Statistical Analysis of Crime Rates in the United States

Tanya Thaker

Abstract

Crime analysis is a key approach to identify and analyze trends with crime. With the aid of statistical analysis, one would be able to analyze previous crimes to discover ways to deal with crimes in the future. With an average unemployment of about 6%, there seems to be clear positive relationship between crime and unemployment rate in the states with the lowest crime rate while there is an unclear relationship in states with highest crime rate.

Introduction

Crime has always been a trending topic in the media. The fact is that we frequently and regrettably misjudge the true crime statistics in the United States. This happens due to many factors not limited to, but including exaggerations by the media and misrepresentation for political gain. According to Walden University[2], Americans often overestimate crime even though crime has been on a decline for the past few years now. It often becomes hard to discern which statements are true and which ones are false and as a result, Americans often overestimate crime even though crime has been on a decline for the past few years now. Thus, it is crucial to analyze the true crime rates, their causes, and the resources needed to ensure the community is a safer place. In this report, we will be investigating how a crucial factor, unemployment, affects the true rate of crime in the United States.

Throughout the paper, we will use the following important terminology: population, sample, and parameter. Population represents a collection of data that is important for the analysis, while a sample is a subset of the population. A parameter refers to a measurement of the population. Furthermore, when referring to the crime rate, the value will refer to the crime rate per 100,000 individuals.

We hypothesize that the average unemployment rate will be about 6% and there will be a correlation between the unemployment rate and average crime rate. This report will provide an insight to not only the true rate of crime in the United States, but also possible solutions to effectively deal with this issue.

Data

The dataset (*Crime rate and Unemployment rate by state 2020*)[12] that will be used throughout this report is called 'Crime rate and Unemployment rate by state' and it comes from Kaggle. It has 1989 observations with 13 variables. These variables include the number of different types of crimes, i.e. assault, homicide, vehicle theft, robbery, burglary and more, for each state. We will also have a new column, true_rate, that will refer to an average of the different crime rates. The dataset(*Crime rate and Unemployment rate by state 2020*)[12] includes crime rates from 1976 to 2014 inclusive. However, in this report, we will only be observing crime data from 2000 to 2014.

In order to make key observations about my research question, I first cleaned the dataset to only include values from 2000 to 2014. The important variables used are:

- *state*: This variable stores the state code.
- *year*: A numerical variable that holds the year.

- *unemployment*: A numerical variable that stores the unemployment rate for each state.
- *violent total*: A numerical variable which tells us the average crime rate of violent crime per 100,000 individuals.
- *Murder*:A numerical variable which tells us the average crime rate of murder per 100,000 individuals.
- *rape*: A numerical variable which tells us the average crime rate of rape per 100,000 individuals.
- *Robbery*: A numerical variable which tells us the average crime rate of robbery per 100,000 individuals.
- *Aggravated assault*:A numerical variable which tells us the average crime rate of aggravated assault crime per 100,000 individuals.
- *property total*:A numerical variable which tells us the average crime rate of property thefts per 100,000 individuals.
- *Burglary*:A numerical variable which tells us the average crime rate of burglary per 100,000 individuals.
- *Larceny theft*:A numerical variable which tells us the average crime rate of larceny theft per 100,000 individuals.
- *vehicle theft*:A numerical variable which tells us the average crime rate of vehicle theft per 100,000 individuals.
- *total_rate*:A numerical variable which tells us the average crime rate of all the different types of crime per 100,000 individuals.

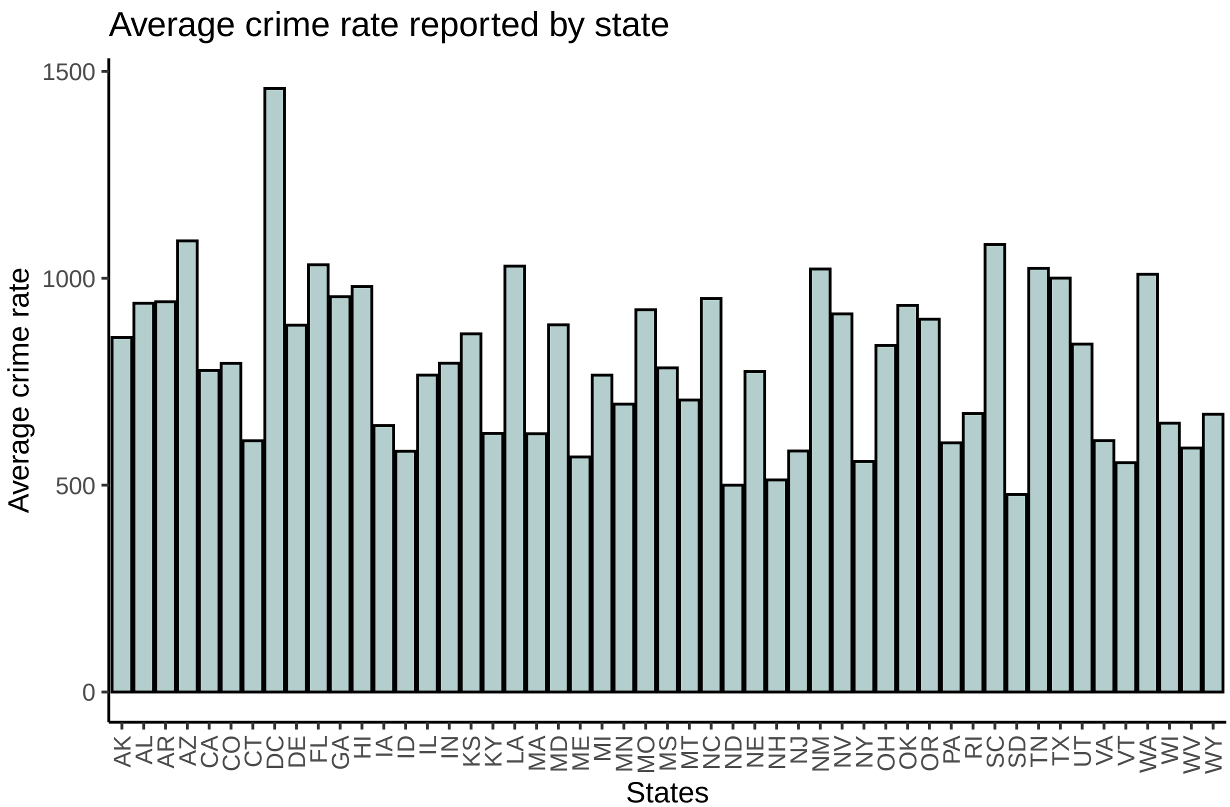

Figure 1

Table 1: States with highest crime rate

state	avg_rate
AZ	1090.304
DC	1458.624
FL	1032.601
LA	1029.314
SC	1081.553

Table 2: States with lowest crime rate

state	avg_rate
ND	499.8919
NH	512.6393
NY	557.2119
SD	477.4504
VT	554.2674

From the table and figures above, a key observation to make would be that the states with the highest average number of crime rates are Washington, Arizona, South Carolina, Louisiana, and Florida, while the states with the lowest average number of crime rates are Nevada, Vermont, New Hampshire, South Dakota and New York. We will be using this observation to find the correlation between unemployment rate and the crime rates for these particular states. This will help us understand the true rate of crime and also how unemployment causes crime.

Methods

Throughout this report, we will be using a number of methodologies in order to make key findings about the crime rate and unemployment rate in the United States. Firstly, there will be a hypothesis test for average unemployment rate. Next, we will perform a goodness of fit test to see if this data does indeed fit a normal distribution. Further, assuming the distribution is normal, we will find a credible interval for the mean unemployment rate.Then, we will find the maximum likelihood estimator for the mean unemployment rate, assuming the data is independent and shows normal distribution. We will also find the confidence interval for the average unemployment rate. Lastly, we will perform a linear regression to find the correlation between the unemployment rate and crime rate.

Hypothesis Test for Mean

I will be conducting a hypothesis test for the average unemployment rate in the states with the highest crime rates. Since the number of observations for this analysis is large and the population mean and variance are known, I will assume that the data is independent and that it follows a standard normal distribution. For this method, a test statistic will need to be calculated and a null hypothesis and alternative hypothesis will be set. Then, the p-value, the probability of observing the test statistic given that the null hypothesis is true, will be calculated. For consistency, we will set our significance level, α, to 0.025. If the calculated p-value is less than the significance level, we will reject the null hypothesis, else, we conclude that we do not have enough evidence to reject the null hypothesis. Based on tradingeconomics.com, the average unemployment rate in the United States was 6 percent as of March 2021. Thus, for this test the null hypothesis and alternative hypothesis will be as follows:

$$H_0 : \mu = 6$$

$$H_a : \mu \neq 6$$

The null hypothesis states that the average unemployment rate in the states with the highest crime rate should be about 6, while the alternative hypothesis states that it will not be equal to 6. In order to make a conclusion about the average unemployment, I will carry out 1000 simulations to calculate the average unemployment rate. In each simulation, I will take a sample of 25 fields from the dataset to calculate the average. Then, I will calculate the p-value using the results from the 1000 simulation. The formula to calculate the test statistic is as follows:

$$\frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}}$$

The test statistic is -0.3905379.

Goodness of Fit Test

In this method, we will be analyzing how well the dataset fits a distribution. From Figure 4 below, observe that the distribution appears to be normally distributed with a right skew. Thus, we will approximate this distribution with a normal distribution.

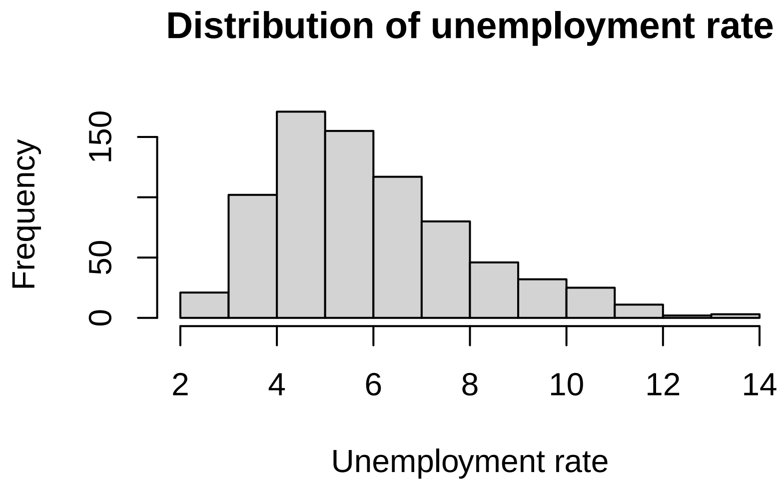

Figure 2

To summarize, our hypothesis will be as follows:

$$H_0 = \text{the data for unemployment fits a normal distribution}$$

$$H_a = \text{the data for unemployment does not fit a normal distribution}$$

In order to gauge how well the data fits the distribution in our hypothesis, we will use a modification of the Kolmogorov-Smirnov test, the Anderson-Darling Goodness of Fit Test (AD-Test). This test is commonly used as a test for normality. We will find the statistics p value based on the value of AD for the data. For consistency, we will set our significance level, α, to 0.025. If the calculated p-value is less than the significance level, we will reject the null hypothesis, else, we can conclude that we do not have enough evidence to reject the null hypothesis.

Bayesian Credible Interval

In this method, we will find an interval that may contain the estimate of the parameter of interest, given the evidence provided by the observed data. We will assume that the data is identical and independent, following a normal distribution with mean 5.9382261 and standard deviation 1.1296052. Since we have little knowledge about the parameter of interest, the average unemployment rate in the states with the lowest crime rate, we will assume that it has a non-informative prior, a normal distribution with a mean of 0 and variance of 1000.

The posterior distribution we found is:

$$N\left(\left(\frac{1}{1000}+\frac{n}{\sigma^2}\right)^{-1}\left(\frac{n}{\sigma^2}\bar{x}\right), \left(\frac{1}{1000}+\frac{n}{\sigma^2}\right)^{-1}\right)$$

where n refers to the number of observations, σ^2 is the sample variance and $\bar{x}$ is the sample mean. With the formula above, we calculated the 2.5th and 97.5th quantiles.

Maximum Likelihood Estimator

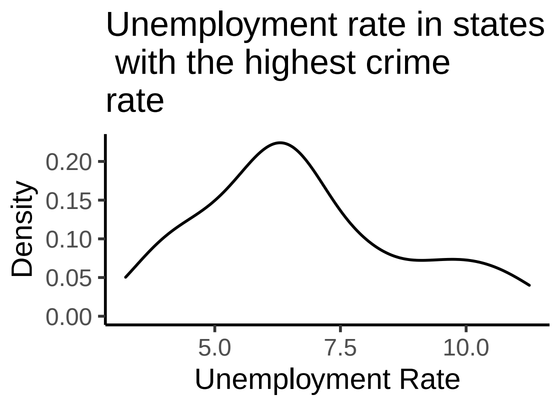

Figure 3

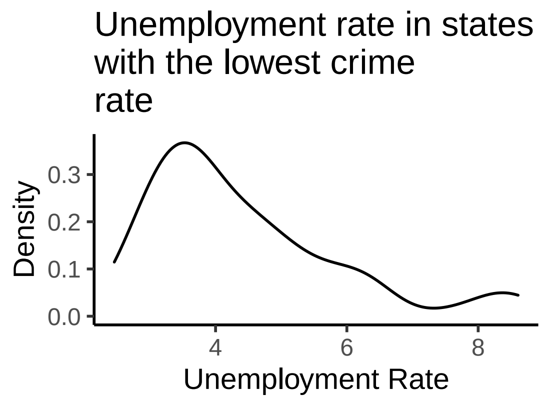

Figure 4

For the purpose of this method, I will assume that the unemployment rate in the states with the highest number of crimes are independent and identically distributed with a normal distribution. Based on the derivation, as seen in the Appendix, we find that the mean estimator for normally distributed data can be expressed as:

$$\mu = \frac{1}{n}\sum_{i=1}^{n}(X_n)$$

which essentially equals,

$$\bar{X}_n$$

which is the sample mean. By finding the mean unemployment rate of the dataset, I will have found a value of average unemployment rate that best describes the dataset.

Confidence Interval

In this method, we will carry out bootstrap sampling to find a 95% confidence interval. The bootstrap sampling method uses re-sampling of an original sample, assuming that the sample is representative of the population, multiple times in order to achieve an estimation of our parameter of interest. The parameter of interest will be the average unemployment rate in the states with the highest crime rate.

For the average unemployment rate in these states, the empirical distribution function can be thought of as the number of bootstrap sample means that are less than or equal to the sample mean. This will allow us to find a confidence interval that provides us with a range of plausible values that may include the true average unemployment rate. We will simulate 1000 iterations of bootstrap samples with a sample size of 25 each.

Linear Regression

The model to be studied is as follows:

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

In the equation above, α is the y-intercept refers to the average crime rate per 100,000 individuals when there is no unemployment. β is the slope, which signifies the increase in the average crime rate when the unemployment rate increases by 1. Finally, U_i accounts for the random fluctuations in the data set[12]. For the purpose of this report, we assume that x_i's are nonrandom.

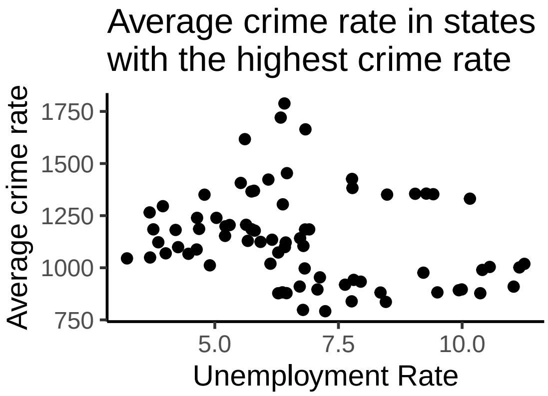

Figure 5

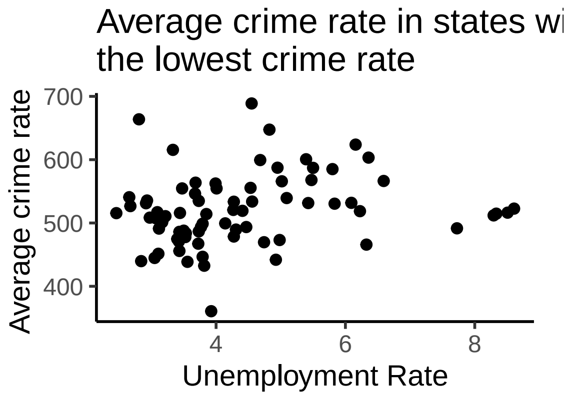

Figure 6

In the model above, the average crime rate per 100,000 individuals is the response variable while the unemployment rate is the explanatory variable.

The model seems appropriate since it not only makes practical and statistical sense, but also numerically since the dependent and independent variables are numerical. Each random fluctuation seems to have the same amount of variability, which implies that U_i are independent and have an expectation of 0. Furthermore, the data points in Figure 5 and 6 seem to have a constant variance and are slightly linear.

We will use a simple linear regression for the states with the highest crime rate(Figure 5) and the lowest crime rate (Figure 6) in order to find the correlation between unemployment and crime rate.

Results

In this section, we will see the results of all of the methods mentioned in the Methods section. This will help us gain insight about the crime rate and unemployment rate in the United States.

Hypothesis Test for Mean

Table 3: P value for the Hypothesis Test

Measure	Value
p value	0.6961389

The p-value for our hypothesis test when the test statistics is -0.3905379 is 0.6961389. In our Methods section, we set the significance level to 0.025. Since the p-value is greater than our significance level, we can conclude that we do not have enough evidence to reject the null hypothesis and thus, we fail to reject the null hypothesis. This essentially implies that the average unemployment rate in the states with the highest number of crimes can possibly be 6 percent.

Goodness of Fit Test

Table 4: P value for the Goodness of Fit Test

Measure	Value
p value	2.2e-16

The p-value calculated from the Anderson-Darling test is 2.2×10^{-16} which is less than our significance level, 0.025. Thus, we can reject the null hypothesis which implies that our data does not fit a normal distribution.

In terms of unemployment rate and average crime rate, this would mean that the two variables are not evenly distributed.

Bayesian Credible Interval

Table 5: Credible Interval

Quantile	Value
2.5	5.628062
97.5	6.248094

The 95% credible interval achieved by the Bayesian method is approximately (5.63, 6.25). This implies that we are 95% confident that this interval contains the true average for the unemployment rate in the states with the lowest crime rate. Since the true average unemployment rate is 5.9382261, we can conclude that our credible interval does indeed include the true parameter of interest value.

Maximum Likelihood Estimator

Table 6: Mean unemployment rate

States	Mean
Highest crime	6.750707
Lowest crime	4.406093

Table 6 shows us that the mean unemployment rate for the states with the highest crime rates is 6.75, while the unemployment rate for the states with the lowest crime rate is 4.4. Through this test, we observe that the states with the lowest crime rates have a lower unemployment rate as compared to the states with the highest crime rates.

Confidence Interval

Table 7: Confidence Interval for average unemployment rate

Quantile	Mean
0.025	5.964423
0.975	7.554830

Table 8: Confidence Interval for average unemployment rate

Quantile	Mean
0.025	3.866159
0.975	5.041408

From Table 7, the values of the confidence interval tells us that we can be 95% confident that the average unemployment rate for the states with the highest crime rate is approximately between 5.96 and 7.55. In other words, this can be written as (5.96, 7.55). While, as seen in Table 8, states with the lowest crime rate have a confidence interval of (3.87, 5.04). Both of these confidence intervals seems reasonable since they include the true parameter value, the average unemployment rate, which are 6.7507067 and 4.4060933 for the states with highest crime rate and lowest crime rate respectively.

Linear Regression

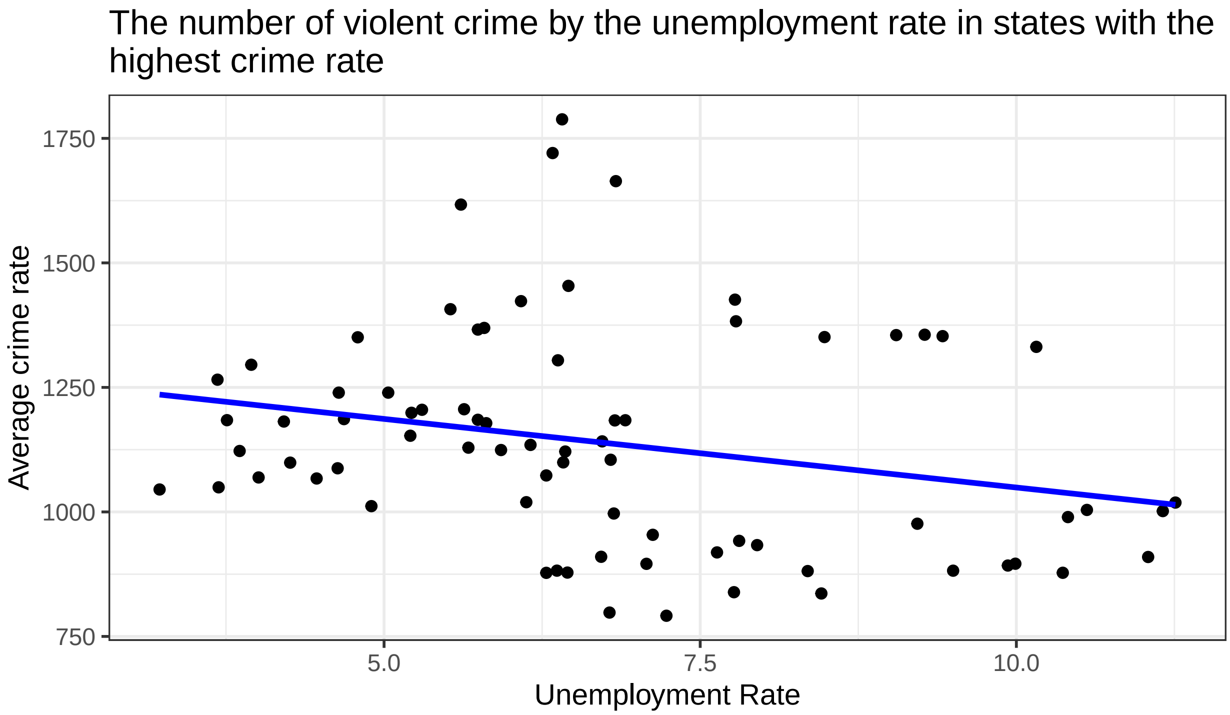

Figure 7

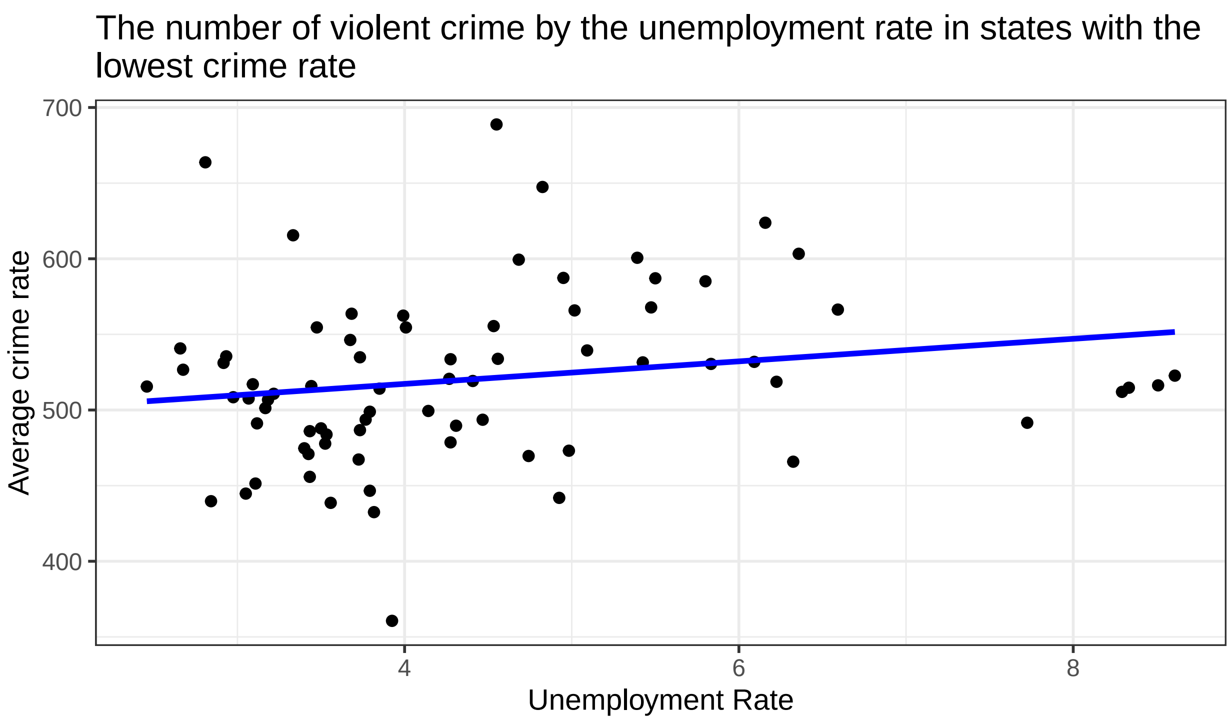

Figure 8

Table 9: Numerical estimates for states with highest crime rate

Intercept	Slope
1324.34	-27.53

Table 10: Numerical estimates for states with lowest crime rate

Intercept	Slope
487.456	7.453

In Figure 7, we see a regression line has been plotted with a slope of -27.53 and a y-intercept of 1324.24. This implies that an increase in the unemployment rate leads to about 28 less crimes on average. The y-intercept value implies that when the unemployment rate is 0, there are about 1324.24 crimes, which would mean that the crime rate does not only depend on the unemployment rate and may depend on other factors, as well. Figure 7 shows that the unemployment rate and crime rate share a negative and linear relationship. This would mean that as the unemployment rate increases, the average crime rate decreases. Practically, this does not make sense.

In Figure 8, we see a regression line that has been plotted with slope 7.45 and y-intercept 487.46. This implies that they share a positive, linear relationship. As the unemployment increases, the average rate of crime per 100,000 individuals increases by approximately 7.45. This result seems to be logical as compared to the results from Figure 7.

Conclusion

Overall, we hypothesized that the average unemployment rate is about a 6%. We also hypothesized that there would be a correlation between the unemployment rate and average crime rate. Throughout this paper, we used several statistical methods like hypothesis testing, goodness of fit, etc. in order to make better conclusions about our hypothesis.

Overall, we have fulfilled the objective of our report by using several analytical methods. We made several key findings about the overall unemployment rate in the United States with an emphasis on how the unemployment rate caused the crime rate in specific states with the highest and lowest crime rates. One major conclusion to highlight is that the unemployment rate in the United States could be about 6% as per the results of our hypothesis test and that the crime rate and unemployment rate are not evenly distributed. Furthermore, through the Maximum Likelihood Estimator, we found that the average unemployment rate in the U.S. is about 6.57 for the 5 states with the highest crime rate and 4.4 for the 5 states with the lowest crime rate. The confidence interval for the states with the highest crime rate was (5.99, 7.56) and the lowest crime rate was (3.88, 5.04). This result conforms with our findings from the Maximum Likelihood Estimator. Finally, we found a positive linear relationship between unemployment rate in the states with the lowest crime rate and a negative relationship in the states with the highest crime rate. This implies that, in the states with lowest crime rates, as unemployment increases, individuals tend to look for other means to either earn income or express their frustration, thus resulting in a higher crime rate. However, our findings for the states with the highest crime rates implies that there may be a presence of other underlying factors that may affect the crime rate. This will be further discussed in the Weaknesses section.

All in all, we have found that the unemployment rate in the U.S. is about 6% and is not evenly distributed with respect to the crime rate. Furthermore, it holds that the states with highest crime rates have almost twice the unemployment rate as states with the lowest crime rate. While there may be a positive linear relationship between unemployment rate and crime rate in states with lowest crime rate, there seems to be a discrepancy with the relationship in the states with the highest crime rate. Given the results from this analysis, a possible solution to lower the crime rate would be to increase human capital in order to ensure that more individuals can be employed.

Based on all our results, we can conclude that our hypothesis seems reasonable since the unemployment rate could be about a 6%. In addition to that, our hypothesis about the correlation about crime rate and unemployment rate seems to be right for the states with the lowest crime rate. For the states with the highest crime rates, there seems to be lack of evidence to identify a correlation.

Weakness

Despite the fact that the dataset[12] originates from a reliable source, there are inevitably some problems with this data. For instance, the data collected ranges from the year 1993 to 2014. Since the dataset is quite outdated, statistically analyzing this dataset to make conclusions on crime rate and unemployment rate may lead to inaccurate inconclusive results. Furthermore, when carrying out analysis, a large amount of data was filtered out in order to achieve our objective which resulted in at most 3.7707391% of the original dataset being used. Furthermore, in several methods, we assumed that the data is normally distributed. This is not true since the goodness of fit test gave us sufficient evidence to reject the fact that the data is distributed normally. In addition to that, the linear relation in Figure 7 seems to be quite surprising. This implies that there may be a lot more underlying factors that affect the crime rate.

In order to achieve more satisfying results, we could assume that the data takes the form of a gamma distribution since the graph looks closer to a gamma distribution as compared to a normal distribution. We could also use a large confidence interval like 98% to increase the likelihood of capturing the true value of the population parameter with a larger sample size to ensure the sample is more representative of the population.

Discussion

The goal of this report was to gain an overall understanding about the unemployment rate and find the relation between the unemployment rate and crime rate, in order for the government to recognize places to allocate their resources.

The results that were observed throughout this paper suggest that the mean unemployment rate is about a 6% and while there may be a positive linear relationship between the unemployment rate and crime rate for states with the lowest crime rate, there seems to be a lack of evidence of a similar relationship for the states with the highest crime rates. However, there were multiple weaknesses in our analysis. Thus, one should reconsider reanalyzing the data to avoid weakness and lead to more conclusive results.

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

Bibliography

1. Allaire J.J., Xie Y., McPherson J., Luraschi J., Ushey K., Atkins A., Wickham H., Cheng J., Chang W., & Iannone R. (2020). rmark down: Dynamic Documents for R. R package version 2.6. https://rmarkdown.rstudio.com
2. Criminal Justice Update: Five Facts About Crime in the U.S. https://www.waldenu.edu/online-masters-programs/ms-in-criminal-justice/resource/criminal-justice-update-five-facts-about-crime-in-the-us
3. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
4. Evans J M., Rosenthal S J. Probability and Statistics - The Science of Uncertainty, Second Edition: http://www.utstat.toronto.edu/mikevans/jeffrosenthal/
5. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)
6. Hadley Wickham, Jim Hester and Winston Chang (2020). devtools: Tools to Make Developing R Packages Easier. https://devtools.r-lib.org/, https://github.com/r-lib/devtools.
7. Juergen Gross and Uwe Ligges (2015). nortest: Tests for Normality. R package version 1.0-4. https://CRAN.R-project.org/package=nortest
8. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
9. Ren K. & Russell K. (2021). formattable: Create 'Formattable' Data Structures. https://renkun-ken.github.io/formattable/, https://github.com/renkun-ken/formattable.
10. Stephanie, (2021). Anderson-Darling Test & Statistic: Definition, Examples. https://www.statisticshowto.com/anderson-darling-test/
11. Thomas Lin Pedersen (2020). patchwork: The Composer of Plots. https://patchwork.data-imaginist.com, https://github.com/thomasp85/patchwork.
12. Vasilyeva L. (2020). Crime rate and Unemployment rate by state. https://www.kaggle.com/lydiavasil/crime-rate-and-unemployment-rate-by-state/metadata
13. Wickham H. et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
14. Wickham H., François R., Henry L. and Müller K. (2021). dplyr: A Grammar of Data Manipulation. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr
15. Wickham H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
16. Wickham H. and Byran J. (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl
17. Xie Y., Allaire J.J. & Grolemund G. (2018). R Markdown: The Definitive Guide. Chapman and Hall/CRC. ISBN 9781138359338. https://bookdown.org/yihui/rmarkdown.
18. Xie Y., Dervieux C., & Riederer E. (2020). R Markdown Cookbook. Chapman and Hall/CRC. ISBN 9780367563837. https://bookdown.org/yihui/rmarkdown-cookbook.

Appendix

MLE Derivatoin

1. Likelihood Function

$$\begin{aligned}
L(\mu \mid X_1, ..., X_n) &= f_{X_1}(x_1)...f_{X_n}(x_n) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}(\frac{X_1-\mu}{\sigma})^2}\right) ... \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}(\frac{X_n-\mu}{\sigma})^2}\right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \left(e^{\frac{-1}{2}(\frac{X_1-\mu}{\sigma})^2 ... \frac{-1}{2}(\frac{X_n-\mu}{\sigma})^2}\right) \\
L(\mu) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \left(e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2}\right)
\end{aligned}$$

2. Loglikelihood Function

$$\begin{aligned}
\ell(\mu) &= log(L(\mu)) \\
&= ln(L(\mu)) \\
&= ln\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \left(e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2}\right)\right] \\
&= ln\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\right] + ln\left(e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2}\right) \\
&= nln[(2\pi\sigma^2)^{-1/2}] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \\
&= \frac{-n}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \\
\ell(\sigma^2) &= \frac{-n}{2}ln(\sigma^2) - \frac{-n}{2}ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2
\end{aligned}$$

3. First Derivative

$$\begin{aligned}
\frac{\partial\ell}{\partial\mu} &= -\frac{1}{2\sigma^2}\cdot(-2(X_1-\mu) - 2(X_2-\mu) - ..2(X_n-\mu)) \\
&= \frac{2}{2\sigma^2}\cdot((X_1-\mu) + (X_2-\mu) + ..(X_n-\mu)) \\
&= \frac{1}{\sigma^2}\cdot\sum_{i=1}^{n}(X_n-\mu)
\end{aligned}$$

Equating the first derivative to 0:

$$
\begin{aligned}
\frac{\partial \ell}{\partial \mu} &= 0 \\
\frac{1}{\sigma^2} \cdot \sum_{i=1}^{n} (X_n - \mu) &= 0 \\
\sum_{i=1}^{n} (X_n - \mu) &= 0 \\
\sum_{i=1}^{n} (X_n) - n\mu &= 0 \\
\sum_{i=1}^{n} (X_n) &= n\mu \\
\frac{1}{n} \sum_{i=1}^{n} (X_n) &= u
\end{aligned}
$$

4. Second Derivative Test

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial \mu^2} &= \frac{1}{\sigma^2} \cdot (-X_1 - X_2 - ...X_n) \\
&= -\frac{1}{\sigma^2} \cdot \sum_{i=1}^{n} (X_n)
\end{aligned}
$$

Using the μ value from the first derivative

$$
= -\frac{1}{\sigma^2} \cdot n\mu
$$

Since σ^2, n, μ will always be greater than zero thus the second derivative will always be less than zero. This proves that the maximum likelihood estimator is concave down and thus yields a maximum.

Posterior Distribution Derivation

Assume that the prior has a distribution as follows $\mu \sim N(u_0, \tau^2)$ and the data, $X_1, X_2, ..., X_n$ also follows a normal distribution with mean μ and variance σ_0^2 which is known. The likelihood function is given by

$$
L(\mu | x_1, ..., x_n) = e^{\frac{-n}{2\sigma_0^2}(\bar{x}-\mu)^2}
$$

Then the posterior density of μ is then proportional to

$$
\begin{aligned}
e^{\frac{-n}{2\tau^2}(\mu-\mu_0)^2} e^{\frac{-n}{2\sigma_0^2}(\bar{x}-\mu)^2} &= e^{\frac{-1}{2\tau^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2) - \frac{n}{2\sigma_0^2}(x_i^2 - 2\mu x_i + \mu^2)} \\
&= e^{\frac{-1}{2}(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2})[\mu^2 - 2(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2})^{-1}(\frac{\mu_0}{\tau_0^2} + \frac{n\bar{x}}{\sigma_0^2})\mu]} \cdot e^{-\frac{\mu_0^2}{2\tau_0^2} - \frac{n\bar{x}^2}{2\sigma_0^2}} \\
&= e^{\frac{-1}{2}(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2})[\mu - 2(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2})^{-1}(\frac{\mu_0}{\tau_0^2} + \frac{n\bar{x}}{\sigma_0^2})]^2} \cdot e^{-\frac{1}{2}(\frac{\mu_0^2}{\tau_0^2} + \frac{n\bar{x}^2}{\sigma_0^2})} \cdot e^{\frac{1}{2}(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2})^{-1}(\frac{\mu_0}{\tau_0^2} + \frac{n\bar{x}}{\sigma_0^2})^2}
\end{aligned}
$$

The expression above we can be written as a function of μ as being proportional to the density of an:

$$
N((\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2})^{-1}(\frac{\mu_0}{\tau_0^2} + \frac{n\bar{x}}{\sigma_0^2}), (\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2})^{-1})
$$

Thus, we can conclude that the posterior distribution is also a normal distribution with a mean

$$\left(\frac{1}{1000} + \frac{n}{\sigma^2}\right)^{-1} \left(\frac{n}{\sigma^2}\bar{x}\right)$$

and a variance of

$$\left(\frac{1}{1000} + \frac{n}{\sigma^2}\right)^{-1}$$

This derivation can be found in *Probability and Statistics - The Science of Uncertainty, Second Edition*[4].