K. J. Somaiya College of Engineering, Mumbai-77
(A Constituent College of Somaiya Vidyavihar University)
Department of Computer Engineering

Batch: h3-1 Roll No.: 16010122323

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, $4! = 4 \times 3 \times 2 \times 1$).

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
```

```
> n_trials <- 25
> prob_success <- 0.5
> random_sample <- rbinom(n = 1, size = n_trials, prob = prob_success)
> cat("Random sample:", random_sample, "\n")
Random sample: 11
>
> values <- c(2,4,6,8)
> pmf_values <- dbinom(x=values, size = n_trials,prob=prob_success)
> cat("PMF at ",values,":",pmf_values,"\n")
PMF at  2 4 6 8 : 8.940697e-06 0.0003769994 0.005277991 0.03223345
> cdf_values <- pbinom(q = values, size = n_trials, prob = prob_success)
> cat("CDF at", values, ":", cdf_values, "\n")
CDF at 2 4 6 8 : 9.715557e-06 0.0004552603 0.007316649 0.05387607
>
> quantiles <- qbinom(p = c(0.2, 0.6, 0.8), size = n_trials, prob = prob_success)
> cat("Quantiles at probabilities 0.2, 0.6, 0.8:", quantiles, "\n")
Quantiles at probabilities 0.2, 0.6, 0.8: 10 13 15
```

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.

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

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

```
> lambda <- 3
> random_sample <- rpois(n = 25, lambda = lambda)
> cat("Random sample:", random_sample, "\n")
Random sample: 2 6 1 5 4 2 3 1 4 2 3 4 3 5 4 7 2 4 1 3 5 5 4 1 0
>
> values <- c(3, 5, 7, 9)
> pmf_values <- dpois(x = values, lambda = lambda)
> cat("PMF at", values, ":", pmf_values, "\n")
PMF at 3 5 7 9 : 0.2240418 0.1008188 0.02160403 0.002700504
>
> cdf_values <- ppois(q = values, lambda = lambda)
> cat("CDF at", values, ":", cdf_values, "\n")
CDF at 3 5 7 9 : 0.6472319 0.9160821 0.9880955 0.9988975
>
> quantiles <- qpois(p = c(0.45, 0.58, 0.93), lambda = lambda)
> cat("Quantiles at probabilities 0.45, 0.58, 0.93:", quantiles, "\n")
Quantiles at probabilities 0.45, 0.58, 0.93: 3 3 6
```

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.

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

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

```
>
> mean_value <- 0
> sd_value <- 1
> random_sample <- rnorm(n = 25, mean = mean_value, sd = sd_value)
> cat("Random sample:", random_sample, "\n")
Random sample: 1.023855 -1.072636 0.9113457 -0.7678739 -0.2790318 -1.344565 0.5
 1.549624 0.6506096 -0.231638 -0.3658568 0.4835206 -0.2781789 0.03049345 -0.651
3557 1.05442 -0.1493378 -0.3144101 -0.763228 -0.6077479
>
> values <- c(-4, -2, 0, 2, 4)
> pdf_values <- dnorm(x = values, mean = mean_value, sd = sd_value)
> cat("PDF at", values, ":", pdf_values, "\n")
PDF at -4 -2 0 2 4 : 0.0001338302 0.05399097 0.3989423 0.05399097 0.0001338302
>
> cdf_values <- pnorm(q = values, mean = mean_value, sd = sd_value)
> cat("CDF at", values, ":", cdf_values, "\n")
CDF at -4 -2 0 2 4 : 3.167124e-05 0.02275013 0.5 0.9772499 0.9999683
>
> quantiles <- qnorm(p = c(0.23, 0.49, 0.86), mean = mean_value, sd = sd_value)
> cat("Quantiles at probabilities 0.23, 0.49, 0.86:", quantiles, "\n")
Quantiles at probabilities 0.23, 0.49, 0.86: -0.7388468 -0.02506891 1.080319
>
> random_sample <- rnorm(n = 5, mean = mean_value, sd = sd_value)
> cat("Random sample:", random_sample, "\n")
Random sample: -1.165099 1.841465 -0.6187034 0.2664425 -1.131697
>
```

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.

Ans:

a.

In R, the binomial distribution is a probability distribution that describes the number of successes in a fixed number of independent Bernoulli trials, each with the same probability of success. The key components involved are:

- **n**: The number of trials.
- **p**: The probability of success on each trial.
- **x**: The number of successes we are interested in.
- **dbinom(x, size = n, prob = p)**: The probability mass function (PMF) in R for the binomial distribution, which gives the probability of observing exactly **x** successes in **n** trials with probability of success **p**.

b.

In R, the binomial distribution differs from other probability distributions in that it specifically models the number of successes in a fixed number of trials with two possible outcomes (success or failure), each with the same probability of success. Other probability distributions in R, like the normal distribution or Poisson distribution, may describe continuous or discrete phenomena with varying parameters.

c.

- **Fixed Number of Trials (n)**: In R, you must have a fixed number of trials specified.
- **Independent Trials**: Each trial must be independent of the others. In R, this means that the outcomes of the trials should not be dependent on each other.
- **Two Possible Outcomes**: Each trial must have only two possible outcomes, usually coded as 0 and 1 for failure and success, respectively, in R.
- **Constant Probability of Success (p)**: The probability of success (**p**) must remain constant for each trial in R. This means that the success probability should not change across trials for the binomial distribution to hold.

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.

Ans:

1. Binomial Distribution Example: Quality Control in Manufacturing

Let's consider a manufacturing plant producing light bulbs. Each light bulb produced can be classified as either defective or non-defective. Suppose the probability of producing a defective light bulb is 0.05, and we are interested in the number of defective bulbs in a sample of 100 bulbs.

Explanation:

- **Fixed Number of Trials**: We have a fixed number of trials (100 light bulbs).
- **Two Possible Outcomes**: Each light bulb can either be defective or non-defective.
- **Constant Probability of Success**: The probability of producing a defective light bulb remains constant at 0.05 for each trial.

The binomial distribution is appropriate because we are interested in the number of successes (defective bulbs) out of a fixed number of independent trials (light bulbs produced), each with the same probability of success (probability of producing a defective bulb).

2. **Poisson Distribution Example: Arrival of Customers at a Store**

Consider a retail store and the number of customers arriving at the store within a given time period, say one hour. The average arrival rate of customers is known to be 10 customers per hour, and we want to model the number of customers arriving in the next hour.

Explanation:

- **Rare Events**: The arrival of customers at the store can be considered a rare event within a specific time interval.
- **Count of Events**: We are interested in counting the number of events (customer arrivals) within a fixed time period (one hour).
- **Independence**: The arrival of one customer does not affect the arrival of another customer within the hour.

The Poisson distribution is appropriate because it models the number of events occurring within a fixed interval of time when the events are rare and occur independently of the time since the last event. In this scenario, the average rate of arrivals (10 customers per hour) helps determine the parameter for the Poisson distribution.

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.

Ans:

a. The normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is symmetric around its mean. It is characterized by its bell-shaped curve, where the majority of the data falls near the mean, with fewer observations toward the tails.

Key characteristics of the normal distribution include:

- Symmetry: The normal distribution is symmetric around its mean, with equal areas under the curve on both sides of the mean.
- Unimodality: It has a single peak at the mean. - Mean, Median, and Mode Equality: In a normal distribution, the mean, median, and mode are equal, which means the distribution is perfectly symmetrical.

b. The standard normal distribution has a mean of 0 and a standard deviation of 1. The z-score measures the number of standard deviations a data point is from the mean, facilitating comparison and standardization across different normal distributions.
c. The normal distribution is commonly observed in phenomena like human height and weight, IQ scores, and measurement errors due to the central limit theorem, which ensures that the distribution of the sum or average of many independent variables tends toward a normal distribution, making it suitable for statistical analysis and inference.

Conclusion:

In conclusion, probability-based statistical modeling provides a powerful approach for analyzing data, making informed decisions, and understanding uncertainties. It enables accurate predictions, risk quantification, and informed strategies across diverse domains.