is independent of the others. The outcome of one trial does not affect the outcome of another.
* Two Possible Outcomes: Each trial has only two possible outcomes - success or failure.
* Constant Probability of Success (p): The probability of success (p) is constant for each trial. It does not change from trial to trial.

If these conditions are met, the random variable representing the number of successes in the fixed number of trials follows a binomial distribution.

1. **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.**

Example Scenario for Poisson Distribution – Call Center Arrivals

Consider a call center where phone calls from customers arrive randomly, and the average number of calls per hour is known to be 10.

Poisson Distribution:

* Average Rate (λ): The average number of calls per hour is 10.
* Time Interval (t): The time interval is one hour.
* Random Variable (X): Represents the number of calls arriving in one hour.

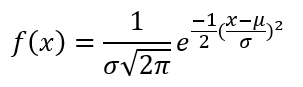
The Poisson distribution is applicable here because:

* The events (call arrivals) occur randomly and independently.
* The average rate of occurrence is known and remains constant.
* We are interested in the number of events in a fixed time interval (one hour).

The probability mass function of the Poisson distribution can be used to calculate the probability of receiving a specific number of calls in a given time period. It is suitable for modelling rare events that occur independently over time.

1. **The normal distribution is a fundamental concept in statistics and probability. Provide a comprehensive description of the normal distribution, covering the following aspects:**
2. **Define the normal distribution and explain its key characteristics.**

The normal distribution, also known as the Gaussian distribution or bell curve, is a continuous probability distribution that is symmetrical around its mean. It is fully characterized by two parameters: the mean (μ) and the standard deviation (σ). The probability density function (PDF) of the normal distribution is given by the formula:



Key characteristics:

* Symmetry: The distribution is symmetric, with the mean, median, and mode all coinciding at the center.
* Bell-Shaped Curve: The majority of data falls near the mean, and as values deviate from the mean, the frequency decreases symmetrically.
* 68-95-99.7 Rule (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.
* Mean and Standard Deviation Define the Shape: Once the mean and standard deviation are known, the entire distribution is characterized.

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

The standard normal distribution is a specific instance of the normal distribution with a mean (μ) of 0 and a standard deviation (σ) of 1. The z-score (or standard score) is a measure of how many standard deviations a particular data point is from the mean of a distribution and is calculated using the formula:

The z-score standardizes values by transforming them into a common scale. In the standard normal distribution, a z-score of 0 corresponds to the mean, negative z-scores represent values below the mean, and positive z-scores represent values above the mean. This standardization facilitates comparisons and allows us to make inferences about the relative position of a data point within a distribution.

1. **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 various real-world situations due to the central limit theorem and the prevalence of random variables with many contributing factors. Some scenarios include:

* Height of Individuals: The height of a population tends to follow a normal distribution. The combination of genetic and environmental factors contributes to this distribution.
* IQ Scores: IQ scores are designed to follow a normal distribution with a mean of 100 and a standard deviation of 15. This facilitates the comparison of an individual's intelligence relative to the population.
* Measurement Errors: Errors in measurement, such as instrument precision or human error, often follow a normal distribution. The central limit theorem supports this, as measurement errors are typically the sum of many independent, small errors.
* Test Scores: In educational testing, scores on standardized tests are often designed to follow a normal distribution. This assumption helps in interpreting and comparing individual performances.
* Financial Returns: Daily stock returns and financial data often exhibit a roughly normal distribution, making statistical analysis and risk assessment easier to perform.

The normal distribution is a suitable model for these scenarios because of the central limit theorem, which states that the sum (or average) of a large number of independent, identically distributed random variables approaches a normal distribution, regardless of the original distribution of the variables. This theorem explains why the normal distribution is pervasive in various fields and why it emerges in situations where multiple independent factors contribute to an observed outcome.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Conclusion:**

In conclusion, the implementation of probability-based statistical models, including Binomial Distribution, Poisson Distribution, and Normal/Gaussian Distribution, using R programming has provided valuable insights into data science and business analytics. This experiment has enhanced our understanding of probability modeling, enabling us to make informed decisions and extract meaningful patterns from diverse datasets.