Analysis of Intimate Partner Violence Against Women Globally

Priyanka Michelle Bangalore - 1005855460

April 19, 2021

Abstract

The gaps between gender-based violence continually increase, with violence against women on the rise. With the stay-at-home regulations due to COVID-19, violence against women at the hands of their intimate partner is at an all time high. Using a number of statistical methods, the purpose of this study is to determine if there is a correlation between age and women subject to intimate partner violence, and residential continent and women subject to intimate partner violence. Using a linear regression model, hypotheses tests, confidence intervals, a maximum likelihood estimator, credible intervals, and a goodness of fit test, it was found that women from some continents are more prone to violence than others, proportions of violence decrease as the age of occurence increases, and that the average proportion of women subject to violence within the data falls within the same interval that the average proportion of women subject to violence in the entire population falls in.

Introduction

In the 1970s, there was a drastic surge in the awareness and prevention of physical and sexual violence against women. In the United States, several coalitions against rape and physical abuse were developed, and some of the first rape crisis centers were established, giving women a safe place in their neighbourhood [17]. In 1976, women protested the streets of the United States and Europe in the Take Back the Night movement – a movement that protested the violence women faced at night on the streets [17]. Women around the world chose to no longer remain silent on the violence and suffering they were subject to at the hands of family, significant others, coworkers, and strangers. In 1990, The Clothesline Project was brought into play. Survivors hung up their t-shirts to shed light on those who often are a statistic ignored by society [17]. The cries for help led activists to begin using their platforms to raise awareness of the violence against women.

In 2009, former United States President Barack Obama officially declared April as Sexual Assault Awareness Month [17]. Since then, several campaigns for women were established, providing women with rape kits, preventative services, safe places, and access to help. Today, there are over 1000 rape and violence crisis centers in the United States and over 700 in Canada [17].

Violence against women has always been a prevalent issue in society. Women's basic human rights often go overlooked, leaving many women with post traumatic stress disorders, poor mental health, and physical traumas that may never heal.

This paper will focus on the violent tragedies faced by women globally in specifically intimate partner relationships.

We hypothesize that women in some residential continents are more prone to intimate partner violence compared to other continents, and that women within a certain age group are more prone to intimate partner violence than women of other age groups.

These hypotheses align with the findings of various other studies. An article published in 2019 by Lia Ryerson found that 6 of the most dangerous countries for women were in Africa, while the other 4 were in Asia [4]. The violence women face in these continents is due to a number of practices found in many developing nations;

child marriages, forced marriages, sex slavery, women trafficking, and honorary practices. Child and forced marriages often place women with their abuser, and in the case of child marriages, at an extremely young age. Women are forced to deal with physical, emotional and sexual abuse at the hands of their spouse, and more often than not, they are unable to leave due to their families' honour at stake.

This analysis aims to highlight the abuses of women's human rights around the world, in hopes of stronger preventative practices and more awareness during Sexual Assault Awareness Month.

The Data section of this paper contains information about the dataset, the cleaning process, important variables, and summary measures.

The Methods section of this paper describes all statistical methodologies that will be used within this analysis and the assumptions associated with them.

The Results section of this paper looks at the coded results and their meanings of the methodologies described in the Methods section.

The Conclusion section of this paper summarizes the outcomes of the Results section and concludes the analysis.

The Bibliography section of this paper contains all full-length citations to the in-text citations in the above sections.

The Appendix section of this paper contains all mathematical derivations.

Data

The following was published by the World Health Organization, last updated on August 20th, 2019. It is a table data set which contains intimate partner violence (sexual and physical) information, such as country and year of incident, from 2005 to 2017 globally [5]. The data was collected via a series of questions asked by interviewers in household surveys [5]. The data depicts the proportion (percentage out of 100) of women ages 15-49 who have experienced physical or sexual violence at the hands of their intimate partner in the 12 months before the survey [5]. The pre-cleaned data consists of more than 550 observations and 4 specifiers [6].

A note from the publisher about the data:

1. The exact number of the surveyed participants was not recorded for public usage [5].

The data was cleaned by filtering out unnecessary columns, creating 5 new variables, renaming variables, filtering out all NA values in the dataset, and rounding all values within the Proportion column.

Columns filtered out: Year_Group and Age.

Columns added: Continent, Age-Group, Prop_Sum_Age, Prop_Sum_Cont, and Prop_Sum_Year.

Age_Group groups ages within a 4 year range, from 15 to 49. The groupings are:

1: Ages between 15 and 19

2: Ages between 20 and 24

3: Ages between 25 and 29

4: Ages between 30 and 34

5: Ages between 35 and 39

6: Ages between 40 and 44

7: Ages between 45 and 49

The new dataset now consists of 7 columns; Year, Country, Age, Proportion, Prop_Sum_Age, Prop_Sum_Cont, Prop_Sum_Year, and Continent.

Important variables to keep in mind:

Age: Age of the female individual subject to violence; *numerical*.

Year: Year of the occurence of violence; *numerical*.

Proportion: Percentage of females who have been subject to violence at the hands of their intimate partner out of 1000∗; *numerical*.

Country: Country of residence of the female; *categorical*.

Continent: Continent of residence of the female; *categorical*.

Prop_Sum_Age: Sum of all proportions within an age range; *numerical*.

Prop_Sum_Cont: Sum of all proportions within a continent; *numerical*.

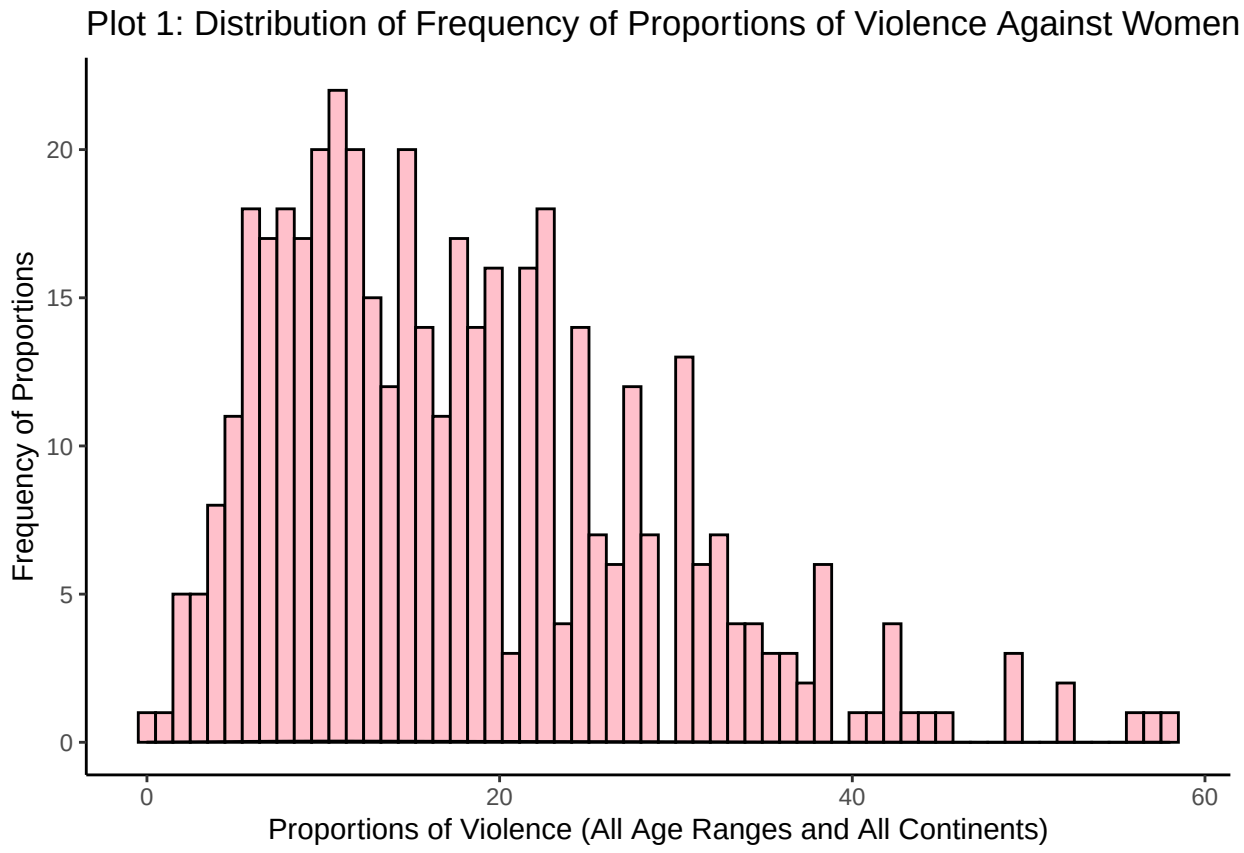
Prop_Sum_Year: Sum of all proportions within a year; *numerical*.

∗ The exact number of the surveyed participants was not provided by the World Health Organization.

Table 1: Values of the Sample Mean, Sample Median and Standard Deviation

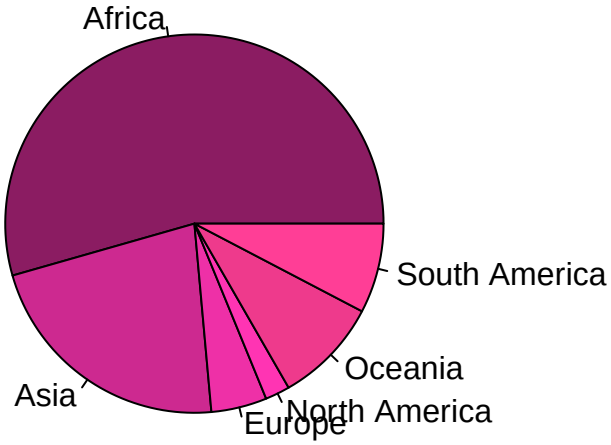
Sample Mean	18.03917
Sample Median	16
Standard Deviation	10.64158

Table 1 above depicts the values of the sample mean, sample median, and standard deviation of the violence proportions in all 6 continents from 2005 to 2017 within the dataset.



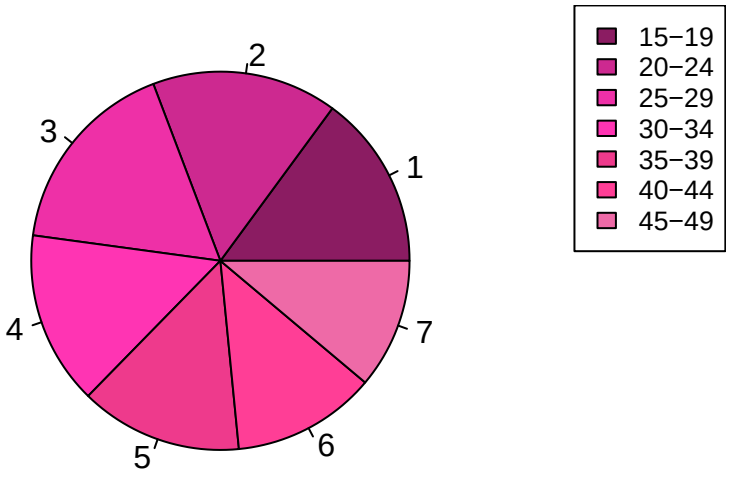
Plot 1 above depicts the frequency of intimate partner violence proportions, regardless of age and continent of occurence. We can see that the majority of proportions lie at around 10-12% of surveyed women. As the proportion increases after 10%, notice the frequency decreases. The data however is not constant; the proportions increase and decrease in a spiratic pattern. Based on the spread of the data, it looks as though the data follows a poisson distribution, with a skewed right tail.

Plot 2: Proportions of Partner Violence and the Continent of Occuren



Plot 2 above of a pie chart displays the continents of occurence of the violence proportions in the data. You can see that Africa has more intimate violence proportions that any of the other 5 continents, accounting for over 50% of the data. Contrastingly, North America has the least intimate violence proportions, accouting for less that 10% of the data.

Plot 3: Proportions of Partner Violence and the Age Group of Occurer



Plot 3 above of a pie chart displays the age group of individuals in accordance with their violence proportions in the data. You can see that the proportions of violence are fairly equal within each age group. By taking a look at the data, we can see that women in age group 7 - ages 45-49 - face the least intimate partner violence, while women in age group 3 - ages 25-29 - face the most intimate partner violence.

Software

All analysis for this report was programmed using `R version 4.0.4`.

Methods

This section of the report will highlight the methodologies that will be used to draw conclusions related to the research topic.

Linear Regression

The model that will be used to analyze the potential relationship between two variables within the Intimate Partner Violence data is a Simple Linear Regression model. A Simple Linear Regression model attempts to create a linear relationship between two variables, depicting their correlation. The model attempts to create the best fitting line for the data, minimizing error and deviation, while following the trend of the data. The ultimate goal of a Simple Linear Regression Model is to be able to use the model to predict future values or values which have not been recorded yet. If the model fits well enough, minimizing deviation, we should be able to use it to predict values outside of the dataset. The 'Simple' in the model name refers to the fact that there is only one independent variable in the model which you will see below.

$$Y_i = \hat{\alpha} + \hat{\beta x_i} + U_i$$

Important components within this model to keep in mind:

Y_i	Dependent Variable
$\hat{\alpha}$	Y-Axis Intercept
$\hat{\beta}$	Slope of Line
x_i	Independent Variable
U_i	Random Error Term

x_i in our first model will be Age_Group - the age groups from 1-7 of all individuals surveyed.

Y_i in our first model will be Prop_Sum_Age - the proportions of violence, categorized by age group.

U_i in our first model accounts for the deviation of our data points from the regression line when the regression line is plotted on top of the data points. The error term accounts for the squared deviations; where the regression line is off from the data [15].

x_i in our second model will be Year - the year of occurence.

Y_i in our second model will be Prop_Sum_Year - the proportions of violence, categorized by year of occurence.

U_i in our second model accounts for the deviation of our data points from the regression line when the regression line is plotted on top of the data points. The error term accounts for the squared deviations; where the regression line is off from the data [15].

$\hat{\alpha}$ and $\hat{\beta}$'s values for our model will be determined in the sections Linear Regression 1 and Linear Regression 2 in Results. Note: the hat above α and β signifies that they are both estimated values, calculated from the data provided.

Confidence Interval

A confidence interval describes how well a sample of data represents the entire population. The interval itself tells us the probability of the true parameter from the population falling within the interval. Often times, datasets are only portions of the entire population, given that it is difficult to accumulate data from a large population like a country or gender. The data that will be used is also a sample of 1000 women who were surveyed, not the entire population of women within the country of residence or within an age group. Thus, a confidence interval will be used in determining how well the data represents all women.

A 95% and 99% confidence interval for mean will be used. 95% and 99% confidence intervals are often the standard choice for confidence intervals. In order to compute the intervals, an empirical bootstrap will be used to sample from the given data and explore variations within the data. For the bootstrap, we know that our data sample is $X_1, ..X_{434}$ is independent and identially distributed. 5000 simulations will be run, and for each simulation, a bootstrap sample will be taken from from $X_1, ..X_{434}$ and used to compute the parameter of interest - mean.

Maximum Likelihood Estimator

The maximum likelihood estimator is a frequentist method of estimating a distributions parameters using given data by maximizing a likelihood function [18]. Using the maximum likelihood estimator, and estimation of the parameter using the data can be made.

Assume the data is a random sample of Poisson random variables with mean λ. I have used the maximum likelihood estimator (MLE) approach to estimate the mean, λ. The MLE of λ is $\bar{x}$. All derivations regarding the MLE can be found in *Section 1: MLE Derivations* of the Appendix.

Hypothesis Test

A hypothesis test is used to determine the credibility of a hypothesis using a sample of data. It allows us to draw a conclusion about a population statistic using a sample statistic. A hypothesis test with a null and alternative hypothesis test will be used to determine if the hypothesis stated below is plausible. A test statistic will first be calculated, and will then be used to determine a p-value.

A test statistics is a number calculated from a hypothesis test. It represents how closely the observed data matches the distribution expected under the null hypothesis of the test [19]. There are various kinds of test statistics. For the sake of this analysis, a t-test will be used to compute the test statistic since the sample size of the data is relatively large (434). A p-value is the probability of obtaining results at least as extreme or more extreme as the observed results of a hypothesis test, under the null hypothesis. A p-value aids in determining whether a null hypothesis should be accepted or rejected, which further relies on a pre-specified cut-off (α).

According to the Central Limit Theorem (CLT), the distribution of the mean converges to a normal distribution as the sample size (n) gets larger, even if the data itself does not follow a normal distribution. Given that the Proportion data $X_1, ..., X_n \overset{iid}{\sim} Poisson(\lambda)$, we can use a normal distribution to determine the test statistic for the hypothesis test. All derivations regarding the test statistic can be found in *Section 3: Hypothesis Test Test Statistic Computation* of the Appendix.

Our hypotheses:

$$H_o : \mu = 20$$

We hypothesize that the population mean of the Proportion data is equal to 20.

$$H_A : \mu \neq 20$$

The alternative hypothesis is that the population mean of the Proportion data is not equal to 20.

The following criteria will be used in determine whether the p-value calculated will cause us to reject or accept the null hypothesis, H_o. We will assume an α value of 0.05, which is standard practice.

$\alpha > P - Value$	Reject H_o
$\alpha < P - Value$	Do Not Reject H_o

Goodness of Fit Test

A goodness of fit test is a method of determining how well given data fits a distribution. A goodness of fit test highlights the discrepancies between the observed values within the test and the expected values under a distribution - which is Poisson in this case [16]. We will be using a goodness of fit test to test if all probabilities within the data are equal.

Similar to the hypothesis test explained above, goodness of fit tests have a null hypothesis, and an alternative hypothesis. Two goodness of fit tests will be run - a Chi-Squared test for age, and a Chi-Squared test for continent. A Chi-squared test is a method of computing a test statistic, which compares two variables to see if they are related or are independent [20]. The formula to determine the test statistic is $X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$, where X^2 is the Chi-squared test statistic, O_i is the observed value, and E_i is the expected value. We will be using code instead to calculate the Chi-squared test statistic.

The null hypotheses in both cases hypothesize that proportions of violence are equal amongst groups, while the alternative hypothesizes that they are not. The respective null and alternative hypotheses are below.

H_o : women of all ages are equally subject to violence. H_A : women of all ages are not equally subject to violence.

H_o : women from all continents are equally subject to violence. H_A : women from all continents are not equally subject to violence.

After the test statistic is calculated, a p-value will be calculated, and the p-value will be used to determine whether we will reject or accept the null hypothesis, H_o. The α value will be 0.05. For a description of what a test statistic and p-value is, or the criteria that will be used in determining the acceptance/rejection of the null hypothesis, look at the *Hypothesis Test* section above.

Bayesian Credible Interval

A Bayesian credible interval is an interval where an unobserved parameter value falls with a probability, which is used to say something about the parameter [21]. Credible intervals have more ideal interpretations compared to frquentist confidence intervals. They are determined using a posterior probability distribution which combines an assumed prior distribution and a likelihood function containing relevant data.

Suppose our data is a random sample of Poisson random variables with mean λ; and the prior distribution of λ is assumed to be Exponential(λ) in hopes of yielding a neutral/non-informative prior. The posterior distribution of λ is Gamma with $\alpha = \bar{X} + 1$ and $\beta^* = \frac{\beta}{n\beta+1}$. Thus, we can use the 0.5, 2.5, 97.5 and 99.5 percentiles of this distribution to derive a range of values, in which λ has a 95% and 99% probability of falling into. All derivations regarding the posterior distribution can be found in *Section 2: Posterior Derivations* of the Appendix.

The Exponential prior was chosen so that given the likelihood, the posterior will also be within the same distribution family as the prior. This is known as a conjugate prior. Working with priors with known conjugate posteriors allows us to simply modify the parameters in the prior distribution instead of computing often tedious integrals [12].

Assume we are interested in finding a 95% and 99% credible interval of the parameter λ. Using $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$, we will find the 2.5, 0.5, 97.5 and 99.5 percentiles to find the $100(1-\alpha)\%$ credible interval for λ, where $\alpha = 1 - CI$. The credible intervals 95% and 99% will be used in accordance with the 95% and 99% confidence intervals described prior.

Results

The following section will highlight the computational results of the methodologies that were explained in the Methods section.

Linear Regression 1

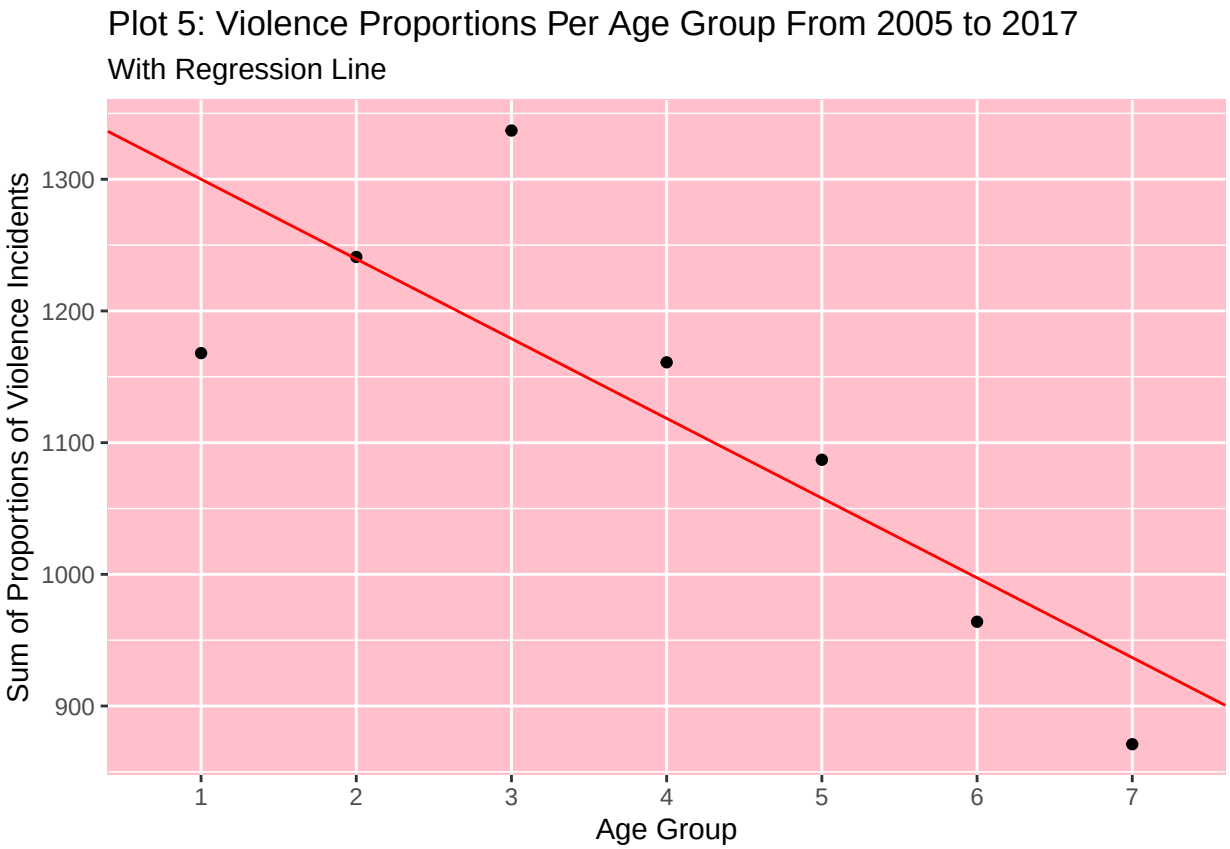
Table 1: Linear Regression Values

Intercept	$\hat{\alpha}$	-60.54
Slope	$\hat{\beta}$	1360.57

$$Y_i = -60.54 + 1360.57x_i + U_i$$

Table 1 above displays the values of the components of our linear regression model; $\hat{\alpha}$ (y-axis intercept) and $\hat{\beta}$ (slope of line). The line below the table is the mathematical model for the linear regression. It is difficult to intuitively interpret $\hat{\alpha}$ and $\hat{\beta}$, but to relate it to this analysis, the value 1360.57 is the proportion of violence in an age group before to age group 1, and the value -60.54 tells us that for every age group, building from the previous, the proportion of violence decreases.

Plot 4 above displays the relationship between the number of proportions of violence and the age group of the female victims in those incidents globally from 2005 to 2017. We can see that the trend of the data is inversely proportional after age group 3 - as the age group increases as the sum of proportions decreases, and is proportional before age group 3 - as the age group increases, so does the sum of proportions. A pattern in the data to notice is that the largest sum of violence proportions, around 1,375 occur within age group 3, between ages 25-29.



In Plot 5 above, we can see the regression line in red, overlaying the scatterplot of the sum of violence proportions per age group. You will notice that the linear regression line does not fit the data well at all; infact it 6/7 of the data points. If you compare the regression line to the data points, you can see the many deviations of the data points from the regression line, which are referred to as errors (they show us how off the model is from the actual data). In this case, it seems better fit to use a nonlinear regression line instead. There is no linear relationship between the proportion of violence incidents and age group, but a curvical one.

In Plot 6 above, we can see the curve of best fit line in red, overlaying the scatterplot of the sum of violence proportions per age group. Compared to Plot 5 with the regression line, we can see that this curve of best fit is better suited for the data, with minimal deviations from the data points (errors). Notice that as the age of occurence increases (age group), the proportions of women subject to violence within that age group decreases, and this is shown with the red line of best fit. With the equation of this curve, we could accurately estimate future values like the sum of proportions for ages above 49.

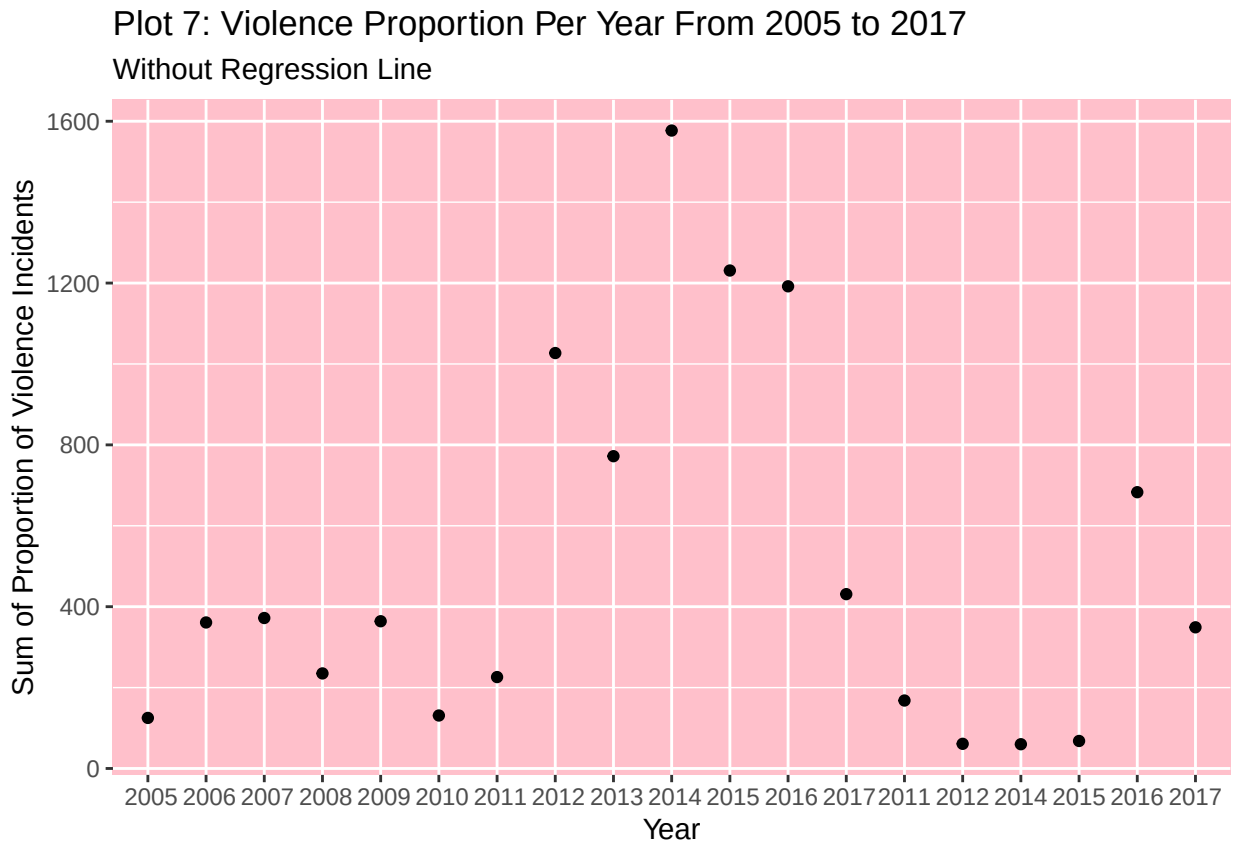
Linear Regression 2

Table 2: Linear Regression Values

Intercept	$\hat{\alpha}$	-94099.72
Slope	$\hat{\beta}$	47.02

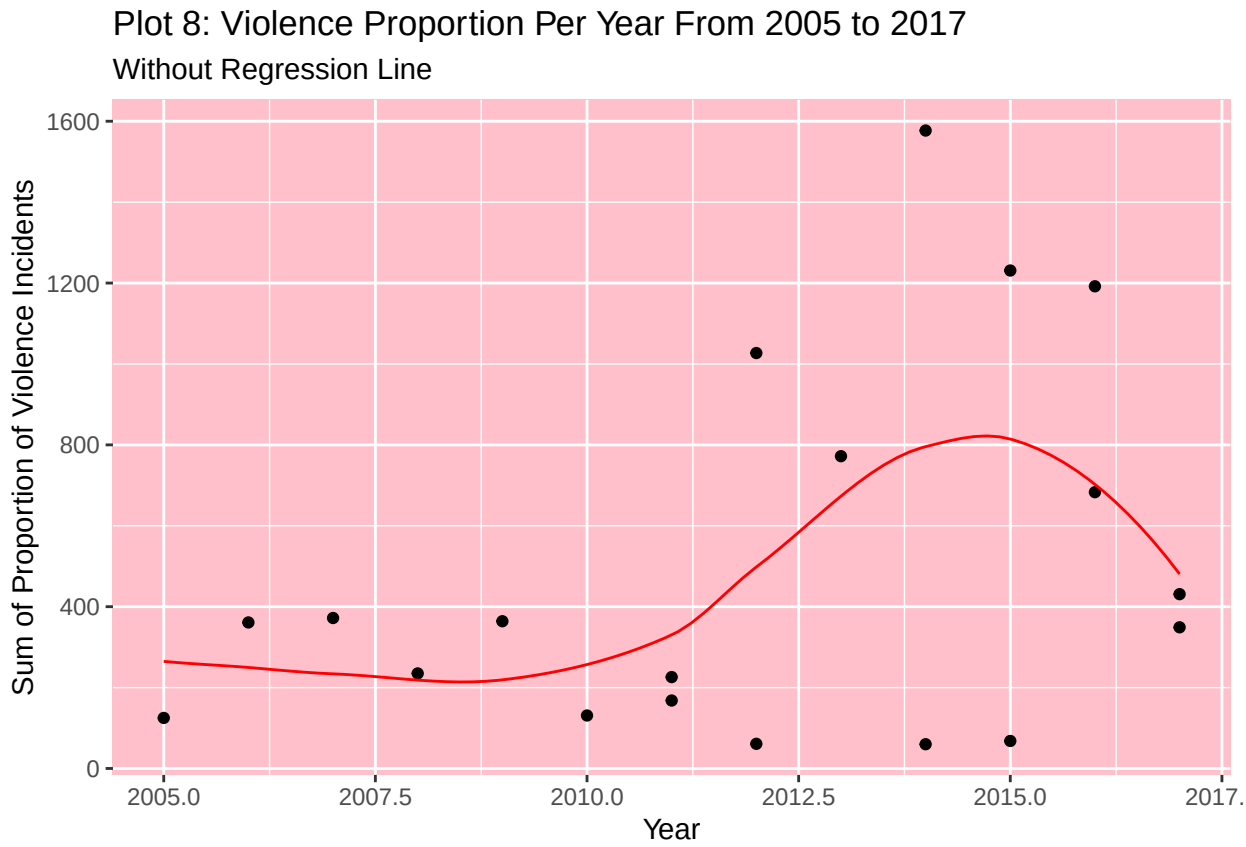
$$Y_i = -94099.72 + 47.02x_i + U_i$$

Table 2 above displays the values of the components of our linear regression model; $\hat{\alpha}$ (y-axis intercept) and $\hat{\beta}$ (slope of line). The line below the table is the mathematical model for the linear regression. It is difficult to intuitively interpret $\hat{\alpha}$ and $\hat{\beta}$, but to relate it to this analysis, the value -94099.72 is the proportion in the year before the first year accounted for in the data - 2004, and the value 47.02, tells us that for every year, building from the previous, the proportion of violence will increase.



Plot 7 above displays the relationship between the number of proportions of violence and the year of occurence globally from 2005 to 2017. We can see that the data points are scattered, with no definitive trend or relationship. We can deduce from the spread alone that a linear regression model will not be best fit to represent the relationship.

Since the y-intercept is so small (-94099.72), it will be difficult to display the linear regression line on the plot of year against proportions per year without making the results difficult to interpret. Thus, a plot with the linear regression model overlaying it will not be constructed, but instead a plot with a curve of best fit, to determine if a non-linear model is better suited to fit the data.



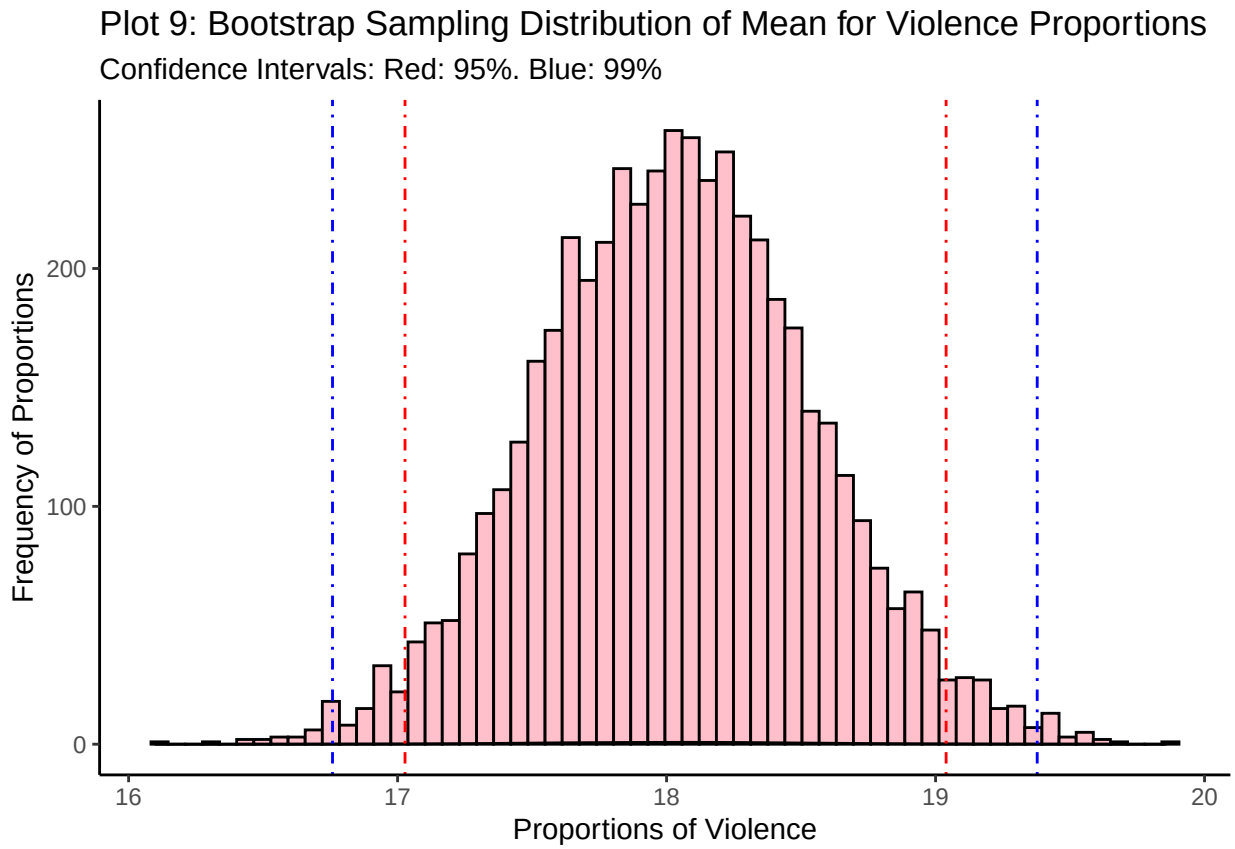
In Plot 8 above, we can see the curve of best fit line in red, overlaying the scatterplot of the sum of violence proportions per year. Compared to Plot 8 with the regression line, we can see that this curve of best fit is slightly better suited for the data, but not a good representation overall. Given that the data is so spiratic over the years, with no true trend, we would not be able to accurately estimate future values like the sum of proportions that would occur in years beyond 2017 using this model.

Confidence Interval

Table 3: Confidence Intervals for Mean

95% Confidence Interval	(17.02759, 19.03923)
99% Confidence Interval	(16.75804, 19.37793)

Table 3 above displays the confidence intervals for mean. A 95% confidence interval means that there is a 95% probability that the true population mean lies within the confidence interval of (17.02759, 19.03923). We are 95% certain that the true population mean of the population lies within the interval. Similarly, the 99% confidence interval of (16.75804, 19.37793) means that there is 99% probability that the true population mean of the population lies within the interval.



Plot 9 above displays the confidence intervals such that the true parameter of the mean would fall within either the 95% bounds or 99% bounds. There isn't a very large difference in the 95% and 99% confidence intervals which implies that the mean falls within a small range most of the time. Notice that the 99% interval is wider than the 95% interval. This is due to the increased confidence level that there is a 99% probability that the true mean of the population lies within the interval.

Maximum Likelihood Estimator

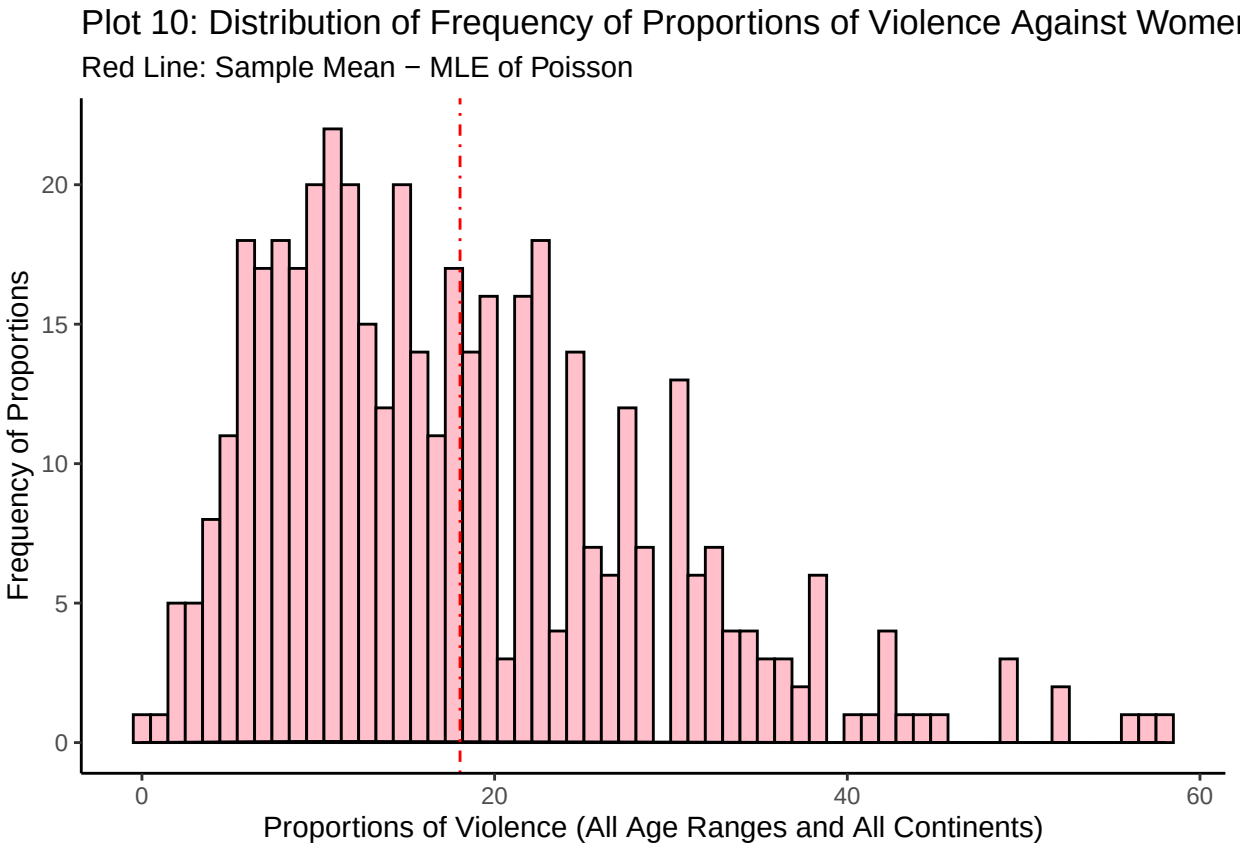


Table 4: Maximum Likelihood Estimator - Sample Mean

Sample Mean	18.03917

Plot 10 and Table 4 above depict the maximum likelihood estimator of the Poisson distributed data - the sample mean $\bar{x}$. Plot 11 shows a better visualization compared to Table 4 as it shows the estimator $\bar{x}$ against the rest of the data. Notice that the sample mean happens to fall within the bounds of both the 95% and 99% confidence interval in the section above.

Hypothesis Test

Table 5: Hypothesis Test of Mean

Test Statistic	-3.838554174
P-Value	0.0001237609

Table 5 above displays the test statistic and p-value of the hypothesis test conducted under the null hypothesis that $\mu = 20$. Since the p-value is less than the 0.05 threshold defined in the Methods section ($\alpha > P - Value$), we must reject the null hypothesis that the population mean of violence proportions is equal to 20 ($\mu = 20$).Thus, we accept the alternative hypothesis that the population mean of violence proportions is not equal to 20 ($\mu \neq 20$).

Goodness of Fit Test

Table 6: Chi Squared Goodness of Fit Test for Age

Test Statistic	-Infinity
P-Value	1

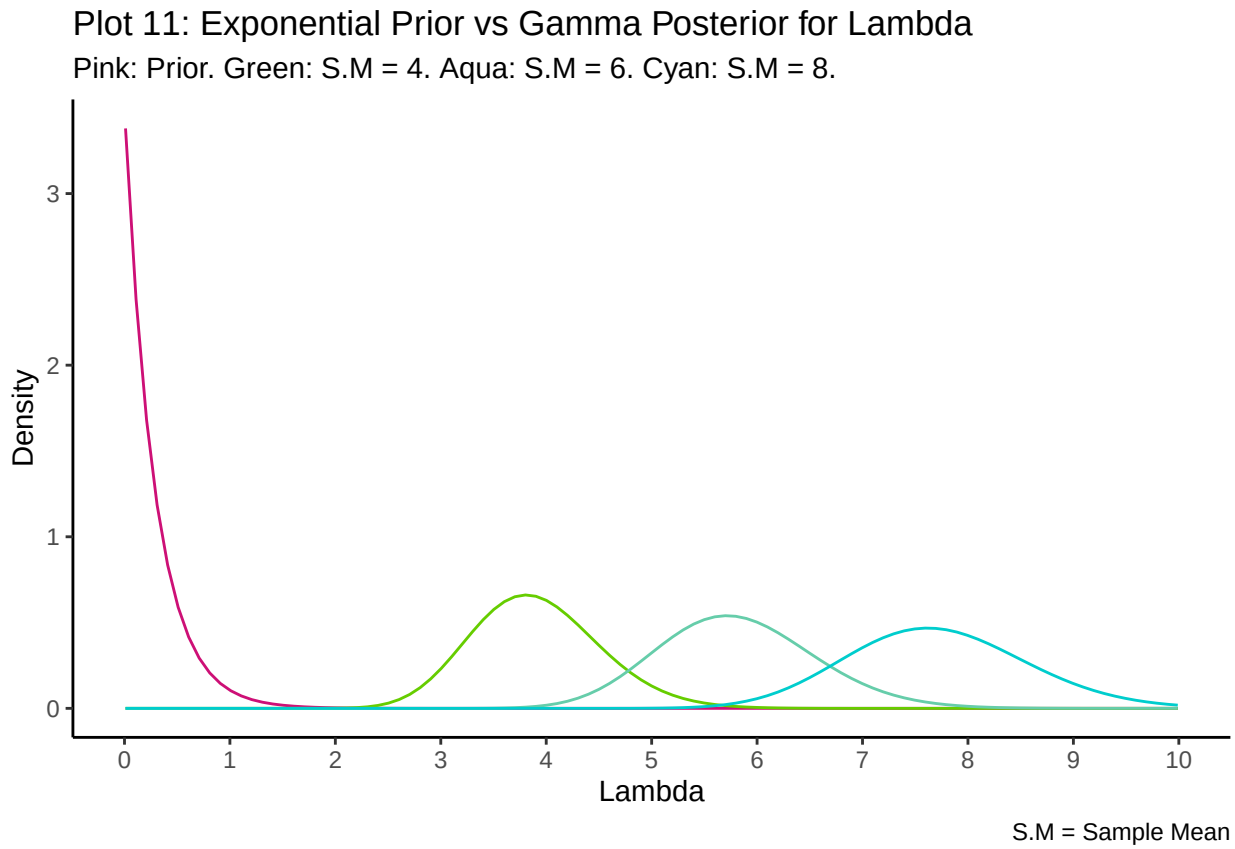
Table 6 above contains the test statistic and p-value of the Chi Squared goodness of fit test conducted for age. The p-value obtained is 1, which is greater than the 0.05 threshold defined in the Methods section ($\alpha > P - Value$). Thus we must accept the null hypothesis that women of all ages are equally subject to violence, and reject the alternative hypothesis that women of all ages are not equally subject to violence.

Table 7: Chi Squared Goodness of Fit Test for Continent

Test Statistic	Nan (0/0)
P-Value	Nan (0/0)

Table 7 above contains the test statistic and p-value of the Chi Squared goodness of fit test conducted for continent. The P-value obtained is undefined (0/0), which means that p_o and $\hat{p}$ in the likelihood ratio were 0. Intuitively, the p-value obtained is 0, which is lesser than the 0.05 threshold defined in the Methods section ($\alpha < P - Value$). Thus we must reject the null hypothesis that women from all continents are equally subject to violence, and accept the alternative hypothesis that women from all continents are not equally subject to violence.

Bayesian Credible Interval



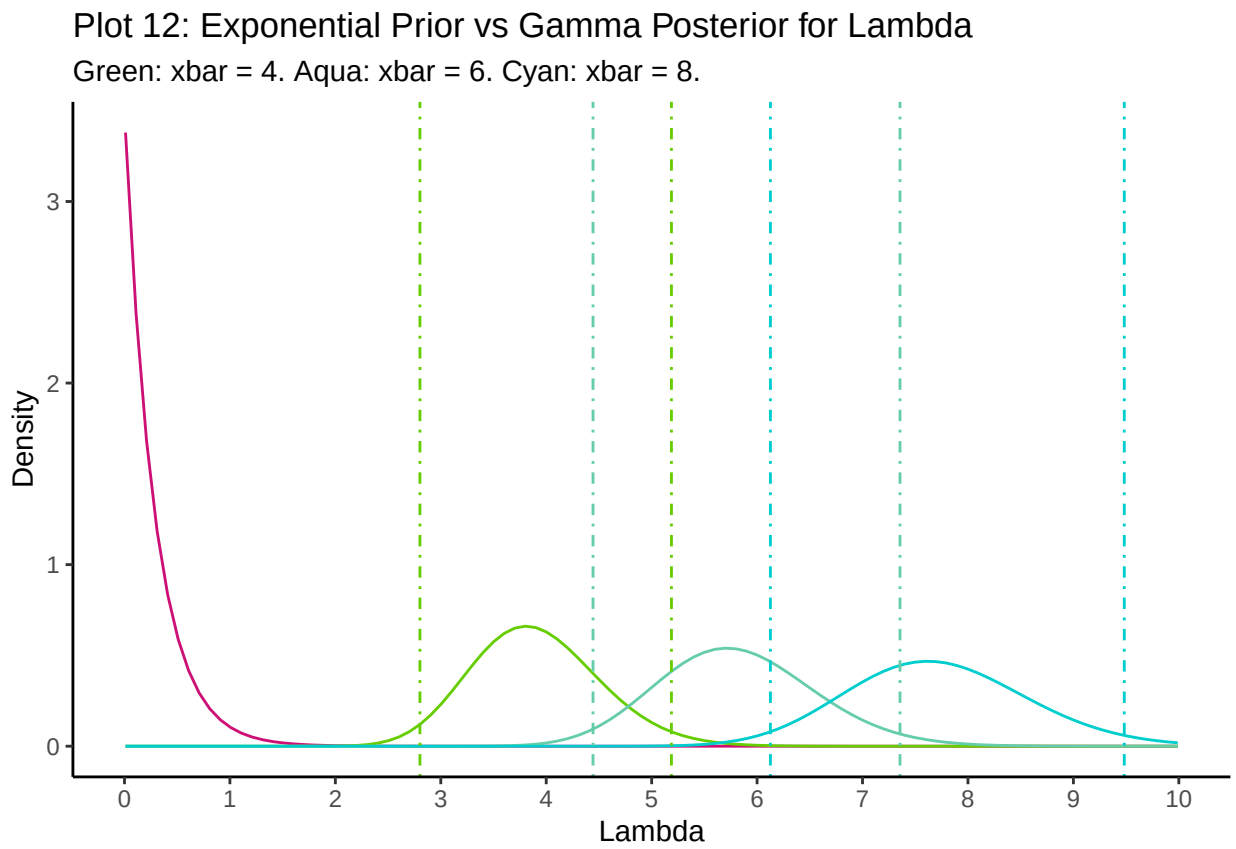
Plot 11 above depicts the exponential prior and gamma posteriors for a sample size of $n = 10$. You can see the drastic difference between the exponential prior curve, and the gamma curves influenced by the data.

For 10 surveyed women (n), with an average of 2 women being subject to violence (β), and a sample of women subject to violence at 4, 6, and 8 out of n women ($\bar{X}$), we can see that the likelihood of the number of women subject to violence decreases as the number of women subject to violence increases ($\bar{X}$). When the sample mean $\bar{X}$ is close to β, the density peak is higher, as observed in the green curve because the sample mean is approaching the true population mean.

Table 8: 95% Credible Interval with Varying Xbar Values

$\beta = 4$	(2.802125, 5.18749)
$\beta = 6$	(4.443825, 7.355418)
$\beta = 8$	(6.126256, 9.482564)

Table 8 above depicts the 95% credible intervals for the $\bar{x}$ values used in plot 11: 4, 6 and 8. Intuitively, these intervals tell us that there is a 95% probability that the parameter of interest λ is between (2.802125, 5.18749) when $\bar{x}$ is 4, between (4.443825, 7.355418) when $\bar{x}$ is 6, and between (6.126256, 9.482564) when $\bar{x}$ is 8. These intervals will be used to take a look at the central 95% of the posterior distribution curves in Plot 11.

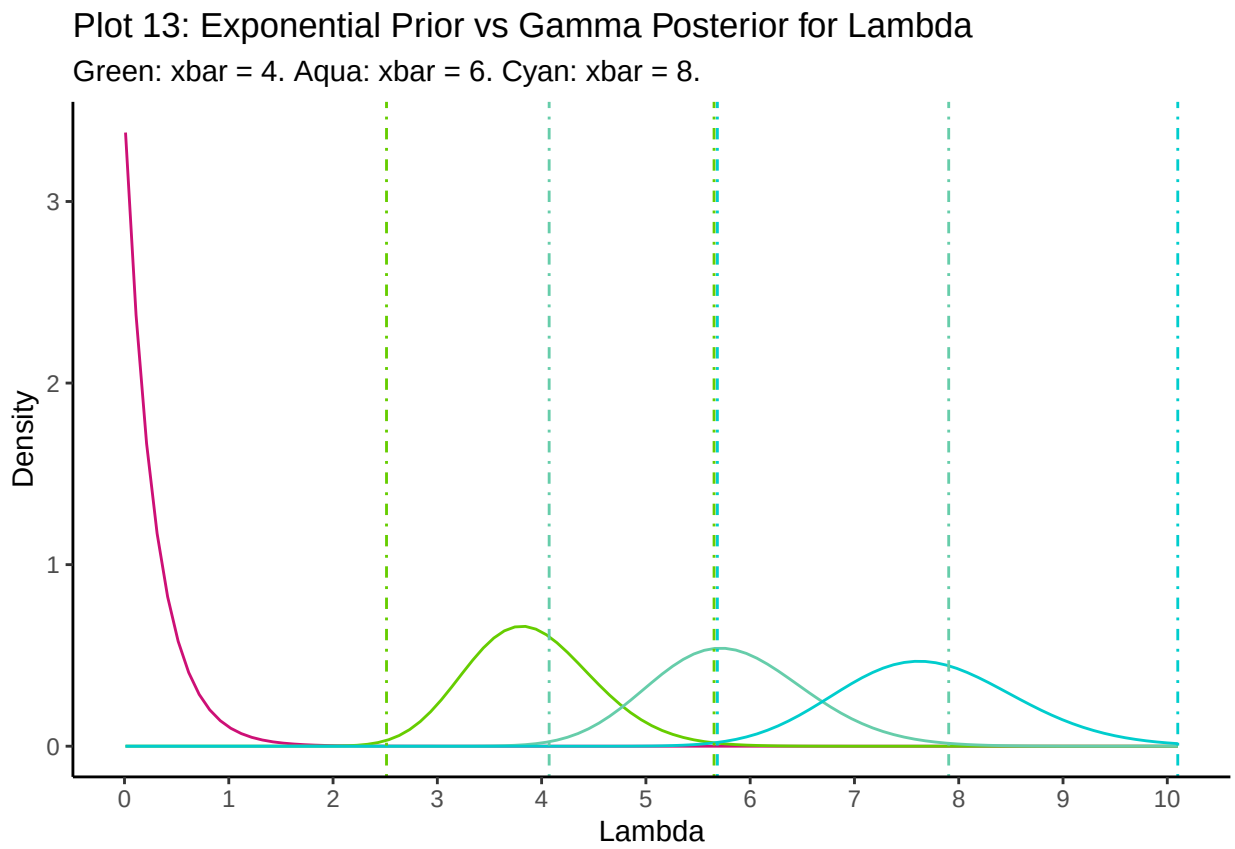


Plot 12 above shows the same curves of the prior and posterior distributions from Plot 11, with the three credible intervals from Table 8 for each value of $\bar{x}$ (4, 6, 8). As you can see, the intervals display the central 95% of each posterior gamma distribution, telling us that there is a 95% probability that the parameter of interest λ falls between each division. Notice the interval gets wider as $\bar{x}$ increases.

Table 9: 99% Credible Interval with Varying Xbar Values

$\beta = 4$	(2.512731, 5.653625)
$\beta = 6$	(4.072404, 7.903786)
$\beta = 8$	(5.685343, 10.10053)

Table 9 above depicts the 99% credible intervals for the $\bar{x}$ values used in plot 11: 4, 6 and 8. Intuitively, these intervals tell us that there is a 95% probability that the parameter of interest λ is between (2.512731, 5.653625) when $\bar{x}$ is 4, between (4.072404, 7.903786) when $\bar{x}$ is 6, and between (5.685343, 10.10053) when $\bar{x}$ is 8. These intervals will be used to take a look at the central 99% of the posterior distribution curves in Plot 11. Notice that the 99% credible intervals in Table 9 are wider than the 95% credible intervals. This is due to the increased confidence that there is a 99% probability that the parameter of interest - the true mean of the population lies within the interval.



Plot 13 above shows the same curves of the prior and posterior distributions from Plot 11, with the three credible intervals from Table 8 for each value of $\bar{x}$ (4, 6, 8). As you can see, the intervals display the central 99% of each posterior gamma distribution, telling us that there is a 99% probability that the parameter of interest λ falls between each division. Notice the interval gets wider as $\bar{x}$ increases.

Conclusions

The purpose of this analysis was to determine if there was a correlation between the proportion of women subject to violence and age and continent. We hypothesized that women in some residential continents are more prone to intimate partner violence compared to other continents, and that women within a certain age group are more prone to intimate partner violence than women of other age groups. In order to test this, we ran 2 linear regression tests, a hypothesis test of the mean, a goodness of fit test, determined the maximum likelihood estimator, and computed both critical and credible intervals.

A linear regression model of proportion and age, and proportion and year were created to determine if the variables have a linear relationship. The first linear regression model for proportions of violence and age did not fit the data, the data was better fit using a curvical model. However, based on the shape of the curvical model, there does seem to be a non-linear correlation between the proportions of violence and the age of occurence. As the age group of women increases, their proportion of violence in the data decreases. The second linear regression model for proportions of violence and age did not fit the data. Based on the scattered data points, it also looks as if there is no relationship between the proportions of violence and year of occurence.

A 95% and 99% confidence interval was calculated using a bootstrap simulation of 5000; an interval where there is a 95% or 99% probability that the true population mean lies in between the bounds. Since the given data is solely a sample of the entire population, a bootstrap of the mean was successfully ran. Both the 95% and 99% confidence intervals computed happened to contain the sample mean, 18.03917 and were narrowed

to a number around 18, suggesting that the population mean could be the same or a very similar value.

A hypothesis test was run to determine if the true population mean of proportion is equal to 20 (null) using a t-test. Since the p-value calculated from the t-test test statistic was less than the α cut-off, the null hypothesis that the true population mean of proportion is equal to 20 was rejected, and the alternative hypothesis that the true population mean of proportion is not equal to 20 was accepted.

The maximum likelihood estimator was calculated under the premise that the data follows a Poisson distribution. The maximum likelihood estimation (value that maximizes the likelihood function) of the Poisson distribution is λ, the sample mean. Applying the estimator to the data to compute an estimation, the sample mean of the data produced was 18.03917.

A goodness of fit test was run to determine if the distribution of proportions for age and continent of occurence were equal. For the first goodness of fit test, we hypothesized that the proportion of women subject to violence is equal for all age groups. After using a chi-squared test to determine the test statistic, the p-value computed was lesser than the α cut-off, allowing us to accept the null hypothesis that the proportion of women subject to violence is equal for all age groups. For the second goodness of fit test, we hypothesized that the proportion of women subject to violence is equal for all continents of occurence. After using a chi-squared test to determine the test statistic, the p-value computed was greater than the α cut-off, forcing us to reject the null hypothesis that the proportion of women subject to violence is equal for all continents of occurence and accept the alternative hypothesis that the proportion of women subject to violence is not equal for all continents.

A 95% and 9% Bayesian credible interval was computed using an assumed Exponential prior, an observed Poisson data distribution, and a computed Gamma posterior. Both credible intervals - similar to the confidence intervals - produced a range of values in which the true population mean could lie. The intervals were similar to that of the frequentist confidence interval, however the credible intervals are influenced by the prior distribution of exponential. The credible intervals gave us a 95% and 99% probability that the true population mean would fall within the bounds of the intervals, in alignment with the inputted of $\bar{x}$, n and β.

Based on the results of the analysis conducted using the above methods, it can be concluded that the proportions of violence are not equal for all continents of occurence. According to Plot 2, it seems that a majority of women subject to violence at the hands of their intimate partner reside in the African continent. This aligns with Lia Ryerson's findings, that some of countries that are most unsafe for women are in Africa. It can also be concluded that the proportions of violence are equal for all age groups. The proportions of violence show no 'favouritism' to a specific age group of women, but that as the age group of occurence increases, the proportions of violence tend to decrease. Additionally, it can be said that based on the confidence intervals calculated, out of 1000 women, 16% to 19% of them have been subject to intimate partner violence (population mean) in the last 12 months.

To reinstate, the purpose of this analysis was to raise awareness to the sexual and physical violence faced by women around the world by shedding light on the confounding factors associated with the proportions of women subject to violence globally.

Weaknesses

In calculating the maximum likelihood estimator and posterior, a weakness faced is the assumption that the data follows a Poisson distribution. Although by plotting, the data is similar in shape and spread to a Poisson distribution, there may have been a better fitting distribution for the data.

Another weakness faced is the assumption of the prior being an Exponential distribution in determining the posterior. The prior represents my assumption of the probability before the influence of data is taken into account.

Another weakness faced is the assumption that the sample mean converges to a normal distribution under the Central Limit Theorem in the hypothesis test test statistic calculation. Although assumptions under the Central Limit Theorem are subjective to the sample size, there is no substantial proof that the assumption made was right.

Additionally, another weakness faced is the fact that the data did not contain information for all countries within a continent. Thus making the data sample less representative of all continents.

Lastly, a weakness faced is the fact that the metadata for the intimate partner violence dataset from the World Health Organization failed to mention the size of surveyed individuals in which the Proportion column was made from. Thus, I chose the lowest power of ten sample size that could be assumed - 1000. Although the actual number of surveyed individuals was not used in any calculations, it would have been informative to know.

Next Steps

In regards to this analysis, moving forward, the violence faced by women in non-intimate partner relationships should be analysed. Often times, women are subject to violence at the hands of their immediate family members or legal guardians. This would be informative in understanding the hardships women endure at the hands of their loved one - not only significant partners but family as well. By comparing, we may be able to find a pattern and narrow down the age groups, continents and countries where women are most subject to violence.

Discussion

April is Sexual Assault Awareness Month. During this time, we must remember and stand by those who have been subject to sexual violence at the hands of those they most trust; family, friends and intimate partners. We must highlight the sexual violence tragedies faced by individuals around the world, and the negative effects COVID-19 has on violence rates. According to a study conducted by Western University, 1 in 10 women in Canada are worried about the possibility of being subject to violence during COVID-19 [9]. 67% of Canadians know a woman who has been subject to physical and/or sexual abuse [10]. These numbers are alarming, and are constantly increasing. Given the current state of the COVID-19 pandemic stay-at-home orders, many women are trapped inside with their abuser, being unable to seek help even via a hotline [11]. The negative effects of COVID-19 on job and school stress is also a dangerous factor in triggering violence in individuals, often experienced by their partners. Alhough the nation wide stay-at-home orders are intended to keep us all safe from the virus, we must acknowledge and equally dangerous virus of violence that many women are now facing being at home.

This analysis aims to educate and raise awareness on the intimate partner violences - both sexual and physical - faced by women around the world over the years. We all have a role to play in ending domestic violence and protecting women in Canada, and globally. For more information on gender based violence, and how you can help, visit https://canadianwomen.org/the-facts/gender-based-violence/.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)
2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)
4. Ryerson, L. (2019, March 8). *The 10 worst countries in the world for women.* Insider. https://www.insider.com/worst-countries-for-women-2018-3
5. World Health Organization. (n.d.). *Indicator Metadata Registry Details.* WHO. Retrieved April 5, 2021, from https://www.who.int/data/gho/indicator-metadata-registry/imr-details/3745
6. World Health Organization. (2019, August 20). *GHO | By category | Intimate partner violence prevalence - Data by country.* WHO. https://apps.who.int/gho/data/node.main.IPV?lang=en

7. R - Pie Charts. (n.d.). *Tutorials Point.* Retrieved April 11, 2021, from https://www.tutorialspoint.com/r/r_pie_charts.htm

8. Orloff, J., & Bloom, J. (2014). *Conjugate priors: Beta and normal* [Slides]. Mit.Edu. https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading15a.pdf. (Last Accessed: March 5, 2021)

9. Department of Communications and Public Affairs, Western University. (n.d.). *Western University.* Uwo.Ca. Retrieved April 19, 2021, from https://www.uwo.ca/

10. Howard, J. (2021, January 13). *Gender Based Violence in Canada | Learn the Facts.* Canadian Women's Foundation. https://canadianwomen.org/the-facts/gender-based-violence/

11. Evans, M., Lindauer, M., & Farrell, M. (2020, December 10). *A Pandemic within a Pandemic — Intimate Partner Violence during Covid-19.* The New England Journal of Medicine. https://www.nejm.org/doi/full/10.1056/NEJMp2024046

12. Orloff, J., & Bloom, J. (2014). *Conjugate priors: Beta and normal* [Slides]. Mit.Edu. https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading15a.pdf. (Last Accessed: March 5, 2021)

13. *Finding the MLE of Poisson in R.* (2019, April 17). Stack Exchange. https://stats.stackexchange.com/questions/403483/finding-the-mle-of-poisson-in-r

14. Taboga, Marco (2017). *Poisson distribution - Maximum Likelihood Estimation*, Lectures on probability theory and mathematical statistics, Third edition. Kindle Direct Publishing. Online appendix. https://www.statlect.com/fundamentals-of-statistics/Poisson-distribution-maximum-likelihood.

15. dataminingincae. (2015, August 30). Video 1: *Introduction to Simple Linear Regression* [Video]. YouTube. https://www.youtube.com/watch?v=owI7zxCqNY0&ab_channel=dataminingincae

16. Kenton, W. (2021, April 18). *Goodness-Of-Fit.* Investopedia. https://www.investopedia.com/terms/g/goodness-of-fit.asp

17. Wikipedia contributors. (2021, April 16). *Sexual Assault Awareness Month.* Wikipedia. https://en.wikipedia.org/wiki/Sexual_Assault_Awareness_Month

18. Katz, A. (n.d.). *Maximum Likelihood Estimation (MLE) | Brilliant Math & Science Wiki.* Brilliant. Retrieved April 19, 2021, from https://brilliant.org/wiki/maximum-likelihood-estimation-mle/

19. Bevans, R. (2021, January 7). *Test statistics explained.* Scribbr. https://www.scribbr.com/statistics/test-statistic/

20. *Chi-Square Statistic: How to Calculate It / Distribution.* (n.d.). Statistics How To. Retrieved April 19, 2021, from https://www.statisticshowto.com/probability-and-statistics/chi-square/

21. Wikipedia contributors. (2020, September 1). *Credible interval.* Wikipedia. https://en.wikipedia.org/wiki/Credible_interval

Appendix

Section 1: MLE Derivation

Keep in mind the probability mass function of the Poisson distribution, $P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$.

$$L(\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Assuming that $X_1, ..., X_n \overset{iid}{\sim} Poisson(\lambda)$:

$$L(\lambda|X_1, ..., X_n) = \Pi_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

Using the loglikelihood:

$$l(\lambda) = \log(\Pi_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!})$$

$$l(\lambda) = \sum_{i=1}^{n} \log(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!})$$

$$l(\lambda) = \sum_{i=1}^{n} x_i \log(\lambda) - \sum_{i=1}^{n} \lambda - \sum_{i=1}^{n} \log(x_i!)$$

$$l(\lambda) = \log(\lambda) \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \log(x_i!)$$

To determine the value of λ that maximizes the equation above, we must take a derivative, set the derivative equal to 0, and solve for λ.

$$\frac{d}{d\lambda} l(\lambda) = \frac{d}{d\lambda} (\log(\lambda) \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \log(x_i!))$$

The derivative with respect to λ:

$$\frac{d}{d\lambda} l(\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n$$

Setting the derivative with respect to λ equal to 0:

$$0 = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n$$

Solving for λ:

$$n\lambda = \sum_{i=1}^{n} x_i$$

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Simplifying the term on the right, we get:

$$\hat{\lambda} = \bar{x}$$

Thus, the value of λ that maximizes the equation is the estimator $\bar{x}$, the sample mean of n observations within the data.

Section 2: Posterior Derivation

Assume $X_1, ..., X_n \overset{iid}{\sim} Poisson(\lambda)$. We will use a Bayesian approach, where $\lambda \sim Exponential(\lambda)$ is the prior distribution, to draw a conclusion about λ. Keep in mind the probability mass function of the Poisson distribution, $P(y) = \frac{e^{-\lambda}\lambda^y}{y!}$, which becomes $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$, and the probability density function of the Exponential distribution, $f(y) = \frac{1}{\beta}e^{\frac{-y}{\beta}}$, which becomes $f(\lambda) = \frac{1}{\beta}e^{\frac{-\lambda}{\beta}}$.

We will be deriving the posterior distribution of λ below, using the Bayesian approach

$$P(\theta|data) = \frac{P(data|\theta)P(\theta)}{P(data)}$$

where $P(\theta|data)$ is the posterior distribution of parameter θ, $P(data|\theta)$ is the likelihood of our hypothesis, and $P(\theta)$ is the prior distribution of parameter θ. The posterior distribution will tell us the likelihood, given the new data. Since our parameter of interest is λ, the equation becomes:

$$P(\lambda|data) = \frac{P(data|\lambda)P(\lambda)}{P(data)}$$

Plugging in what we know about λ and the data:

$$P(\lambda|X_1, ..., X_n) = \frac{P(X_1, ..., X_n|\lambda)P(\lambda)}{P(X_1, ..., X_n)}$$

Since we know that $X_1, ..., X_n$ are indepedent and identically distributed variables, we can simplify $P(X_1, ..., X_n|\lambda)$ to $P(\sum_{i=1}^n x_i|\lambda)$.

$$P(\lambda|X_1, ..., X_n) = \frac{P(\sum_{i=1}^n x_i|\lambda)P(\lambda)}{P(\sum_{i=1}^n x_i)}$$

Pluging in the values for λ:

$$P(\lambda|X_1, ..., X_n) = \frac{(\frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{\sum_{i=1}^n x_i!})(\frac{1}{\beta}e^{\frac{-\lambda}{\beta}})}{P(\sum_{i=1}^n x_i)}$$

Since we are solely interested in the posterior distribution of λ, we can treat all other terms as the normalizing constant, C.

$$P(\lambda|X_1, ..., X_n) = C(e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}e^{\frac{-\lambda}{\beta}})$$

Simplifying the term on the right side, we get:

$$P(\lambda|X_1, ..., X_n) = C(e^{-n\lambda-\frac{\lambda}{\beta}}\lambda^{\sum_{i=1}^n x_i})$$
$$P(\lambda|X_1, ..., X_n) = C(e^{-\lambda(n+\frac{1}{\beta})}\lambda^{\sum_{i=1}^n x_i})$$

To determine the posterior distribution of λ, we want to find a distribution which resembles the form $e^{-variable}variable^{number}$. Notice, the gamma probability distribution is

$$f(y) = [\frac{1}{\Gamma(\alpha)\beta^\alpha}]e^{\frac{-y}{\beta}}y^{\alpha-1}$$

where $\frac{-y}{\beta}$ is equal to $n + \frac{1}{\beta}$, where $\beta = \frac{\beta}{n\beta+1}$ and $\alpha - 1$ is equal to $\sum_{i=1}^n x_i$, where $\alpha = \sum_{i=1}^n x_i + 1$.

Something to notice is that the Exponential distribution is a special case of the Gamma distribution when $\alpha = 1$. This is shown below.

$$Gamma(y) = [\frac{1}{\Gamma(\alpha)\beta^{\alpha}}]e^{\frac{-y}{\beta}}y^{\alpha-1}$$

$$f(y) = [\frac{1}{\Gamma(1)\beta^{1}}]e^{\frac{-y}{\beta}}y^{1-1}, \alpha = 1$$

$$f(y) = \frac{1}{\beta}e^{\frac{-y}{\beta}}$$

$$Exponential(y) = \frac{1}{\beta}e^{\frac{-y}{\beta}}$$

This means that the prior and posterior distributions of λ are in the same distribution family, making them conjugate distributions.

$$\lambda \sim Gamma(\alpha = \sum_{i=1}^{n} x_i + 1, \beta^* = \frac{\beta}{n\beta + 1})$$

Which is equivalent to:

$$\lambda \sim Gamma(\alpha = \bar{X} + 1, \beta^* = \frac{\beta}{n\beta + 1})$$

Note: β^* represents the posterior parameter in $Gamma(\alpha, \beta)$, while β represents the prior parameter in $Exponential(\beta)$.

By taking a look at the shape and scale parameters of Gamma - α and β - notice that α is a function of the sample mean ($\bar{X}$), which is reliant on the sample size (n), and β^* is a function of the sample size (n) and the prior parameter β. The Gamma posterior is the Exponential prior distribution of λ after collecting relevant data ($\bar{X}$), being the likelihood.

For a better understanding of the posterior parameters:

- n: amount of data, the number of years, locations and ages of violence data observed and collected - *sample size*.
- β^*: true average number of women subject to intimate partner violence; mean of the prior distribution's mean - *true mean*.
- $\bar{X}$: sample average number of women subject to intimate partner violence - *sample mean*.

We can deduce from this that the shape parameter α - which depicts the fundamental shape of a Gamma distribution - is a function of the data. α relies on the sample mean of women subject to violence ($\bar{X}$), which relies on the sample size of the data (n). The scale parameter β^* - which depicts the stretch or compression of a Gamma distribution - is a function of the sample size of the data (n).

Section 3: Hypothesis Test Test Statistic Computation

According to the Central Limit Theorem (CLT), the distribution of the mean converges to a normal distribution as the sample size (n) gets larger, even if the data itself does not follow a normal distribution. Given that the Proportion data $X_1, ..., X_n \overset{iid}{\sim} Poisson(\lambda)$, we can use a normal distribution to determine the test statistic for the hypothesis test.

A t-test will be used to compute the test statistic since the sample size is relatively large (434).

$$t = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

Plugging in the hypothesized sample mean ($\mu = 20$), the sample mean ($\bar{X} = 18.03917$), the standard deviation ($\sigma = 10.64158$), and the sample size ($n = 434$), the equation becomes:

$$t = \frac{18.03917 - 20}{\frac{10.64158}{\sqrt{434}}}$$

$$t = \frac{-1.96083}{0.510825147}$$

$$teststatistic = -3.838554174$$

Thus, the test statistic under the null hypothesis $\mu = 20$ is -3.838554174.