MA677 MidtermProject

Yueqi(Charlene) Jin

2024-03-27

Based on the requirement of Prof. Haviland, I choose to answer the most interesting three problems from five topics: Irrigation Circles Problem, Order Statistics and Markov Chains, which resonate with me. And Prof.Haviland ever said:"Curiosity is the leavening of education. So, be curious and rise to the occasion!". I agreed with his opinion so much!

1.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

2.Order statistics

Finding the k-th smallest or largest element in a given array.Here is following steps:

1. **Choose a Pivot**: An element is randomly selected from the array to serve as the pivot. The choice of a good pivot is crucial for achieving an average time complexity of $O(n)$.
2. **Partition**: The array is rearranged such that all elements less than the pivot precede it, while all greater elements follow it. This step places the pivot in its correct, sorted position.
3. **Recursive Selection**: With the pivot in place:
   - If the pivot is the k-th smallest element, the search concludes.
   - If the pivot's position is higher than k, the search continues for the k-th smallest element in the left sub-array.
   - If the pivot's position is lower than k, the search proceeds in the right sub-array for the k-th smallest element, adjusting k to account for the excluded left portion of the array.
4. **Termination**: The algorithm terminates when a sub-array of one element is reached, identifying the sought-after Order Statistic.

Uniform Distribution

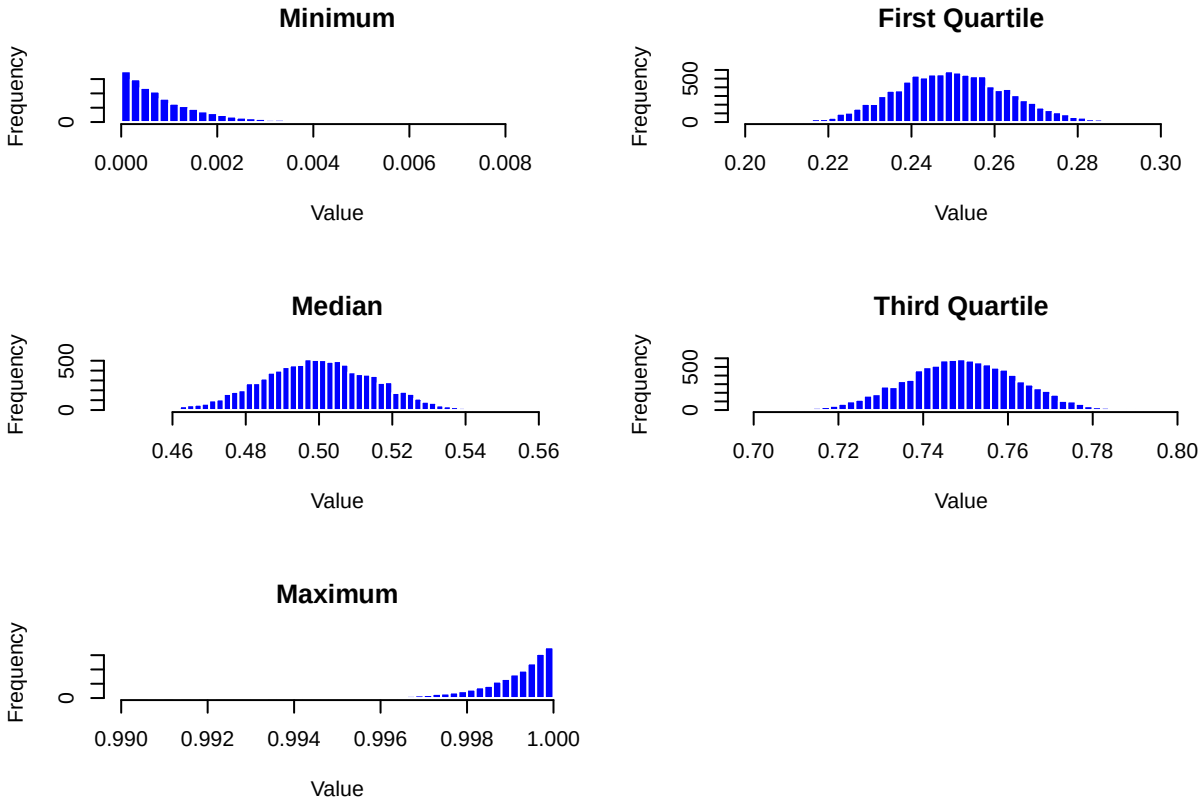
When consider the uniform distribution, denoted as $U(a, b)$. The PDF and CDF can be described as follows:

1. **Probability Density Function (PDF)**: The PDF describes the likelihood that a random variable falls within a specific range of values. For a uniform distribution $U(a, b)$, the PDF is expressed as: $f(x) = \frac{1}{b-a}$, for $a \leq x \leq b$
2. **Cumulative Distribution Function (CDF)**: The CDF calculates the probability that a random variable X is less than or equal to a certain value x. For $U(a, b)$, the CDF is expressed as: $F(x) = \frac{x-a}{b-a}$, for $a \leq x \leq b$

For the k-th order statistic $X_{(k)}$ derived from a sample of size n, the Probability Density Function (PDF) is expressed as: $f_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1}[1 - F(x)]^{n-k}f(x)$

When applying the specific CDF and PDF of the uniform distribution $U(a, b)$ to this formula, the PDF of $X_{(k)}$ is refined to:$f_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!}\left(\frac{x-a}{b-a}\right)^{k-1}\left(1 - \frac{x-a}{b-a}\right)^{n-k}\frac{1}{b-a}$

Simulation and Plot of Uniform Distribution



Exponential Distribution

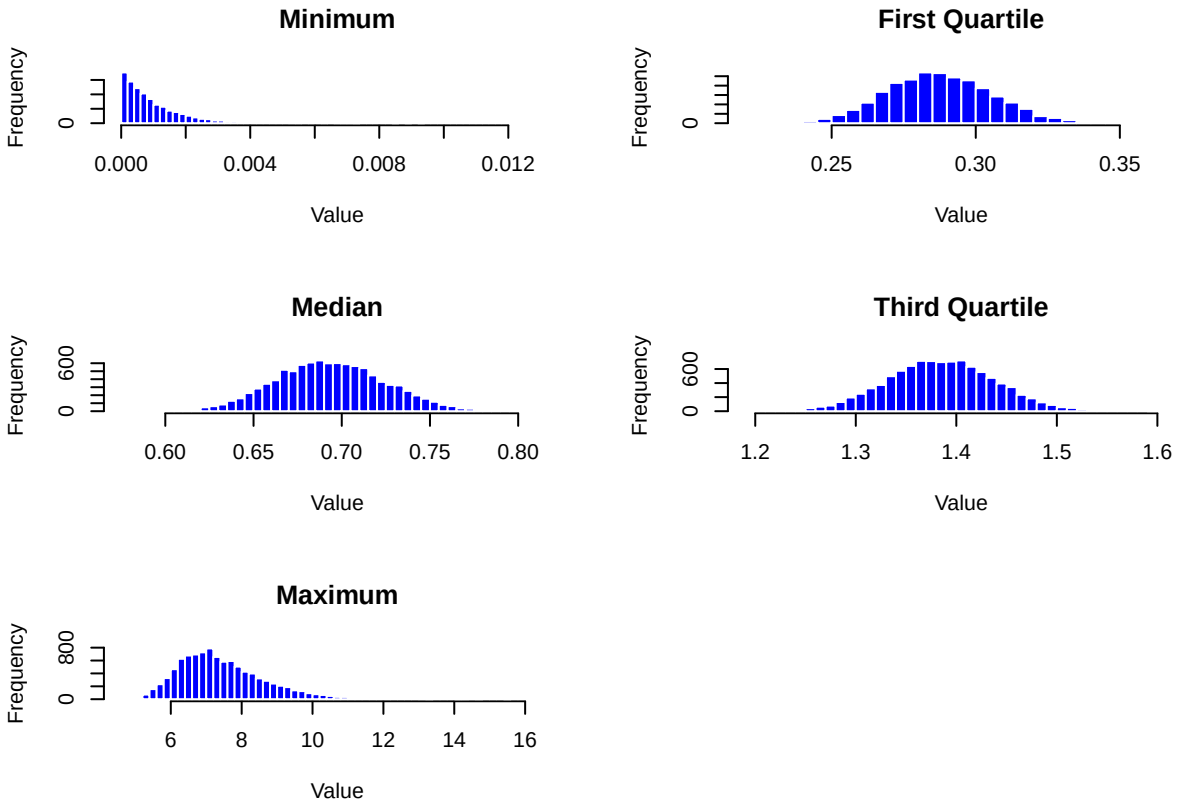
When consider the exponential distribution with rate λ, the PDF and CDF can be described as follows:

1. **Probability Density Function (PDF)**: $f(x) = \lambda e^{-\lambda x}$, for $x \geq 0$
2. **Cumulative Distribution Function (CDF)**: $F(x) = 1 - e^{-\lambda x}$, for $x \geq 0$

The PDF of the k-th order statistic $X_{(k)}$, the Probability Density Function (PDF) is expressed as:

Probability Density Function (PDF): $X_{(k)}$: $f_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!}[1 - e^{-\lambda x}]^{k-1} e^{-\lambda x[n-k]} \lambda e^{-\lambda x}$

Simulation and Plot of Exponential Distribution

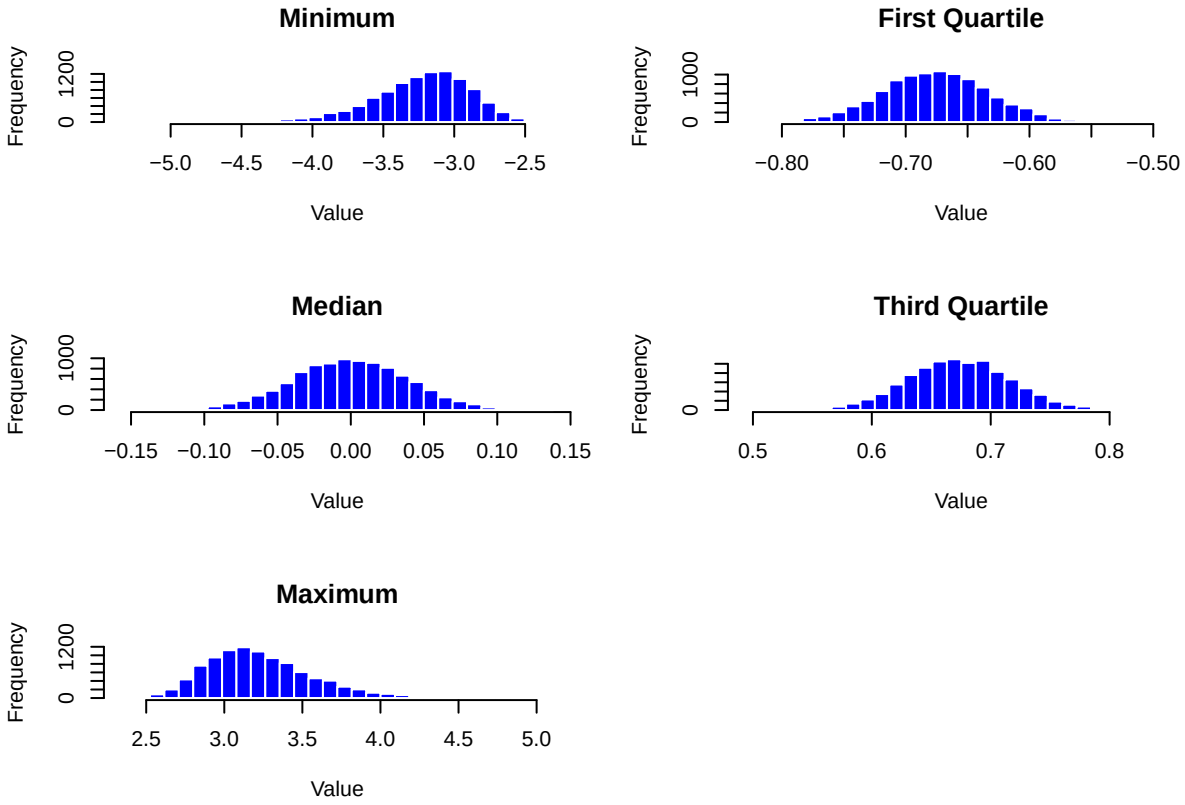


Normal Distribution

When consider the normal distribution with two parameters: the mean, u, and the variance, σ^2. , the PDF and CDF can be described as follows:

1. **Probability Density Function (PDF)**: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
2. **Cumulative Distribution Function (CDF)**: Although CDF for the normal distribution cannot be articulated through elementary functions, it is nonetheless thoroughly defined and symbolized by the Φ function in the case of the standard normal distribution. The goal of the CDF is to measure the probability that a given random variable X will assume a value less than or equal to some specified value x.

Simulation and Plot of Normal Distribution



3.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

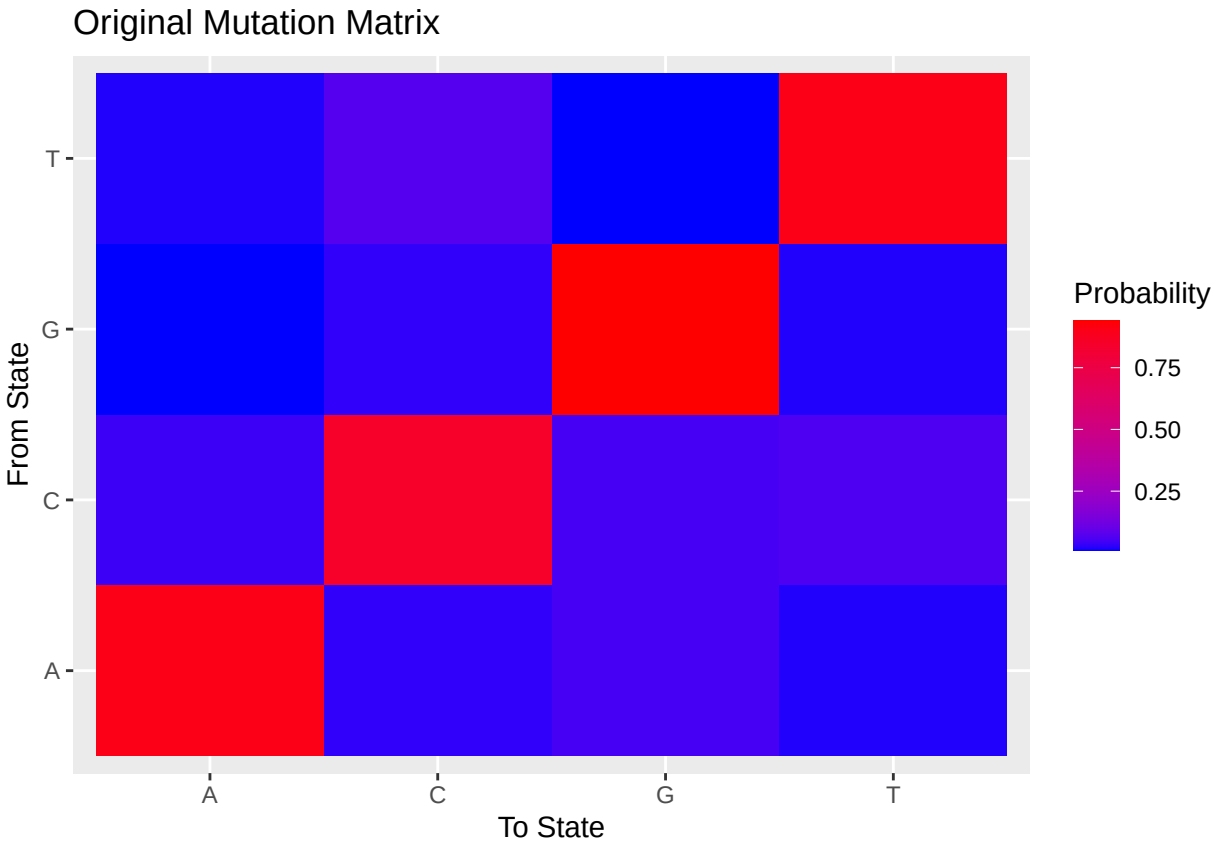
C to A: 0.04, C to C: 0.85, C to G: 0.05, C to T: 0.06

G to A: 0.01, G to C: 0.03, G to G: 0.94, G to T: 0.02

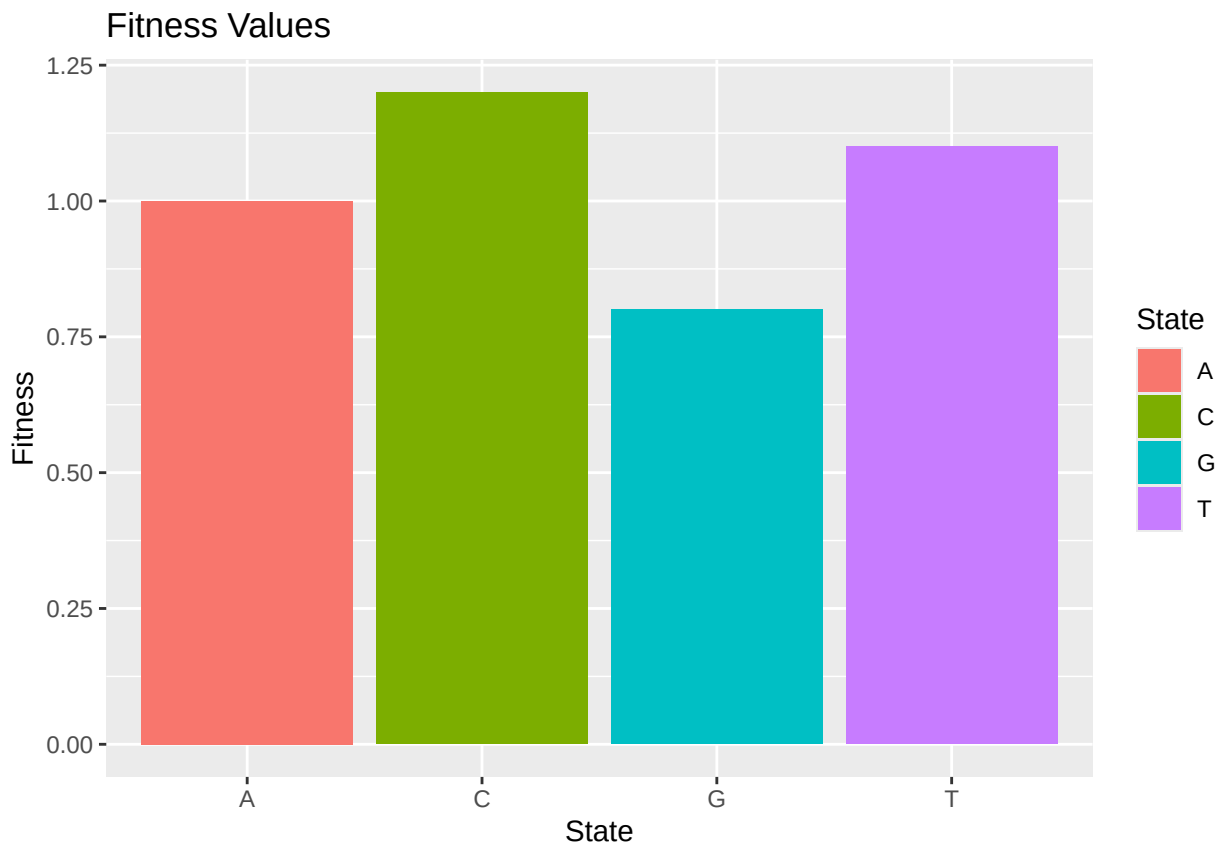
T to A: 0.02, T to C: 0.07, T to G: 0.01, T to T: 0.9

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.




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




```
##  [1] "A" "A" "A" "A" "G" "G" "G" "G" "G" "G" "C" "C" "C" "C" "C" "T" "T" "T" "T"
## [20] "C"
```