Statistical Theory and Modeling - Home Assignment - Part 1

Group number - 20

Name: Aparna Ramesh Pai
Email: aparnapai247@gmail.com

Name: Chaithra Satheesh
Email: s.chaitra12@gmail.com

Name: Chinthaka Chamil Prasanga Amarbandu
Email: guam8978@student.su.se

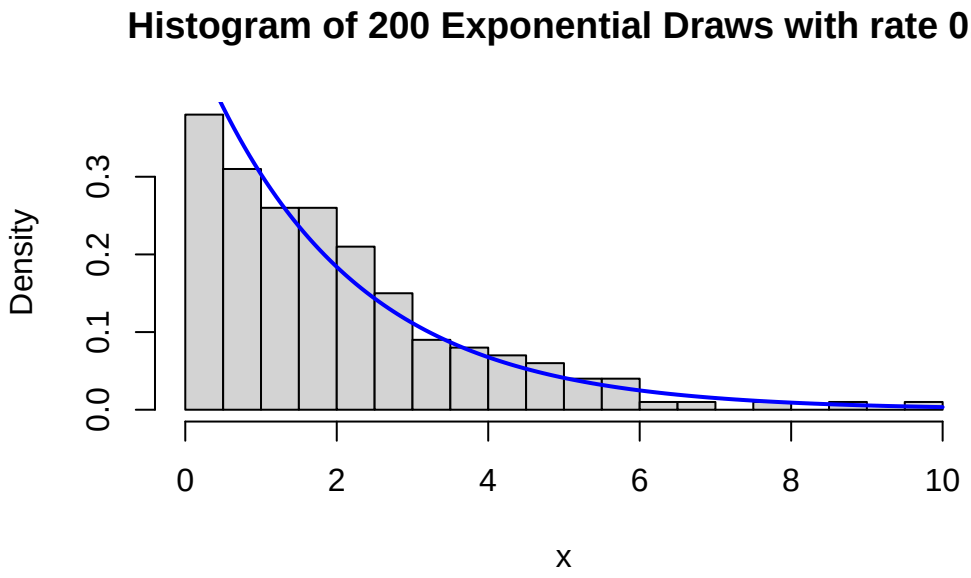
Problem 1 - Exponential distribution and Numerical integration

1a. Simulate $n = 10000$ random numbers from the exponential distribution with rate $\lambda = 2$ to verify that R is using the rate parameterization.

The exponential distribution with rate $\lambda = 2$ has a theoretical mean, $\mu = \frac{1}{\lambda} = \frac{1}{2} = 0.5$

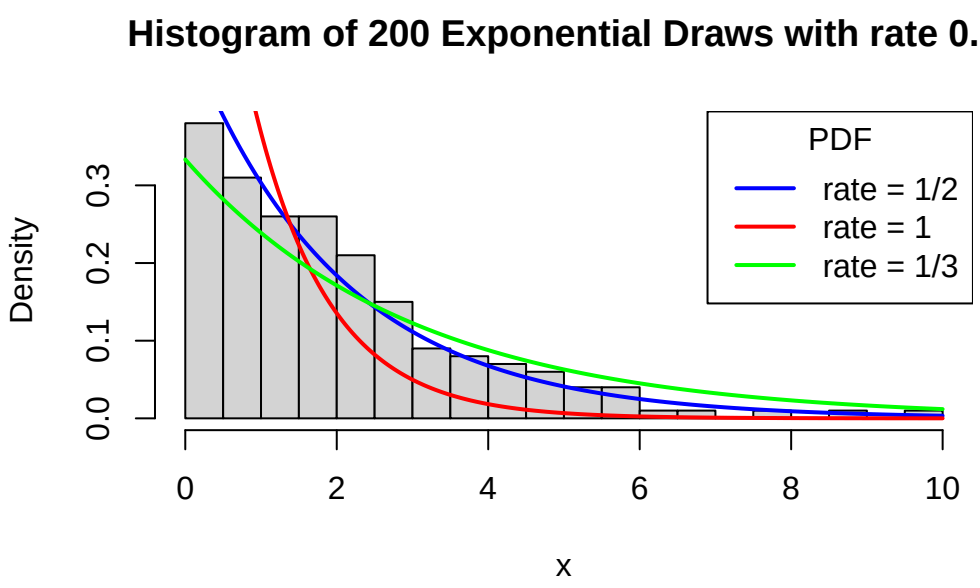
The sample mean is 0.5030445 which is close to the theoretical value of 0.5, thus confirming rate parameterization in R.

1b. Simulate 200 random numbers from the $X \sim \text{Expon}(\beta = 2)$ distribution. Plot histogram of draws (using 30 bins) and overlay the theoretical pdf for $\text{Expon}(\beta = 2)$ distribution as a curve.



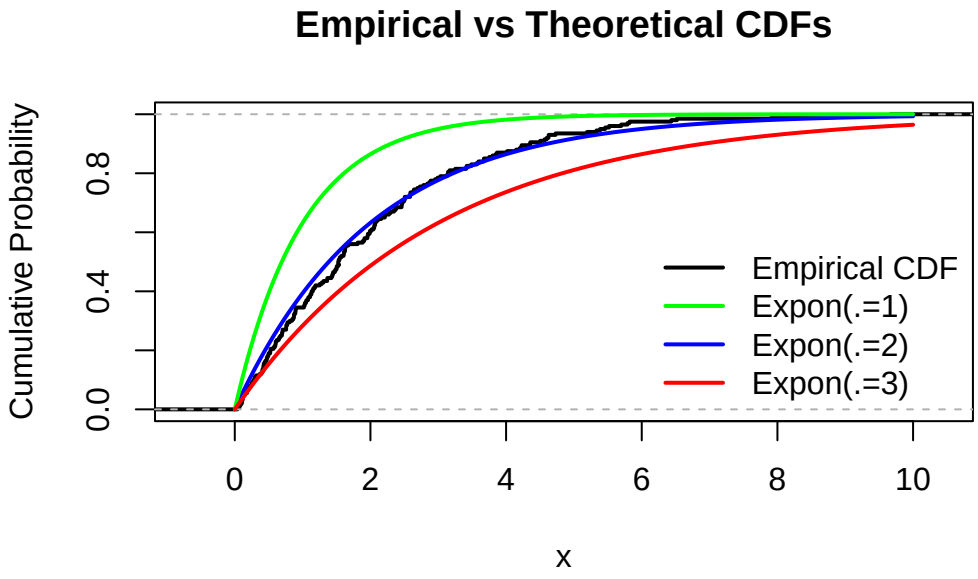
The Histogram closely follows the shape of the theoretical curve, confirming that the simulated data follows the expected behavior.

1c. Overlay two more pdf curves: one for $\text{Expon}(\beta = 1)$ and other for $\text{Expon}(\beta = 3)$. Which of the three pdf curves fit the histogram data best? Why?



The PDF curve for Expon($\beta = 2$) distribution fits the histogram of data simulated from $X \sim$ Expon($\beta = 2$) distribution best, as they both have the same rate: $1/2$. The curve with rate 1 ($\beta = 1$) falls faster, underestimating the frequency of larger values, while the curve with rate $1/3$ ($\beta = 3$) falls more slowly and overestimates the probability of larger values.

1d. Plot the empirical cdf for $n = 200$ observations simulated in 1b. Overlay the cdf from 3 distributions: Expon($\beta = 1$), Expon($\beta = 2$), Expon($\beta = 3$). Which distribution fits best?



The cdf from distribution Expon($\beta = 2$) matches the empirical CDF best, thus confirming previous conclusion.

1e. Compare sample median from $n = 200$ observations to the theoretical medians for each of the above distributions. Explain both how: **- a sample median is defined - how a median of a statistical distribution is defined.**

Sample median is the middle value in the sorted sample.

Theoretical median for an exponential distribution is $\beta * \log(2)$.

- Sample median = 1.5278397
- Median ($\beta = 1$)= 0.6931472
- Median ($\beta = 2$)= 1.3862944
- Median ($\beta = 3$)= 2.0794415

The sample median is closest to the theoretical for $\beta = 2$.

1f. Verify by numerical integration that Expon($\beta = 2$) density in R fulfills the required property of any density $\int_{-\infty}^{\infty} f(x)dx = 1$.

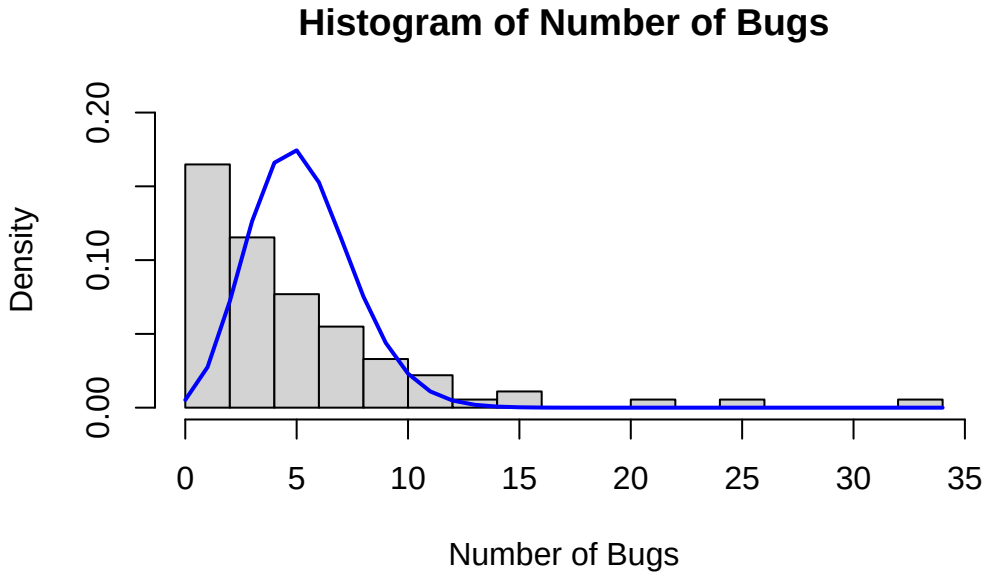
The rectangle sum approximation resulted in a value of 1.000125, which verifies the property of Expon($\beta = 2$) density that any probability density function must integrate to 1.

1g. Compute the expected value of the exponential distribution with $\beta = 2$ using numerical integration. Verify this result using built in integration routine.

The expected value of exponential distribution with $\beta = 2$ using numerical integration is 1.9998958 and using the built in function *integrate* is 2, which are both close/equal to the theoretical mean $E(X) = \beta = 2$.

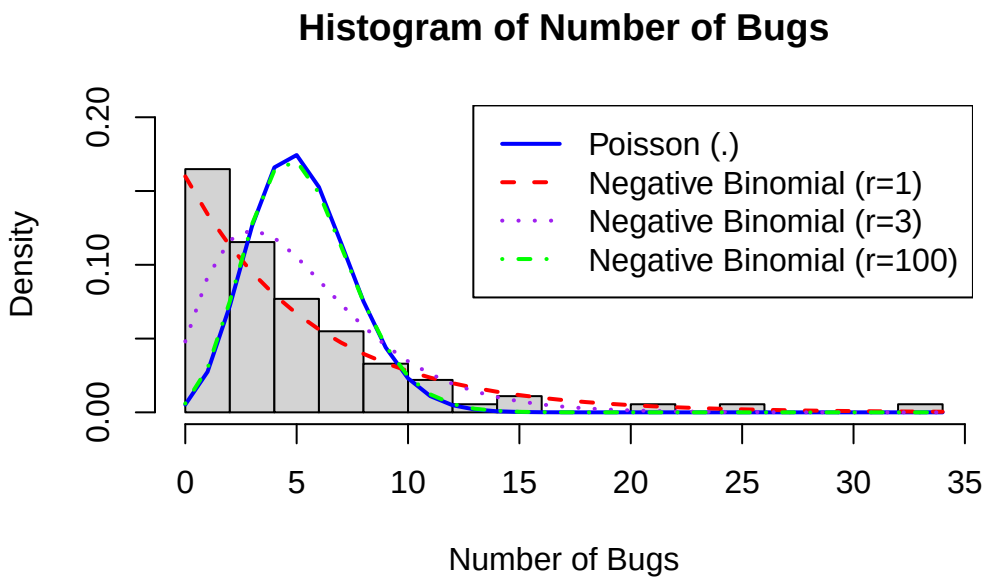
Problem 2 - Exponential distribution and Numerical integration

2a. Load dataset and store number of bugs in vector y. Plot a histogram of data and overlay the density of poisson distribution with $\lambda = \bar{y}$.



The Poisson model doesn't fit the data well. Poisson model assumes $Mean = Variance = \lambda$, which isn't the case with this data. It has a variance of 29.2576313 which is much higher than the mean (5.2527473). Thus the Poisson model underestimates the spread of the data.

2b. Add probability function from negative binomial model for three different r values: $r = 1$, $r = 3$, and $r = 100$. Which of these models do you prefer? Why? Which of the negative binomial models is closest to the Poisson model? Why?



Probability density of the negative binomial model for $r = 1$ fits the data well. This is because the parameter r introduces an extra parameter in Poisson, allowing the variance to exceed the mean.

Probability density of the negative binomial model for $r = 100$ resembles the poisson model since Negative binomial approaches Poisson as $r \to \infty$.

Problem 3 - Transforming a Variable

3a. Simulating the Distribution of $Y = \exp(X)$

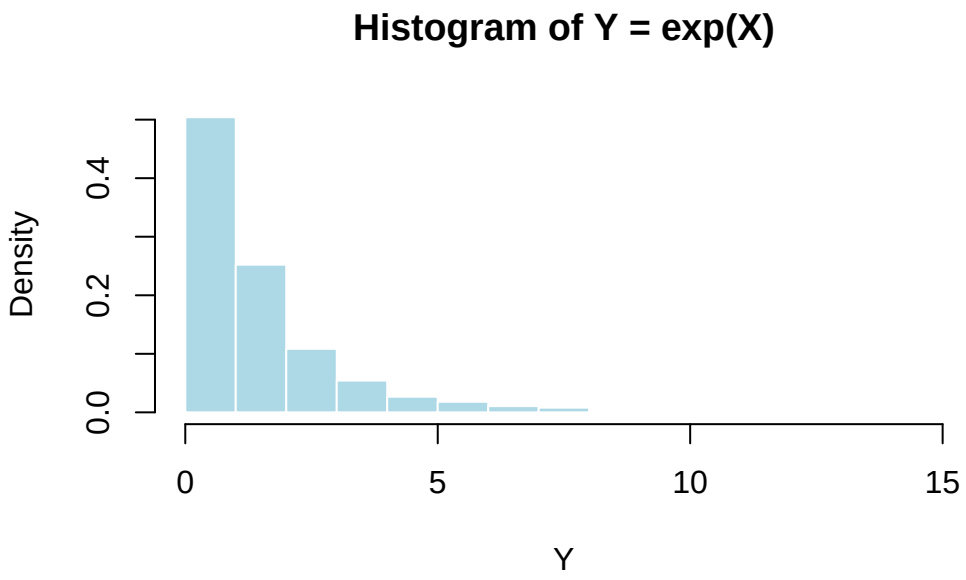
We start with the assumption that

$$X \sim N(0, 1)$$

We are interested in the distribution of

$$Y = \exp(X)$$

In order to visualize this distribution, we simulate 10,000 draws from the normal distribution, transform them using the exponential function, and plot a histogram. It shows the simulated distribution of the random variable $Y = \exp(X)$.



If frequency distribution that approximates the shape of the log normal distribution Because

$$If \quad X \sim N(\mu, \sigma^2) \quad Then$$
$$Y = \exp(X) \sim LogNormal(\mu, \sigma^2) \qquad Where \quad \mu = 0, \sigma = 1$$

The histogram shows a strong right skewness with a peak near 1 and a long right tail. This matches the known shape of the log-normal distribution, which confirms that the exponential transformation of a normal variable results in a log-normal distribution.

3b. Deriving the PDF of Y and Overlaying the Theoretical Curve

We now derive the probability density function of $Y = exp(X)$ using the method of transformation. The theoretical PDF is worked out below.

Let,

$$X \sim N(\mu, \sigma^2)$$

For a strictly increasing transformation, $Y = g(X) = exp(X)$ the PDF of y is given by,

$$f_y(y) = f_x(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)|$$

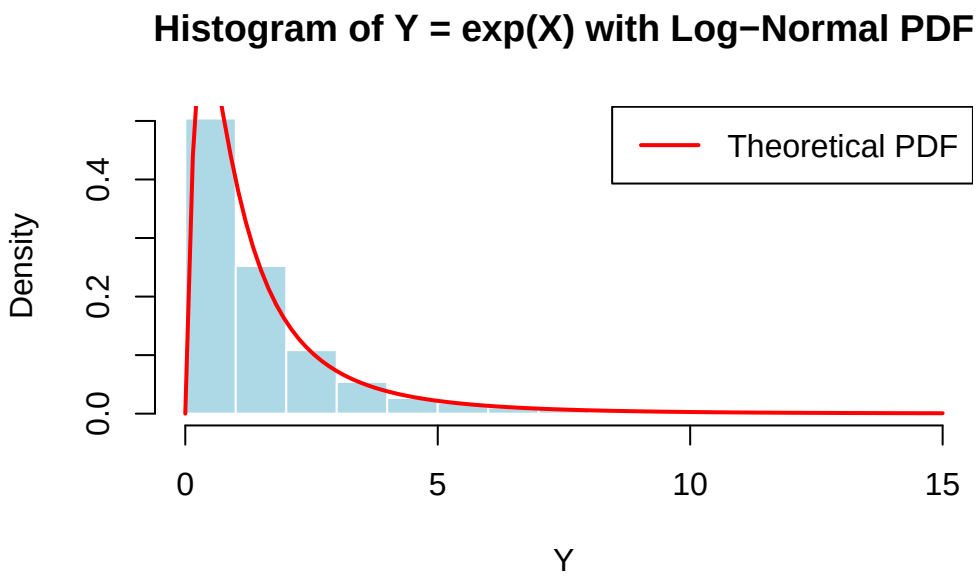
$$Here, \quad g^{-1}(y) = log(y) \quad and \quad \frac{d}{dy}g^{-1}(y) = \frac{1}{y} \qquad (Since \quad f_x(x) \quad is \quad the \quad normal \quad density)$$

$$f_x(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}(x-\mu)^2)$$

$$We \quad substitute \quad 2 = log(y) with \mu = 0, \quad hence \quad we \quad get$$

$$f_y(y) = \frac{1}{\sqrt{2\pi y}} exp(-\frac{1}{2}(log_y)^2)$$

Now , we shall overlay this theoretical PDF onto the histogram. We obtain the following graph.



The overlaid red curve represents the theoretical probability distribution function of a log normal distribution.

3c. Monte Carlo Estimation and Convergence of E[Y]

In this part, we estimate the expected value of Y = exp(X) , where X N (0,1) using Monte Carlo simulation. Since Y follows a log-normal distribution, we know from theory that,

$$E(X) = exp(\mu + \frac{\sigma^2}{2})$$

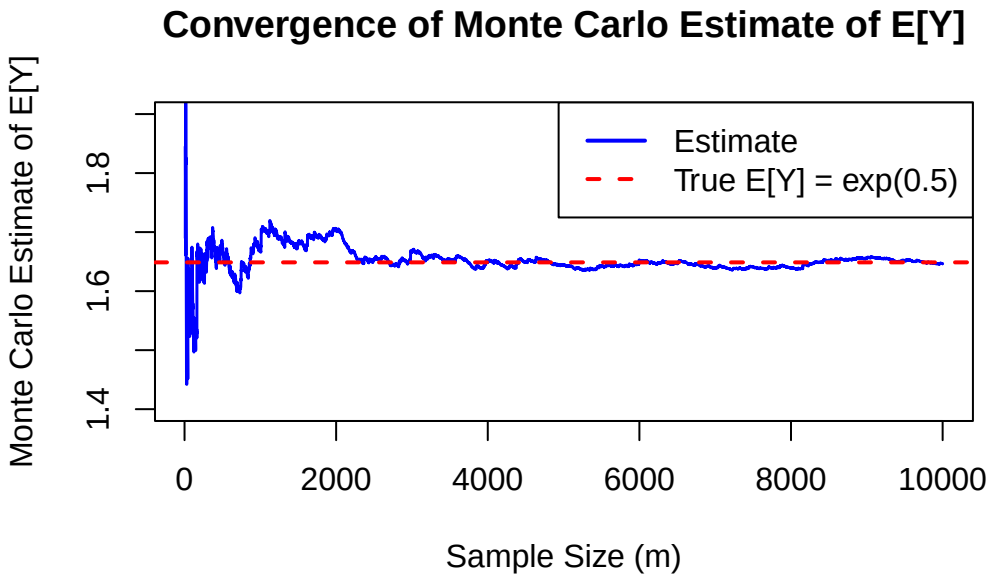
$$= exp(0 + \frac{1}{2})$$

$$= exp(0.5) \approx 1.6487$$

To approximate this expectation, we simulate 10,000 random draws of X from the standard normal distribution, compute $Y = exp(X)$ and then calculate the sample mean

$$\hat{\mu}_m = \frac{1}{m} \sum_{i=1}^{m} Y_i$$

We compute $\hat{\mu}_m$ for increasing values of $m = 10, 20, 30.....10,000$ and plot these estimates against m



The resulting convergence graph shows that the estimates fluctuate for small sample sizes but eventually stabilize around the true value of $exp(0.5)$ confirming the convergence predicted by the Law of Large Numbers. The red dashed line in the plot represents the true expectation $E[Y] = exp(0.5)$ and the Monte Carlo estimates approach this line as the sample size increases.