School of Electrical and Information Engineering
University of the Witwatersrand
ELEN3007: Probabilistic Systems and Signal Analysis

Michael Moore
2118213

Dr. Hunt
31 October 2022

Assignment 1: Data Analysis

Introduction:

When experiments are conducted, and data is collected, the probability of specific outcomes can be determined. When plotting these probabilities as a histogram, the unique probability density function of the data can be seen. To determine the mathematical expression of the PDF, an approximation is made, and one of the known probability density functions that best fits the data is chosen. In this report, data consisting of the peak currents delivered by lightning strikes in Johannesburg will be analysed, and three different PDFs will be selected to fit the data best. Test statistics between the data and the approximations will be performed to determine the quality of fit and the best PDF will be chosen. Due to most distributions' inability to work with negative values, only the absolute value of the data will be used.

Probability density functions and cumulative density functions:

The first PDF chosen was the Gaussian distribution. The reason for this choice was due to it being able to model natural phenomena with good accuracy and only being dependent on the mean and standard deviation of the data (1).
The second PDF selected to fit the data was the Gamma distribution. The selection of this probability density function was due to its ability to accurately model data skewed to the right, which is the case with the recorded data (2).
The third PDF used to model the data was the Rayleigh distribution. This PDF was selected because it is skewed to the right and thus would make it better at fitting the given data (3).
Please refer to Appendix A for the equations of the selected probability density functions.

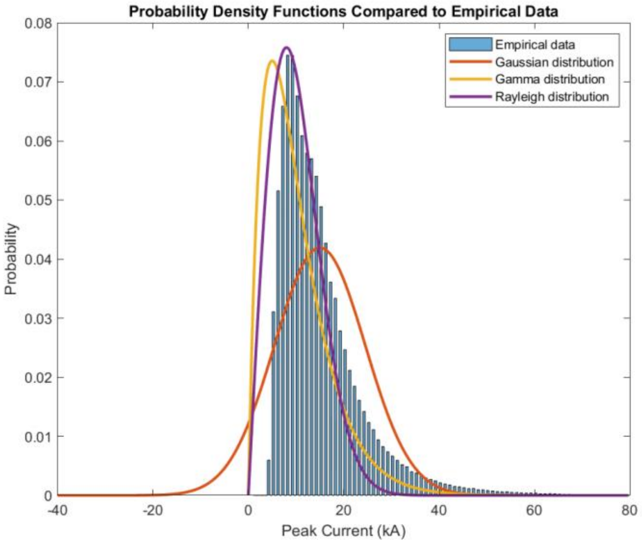

Figure 1: Comparison between different approximate PDFs and the empirical data.

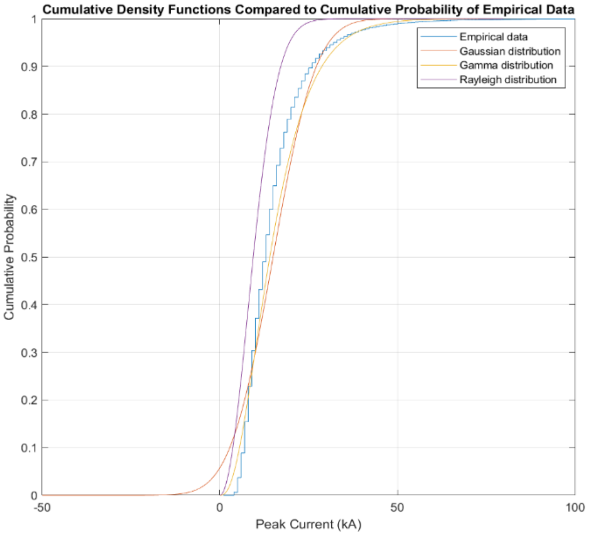

Figure 2: Comparison between different approximate CDFs and the CDF of the recorded data.

In *Figure 1*, the three different PDFs were plotted on top of the recorded data, and using graphical analysis, it can be seen that the Rayleigh distribution most closely matched the data's distribution. The Gamma PDF provided a good approximation; however, it deviated more toward the top of the plot as compared to the Rayleigh distribution. The Gaussian distribution possessed the same mean and standard deviation as the empirical data; however, it did not accurately model the given data.
In *Figure 2*, the cumulative density functions of the three approximations as well as the empirical data are plotted. From this plot, it can be seen that most of the currents lie between values of 0 – 50kA and that the Gamma distribution's CDF most accurately matches the empirical data.

Table 1: Comparison of the test statistics obtained from each distribution.

PDF	**RMSE**	**SSE**	**R^2**
Gaussian	0.0202	0.1017	-1.4740
Gamma	0.0175	0.0764	0.3459
Rayleigh	0.0218	0.1182	0.1148

The RMSE or root mean square error is the standard deviation of the approximation errors (4). Thus, when comparing the empirical data to the different probability density functions, a lower RMSE would mean smaller errors and thus, a better fit. From the data presented in *Table 1*, the Gamma distribution was the best fit for the data when considering the RMSE.

The SSE or residual sum of squares is the sum of the squares of the errors between two data sets (5). When the probability density functions were compared to the recorded data, there were errors between the true values and the approximations. The smaller the value of the total error, the better the fit, and thus the Gamma distribution would be the best in terms of the SSE.

The coefficient of determination is a statistical measure that explains the variance of a dependent variable due to the variance of an independent variable (5). When using it to analyse the goodness of fit between two data sets, the higher the value, the better the fit, and thus the Gamma distribution is the best fit for the data. The reason for the Gaussian distribution having a negative coefficient of determination is due to it being spread across both negative and positive values, as seen in *Figure 1*, while the data is only positive, and thus this distribution is not an accurate model for the data.

Positive and negative lightning distributions:

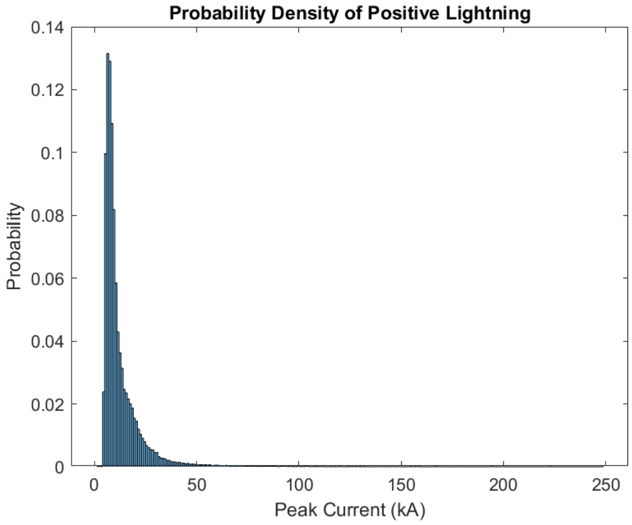

Figure 3: Probability density of positive lightning data.

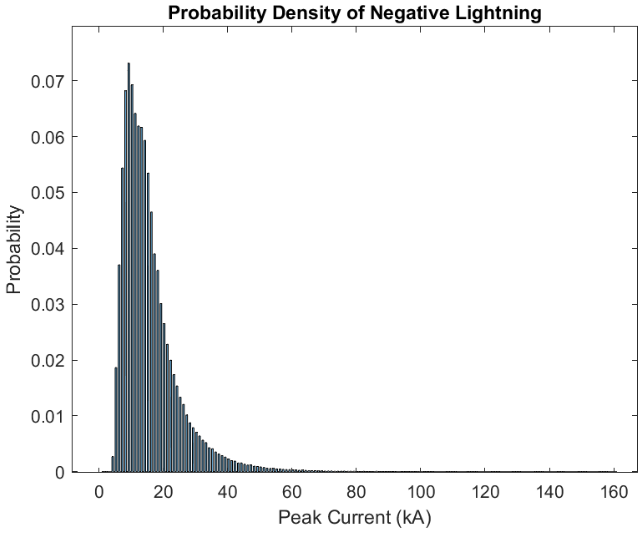

Figure 4: Probability density of negative lightning data.

The distributions of both positive and negative lightning are shown in *Figures 3 & 4*, respectively. Positive lightning has a greater probability of having currents closer to zero, whereas negative lightning has a greater spread of currents. The distribution of the negative lightning shows currents as high as 60kA; however, the positive lightning only has peak currents that go up to 50kA. Both distributions are similar in shape, however, and both data sets are skewed to the right.

Conclusion:

Three different probability density functions were selected and compared to the empirical data through the use of test statistics. From the data obtained it can be concluded that the Gamma distribution most accurately models the data with the lowest RMSE and SSE and the highest R^2.

References:

1. **Bhandari, Pritha.** Scribbr. *Normal Distribution | Examples, Formulas, & Uses.* [Online] 23 10 2020. [Cited: 25 10 2022.] https://www.scribbr.com/statistics/normal-distribution/.
2. **Kim, Aerin.** Towards Data Science. *Gamma Distribution — Intuition, Derivation, and Examples.* [Online] 19 10 2019. [Cited: 25 10 2022.] https://towardsdatascience.com/gamma-distribution-intuition-derivation-and-examples-55f407423840.
3. **Statistics How To.** Statistics How To. *Rayleigh Distribution: Definition, Uses, Mean, Variance.* [Online] [Cited: 25 10 2022.] https://www.statisticshowto.com/rayleigh-distribution/.
4. —. Statistics How To. *RMSE: Root Mean Square Error.* [Online] [Cited: 25 10 2022.] https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/.
5. **Statology.** Statology. *A Gentle Guide to Sum of Squares: SST, SSR, SSE.* [Online] 22 02 2021. [Cited: 25 10 2022.] https://www.statology.org/sst-ssr-sse/.

Appendix A

Gaussian distribution:

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

Gamma distribution:

$a = 2$
$b = 5$

$$y = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{\frac{-x}{b}}$$

Rayleigh distribution:

$b = 8$

$$y = \frac{x}{b^2} e^{\frac{-x^2}{2b^2}}$$