Analyzing Characteristics of Mass Shootings in the USA, 1966-2021.

Adele Lauzon - 1005695801

April 19, 2021

Abstract

This report analyzes characteristics of mass shootings in the United States between August 1, 1966, and March 31, 2021, and unpacks the "troubled-young-white-man" narrative that often surrounds mass shooters (Paterson, 2019). Combining two datasets of mass shootings, the effects of temporal factors and mass-shooter demographics on the results of shootings are analyzed using a variety of different methodologies, including but not limited to linear regression, hypothesis testing, and goodness-of-fit testing. Among other results, it is found that shootings likely are equally probable throughout the week, mass shooters are likely in their late twenties/early thirties, and inflict on average between 5 and 6 deaths. Further research is necessary, but a strong understanding of those who are at risk of causing mass shootings and the kind of havoc they wreak may motivate impactful legislation to reduce the occurrence of these tragic events.

Introduction

Since 1966, there have been over 300 mass shootings in the United States, resulting in 1,294 deaths (Berkowitz & Alcantara, 2021). While mass shootings account for a small fraction of the yearly death toll, they are uniquely terrifying–unlike other acts of violence, victims of mass shootings are often sought out simply for where they happen to be at a given moment. Reading news stories, it can seem impossible to draw commonalities between these incidents beyond their devastating outcomes–men, women, children, grocery stores, movie theaters, schools–it appears as though no one and no place is immune to gun violence.

Unfortunately, it seems as though mass shootings are happening more and more often. The 5 deadliest shootings in recorded history–ranging from 26 to 58 deaths–all took place after 2007 ("Mass Shootings in the United States", 2021). In fact, mass shootings have tripled in frequency in recent years (Cohen, Azrael, & Miller, 2014).

Though these acts of violence may appear senseless and arbitrarily random, it is possible that understanding past shootings may aid in preventing future ones. Asking important questions–like what aspects of the shooter's life motivated them to kill, how they accessed their weapons, etc.–can motivate future policies that may reduce the probability of similar events recurring.

This report will aim to answer the question of how temporal aspects (day of the week) and demographic characteristics of the shooter (age, race, sex, and mental health of the shooter) impact the outcome of the shooting (number of fatalities, total victim count, and mortality rate). Specifically, we will unpack the "troubled young white man" narrative that often surrounds mass shooters (Siemaszko, 2021).

It is hypothesized that the typical shooter is a young white man with a history of mental illness, and that shooters that meet this criteria perpetrate shootings with the highest death toll and victim counts. Additionally, it is predicted that mass shootings result in an average of 5 deaths per incident (de Jager, Goralnick, & McCarty, 2018).

The following terms will be used throughout this report:

A *Mass Shooting* is defined as an incident of gun violence with at least four victims at one or more locations that are near one another (Berk, 2018).

A *Semi-Automatic* weapon is one fires once per trigger pull, but does not require the user to reload the gun between each shot (Merriam-Webster). The purchase of semi-automatic weapons is legal in most states ("Assault weapons legislation in the United States", 2021).

Data

This report uses two datasets–the Stanford Mass Shootings in America (MSA) dataset from the Stanford University Geospatial Center, and the US Mass Shootings, 1982-2021 dataset from nonprofit news organization Mother Jones. Both projects began in 2012–the MSA in response to the Sandy Hook Elementary shooting (the fourth deadliest shooting on record), and the Mother Jones dataset in response to a movie theatre shooting in Aurora, Colorado. The information in both databases was collected using online news coverage of mass shootings. The MSA starts with the University of Texas Tower shooting in 1966, but the project was archived in 2016 due to the resources required for adequate maintenance; in contrast, the earliest incident in the Mother Jones dataset is from 1982, but the project is continuously updated. Therefore, in order to get the most complete dataset possible, this report combined the two, using all of the entries in the MSA, and then using entries from 2016 onward from Mother Jones. This new dataset spans from 1966 to 2021, and contains 262 mass shooting incidents.

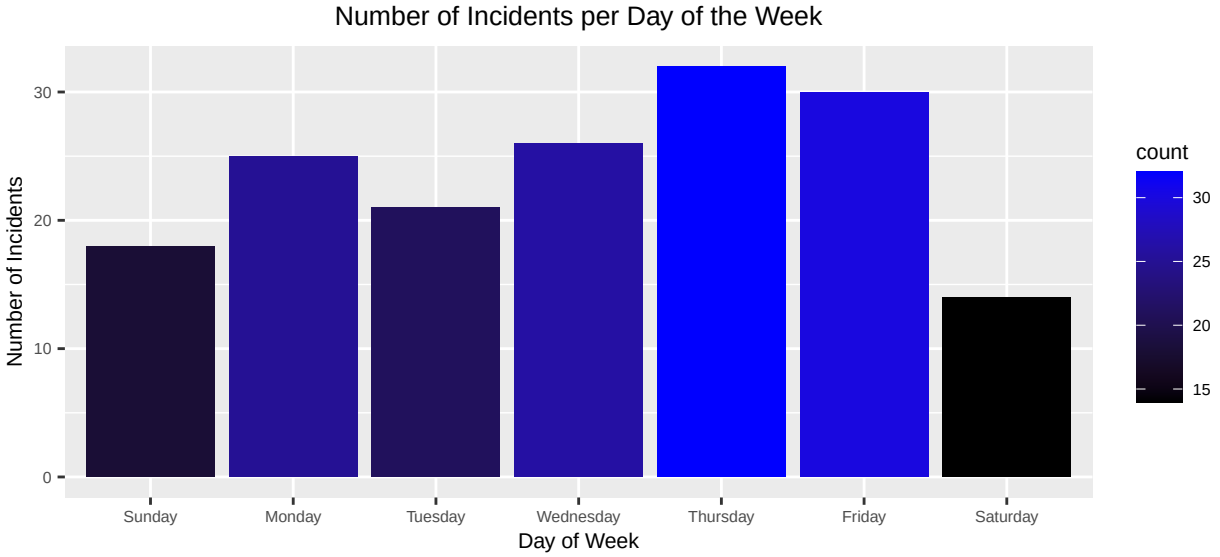
This new dataset required some cleaning; firstly, to remain consistent with the current definition of mass shooting, all incidents with fewer than 4 victims were removed. Additionally, all incidents with more than one shooter were removed. Finally, all entries that had missing values for any of the selected variables (see below) were also removed for more seamless calculations. This final dataset has 166 entries.

The following variables were extracted from the MSA and the Mother Jones project, and therefore are present in the dataset used for this report.

- **Fatalities:** Discrete numeric variable representing the total number of people who died due to the shooting.*
- **Total Victims:** Discrete numeric variable representing the total number of victims of the shooting, including survivors and the deceased.*
- **Mortality Rate:** Continuous numeric variable representing the mortality rate of the shooting, equal to $\frac{\text{Fatalities}}{\text{Total Victims}}$.*
- **Age:** Discrete numeric variable representing the age of the shooter.
- **Race:** Categorical variable representing the race of the shooter, taking on one of the following: Asian, Black, Indigenous, Latino, White, Unknown, or Other.
- **Day of Week:** Categorical variable representing the day of the week on which the shooting occurred.
- **Semi-automatic:** Variable representing whether semi-automatic weapons were used in the incident, taking on Yes if at least one semi-automatic weapon was used, and No if none were used, and $Unknown$ if the answer is unknown.

* *Note:* Perpetrators who were wounded or died as a result of the shooting are not included in **Fatalities**, **Total Victims**, or **Mortality Rate**.

Categorical Summaries



The bar chart above details the number of shootings per day of the week. Thursday has the most incident occurrences (32), and Saturday has the least (14). Observe that from Sunday to Friday, there appears to be a slight, gradual increase in the number of incidents; however, there is a sharp decline between Friday and Saturday.

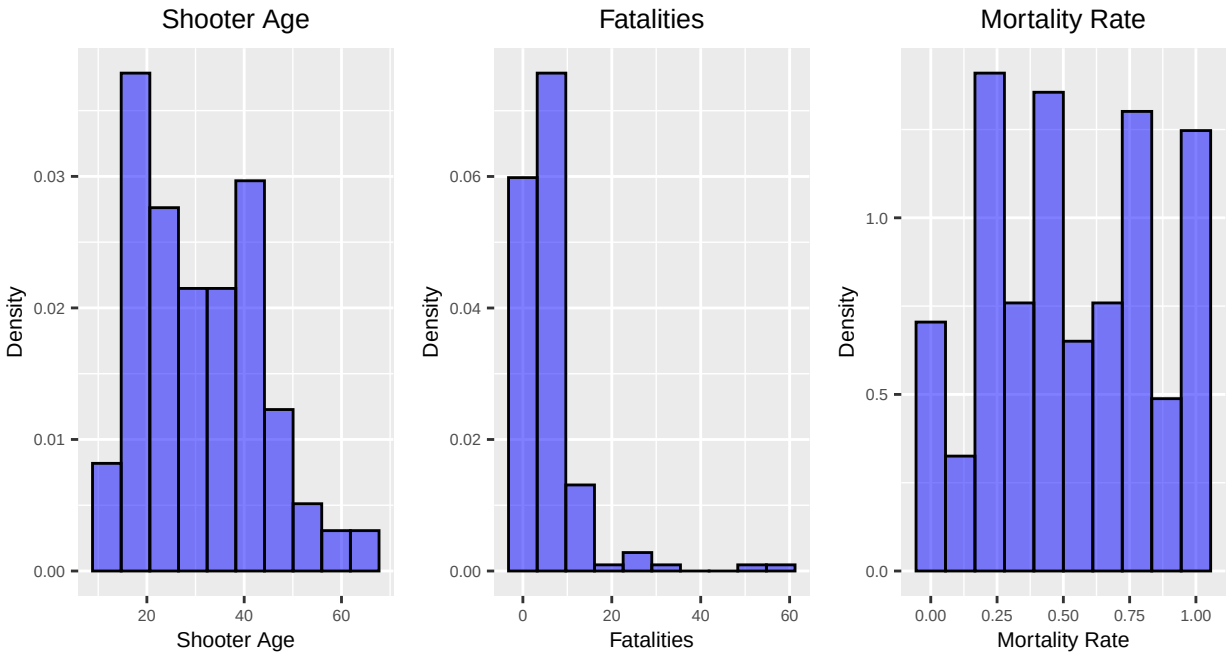
Numerical Summaries

Below is a table depicting numerical summaries for each of the numeric variables:

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
Age	13.00	20.00	29.00	31.09	40.00	66.00
Total Victims	4.00	5.00	7.00	14.78	11.00	604.00
Fatalities	0.000	2.000	4.500	5.934	7.000	58.000
Mortality Rate	0.0000	0.2500	0.5000	0.5371	0.7944	1.0000

There are a few takeaways from these broad summary statistics:

- The youngest shooter in this dataset was only 13 years old, and the oldest was in his sixties (66). The mean age of shooters (31.09) is a little bit higher than the median (29), indicating that the distribution of this variable is likely skewed to the right.
- The spread for the total victims is quite large, spanning 4 victims to 604 victims. It appears as though this variable is right-skewed, seeing as the bottom 50% of incidents had between 4 and 7 total victims, but the second 50% had between 7 and 604. In other words, the upper 50% of incidents has a much larger spread than the bottom 50%. This skew is corroborated by the fact that mean (14.78) is over twice the the value of the median (7). Additionally, it follows from the definition of mass shooting (incident with at least 4 victims) that the minimum value for this variable is 4.
- Fatalities is also slightly right skewed, with a mean (5.934) slightly higher than the median (4.500). Additionally, the spread from 0 to 58 indicates that there were incidents with no fatalities, and the incident with the highest death toll took 58 lives.
- The Mortality rate spreads from 0 to 1, which is understandable seeing as it is a proportion. Additionally, the median of 0.5 indicates that in 50% of the incidents, at most 50% of victims passed away.



The above plots are density histograms for Shooter Age, Fatalities, and Mortality Rate. Notice the following:

- As previously mentioned, *Shooter Age* appears to be unimodal and right-skewed, with most of the data falling between 15 and 45. That range also captures the mean (31.09) and the median (29), which is logical.
- *Fatalaties* appears to be unimodal and right-skewed, more heavily than *Shooter Age*. While there were some more devastating events with 45 deaths or higher, most incidents resulted in between 0 and 15 deaths. Again, this range captures the mean (5.934) and the median (5).
- *Mortality Rate* may follow a uniform distribution, as there does not seem to be a distinct shape portrayed in the histogram, and the data appears to be evenly spread across its support (0 to 1). There does not appear to be any detectable skew.

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

Methods

There are six methods in this report that analyze the demographics of mass shooters and typical outcomes of mass shootings. The methods are:

- *Linear Regression.* Linear regression uses one variable to predict the results of another. This report uses Shooter Age to predict Fatalities.
- *Confidence Interval of the Mean.* This method uses bootstrapping to construct an interval of possible values of the population parameter. This report creates a 95% Confidence Interval for Mean Shooter Age.
- *Maximum Likelihood Estimator.* This is a method of estimating the parameters of the probability distribution of a certain variable. This report applies the MLE method to Mortality Rate.
- *Hypothesis Test of the Mean.* A hypothesis test is a method of determining whether a proposed population mean is probable, given our sample data. This report tests the hypothesis that the average number of fatalities in mass shootings with semi-automatic weapons is 4.25 (de Jager, Goralnick, & McCarty, 2018).
- *Goodness of Fit Test.* A goodness of fit test allows us to determine whether data follows a particular distribution; particularly, a Chi-Square goodness of fit test uses categorial data frequencies to determine how well a proposed distribution fits the sample. This report applies the Chi-Square GOF test to the variable Day (day of week of the shooting).
- *Bayesian Credible Interval.* A Bayesian Credible Interval is a method of creating an interval of probable values for a given variable. The difference between a 95% Bayesian Credible Interval and 95% Confidence interval comes down to the unique statistical definitions of probability and confidence–we can say that there is a 95% probability that the population parameter falls within our Bayesian interval, but we say that we are 95% confident that the population parameter falls within our confidence interval.

Linear Regression

As previously mentioned, one of the goals of this report is to identify relationships between demographic characteristics of shooters with the outcome of the incidents they cause. One way we can do this is through a Simple Linear Regression Model, using Shooter Age to predit Fatalities.

We will assume that for each observation, Shooter Age is nonrandom and Fatalities is a realization of a random variable Y_i satisfying

$$Y_i = \alpha + \beta x_i + U_i$$

for $i = 1, 2, ..., n$, where $U_1, ..., U_n$ are independent random variables with $E[U_i] = 0$ and $Var(U_i) = \sigma^2$.

Therefore, the line $y = \alpha + \beta x$ is our **Regression Line**.

Let's clarify what each of the symbols in the Simple Linear Regression Model means:

- α is the *y-intercept* of our regression line. In the context of our data, this means that when Shooter Age is 0, the model predicts that Fatalities will be α.
- β is the *slope* of our regression line. In the context of our data, this means that for every 1-year increase in Age, the model predicts that there will be an increase of β in the Fatalities.
- x_i is the independent variable; in our model, this is Shooter Age.
- U_i is the random fluctuation centered around the regression line $y = \alpha + \beta x$. It can be interpreted as the *error term.*
- Y_i is the random variable that we are estimating. In this case, this is Fatalities.

In a controlled setting with all other variables held constant, it is possible that Shooter Age and Fatalities may be correlated in some linear manner. Younger people tend to be more immature, lack impulse control, and may consume more violent media, which might result in a higher death toll in the incidents they perpetrate (Lankford & Hoover, 2018). The Results section will explore the strength of this relationship.

Confidence Interval

Recall from the **Data** section that the mean shooter age from this sample is 31.09. If we are interested identifying the true average age of mass shooters, we can construct a plausible range of ages using a Bootstrapped Confidence Interval.

Bootstrapping is a method of ascertaining estimations about a population by using a sample. In essence, bootstrapping is repeated sampling with replacement. Assuming that the sample is indeed representative of the population at large, bootstrapping simulates sampling from the population. The process occurs as such: a bootstrapped sample of the same size as the original sample (in this case, 166) is selected randomly, with replacement, *from the original sample*. The desired statistic–in this case, the mean age of shooters–is calculated for this boostrapped sample. Then, the process is repeated at a minimum of 1,000 times. From this, we obtain the boostrap sampling distribution. We then can calculate the 2.5th and 97.5th percentiles from this distribution, in order to create a *95% Confidence Interval*. We can then conclude with 95% confidence that the population parameter lies somewhere in this range of values.

Maximum Likelihood Estimator

Here, we will calculate a Maximum Likelihood Estimator for the continuous variable Mortality Rate. Assume Mortality Rate is a random sample of identically and independently distributed Uniform random variables, each following a $Unif(0, \theta)$ distribution. The maximum likelihood estimator (MLE) approach will be used to estimate θ.

The principle of maximum likelihood is that we choose our parameters of interest (in this case, θ) such that our data are *most likely*–in other words, we choose the value of θ that *maximizes* the likelihood function $L(\theta)$. *

The MLE of θ is the maximum of our sample Mortality Rates. All derivations regarding the MLE can be found in Section 1 of the Appendix.

* Recall that the Likelihood function is equal to $L(\theta) = f_\theta(x_1)...f_\theta(x_n)$, where n is the number of observations in our sample, and $f_\theta(x_i)$ is the probability density function of the variable of interest.

Hypothesis Test

A 2018 study found that on average, 4.25 people die in active shooter incidents involving semi-automatic weapons (de Jager, Goralnick, & McCarty, 2018).

From our sample, 123 out of 166 incidents involved at least one semi-automatic weapon. Looking only at these incidents, the mean number of fatalities per incident is $\bar{X} = 6.4553$, and the standard deviation is $s = 6.6397$.

We can use our modified sample (including only incidents involving semi-automatic weapons) to conclude if the claim that the mean number of fatalities in active shooter incidents is 4.25 is still reasonable in 2021.

We can perform a *hypothesis test* to test the claim. We will assume that the sample mean is normally distributed, and that $n = 123$ is considered small.

We will perform a hypothesis test with the following hypotheses:

$H_0 : \mu = 4.25$, and

$H_A : \mu \neq 4.25$,

where μ is the *parameter*, or the population mean.

Our test statistic will be equal to $\frac{\bar{X}-\mu}{s/\sqrt{n}}$, and it will follow a t distribution with 122 degrees of freedom.

Goodness of Fit Test

Knowing whether or not shootings occur with similar frequencies regardless of the day of the week may impact future policies and resource distribution; for example, if it is found that shootings happen more frequently on the weekend, then security at large public places (malls, movie theaters, grocery stores, etc.) can be increased during those days.

One way we can go about answering this question–if shootings occur with the same frequency throughout the week–is through a Goodness of Fit Test to see if the data follows a Uniform Distribution (equal probability). While this will not tell us exactly what days are high-risk, it will tell us if there is at least one day of the week during which shootings do not occur with the same frequency as other days.

Specifically, we will use a Chi-Square Goodness of Fit Test. We will assume that the assumptions for Chi-Square test are met: the observations are independent, the categories are mutually exclusive (no incident can simoultaneously happen on one day and another), and there must be at least 5 frequencies in each group (see table above) ("Chi-Square Goodness of Fit Test", 2021).

The following hypotheses will be used:

H_0: For all days of the week, shootings occur with the same frequency.

In other words, $p_{Sunday} = \frac{1}{7} = p_{Monday} = \frac{1}{7} = ... = p_{Saturday} = \frac{1}{7}$.

H_A: For all days of the week, shootings do NOT occur with the same frequency.

In other words, at least one of $p_{Sunday}, p_{Monday}, ..., p_{Saturday} \neq \frac{1}{7}$.

Our test statistic is equal to $-2log(\frac{L(p_0)}{L(\hat{p})})$, which follows a χ^2 distribution with 6 degrees of freedom (df= number of categories (days of the week) - 1).

Note that $L(p_0)$ is the likelihood of the probabilities of under H_0, and $L(\hat{p})$ are the sample proportions.

Bayesian Credible Interval

It would be useful to have a probable range of values for the average number of fatalities per mass shooting incident. This information may influence resource distribution after mass shooting incidents in terms of supporting the victim's families, hospital amenities, etc. We may construct such an interval by calculating a Bayesian Credible Interval for a particular paremter, which tells us the probability such that the parameter falls in a certain range.

Recall the discrete random variable **Fatalities**, which represents the number of fatalities per mass shooting incident. We will assume that this variable follows a Poisson(λ) distribution.

Assume we are interested in finding a 95% credible interval of the paramter λ–in other words, we are interested in finding a credible interval for the true mean number of deaths per shooting. Suppose our data is a random sample of Poisson random variables with rate λ and fixed variance λ. Additionally, assume the prior distribution of λ is Exponential($\beta = 5$) distribution.*

The posterior distribution of λ is Gamma($\alpha = 1 + \sum_{i=1}^{n} X_i$, $\beta = \frac{5}{5n+1}$). Thus, we can use the 2.5th and 97.5th percentiles of this distribution to derive a range of values, in which λ has 95% probability of falling into. All derivations regarding the posterior distribution can be found in Section 2 of the Appendix.

* Note: We choose an Exponential distribution because the conjugate prior of the poisson distribution is the gamma distribution, which is a variation of the exponential distribution. The support of the exponential distribution matches the support of the poisson distribution. The parameter $\beta = 5$ was chosen as it is expected that the number of fatalities per mass shooting will be on average about 5 (de Jager, Goralnick, & McCarty, 2018).

Results

This section discusses the results from the aforementioned methodologies. Particularly, this section unpacks the "troubled young white man" idea of mass shooters, and answers question regarding the average age of mass shooters, the number of fatalities per shooting, and if shootings are equally likely throughout the week.

Linear regression

We find through our calculation that $\hat{\alpha}$ and $\hat{\beta}$ are as follows:

Component	Value
$\hat{\alpha}$	2.2419
$\hat{\beta}$	0.1187

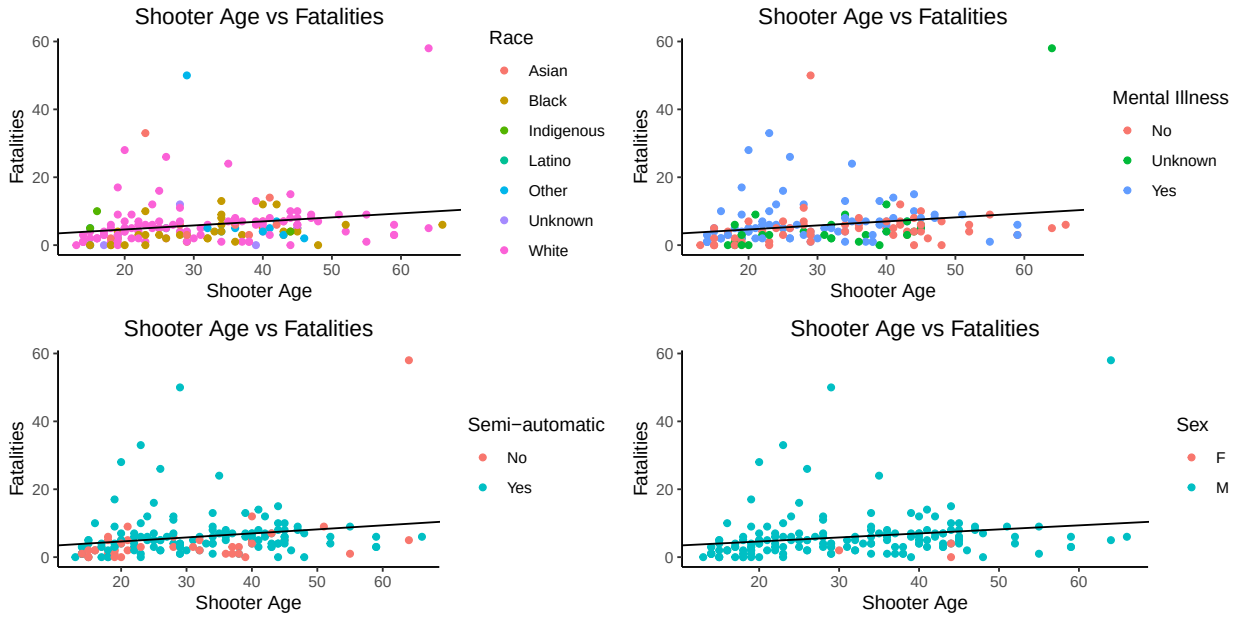
Putting the pieces together, we can write our regression line as

$$y = 2.2419 + 0.1187x.$$

Technically, this means that when the shooter is 0 years old, there are approximately 2.2419 deaths. In the context of our data, is nonsensical and not useful (none of the shooters are age 0, and it is highly unlikely that such an event would occur). However, what is more relevant is the *slope* of 0.1187. This means that for every 1-crime increase in the Shooter Age reported, there is corresponding increase of 0.1187 Fatalities. This is a pretty small number, indicating that there may not be a significantly strong relationship between these two variables.

Let's do an example, using the 1991 University of Iowa shooting. The shooter was 28; our model therefore predicts that the number of fatalities would be $0.1187(28) + 2.2419 = 5.5655$. In reality, there were 6 fatalities due to this shooting. For this observation, the model seems reasonably accurate.

Below are four plots, all using Shooter Age to predict Fatalities. Each graph is "colored" by a different demographic variable of the shooter–race, whether the shooter had a history of mental illness struggles, the sex of the shooter, and whether the shooter used semi-automatic weapons.



Evidently, there does not seem to be a particularly strong relationship between the age of the shooter and the number of deaths that they inflict–age does not seem to be the best predictor of impact. It does seem that

most incidents are concetrated in the 10-45 age range, but there is not a strong linear trend. The "troubled young white man" trope would perhaps predict an inverse relationship between age and fatalities, but that does not seem probable given the graphs above. This calls into question "young" aspect of the aforementioned trope that often surrounds mass shooters (Siemaszko, 2021), and forces us to look into other demographics of perpetrators.

Sex is known to influence an individual's risk for violence and aggression–males tend to be significantly more violent than women (Costanzo et al., 2014). This seems to be corroborated by the graph at the bottom right. In this entire sample, only three incidents were caused by women, and all had relatively low death tolls. Most shootings, as well as the deadliest shootings, were caused by men.

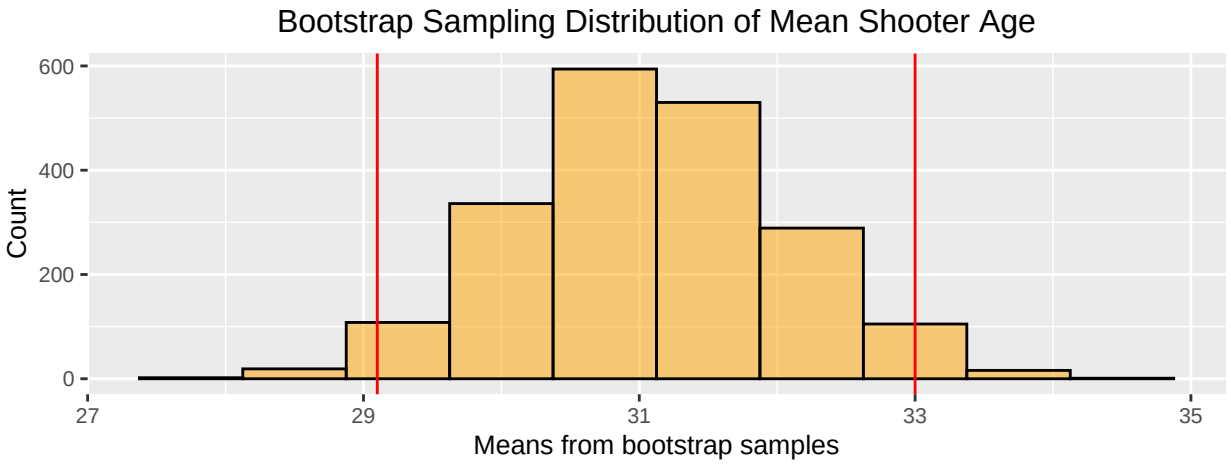
Another piece in the narrative that often surrounds mass shooters is race, and these shooters are often thought of to be white. It has been said that mass shootings are disproportionately committed by white men (Martin & Bowman, 2021); we already know that most shooters are men, and the graph at the top left tells us that most of those men are white. Therefore, whiteness seems may be a contributor to an individual's propensity to be a mass shooter.

Now looking into the "troubled" aspect of the narrative, we can see from the graph on the top right that while there are a number of incidents where mental illness was present in the shooter, there are also a number of cases where there was no history of mental illness or it remains unknown. It does appear from the graph that individuals with mental illnesses account for most of the incidents with the highest death counts. However, significant research shows that the vast majority of indiviuals with mental illnesses are not susceptible to violence (Costanzo et al., 2014). Therefore, more research should be done into the specific circumstances surrounding the mental health of mass shooters.

Lastly, there is the question of semi-automatic weapons. The graph on the bottom left indicates that the majority of incidents do involve semi-automatic weapons, and most of the high-death-count incidents involve these types of guns; however, the deadliest shooting did not involve semi-automatic weapons.

In summary, while shooter age does not seem to be a reliable predictor of the death toll, other static characteristics of the shooter (sex, race, history of mental illness) as well as their choice of weapon may influence the impact of their violence.

Confidence Interval



The plot above was generated using the bootstrap process described in **Methods**, and represents the *bootstrap sampling distribution*. As you can see, it appears to be unimodal and symmetric, centered around 31, and with most of the data spread between about 28 and 34.

The 2.5th and 97.5th quantiles have been calculated to be 29.1 and 33, respectively.

This means that 95% of the observations in this distribution fall between 29.1 and 33, and we can be 95%

confident that the true mean age for mass shooters falls within this range. These results seem plausible, given the range of values found in the original sample, and assuming that the sample is representative of the population.

Identifying the true average age of mass shooters has massive implications, both for how we socialize our struggling youth and how we support our struggling adults. Research shows that young perpetrators often blame their schools for their suffering, making fellow students and teachers the targets of their violence. On the other hand, older offenders tend to cast their anger more broadly, attacking their workplaces, religious institutions, etc. Young offenders also often obtain their weapons illegaly or through family members, whereas older offenders tend to purchase weapons themselves. Given these insights, identifying the age range most at risk of committing a mass shooting will funnel resources into the appropriate categories and guide gun-related legislation in the appropriate directions as to mitigate risk.*

* This paragraph relies on research done by Lankford & Hoover, 2018.

Maximum Likelihood Estimator

See Section 1 of the Appendix for this derivation.

The MLE for θ is $\hat{\theta} = max(X_1, ..., X_n)$.

The largest Mortality Rate value in our sample is 1.0.

Therefore, $\hat{\theta} = 1$.

These results seem reasonable, as we know Mortality Rate may not exceed 1, and there were incidents where all victims did pass away. However, these results are contingent upon the assumption that Mortality Rate is uniformly distributed. The limitations of this assumption are discussed in **Conclusions**.

Hypothesis Test

Recall the hypotheses:

$H_0 : \mu = 4.25$, and

$H_A : \mu \neq 4.25$,

as well as the test statistic, $\frac{\bar{X}-\mu}{s/\sqrt{n}} = \frac{6.4553-4.25}{6.6397/\sqrt{123}} = 3.6835$, which follows a t distribution with 122 degrees of freedom.

The results of this test give us a p-value of 0.0003. Using a significance value of $\alpha = 0.05$, there is evidence *against* the claim that the true mean number of fatalities in active shooter incidents involving semi-automatic weapons is 4.25. The implications of these findings are largely that more research may be necessary to identify the true mean number of fatalities. Having an accurate idea of this parameter is imperative to appropriate resource distribution and funding allocation to communities and families affected by mass shootings.

Please note that these results are not absolute, and there is a possibility that a Type 1 error has been made. Type 1 errors are those that occur when the null hypothesis has been rejected, but it is true in reality. The limitations of this method are discussed in **Conclusions**.

Goodness of Fit Test

The following table outlines the total number of incidents per day of the week:

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Count	18	25	21	26	32	30	14

Recall the hypotheses:

H_0: For all days of the week, shootings occur with the same frequency. In other words, $p_{Sunday} = \frac{1}{7} = p_{Monday} = \frac{1}{7} = ... = p_{Saturday} = \frac{1}{7}$.

H_A: For all days of the week, shootings do NOT occur with the same frequency. In other words, at least one of $p_{Sunday}, p_{Monday}, ..., = p_{Saturday} \neq \frac{1}{7}$.

Observe that the test statistic is equal to $-2log(\frac{L(p_0)}{L(\hat{p})}) = 10.1506$, which follows a χ^2 distribution with 6 degrees of freedom.

This gives us a p-value of 0.0908. Using a significance value of $\alpha = 0.05$, there is NOT sufficient evidence against the claim that shootings happen with the same frequency during each day of the week. In other words, with all other variables held constant, it is possible that every day of the week faces the same risk of a mass shooting occurring. Therefore, resources should likely be distributed evenly, regardless of the day of the week.

Bayesian Credible Interval

The posterior of λ is $Gamma(\alpha = 1 + \sum_1^n X_i, \beta = \frac{5}{5n+1})$, and $n = 166$.

So, $\lambda \sim Gamma(\alpha = 985, \beta = \frac{5}{831})$.

Our calculations yeild the following:

Parameter	2.5th Percentile	97.5th Percentile
λ	5.5622	5.6833

Therefore, there is a 95% probability that λ, or the true average number of fatalities per incident is between 5.477 and 5.596. As mentioned in the **Hypothesis Test** section, an understanding of the true average death toll per incident is critical to sufficient support and resource distribution granted to communities and families affected by mass shootings.

Conclusions

Recall the hypotheses mentioned at the beginning of this report:

- The typical shooter is a young white man with a history of mental illness, and that shooters that meet this criteria perpetrate shootings with the highest death toll and victim counts.
- Mass shootings result in an average of 5 deaths per incident.

The methods that were used to analyze the characteristics of mass shooters and to investigate the above hypotheses and their results are recapped below:

- Linear Regression (Shooter Age vs Fatalities)

The relationship between age and fatalities was positive and weak. Regardless of the strength of the relationship, the lack of an inverse relationship between age and fatalities calls into question the idea that young people are at fault for the deadliest shootings, and invites further research into other characteristics of shooters that posit them at a higher risk of violence via mass shooting (sex, history of mental illness, access to weapons, and race).

- Bootstrapped Confidence Interval for Mean Shooter Age

This method yielded (29.1, 33) as the 95% Confidence Interval for Mean Shooter Age. Therefore, we can be 95% confident that the true average age of mass shooters falls somewhere in this interval. Again, this interval is not particularly young, challenging the common idea that mass shooters tend to be in their late teens or early twenties.

- Maximum Likelihood Estimation for Mortality Rate

The results of this method found that the Maximum Likelihood Estimator for θ, or the upper bound for the uniformally distributed variable Mortality Rate, was equal to 1, or the maximum of our sample observations. These results are reasonable, as Mortality Rate is a proportion, and unfortunately, there were incidents in the sample where all victims of the shooting did pass away.

- Hypothesis Test of Mean Fatalities (in shootings with semi-automatic weapons)

The results of this method found that there is evidence against the claim that the average number of fatalities in shootings with semi-automatic weapons is 4.25 (with $p = 0.0003$). Given the danger that these weapons pose and their relative accessibility in the United States, further research should be done to identify the true average death count of incidents where semi-automatic weapons are used.

- Goodness of Fit Test for Day of the Week

The results of this method found that there is NOT evidence against the claim that shootings happen with the same frequency throughout the week (with $p = 0.0908$). Therefore, resources (security measures, surveillance, staff) should be distributed evenly throughout the week.

- Bayesian Credible Interval for Mean Fatalities per incident

This method yielded (5.5622, 5.6833) as the 95% Credible Interval for Mean Fatalities. Therefore, we can say that there is a 95% probability that the true average number of deaths falls somewhere in this interval. This information can guide policies for appropriate resource allocation and distribution for communities affected by mass shootings (funeral preparations, hospital resources, etc).

The big takeaways from the above results are that young people may not be automatically predisposed to causing the most devastating mass shootings, and that other factors may influence an individual's risk of perpetration. In fact, the average age for mass shootings may very well fall between 29.1 and 33. Therefore, violence risk prevention should not just occur in schooling of young people, but should be carried over as they enter adulthood. Workplaces, community centers, universities, and other spaces which adults in their late twenties and early thirties may occupy should implement support systems and anti-violence programming.

Additionally, the average number of deaths resulting from mass shootings is likely around 5 or 6. This information can be used to make yearly predictions for the number of lives lost due to mass shootings, which might inspire voters to support gun control measures. Voter support could then eventually motivate policy makers to draft legislation that reduces the rate at which these events occur. For example, having a concrete idea of how many people might die in a given year due to mass shootings may inspire stricter gun laws, the funding and implementation of anti-violence programming, etc. In tandem with statistics about fatalities influencing mass-shooting prevention, it can also improve mass-shooting response by ensuring that there are adequate funds and resources set aside to support communities that do fall victim to gun violence.

Weaknesses

There are a few weaknesses and limitations regarding this report that should be kept in mind. Firstly, in the derivation of the MLE for Mortality Rate, it was assumed that the variable was uniformally distributed. This assumption was based off of a histogram of the variable, not rigoruous numerical analysis. Therefore, it is entirely possible that this variable does not follow this distribution, which would impact the legitimacy of those particular results. Additionally, the prior distribution of λ in the calculation of the Bayesian Credible Interval was based off of previous beliefs and assumptions about the mean number of fatalities in mass shooting incidents. These beliefs were gained from limited available research, which may not capture the whole story or may be flawed. It is possible that a more sophisticated approach was appropriate. Lastly, in both the Goodness of Fit test and the Hypothesis Test of the mean, there is a potential for error. In the GOF test, since we failed to reject the null hypothesis, it is possible that a Type II error was committed (where the Null is actually False, not True). In the Hypothesis Test, where we rejected the null hypothesis, it is possible that a Type I error was committed (where the Null is actually True).

Considering the sample itself, the information included in this dataset was all found from news stories and secondary sources. It is possible that details about the incidents were misreported or exaggerated for journalistic effect. All of the aforementioned possible weaknesses or errors should be considered when assessing the results of this report.

Next Steps

Given that the null hypothesis of 4.25 as the mean number of fatalities in incidents involving semi-automatic weapons was rejected in the Hypothesis Test of the mean, it is advised that further research be done in that area. Additionally, as age was found to be a weak predictor of impact, other demographics of shooters (sex, history of mental illness, access to weapons, and race) and how they influence impact should be more rigorously analyzed.

Discussion

Mass shootings are truly an epidemic in the United States. In the past month, three mass shootings have occurred: Boulder, Colorado, March 22–10 deaths; Orange, California, March 31–4 deaths; Indianapolis, Indiana, April 15–9 deaths ("US has been wracked", 2021). Yet, still very little progress has been made in reducing the rate at which these events occur. It is possible that the strongest tool we have in making change is data and statistics. If we are able to communicate those who are at risk of committing these acts of violence and the typical outcomes of such events to American voters, perhaps this will cause a domino effect of pressure that results in policy makers enacting effective legislation. This report should certainly not be the last step in this effort–continuous action and pressure may be the best advantage in causing change.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: April 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: April 15, 2021)
4. Paterson, L. (2019) *Many Mass Shooters Share A Common Bond: Male Grievance Culture.* American University Radio. https://wamu.org/story/19/08/13/many-mass-shooters-share-a-common-bond-male-grievance-culture/ (Last Accessed: April 18, 2021)
5. Berkowitz, B. & Alcantara, C. (2021) *The terrible numbers that grow with each mass shooting.* The Washington Post. https://www.washingtonpost.com/graphics/2018/national/mass-shootings-in-america/ (Last Accessed: April 18, 2021)
6. *Mass shootings in the United States.* Wikipedia. https://en.wikipedia.org/wiki/Mass_shootings_in_the_United_States (Last Accessed: April 18, 2021)
7. Cohen, A., Azrael, D., & Miller, M. (2014) *Mass shootings becoming more frequent.* Harvard School of Public Health. https://www.hsph.harvard.edu/news/hsph-in-the-news/mass-shootings-becoming-more-frequent/ (Last Accessed: April 18, 2021)
8. Siemaszko, C. (2021) *After two mass shootings, Americans ask: Is this what a return to normal looks like?.* NBC. https://www.nbcnews.com/news/us-news/after-two-mass-shootings-americans-ask-what-return-normal-looks-n1261841 (Last Accessed: April 18, 2021)
9. de Jager, E., Goralnick, E. & McCarty, J.C. (2018) *Lethality of Civilian Active Shooter Incidents With and Without Semiautomatic Rifles in the United States.* The Journal of the American Medical Association. https://jamanetwork.com/journals/jama/fullarticle/2702134 (Last Accessed: April 18, 2021)

10. Berk, R. (2018) *What is a Mass Shooting? What Can Be Done?*. The University of Pennsylvania. https://crim.sas.upenn.edu/fact-check/what-mass-shooting-what-can-be-done (Last Accessed: April 18, 2021)

11. *Semiautomatic* (2021) Merriam-Webster https://www.merriam-webster.com/dictionary/semiautomatic (Last Accessed: April 18, 2021)

12. *Assault weapons legislation in the United States*. Wikipedia. https://en.wikipedia.org/wiki/Assault_weapons_legislation_in_the_United_States#:~:text=The%20Public%20Safety%20and%20Recreational,a%2010%2Dyear%20sunset%20provision. (Last Accessed: April 18, 2021)

13. *Mass Shootings in America* (2016) Stanford Geospatial Center. [https://library.stanford.edu/projects/mass-shootings-america]https://library.stanford.edu/projects/mass-shootings-america) (Last Accessed: April 18, 2021)

14. *US Mass Shootings, 1982–2021: Data From Mother Jones' Investigation* (2021) Mother Jones. https://www.motherjones.com/politics/2012/12/mass-shootings-mother-jones-full-data/ (Last Accessed: April 18, 2021)

15. *Chi-Square Goodness of Fit Test* (2021) Statistics Knowledge Portal. https://www.jmp.com/en_ca/statistics-knowledge-portal/chi-square-test/chi-square-goodness-of-fit-test.html#:~:text=The%20Chi%2Dsquare%20goodness%20of%20fit%20test%20is%20a%20statistical,representative%20of%20the%20full%20population. (Last Accessed: April 18, 2021)

16. Costanzo M., Krauss, D., Schuller, R., & McLachlan, K. (2014). *Forensic and Legal Psychology: Psychological Science Applied to Law.* Worth Publishers. (Last Accessed: April 18, 2021)

17. Martin, M. & Bowman, E. (2021) *Why Nearly All Mass Shooters Are Men*. National Public Radio. https://www.npr.org/2021/03/27/981803154/why-nearly-all-mass-shooters-are-men (Last Accessed: April 18, 2021)

18. Lankford, A. & Hoover, K.B. (2018) *Do the Ages of Mass Shooters Matter? Analyzing the Differences Between Young and Older Offenders* Violence and Gender. https://www.liebertpub.com/doi/10.1089/vio.2018.0021 (Last Accessed: April 18, 2021)

19. *US has been wracked with several mass shootings in 2021* (2021) The Associated Press. https://apnews.com/article/mass-shootings-2021-indianapolis-atlanta-boulder-1202404f7c9c7c8c4836fd9b05c55561 (Last Accessed: April 18, 2021)

Appendix

Section 1

The Likelihood Function of θ is equal to $L(\theta) = f_\theta(x_1)...f_\theta(x_n)$.

Recall that $X_1, ..., X_n \sim Unif(0, \theta)$. Therefore, $L(\theta) = f_\theta(x_1)...f_\theta(x_n) = \frac{1}{\theta}...\frac{1}{\theta} = (\frac{1}{\theta})^n$.

However, this is only true if $0 \leq x_1 \leq \theta$, $0 \leq x_2 \leq \theta$, ... and $0 \leq x_n \leq \theta$.

Therefore, we can say that $L(\theta) = (\frac{1}{\theta})^n$ if $\theta \geq max(X_1, ..., X_n)$, and 0 otherwise.

Notice that $(\frac{1}{\theta})^n$ is a decreasing function; therefore, given that its domain is $\theta \geq max(X_1, ..., X_n)$, it would achieve its maximum at $\hat{\theta} = max(X_1, ..., X_n)$.

The largest Mortality Rate value in our sample is 1.0.

Therefore, $\hat{\theta} = 1$.

Section 2

Recall that we are assuming our random variables $X_1, ..., X_n \overset{iid}{\sim} Poisson(\lambda)$, and that $\lambda \sim Exp(\beta = 5)$.

Therefore, the *prior* distribution is $f(x) = \frac{1}{5}e^{-\frac{\lambda}{5}}$.

Now, let's find the posterior of λ.

$$\begin{aligned} P(\lambda|data) &= \frac{P(data|\lambda)P(\lambda)}{P(data)} \\ P(\lambda|X_1, ..., X_n) &= \frac{P(X_1, ..., X_n|\lambda)P(\lambda)}{P(X_1, ..., X_n)} \\ &= \frac{(\frac{e^{-\lambda}\lambda^{X_1}}{X_1!})...(\frac{e^{-\lambda}\lambda^{X_n}}{X_n!})(\frac{1}{5}e^{-\frac{\lambda}{5}})}{P(X_1, ..., X_n)} \end{aligned}$$

We are only interested in the posterior distribution of λ. Therefore, we will treat all terms without λ as part of a normalizing constant, C.

$$\begin{aligned} P(\lambda|X_1, ..., X_n) &= \frac{(\frac{e^{-\lambda}\lambda^{X_1}}{X_1!})...(\frac{e^{-\lambda}\lambda^{X_n}}{X_n!})}{P(X_1, ..., X_n)} \\ &= C \cdot \lambda^{X_1+...+X_n} e^{-n\lambda - \frac{1}{5}\lambda} \\ &= C \cdot \lambda^{X_1+...+X_n} e^{-\lambda(\frac{5n+1}{5})} \end{aligned}$$

Notice that this is a $Gamma(\alpha = 1 + \sum_1^n X_i, \beta = \frac{5}{5n+1})$ distribution.

So, the posterior of λ is $Gamma(\alpha = 1 + \sum_1^n X_i, \beta = \frac{5}{5n+1})$.