FISHES ANALYSIS

Derrick Asante

2024-04-29

INTRODUCTION

In this study, we are going to use the data set fishes ; data on the mass of 250 fishes.Our goal is to to analyze this data and determine whether or not this data are well described by a probability distribution.Knowing the distribution that the data best fits will help us to make inferences and draw perfect conclusions. Probability distributions such as the normal and gamma Ndistributions are used to model the data to find the one that best suit the data.

DATASET

The data set used in this study is fishes dataset ; data on the body mass of 250 fishes .

OBJECTIVES

In this report, we seek to solve the following problems.

- Perform exploratory data analysis and determine the probability distribution that best fit the data.
- Draw a histogram of the dataset and overlay distributions.
- Find the distribution that best fits the dataset.
- Predict expected observations using the probability distribution selected.
- Perform goodness of fit (GOF) test.

PRELIMINARY ANALYSIS

Data Preprocessing

To begin, we first understand and preprocess the data.

```
function (x, df1, df2, ncp, log = FALSE)
```

The above shows the mass of 250 fishes.

Descriptive Statistics

```
[1] 4.194275
```

```
The Mean  is  4.194275
```

```
[1] 3.829713
```

```
The Median  is  3.829713
```

```
[1] 4.779735
```

```
The variance  is  4.779735
```

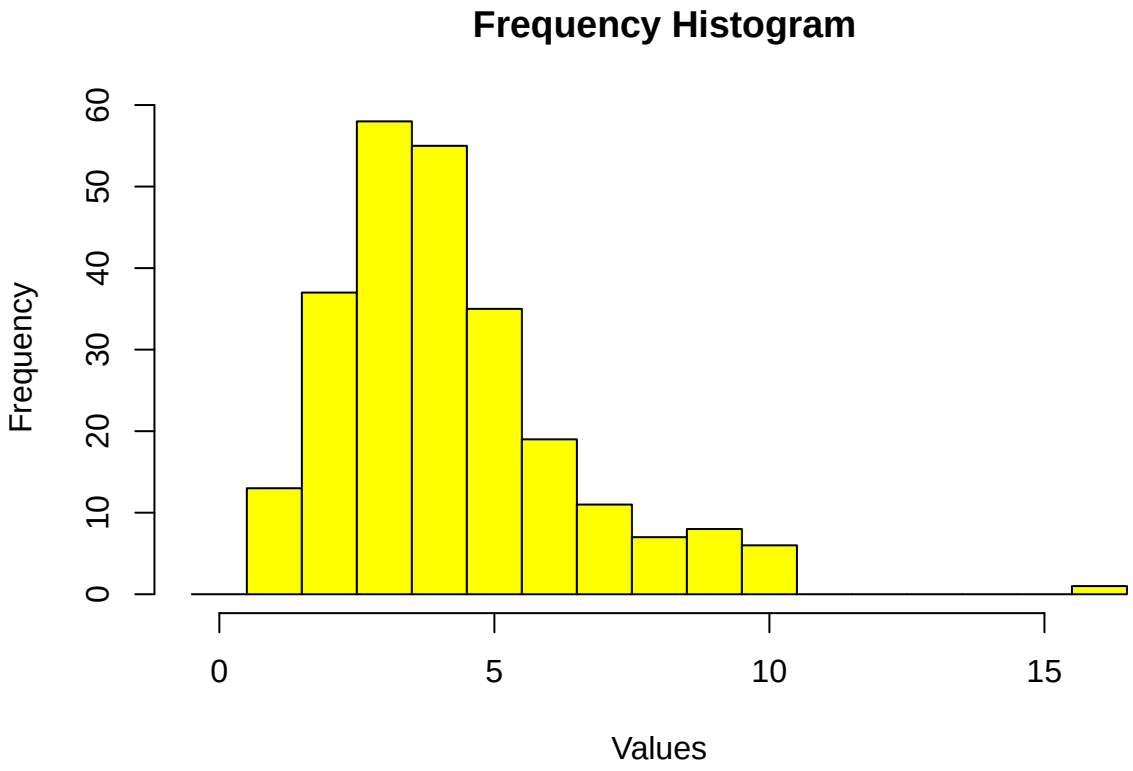
```
[1] 1.29077
```

```
The skewness  is  1.29077
```

```
[1] 2.723529
```

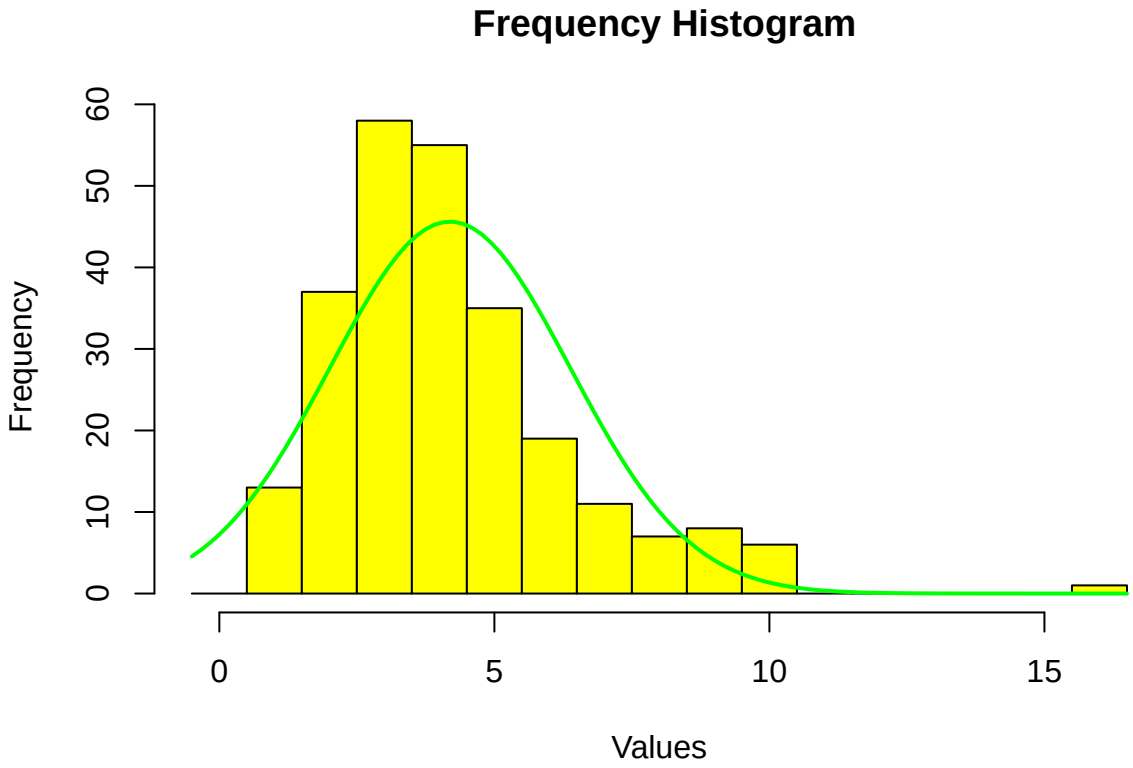
```
The kurtosis  is  2.723529
```

Graphical Summaries

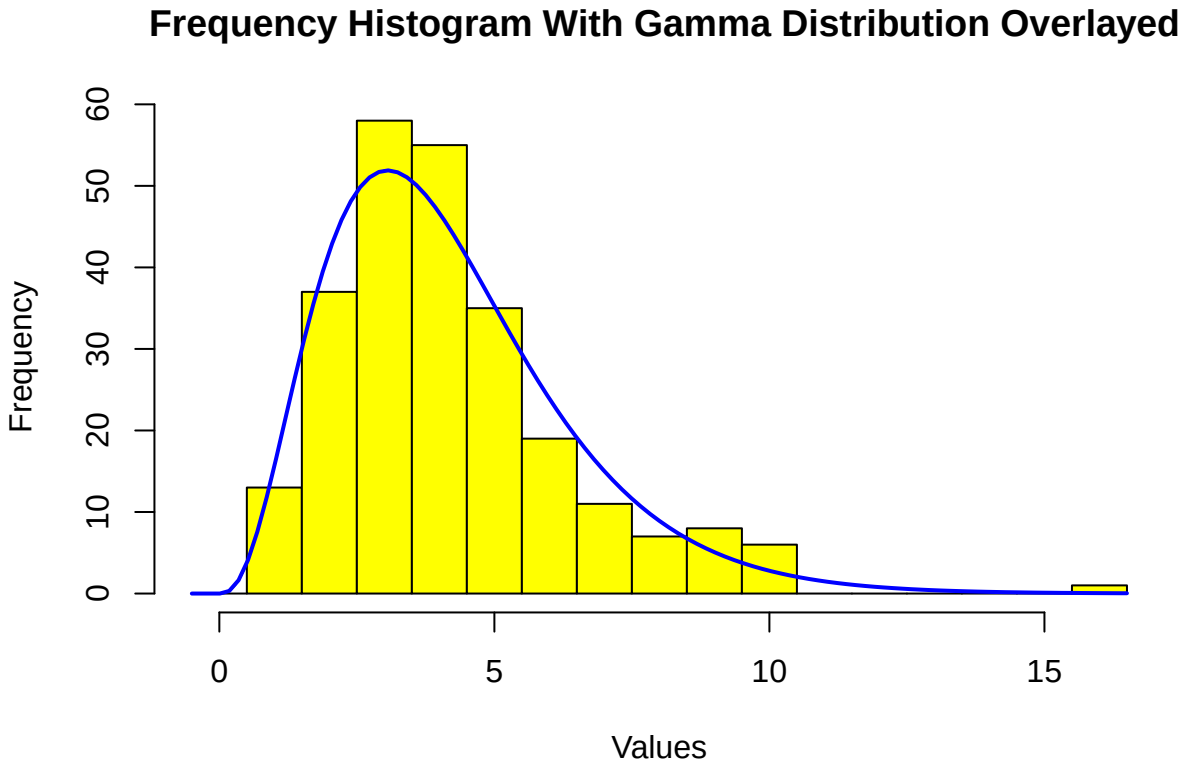


From the histogram above we can see that the data is highly skwewed to the right.

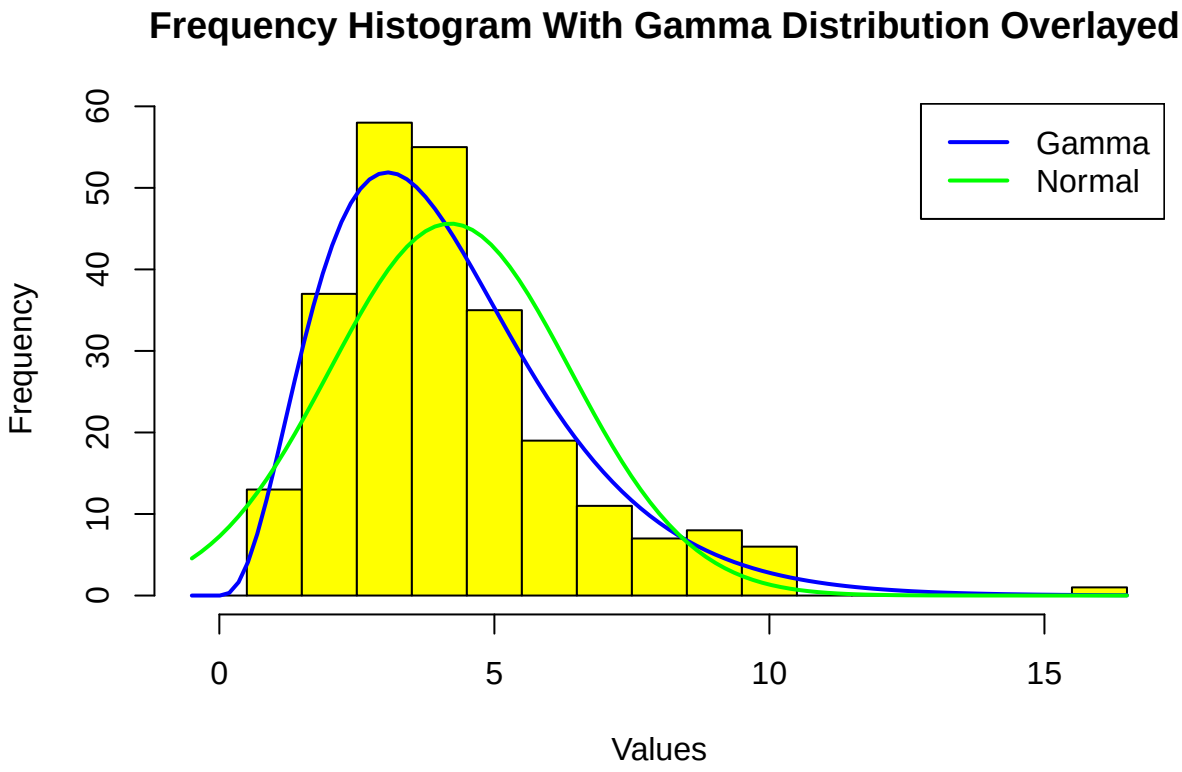
Overlaying Distributions (Normal Distribution)



Overlaying Distributions (Gamma Distribution)



Overlaying Distributions (Both Distribution)



HYPOTHESIS TESTING FOR SKEWNESS

The null hypothesis H_0 is stated as follows:

$$H_0 : \gamma = 0$$

The alternative hypothesis H_a is stated as follows:

$$H_a : \gamma \neq 0$$

The test statistic is calculated as:

$$\text{test statistic} = \frac{\gamma}{SE(\gamma)}$$

$$T = \frac{\sqrt{n} \times \gamma}{\sqrt{6}}$$

where γ is the sample skewness and n is the sample size.

```
The test_statistics  is  8.331883
```

Since the test statistic is greater than 1.96 we reject th null hypothesis and conclude that the data is not symmetric.

ESTIMATION

Now that we have concluded that our data is not symmetric. We will like to fit distributions to the data. We will fit the gamma and normal distribution

Normal Distribution

The probability density function (PDF) of the normal distribution is given by:

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

Where: $f(x|\mu,\sigma)$ is the probability density function.

x is the random variable.

μ is the mean of the distribution.

σ is the standard deviation of the distribution.

In this case we first estimate two parameters ; the mean and standard deviations

```
The  mean is  4.194275
```

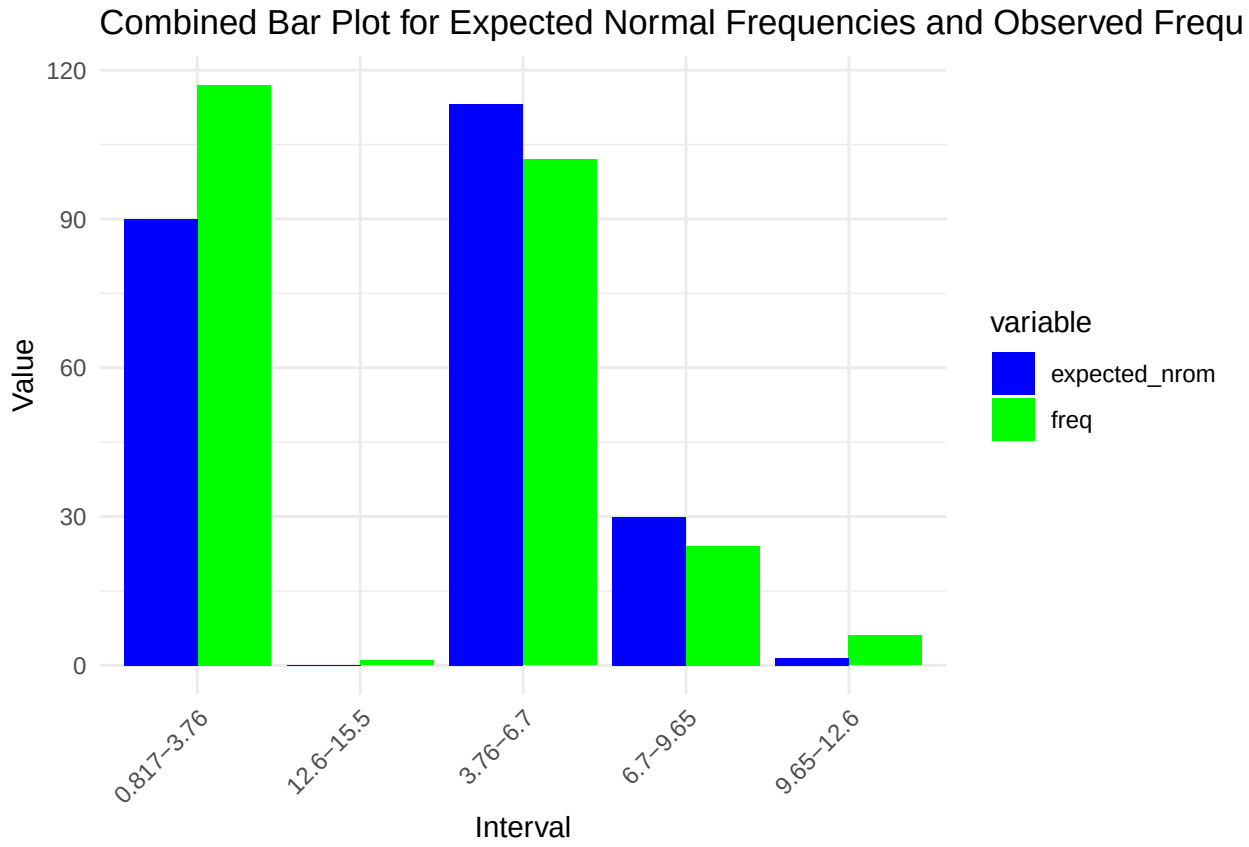
```
The  standard deviation is  2.186261
```

Now that we have been able to calculate the mean and standard deviation for the parameters, we can find the fitted probabilities and frequencies. The idea is to create groups and use these lower and upper bounds to integrate for each categories so that we find the expected frequencies.

Estimated Expected frequencies

intervals.new	freq	expected_normal
0.817-3.76	117	90.018
3.76 - 6.7	102	113.214
6.7-9.65	24	29.896
9.65-12.6	6	1.557
12.6-15.5	1	0.015

Normal Fit vrs Observed Frequencies



Goodness of fit test (Kolmogorov-Smirnov)

Kolmogorov-Smirnov test helps us to decide whether the estimated expected frequency from the normal distributon fits our data.

Hypothesis

(H_0): "The sample data follows the specified distribution."

(H_1) : "The sample data does not follow the specified distribution."

```
    Asymptotic one-sample Kolmogorov-Smirnov test

data:  df1
D = 0.10448, p-value = 0.008522
alternative hypothesis: two-sided
```

The P-value is 0.008522. Since this p-value is less than 0.05, we reject our null hypothesis and conclude that the observed frequency distribution and the estimated expected frequency distribution are the not same. That is, the normal distribution does not fit our data.

GAMMA DISTRIBUTION

The probability density function (PDF) of the gamma distribution is given by:

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$$

Where: $f(x;\alpha,\beta)$ is the probability density function.

x is the random variable.

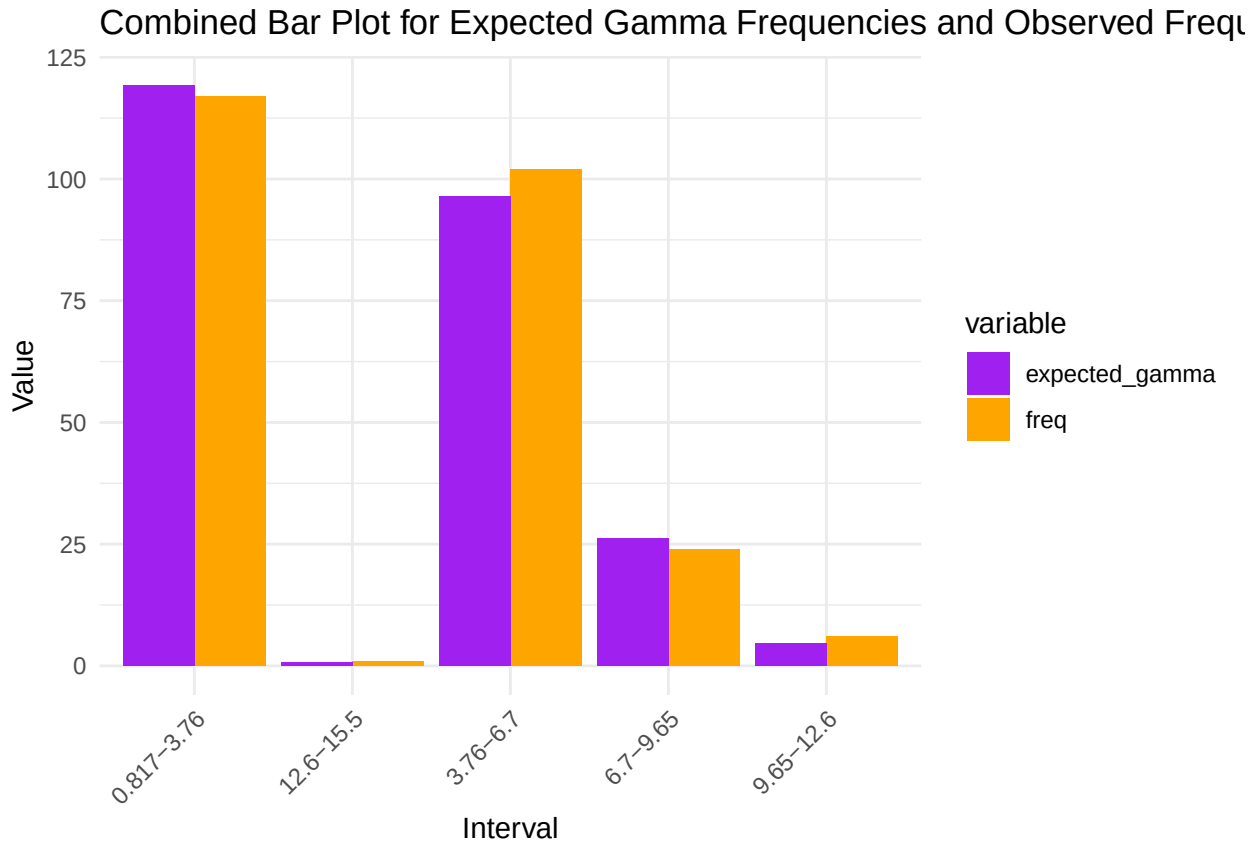
α is the shape parameter.

β is the rate parameter.

$\Gamma(\alpha)$ is the gamma function.

intervals.new	freq	expected_gamma
0.817-3.76	117	119.192
3.76 - 6.7	102	96.448
6.7-9.65	24	26.164
9.65-12.6	6	4.661
12.6-15.5	1	0.667

Gamma Fit vrs Observed Frequencies (Barplot)



The gamma looks like a good fit for the data.

Goodness of Fit Test ((Kolmogorov-Smirnov))

Null Hypothesis (H_0):

$$H_0 : F(x) = F_{\text{gamma}}(x; \alpha, \beta)$$

Alternative Hypothesis (H_a):

$$H_a : F(x) \neq F_{\text{gamma}}(x; \alpha, \beta)$$

```
    Asymptotic one-sample Kolmogorov-Smirnov test

data:  count_data
D = 0.042585, p-value = 0.7551
alternative hypothesis: two-sided
```

The P-value is 0.7551. Since this p-value is greater than 0.05, we fail to reject our null hypothesis and conclude that the observed frequency distribution and the estimated expected frequency distribution from the gamma distribution are the same. That is, the gamma distribution fits our data.

CONCLUSION

We can conclude that the gamma distribution is a perfect fit for our data.The data is also skewed .