MA677 MidtermProject

Yueqi(Charlene) Jin

2024-03-27

Based on the requirement of Prof. Haviland, I've selected what I believe to be the three most engaging problems out of the five presented topics: Order Statistics, Markov Chains and Irrigation Circles Problem. And Prof. Haviland once remarked, "Curiosity is the leavening of education. So, be curious and rise to the occasion!" This sentiment deeply resonates with me. I agreed so strongly with his opinion! Let me discuss the three most interesting problems as follows!

1.Order statistics

Order statistics play a crucial role in statistical analysis and probability theory, providing insights into the behavior of samples from a population. Here, we'll delve into the concept of order statistics, outline the steps involved in analyzing them, introduce key formulas, and demonstrate how to derive the distributions of order statistics for uniform, exponential, and normal distributions.

What are Order Statistics?

Order statistics are the statistics obtained by arranging a sample of observations in ascending order. The k-th order statistic of a sample size n is the k-th smallest value in the sample. For instance, the first order statistic ($k = 1$) is the minimum value, while the n-th order statistic is the maximum value in the sample. These statistics help in understanding the distribution and range of the data.

Key Formulas and Properties

The probability density function (PDF) of the k-th order statistic Y_k from a sample of size n with continuous PDF $f(x)$ and cumulative distribution function (CDF) $F(x)$ is given by:

$$f_{Y_k}(y) = \frac{n!}{(k-1)!(n-k)!}[F(y)]^{k-1}[1-F(y)]^{n-k}f(y)$$

This formula highlights the relationship between the distribution of the k-th order statistic, the original distribution of the sample, and the binomial coefficients that account for the positions of order statistics within the sample.

Derivation of Order Statistics for Specific Distributions

Uniform Distribution Overview

The Uniform Distribution is defined over an interval `[a, b]` and is characterized by having a constant probability density function (PDF) and a linear cumulative distribution function (CDF) within this interval. Specifically, the PDF and CDF of a Uniform Distribution are defined as follows:

- **PDF**: $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$
- **CDF**: $F(x) = \frac{x-a}{b-a}$ for $a \leq x \leq b$

The probability density function (PDF) of the k-th order statistic from a sample of size n is given by a general formula involving the original distribution's PDF and CDF. For the Uniform Distribution, this formula becomes:

$$f_{Y_k}(y) = \frac{n!}{(k-1)!(n-k)!} \left(\frac{y-a}{b-a}\right)^{k-1} \left(1 - \frac{y-a}{b-a}\right)^{n-k} \frac{1}{b-a}$$

Upon simplification, the PDF of the k-th order statistic for a Uniform Distribution over `[a, b]` is:

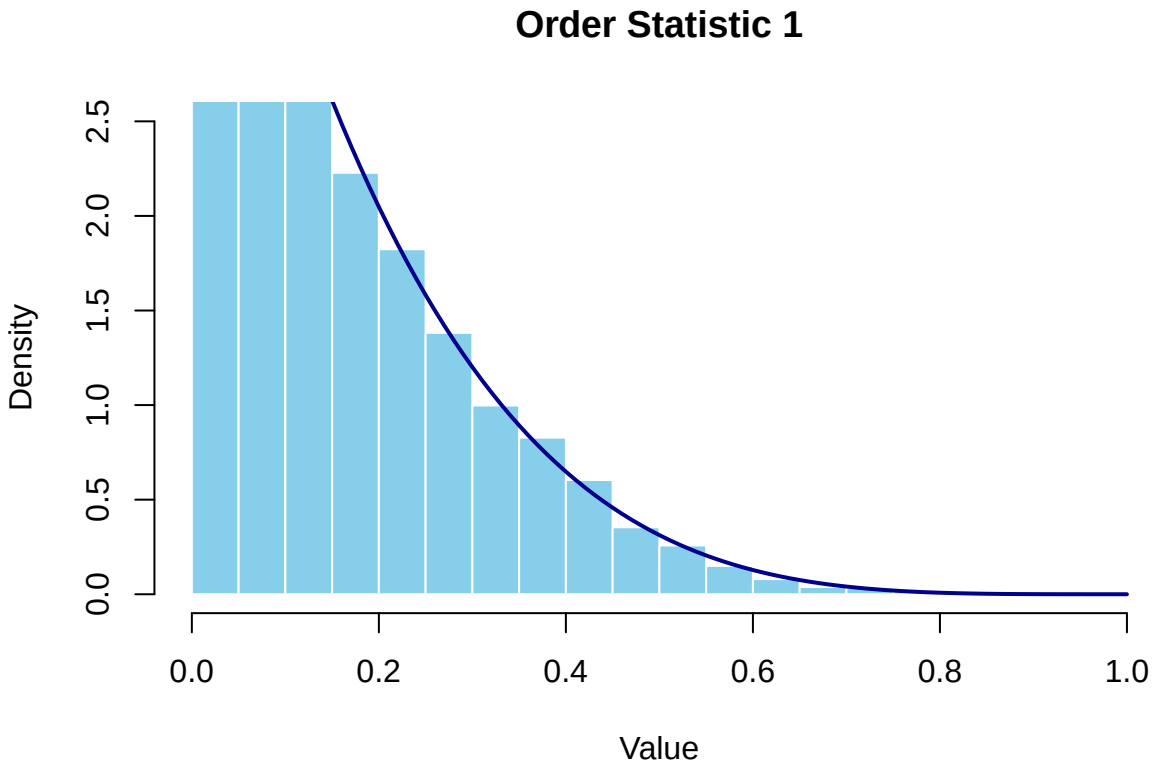
$$f_{Y_k}(y) = \frac{n!}{(k-1)!(n-k)!} \frac{(y-a)^{k-1}(b-y)^{n-k}}{(b-a)^n}$$

This formula indicates how the distribution of the k-th order statistic is influenced by the sample size n, the rank k, and the boundaries of the original Uniform Distribution a and b.

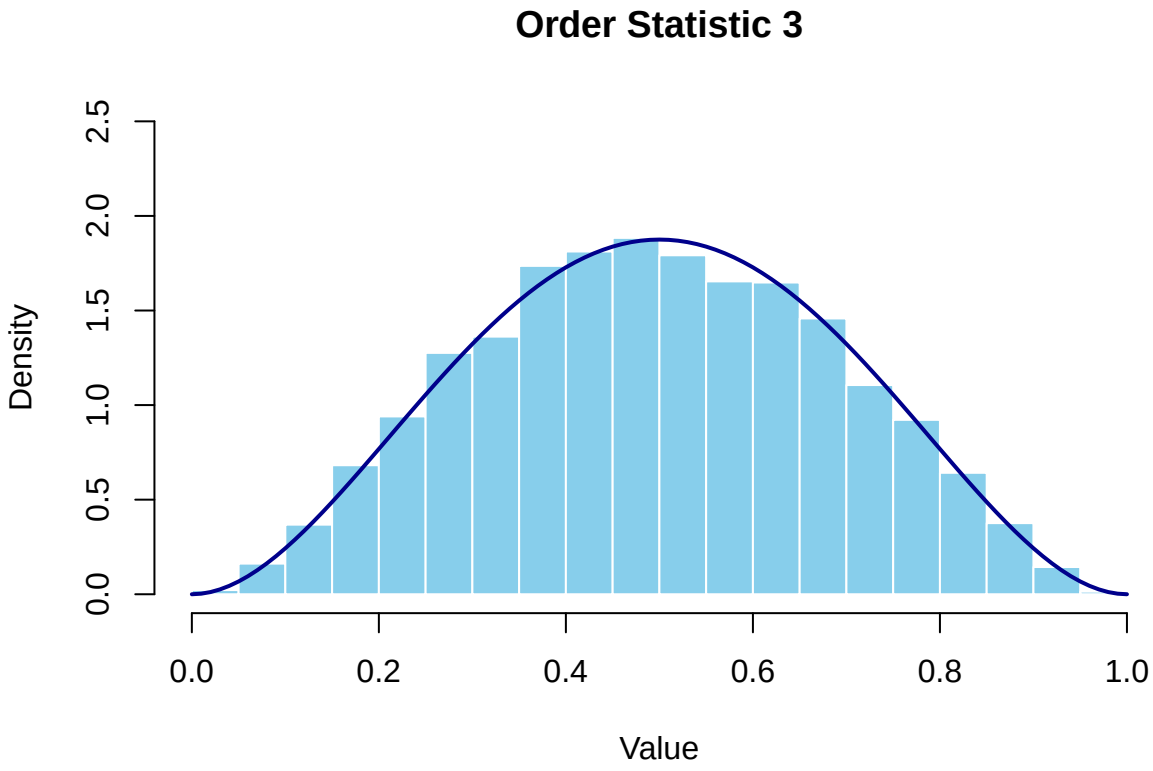
Simulation and Plot of Uniform Distribution

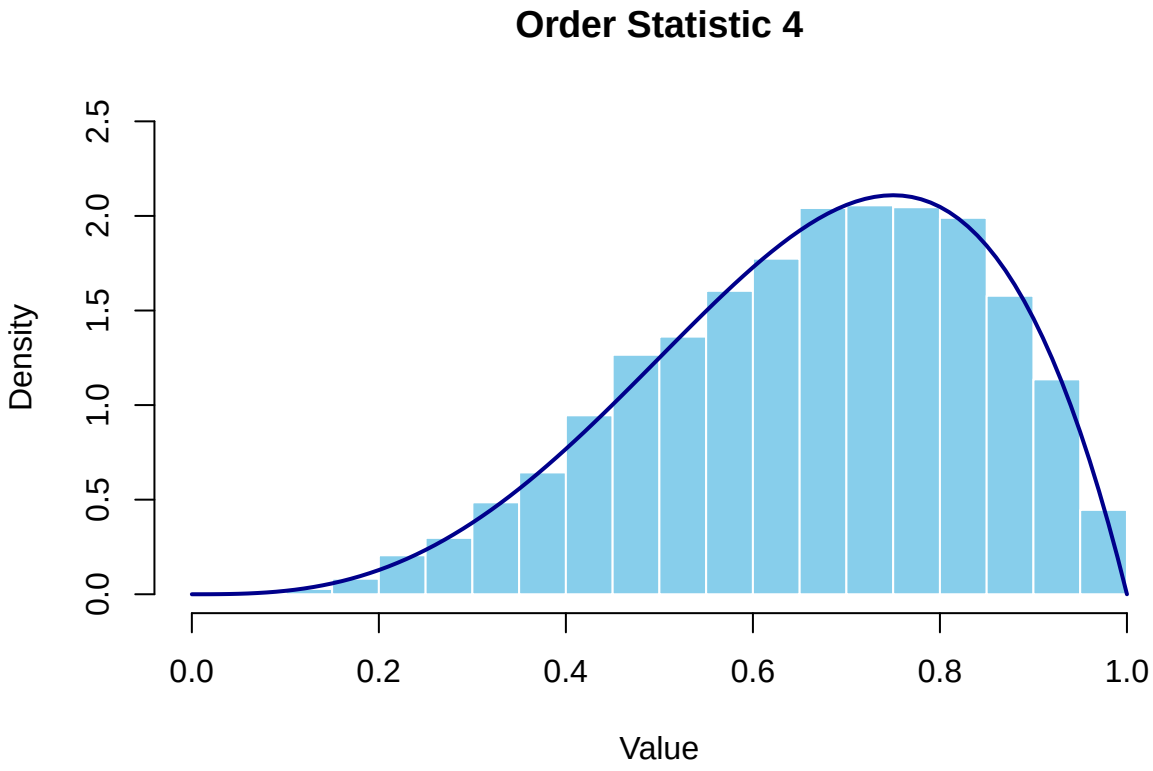
Results from Plot of Uniform Distribution as follows:

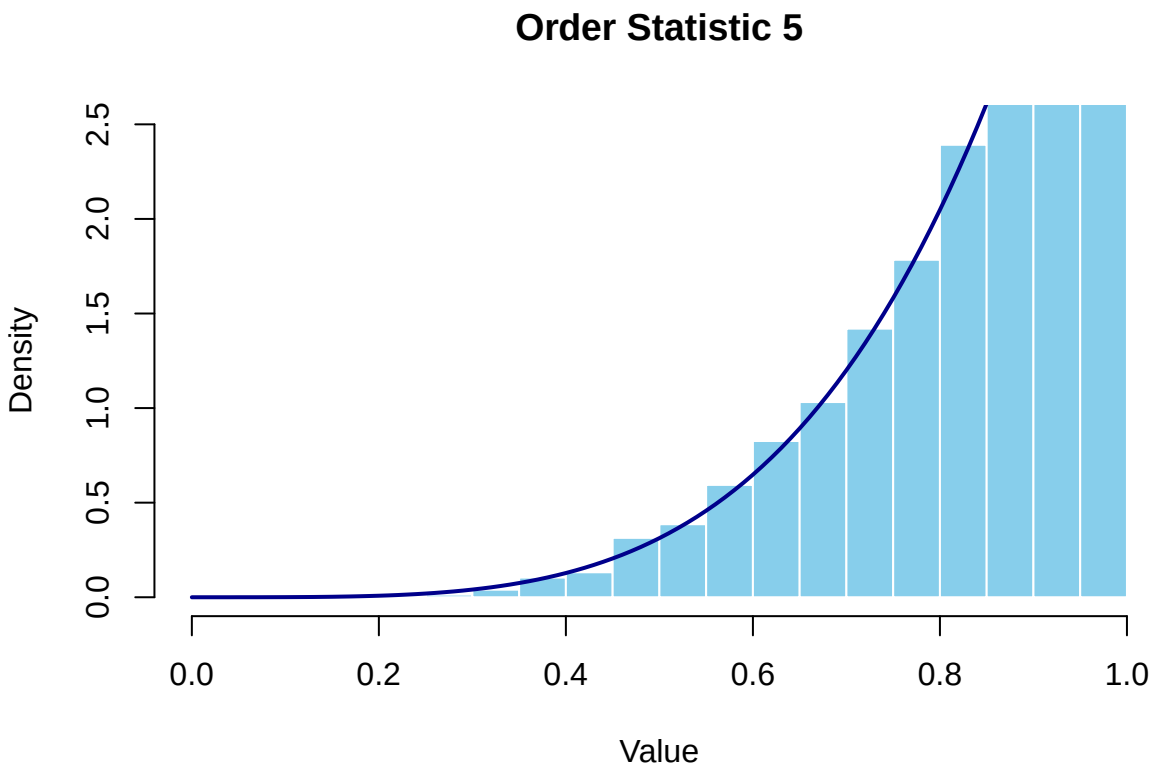
- **First Order Statistic (Minimum):** The distribution of the minimum values would skew towards the lower end of the interval, meaning that in a uniformly distributed sample, the minimum values are more likely to be close to 0.
- **Second and Fourth Order Statistics:** These intermediate order statistics would exhibit distributions that start to show a more bell-shaped pattern. Their distributions are Beta distributions with parameters that move them away from the edges of the interval towards the middle. For a uniform distribution, these are Beta(2, n-1) and Beta(n-1, 2), respectively, which are symmetric for the second and fourth order statistics.
- **Third Order Statistic (Median):** For the median in samples of size five, the distribution would appear fairly uniform across the interval, but with a slight tendency towards the center. This is consistent with the Beta(3, 3) distribution, which is more evenly spread across the interval than the extreme order statistics.
- **Fifth Order Statistic (Maximum):** The distribution of the maximum values would be skewed towards the upper end of the interval, indicating that the maximum values are more likely to be close to 1.


Order Statistic 1
Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
Value


Order Statistic 2
Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
Value


Order Statistic 3
Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
Value


Order Statistic 4
Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
Value



Exponential Distribution Overview

The Exponential Distribution is widely used in the field of statistics to model the time between events in a Poisson process, where events occur continuously and independently at a constant average rate. It is characterized by the rate parameter λ, which is the reciprocal of the mean. The PDF and CDF of an Exponential Distribution are given by:

- **PDF**: $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$
- **CDF**: $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$

For the Exponential Distribution, the derivation of the PDF of the k-th order statistic involves integrating the joint probability density of the k-th smallest value in a sample of size n. Substituting the Exponential Distribution's PDF and CDF into the general formula for the k-th order statistic's PDF yields:

$$f_{Y_k}(y) = \frac{n!}{(k-1)!(n-k)!} \lambda e^{-\lambda y} \left(1 - e^{-\lambda y}\right)^{k-1} \left(e^{-\lambda y}\right)^{n-k}$$

Simplifying this, we obtain:

$$f_{Y_k}(y) = \frac{n!}{(k-1)!(n-k)!} \lambda^k e^{-\lambda y} (1 - e^{-\lambda y})^{k-1} e^{-\lambda(n-k)y}$$

Further simplification gives:

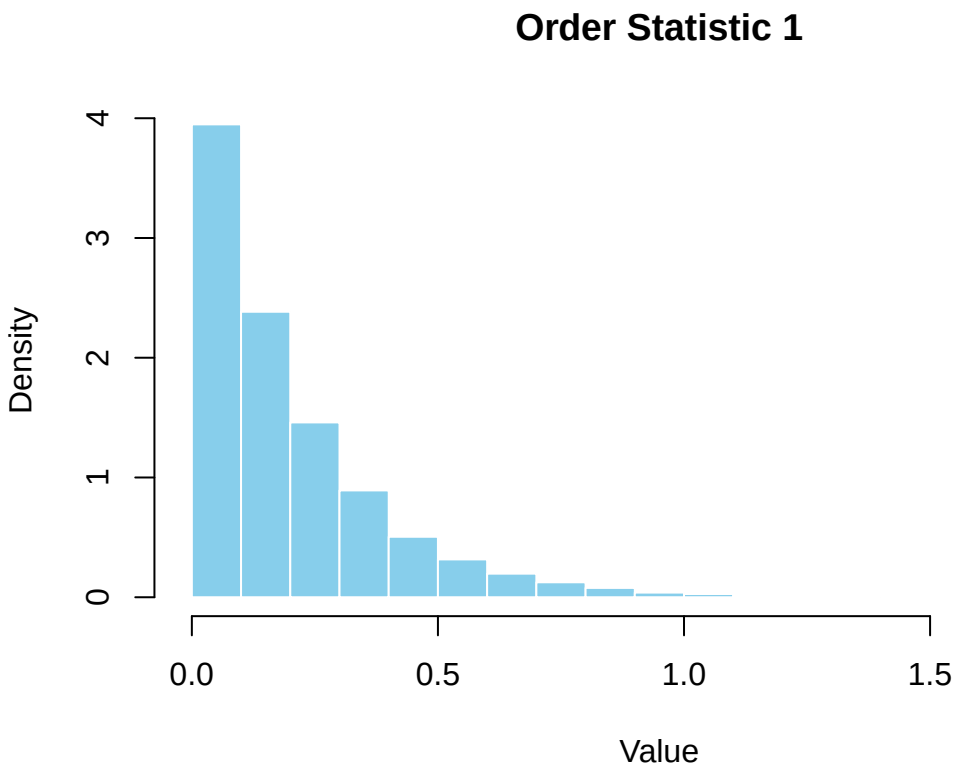
$$f_{Y_k}(y) = \frac{\lambda(\lambda y)^{k-1} e^{-\lambda n y}}{(k-1)!} \text{ for } y \geq 0$$

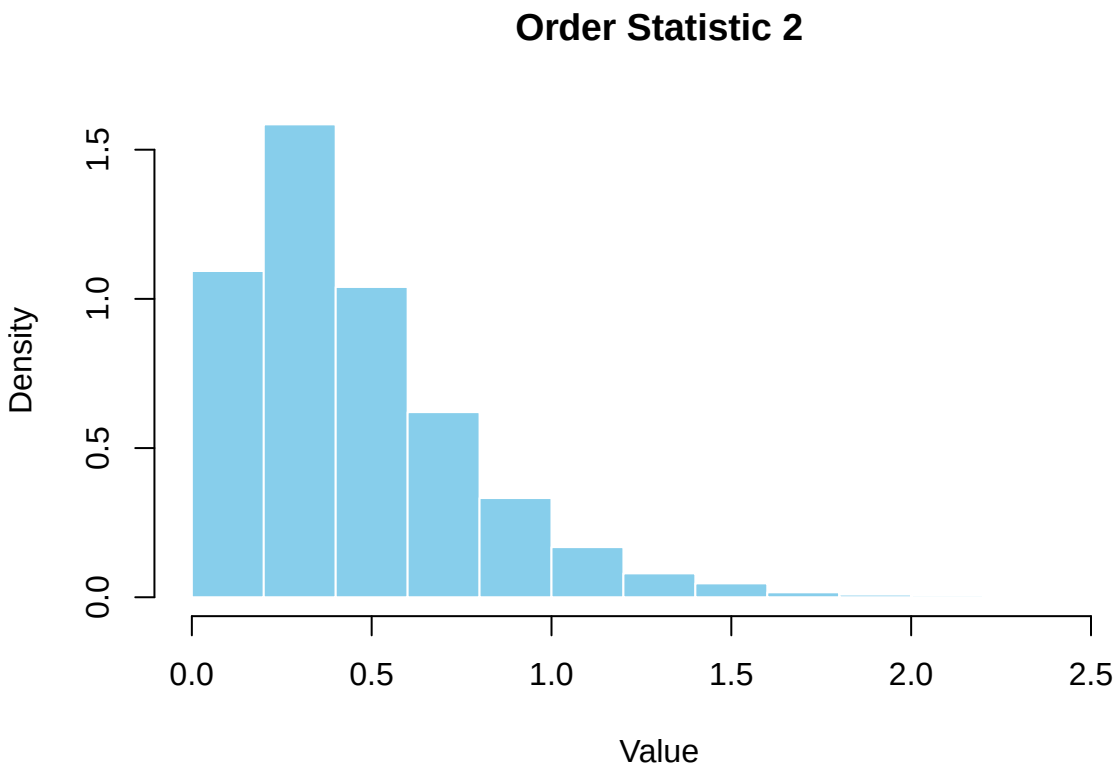
This formula shows how the PDF of the k-th order statistic from an Exponential Distribution depends on the rate parameter λ, the sample size n, and the order k.

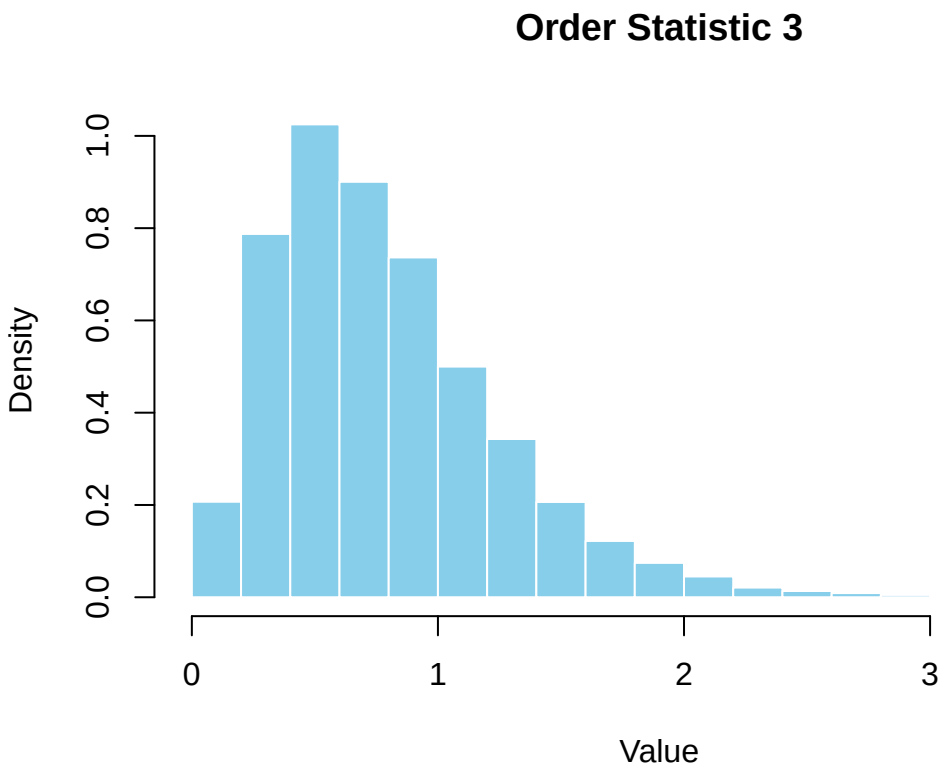
Simulation and Plot of Exponential Distribution

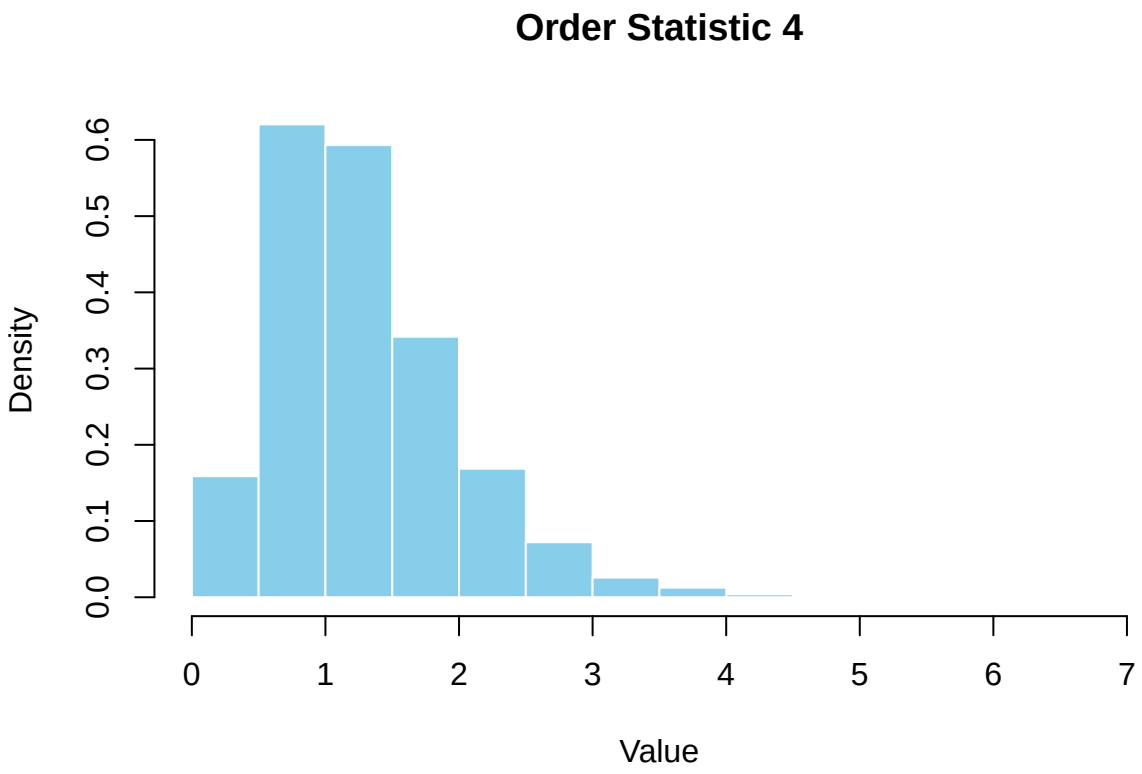
Results from Plot of Exponential Distribution as follows:

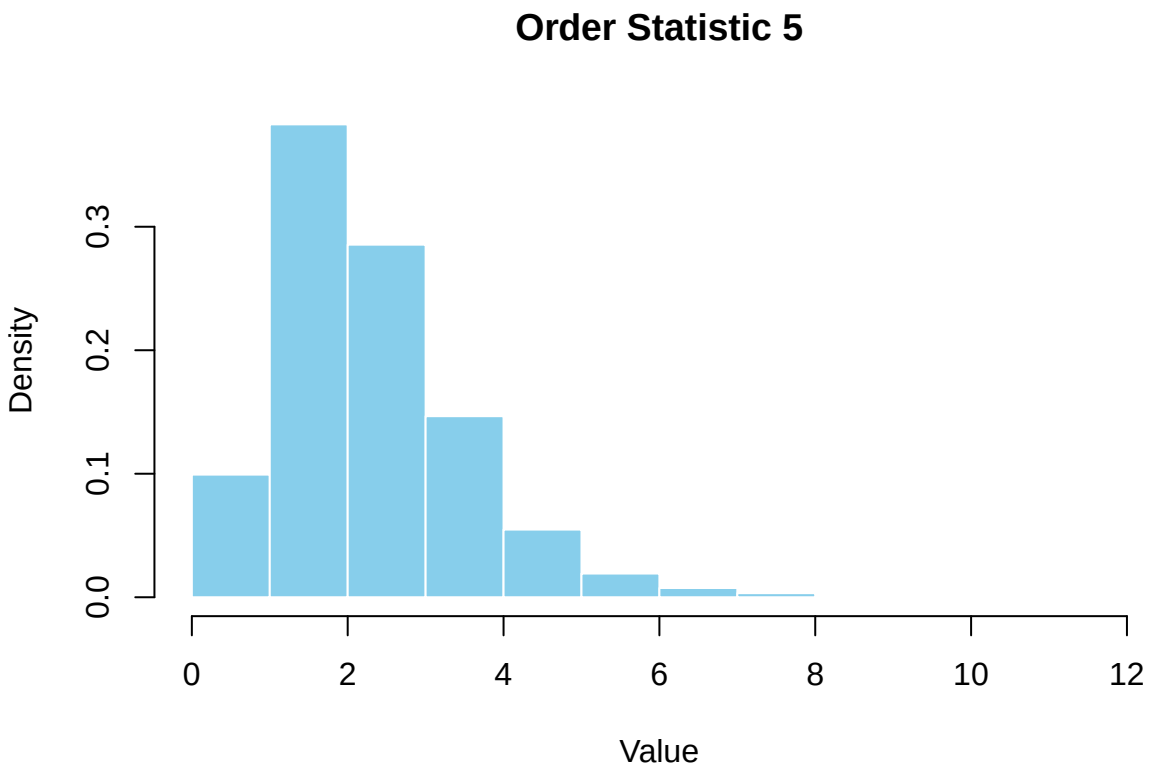
- **Shifting Distributions:** Unlike the Uniform distribution, histograms of order statistics from an Exponential distribution will exhibit different shapes. This reflects the nature of the Exponential distribution, where smaller values are more common, and larger values become increasingly rare.
- **Right-Skewness:** Since the Exponential distribution is right-skewed, the histograms of its order statistics typically show a right-skewed distribution as well. This means that as you move to higher order statistics (from the 1st to the 5th in your case), the distribution shifts towards larger values.
- **Increasing Variance:** The variance of the order statistics generally increases with the order. The first order statistic (the minimum) tends to be closely clustered near the lower end (close to zero for the Exponential distribution), while higher order statistics (like the 5th in a sample of 5) have a wider spread, indicating a higher variance.
- **Absence of Theoretical Overlay:** In your histograms, the absence of a theoretical density curve overlay (which was possible with the Beta distribution for the Uniform distribution case) means that the comparison with the theoretical distribution is not readily visible. For the Exponential distribution, deriving the theoretical distribution of order statistics is more complex and not as straightforward as the Beta distribution for the Uniform case.


Order Statistic 1
Density
0
1
2
3
4
0.0
0.5
1.0
1.5
Value


Order Statistic 2
Density
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
2.0
2.5
Value


Order Statistic 3
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
Value


Order Statistic 4
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
2
3
4
5
6
7
Value



Normal Distribution Overview

The Normal Distribution, also known as the Gaussian distribution, is one of the most important probability distributions in statistics, used to model a wide range of phenomena. It is characterized by two parameters: the mean (μ) and the standard deviation (σ), which determine the distribution's center and width, respectively. The PDF of a Normal Distribution is given by:

- **PDF**: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ for $-\infty < x < \infty$

Deriving the PDF for the order statistics of a Normal Distribution directly is challenging due to the complexity of the Normal Distribution's PDF. Instead, properties of order statistics for the Normal Distribution are often studied through numerical methods or special cases.

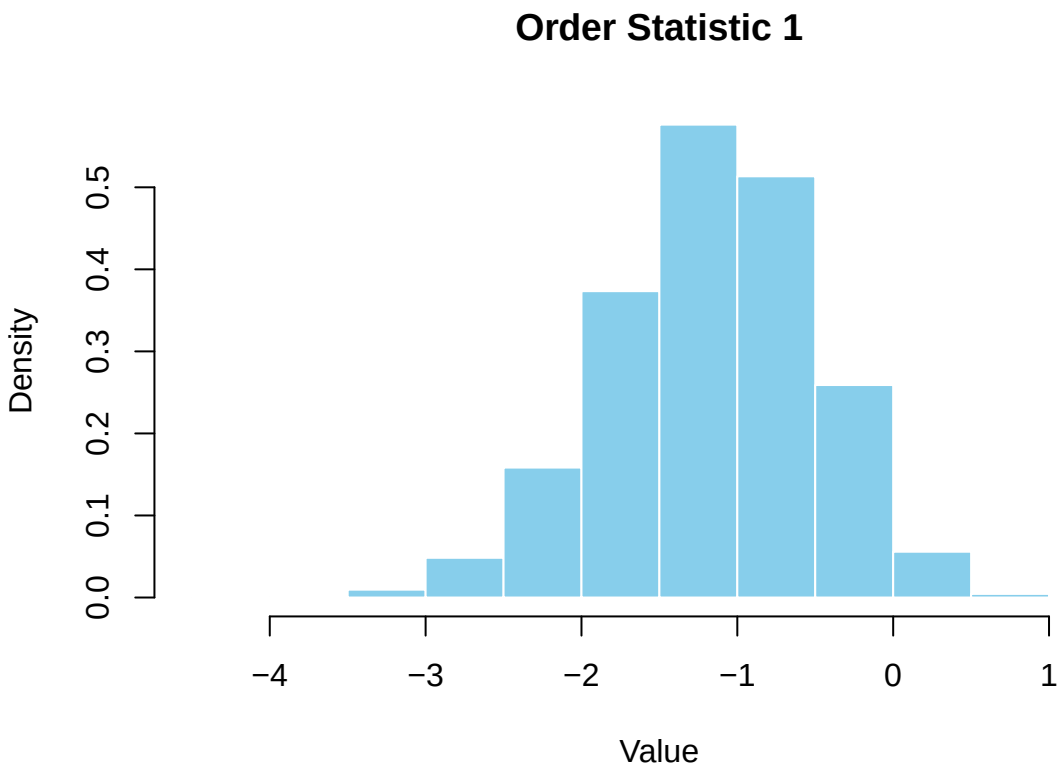
One important property is that the order statistics of a Normal Distribution are not normally distributed themselves, except for specific cases like the maximum or minimum of a very small sample size. Instead, their distribution tends to be more complicated and usually requires numerical methods to analyze fully.

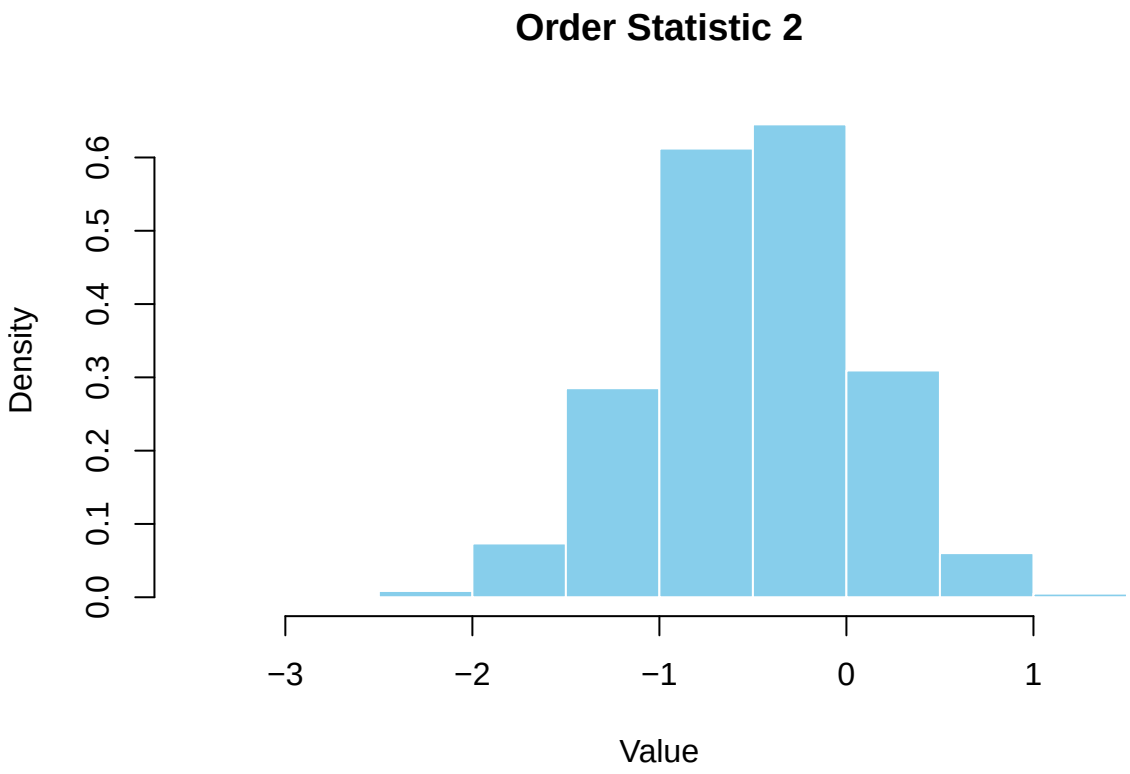
Simulation and Plot of Normal Distribution

Results from Plot of Normal Distribution as follows:

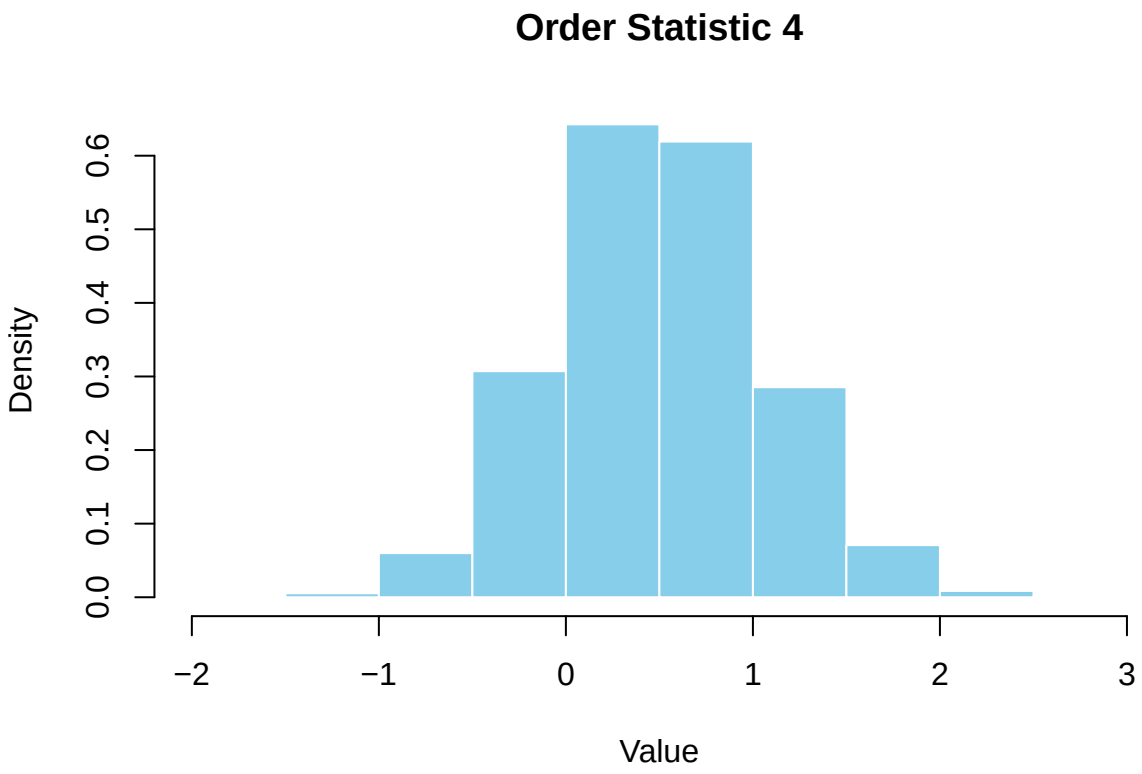
- **Symmetrical Distribution:** Unlike the Exponential distribution, the histograms for order statistics from a Normal distribution are likely to be more symmetrical, especially if the sample size is large.

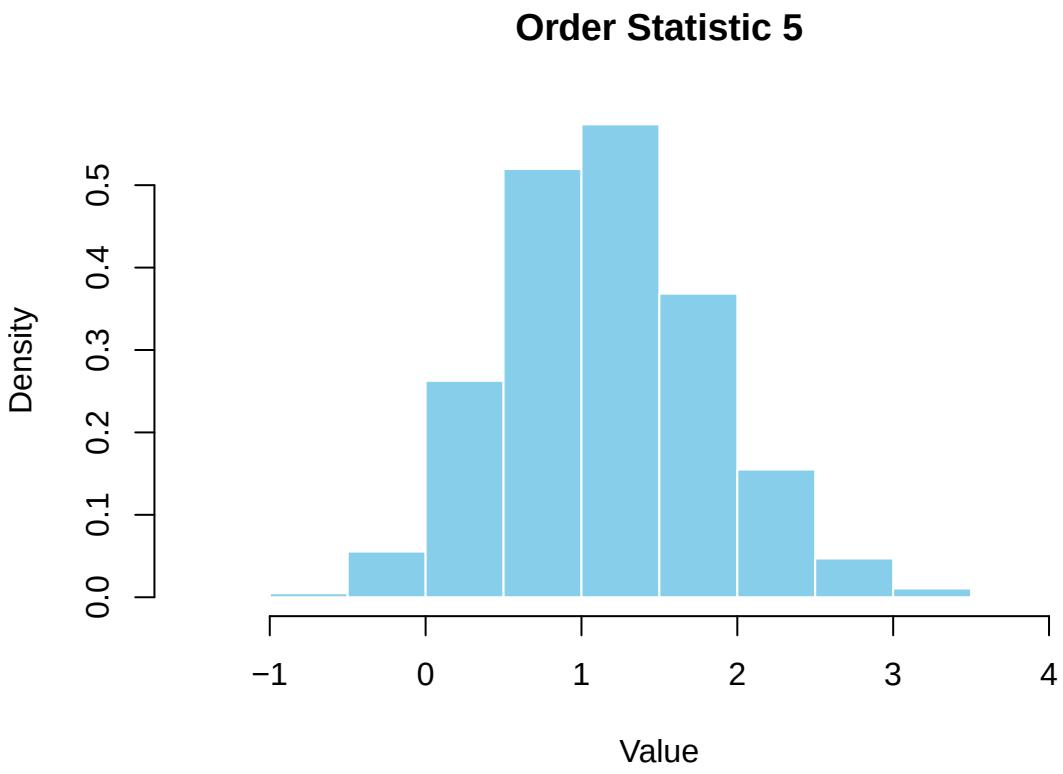
- **Central Tendency:** As the order increases, the central value of the histograms will move away from the mean, but they will generally maintain symmetry.
- **Variability:** The minimum and maximum values (first and last order statistics) tend to have the greatest spread, while the median (or middle order statistic) will have less spread.




Order Statistic 2
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−3
−2
−1
0
1
Value


Order Statistic 3
Density
0.0
0.2
0.4
0.6
−2
−1
0
1
2
Value


Order Statistic 4
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−2
−1
0
1
2
3
Value



2.Markov Chain

Markov Chain is mathematical system of transitions from one state to another on a state space, which is used to model stochastic systems that obey a defined set of rules in the current state. In the field of genetics, Markov chain is able to model sequences of alleles (different forms of genes) that change over generations as a means of predicting, under certain assumptions, the genetic makeup of future generations.

Application in Genetics using Markov Chain

Consider modeling the evolution of genomic sequences over time under the influence of both mutation and natural selection.This application I select will incorporate a more detailed scenario where the probabilities of mutation between nucleotides **(A, C, G, T)** depend on the current state of a sequence and its fitness landscape, which in turn influences the selection process.

- **A to A:** 0.9, A to C: 0.03, A to G: 0.05, A to T: 0.02
- **C to A:** 0.04, C to C: 0.85, C to G: 0.05, C to T: 0.06
- **G to A:** 0.01, G to C: 0.03, G to G: 0.94, G to T: 0.02
- **T to A:** 0.02, T to C: 0.07, T to G: 0.01, T to T: 0.9

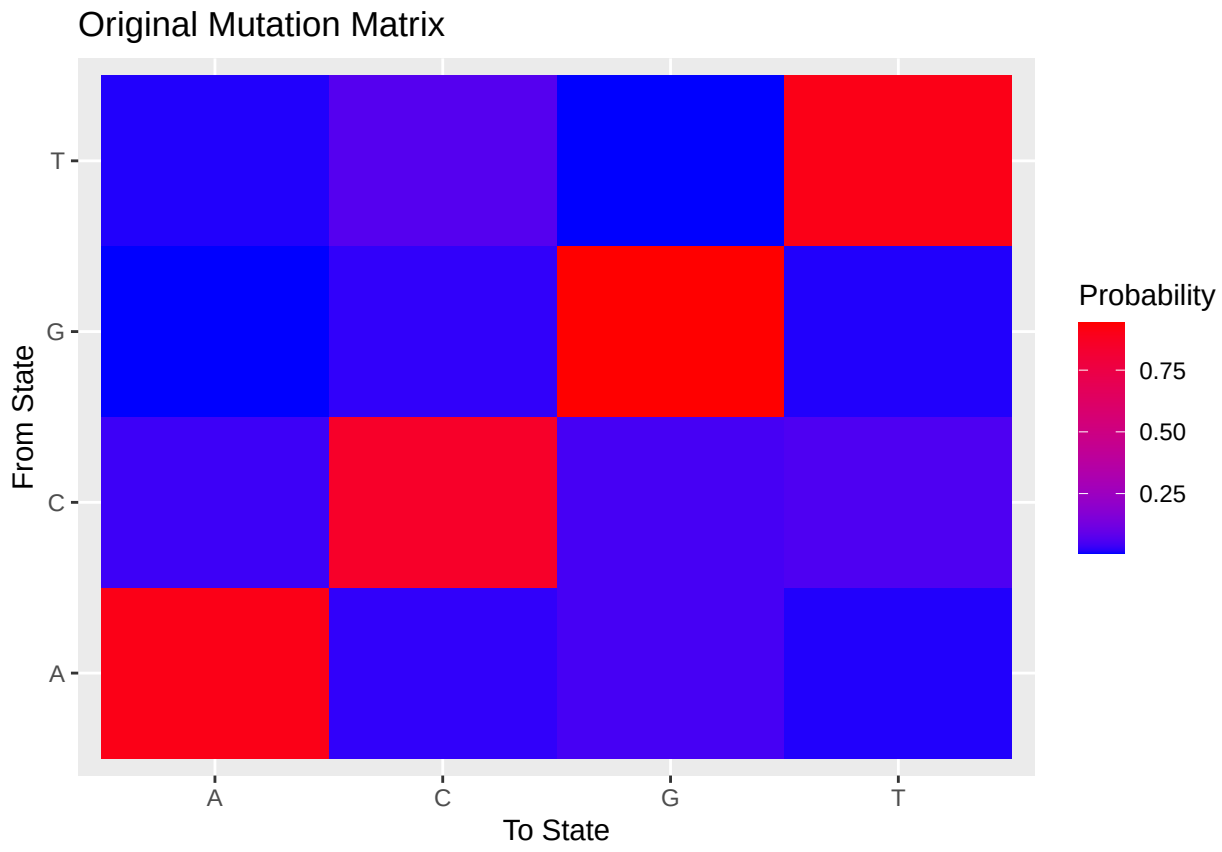
Each sequence has a fitness value. Sequences with higher fitness are more likely to be passed on to the next generation. For simplicity, assign a fitness value to each nucleotide, e.g., A: 1.0, C: 1.2, G: 0.

The plot is a good visual representation of the Markov chain simulation of nucleotide changes over time, with colored blocks indicating the state of the nucleotide at each generation.

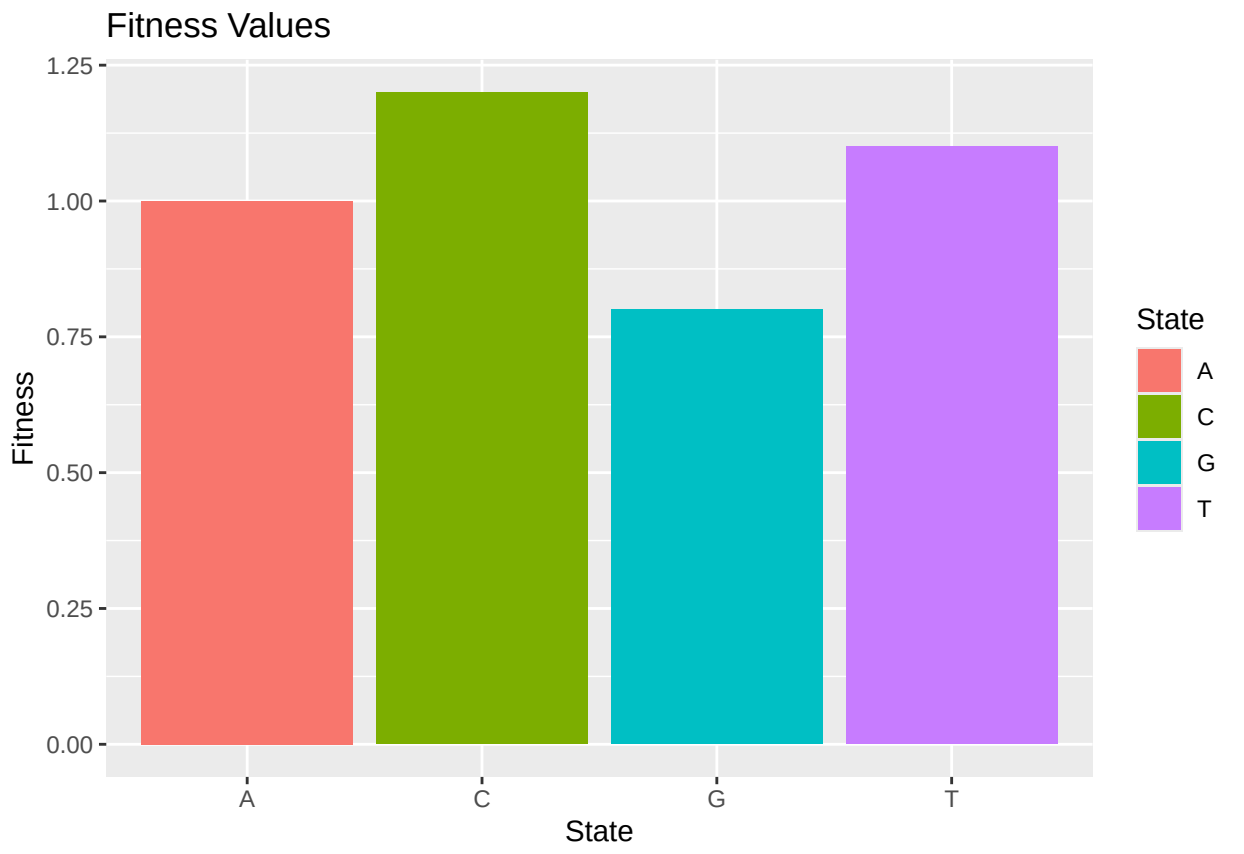
Results

- **Original Mutation Matrix Heatmap:** This plot shows the probabilities of mutation from one nucleotide to another without any adjustments for fitness. The color intensity represents the probability, with red indicating higher and blue indicating lower probabilities. For instance, the probability of "A" staying the same ("A" to "A") is very high (shown by the dark red color), whereas mutations from "A" to "G" or "T" are less likely (shown by the lighter blue color).
- **Adjusted Mutation Matrix Heatmap:** After adjusting for fitness values, the mutation probabilities change.You may notice that the tiles representing transitions to "C" might be darker compared to the original matrix, as "C" has the highest fitness value (1.2), making these transitions more probable after the adjustment. These adjustments are meant to simulate a more realistic scenario where nucleotides with higher fitness are favored.
- **Fitness Values Bar Chart:** This chart displays the fitness values for each nucleotide. The tallest bar represents "C" with the highest fitness value (1.2), followed by "T" (1.1), "A" (1.0), and the shortest bar represents "G" with the lowest fitness value (0.8). These fitness values affect the transition probabilities in the adjusted mutation matrix.

The results suggest that the fitness values have been successfully factored into the mutation probabilities.




Adjusted Mutation Matrix
From State
T
G
C
A
A
C
G
T
To State
Probability
0.75
0.50
0.25



3.Irrigation Circles Problem

Statistical Method

- **Calculate the Mean Speed** The average speed is calculated by dividing the circumference of the irrigation circle by the individual rotation times to give a speed for each rotation time. The average of these speeds provides a center value that is indicative of the overall speed at which the rotating arm is moving around the pivot axis.
- **Standard Deviation and Standard Error** The standard deviation measures the variation in these calculated velocities, providing insight into the extent to which the velocities differ from the average velocity. The standard error of the mean (SEM) is then calculated by dividing the standard deviation by the square root of the number of velocities, thus providing a measure of average velocity accuracy as an estimate of the true average velocity of the rotating arm.
- **90% Confidence Interval for the Mean Speed** Using a t-distribution appropriate for the size of the sample, determine the 90% confidence interval for the average velocity. This interval represents the range within which the actual average speed of the rotating arm is expected to be 90% certain. The calculation takes into account the variability in speed and sample size and provides a statistical estimate of the possible deviation from the observed mean.
- **Conversion from Rotation Time to Speed** The conversion process involves the direct calculation of velocities based on recorded rotation times and known circumferences of irrigation circles, including arm lengths and end-gun extension lengths. The method focuses directly on velocities and eliminates the need to calculate rotation time confidence intervals, thus simplifying the process and providing

practically relevant information directly to farmers and data scientists. The calculated average velocities and their confidence intervals provide valuable insights into the efficiency and performance of irrigation systems.

So, here are the results for Irrigation Circles Problem after running the R coding:

- **Mean Speed:** 62.3546 feet per hour
- **Standard Deviation:** 3.307321 feet per hour
- **90% Confidence Interval:** [61.40931 , 63.2999] feet per hour