A Statistical Analysis of United States Gun Violence Data

Monisha Prakash
Professor Deghani

IE 6200 Engineering Probability and Statistics

5 December 2018

1. Abstract

The number of Americans getting killed by guns are quite significant and this statistic is increasing day by day. The need to study the effect of gun violence and take action to reduce it is of paramount importance since its effects shapes the lives of several Americans who witness it, know someone who was shot or even living in constant fear of being the next victim. Hence, to address this problem we need to have some profound insight on the statistics of gun violence incidents and its details so that we can identify some underlying patterns which can help us solve the issue. In our project, we have used a large dataset from a non-profit organization called Gun Violence Archive, which has provided us with accurate data of more than 260,000 gun violence incidents, with detailed information about each incident from the Jan, 2013 to March, 2018. We will be using descriptive statistical analysis to study the effect of strict gun control laws on gun incidents and deaths. Stratified random sampling will be used to separate various states in the US into lenient and strict categories and then randomly select one state from each group. These sample populations can then be compared using descriptive analysis. We will be using the inferential statistics analysis to check for the claim (made by Children's Hospital of Philadelphia Research Institute), that in recent years, the average number of victims in the age group of 10-24 years old, who are either killed or injured due to gun violence is equal to 13 per day. We will be using the Z statistic in our analysis with a confidence level of 95% to verify whether this claim holds good or not. We will be using python to work with the data set and to compute the necessary statistics for the analysis. Finally, we also use the correlation analysis to understand the relation between the number of killings and the age group of the subjects who participated in the crime in each of these incidents. A scatter plot with trendline will also be implemented to test the relationship for the goodness of data fit and this analysis will also be implemented using python as the environment. The test results and the analysis will definitely shed some light on understanding the gun violence problem in America and it can be used in a positive way for reshaping its gun culture in the near future.

Source of data set: https://www.gunviolencearchive.org

2.Introduction

Gun violence is an extremely visible issue in the United States. Headlines about mass shootings and gun incidents have become an unfortunate part of daily news. Meanwhile, gun control is often a highly debated and controversial topic, leading to individual states with various levels of gun control legislation. While there are not necessarily easy solutions to resolve this societal problem, there is a clear need to accurately collect and understand data from gun incidents in order to recognize patterns to hopefully aid in the process of minimizing casualties from gun violence.

In this report, we will analyze raw data collected from a database of gun incidents, managed by the Gun Violence Archive, a non-profit organization whose objective is to verify and post accurate information regarding US gun incidents for public viewing or use. This database contains information of over 260,000 gun incidents from January 2013 to March 2018. Specific information such as the time, date, and location of the incident, number of casualties, gun type, and participant ages are shown in the database. For this report, this data will be further processed and analyzed using Microsoft Excel, Minitab, and Python to perform statistical analysis.

The objective of this report is to gather and compile gun incident data and perform both descriptive and inferential statistical analysis in order gain a better understanding of the effects of gun control laws and relationships between age and gun incidents. These analysis will be presented in two parts in order to provide a detailed study of two different aspects of gun incidents.

3. Data Cleaning

We obtained the data set to be used for our project from https://www.gunviolencearchive.org and it had a huge volume of data which required a lot of cleaning and preprocessing to bring it to the format on which we could do our analysis for hypothesis testing. The preprocessing of our Gun Violence in US data set required a lot of steps for which we used the following tools,

1. Microsoft Excel – for major initial data manipulations (including Text to columns)
2. Notepad/Text Edit – for applying delimiter to date format
3. Python – merging and sorting
4. MySQL Workbench – Used for adding an aggregate function since data was huge
5. R Studio – Data cleaning, removing NAs and filtering (apart from hypothesis testing)

Since our original dataset was very huge with more that 245,000 rows, we had to use multiple tools at our disposal for arriving at our desired result for which we had invested a lot of time since the data was irregular.

Details of the Gun Violence in US dataset:

The different fields of our data and their data type are as follows:

field	type	description
incident_id	int	gunviolencearchive.org ID for incident
date	str	date of occurrence
state	str	
city_or_county	str	
address	str	address where incident took place
n_killed	int	number of people killed
n_injured	int	number of people injured
incident_url	str	link to gunviolencearchive.org webpage containing details of incident
source_url	str	link to online news story concerning incident
incident_url_fields_missing	bool	ignore, always False
congressional_district	int	
gun_stolen	dict[int, str]	key: gun ID, value: 'Unknown' or 'Stolen'
gun_type	dict[int, str]	key: gun ID, value: description of gun type
incident_characteristics	list[str]	list of incident characteristics
latitude	float	
location_description	str	description of location where incident took place
longitude	float	
n_guns_involved	int	number of guns involved
notes	str	additional notes about the incident
participant_age	dict[int, int]	key: participant ID
participant_age_group	dict[int, str]	key: participant ID, value: description of age group, e.g. 'Adult 18+'

participant_gender	dict[int, str]	key: participant ID, value: 'Male' or 'Female'
participant_name	dict[int, str]	key: participant ID
participant_relationship	dict[int, str]	key: participant ID, value: relationship of participant to other participants
participant_status	dict[int, str]	key: participant ID, value: 'Arrested', 'Killed', 'Injured', or 'Unharmed'
participant_type	dict[int, str]	key: participant ID, value: 'Victim' or 'Subject-Suspect'
sources	list[str]	links to online news stories concerning incident
state_house_district	int	
state_senate_district	int	

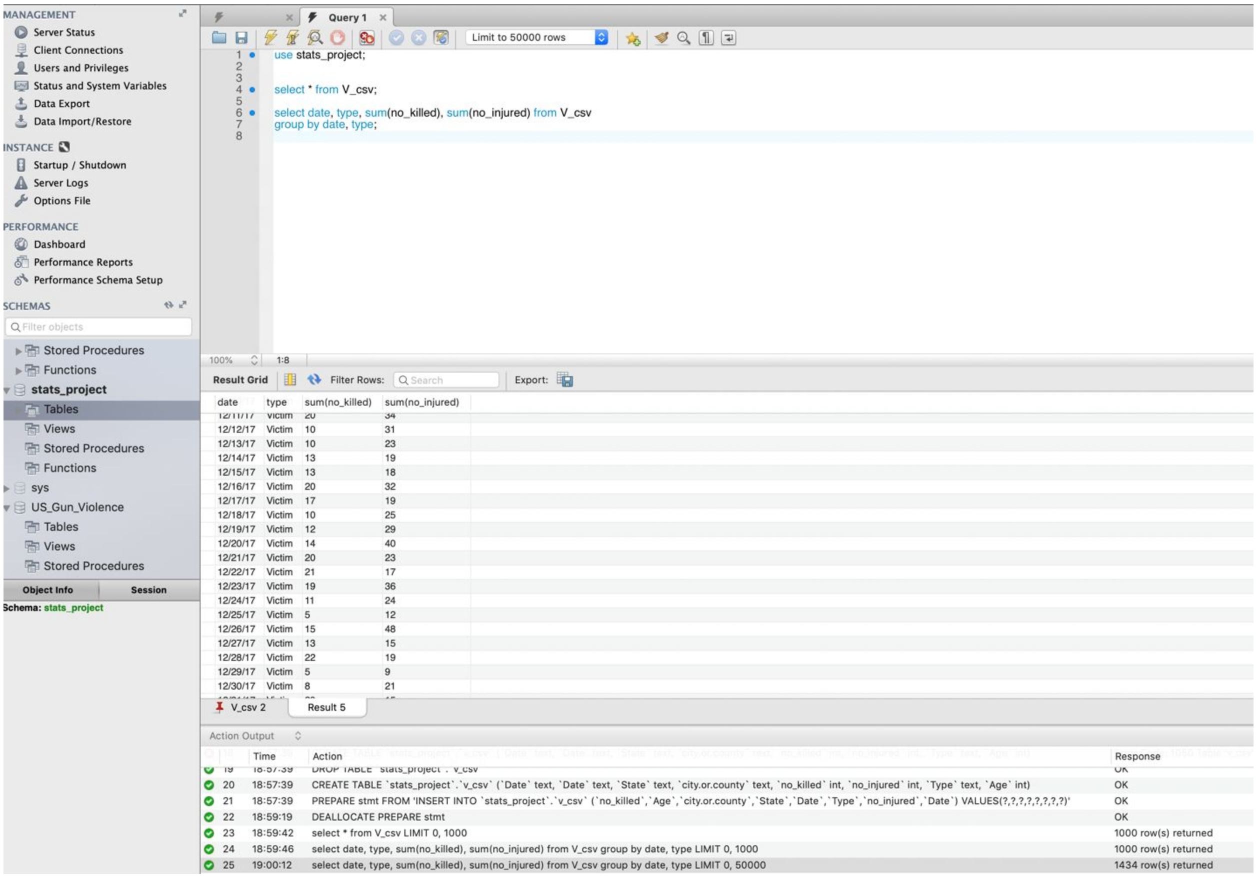
Important notes regarding the data format:

- Each list is encoded as a string with separator ||. For example, "a||b" represents ['a', 'b'].
- Each dict is encoded as a string with outer separator || and inner separator ::. For example, 0::a, 1::b represents {0: 'a', 1: 'b'}.
- The "gun ID" and "participant ID" are numbers specific to a given incident that refer to a particular gun/person involved in that incident. For example, this:
- participant_age_group = 0::Teen 12-17||1::Adult 18+
- participant_status = 0::Killed||1::Injured
- participant_type = 0::Victim||1::Victim

corresponds to this:

	Age Group	**Status**	**Type**	
Participant #0	Teen 12-17	Killed	Victim	
Participant #1	Adult 18+	Injured	Victim	

MySQL workbench:



```
use stats_project;


select * from V_csv;

select date, type, sum(no_killed), sum(no_injured) from V_csv
group by date, type;
```

date	type	sum(no_killed)	sum(no_injured)
12/11/17	Victim	20	34
12/12/17	Victim	10	31
12/13/17	Victim	10	23
12/14/17	Victim	13	19
12/15/17	Victim	13	18
12/16/17	Victim	20	32
12/17/17	Victim	17	19
12/18/17	Victim	10	25
12/19/17	Victim	12	29
12/20/17	Victim	14	40
12/21/17	Victim	20	23
12/22/17	Victim	21	17
12/23/17	Victim	19	36
12/24/17	Victim	11	24
12/25/17	Victim	5	12
12/26/17	Victim	15	48
12/27/17	Victim	13	15
12/28/17	Victim	22	19
12/29/17	Victim	5	9
12/30/17	Victim	8	21

```
     Time      Action                                                                                                                                          Response
19   18:57:39  DROP TABLE `stats_project`.`v_csv`                                                                                                             OK
20   18:57:39  CREATE TABLE `stats_project`.`v_csv` (`Date` text, `Date` text, `State` text, `city.or.county` text, `no_killed` int, `no_injured` int, `Type` text, `Age` int)    OK
21   18:57:39  PREPARE stmt FROM 'INSERT INTO `stats_project`.`v_csv` (`no_killed`,`Age`,`city.or.county`,`State`,`Date`,`Type`,`no_injured`,`Date`) VALUES(?,?,?,?,?,?,?,?)'    OK
22   18:59:19  DEALLOCATE PREPARE stmt                                                                                                                        OK
23   18:59:42  select * from V_csv LIMIT 0, 1000                                                                                                              1000 row(s) returned
24   18:59:46  select date, type, sum(no_killed), sum(no_injured) from V_csv group by date, type LIMIT 0, 1000                                                 1000 row(s) returned
25   19:00:12  select date, type, sum(no_killed), sum(no_injured) from V_csv group by date, type LIMIT 0, 50000                                                1434 row(s) returned
```

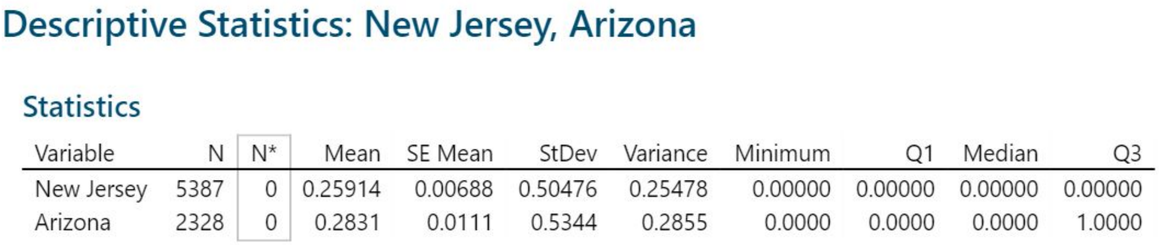
MySQL the aggregate function SUM helped us to add the no of killings that occurred on multiple incidents but on the same date. This resolved the problem of having a lot of rows for the same dates as it was not possible to proceed with the Z test on R before we could group all the data for the number of killings together.

4. Descriptive Statistics:

An Analysis of the Effectiveness of State Gun Control Laws

The primary problem that this analysis sets out to answer is whether states with strict gun control laws have less dangerous gun incidents, specifically involving mass shooting incidents. In particular, we will be looking at the parameter of the number of people killed per incident as a measure of gun control effectiveness. While this is a simplified approach which does not account for many possible variables, looking at the number of killed per incident should still serve as a good representative measure. Strict gun control laws often include policies for extensive buyer and seller background checks, restrictions on the types and number of guns that can be purchased, age requirements, and policies regarding open carry.

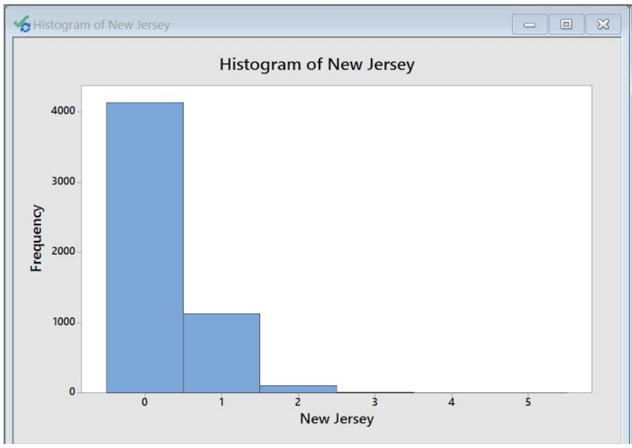
A stratified random sampling method was used to obtain two target population samples, one representative of states with strict gun control laws and the other representative of states with lenient gun control laws. Because of the relative size of the data, this sampling was the most efficient in order to obtain smaller representative samples that can be directly compared using statistical methods. Using outside sources ranking states gun control legislation, two stratified lists of the fifteen most strict gun law states and the fifteen least strict gun law states were established. The lists were then numbered and a random number is generated to select one state from each list. This sampling method resulting in picking Arizona as a representative sample of a lenient gun control states and New Jersey as representative of a strict gun control states. The database of gun incidents was then sorted and Minitab 18 was used to obtain basic descriptive statistics of the data.

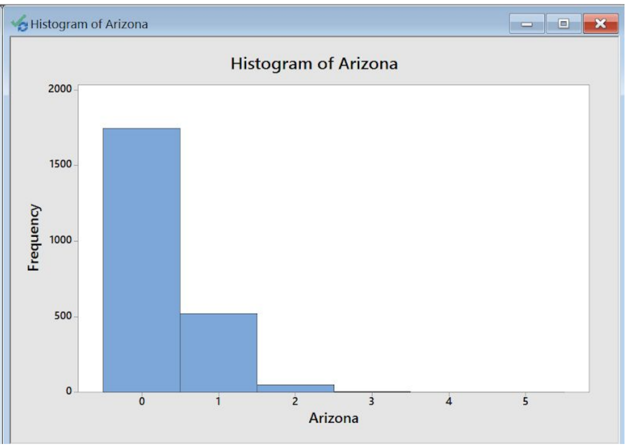


Descriptive Statistics: New Jersey, Arizona

Statistics

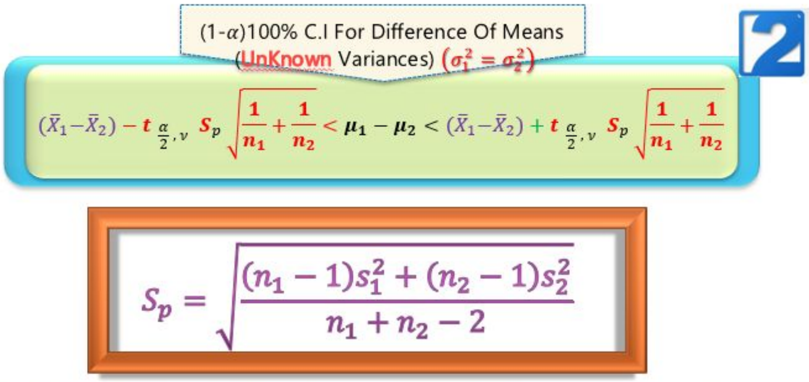
Variable	N	N*	Mean	SE Mean	StDev	Variance	Minimum	Q1	Median	Q3
New Jersey	5387	0	0.25914	0.00688	0.50476	0.25478	0.00000	0.00000	0.00000	0.00000
Arizona	2328	0	0.2831	0.0111	0.5344	0.2855	0.0000	0.0000	0.0000	1.0000





As can be seen from both the descriptive and graphical summary of the data, it appears that there is a small difference in the mean value of casualties. New Jersey, with strict gun control laws, has a mean of .259 casualties per incident, while Arizona, with lenient gun control laws, has a mean of .283 casualties per incident. The histograms and variance also reveal that the distribution of the two sample populations is also quite similar, despite the difference in sample size.

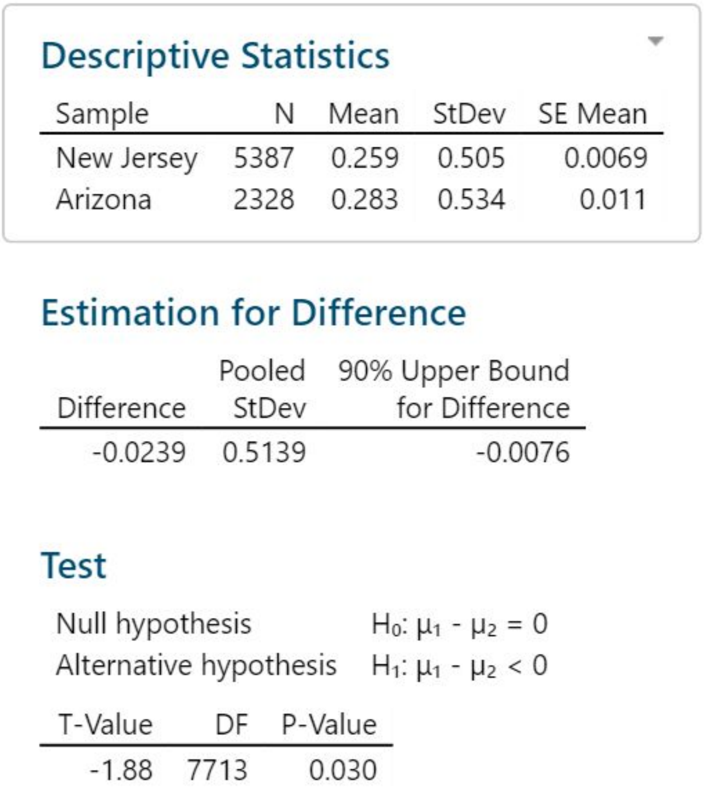
According to the Central Limit Theorem, if n > 30, then it can be safely assumed that the sampling distribution of the mean is approximately normal. Along with the assumption that the two random samples are independent, then the confidence interval of the difference of means can be calculated. In this case, the population variances are unknown, but are assumed to be approximately equal. Based on this information, Minitab can once again be used to calculate the 90% Confidence Interval, following the formula given below:



(1-α)100% C.I For Difference Of Means (UnKnown Variances) $(\sigma_1^2 = \sigma_2^2)$

$$(\bar{X}_1-\bar{X}_2) - t_{\frac{\alpha}{2},v}\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1-\bar{X}_2) + t_{\frac{\alpha}{2},v}\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$S_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$



Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
New Jersey	5387	0.259	0.505	0.0069
Arizona	2328	0.283	0.534	0.011

Estimation for Difference

Difference	Pooled StDev	90% Upper Bound for Difference
-0.0239	0.5139	-0.0076

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 < 0$

T-Value	DF	P-Value
-1.88	7713	0.030

These results show that with 90% confidence, the difference between the mean number of casualties between states with strict gun control laws is less than states with lenient gun control laws. The results of the hypothesis test shows that the p-value of the null hypothesis that the difference between population means equals zero is only 0.03. The low p-value is indicative that this hypothesis must be rejected and the alternative hypothesis that the difference between the population means is less than zero is accepted.

As a conclusion, using inferential statistics, this analysis shows that there is indeed a statistically significant difference in the mean number of casualties per incident in states with different levels of gun control. It seems that stricter gun control laws do indeed help to reduce the number of casualties stemming from gun incidents. The results from this analysis, while not particularly unexpected, does appear to counter popular claims that stricter gun control laws have no effect and are not worth the expense of implementing.

5.INFERENTIAL STATISTICS:

Claim:

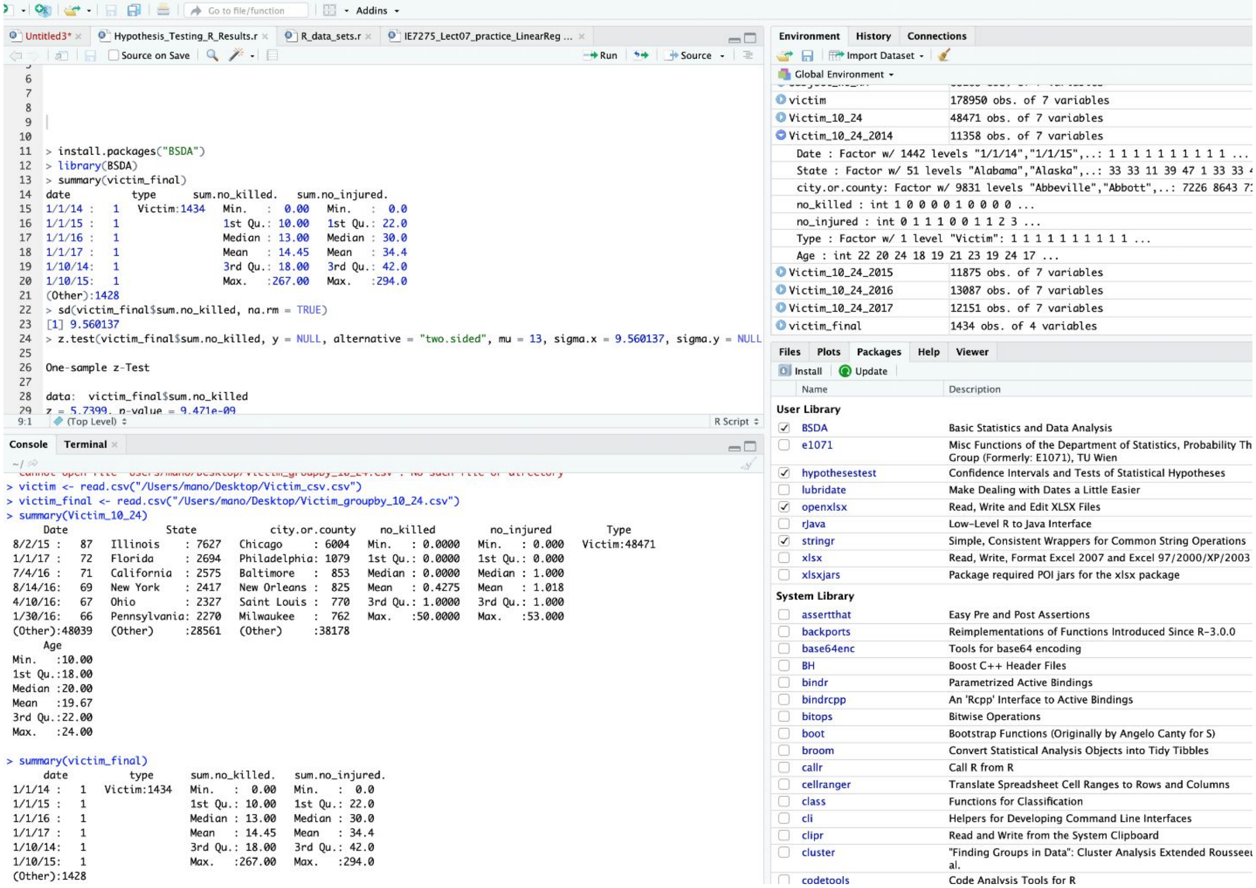
Children's Hospital of Philadelphia Research Institute claimed that in recent years, the average number of victims in the age group of 10-24 years old, who are either killed or injured due to gun violence is equal to 13 per day.

Our hypothesis is devised to verify if the claim hold out to be true. In the digital age, a key role of hypothesis testing is to identify false information based on data.

Hypothesis:

H1: $\mu \neq 13$
H0: $\mu = 13$

Having filtered the data for victims within the age group of 10-24, the summary of the data for people killed due to gun violence (obtained in R) provides the following parameters



```
> install.packages("BSDA")
> library(BSDA)
> summary(victim_final)
date           type       sum.no_killed.   sum.no_injured.
1/1/14 :    1   Victim:1434   Min.    :  0.00   Min.    :  0.0
1/1/15 :    1                 1st Qu.: 10.00   1st Qu.: 22.0
1/1/16 :    1                 Median : 13.00   Median : 30.0
1/1/17 :    1                 Mean   : 14.45   Mean   : 34.4
1/10/14:    1                 3rd Qu.: 18.00   3rd Qu.: 42.0
1/10/15:    1                 Max.   :267.00   Max.   :294.0
(Other):1428
> sd(victim_final$sum.no_killed, na.rm = TRUE)
[1] 9.560137
> z.test(victim_final$sum.no_killed, y = NULL, alternative = "two.sided", mu = 13, sigma.x = 9.560137, sigma.y = NULL

One-sample z-Test

data:  victim_final$sum.no_killed
z = 5.7399, p-value = 9.471e-09
```

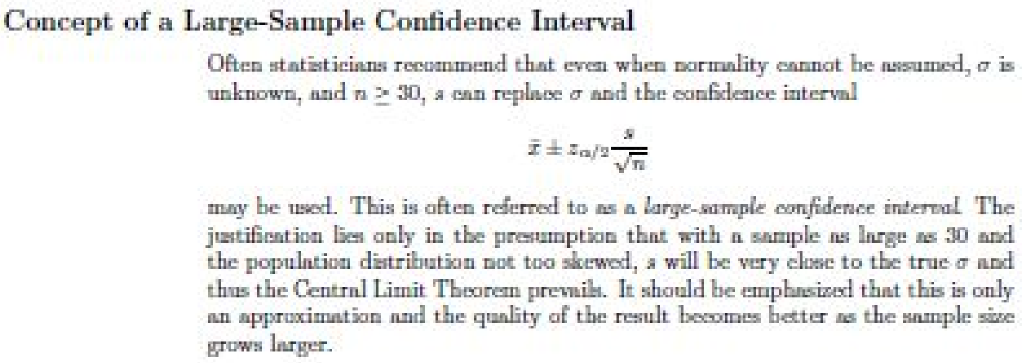
```
> victim <- read.csv("/Users/mano/Desktop/Victim_csv.csv")
> victim_final <- read.csv("/Users/mano/Desktop/Victim_groupby_10_24.csv")
> summary(Victim_10_24)
      Date                 State             city.or.county     no_killed          no_injured          Type      
 8/2/15 :   87   Illinois    : 7627   Chicago     : 6004   Min.   : 0.0000   Min.   : 0.000   Victim:48471  
 1/1/17 :   72   Florida     : 2694   Philadelphia: 1079   1st Qu.: 0.0000   1st Qu.: 0.000                 
 7/4/16 :   71   California  : 2575   Baltimore   :  853   Median : 0.0000   Median : 1.000                 
 8/14/16:   69   New York    : 2417   New Orleans :  825   Mean   : 0.4275   Mean   : 1.018                 
 4/10/16:   67   Ohio        : 2327   Saint Louis :  770   3rd Qu.: 1.0000   3rd Qu.: 1.000                 
 1/30/16:   66   Pennsylvania: 2270   Milwaukee   :  762   Max.   :50.0000   Max.   :53.000                 
 (Other):48039   (Other)     :28561   (Other)     :38178                                                    
      Age       
 Min.   :10.00  
 1st Qu.:18.00  
 Median :20.00  
 Mean   :19.67  
 3rd Qu.:22.00  
 Max.   :24.00  

> summary(victim_final)
      date           type       sum.no_killed.   sum.no_injured.
 1/1/14 :    1   Victim:1434   Min.   :  0.00   Min.   :  0.0   
 1/1/15 :    1                 1st Qu.: 10.00   1st Qu.: 22.0   
 1/1/16 :    1                 Median : 13.00   Median : 30.0   
 1/1/17 :    1                 Mean   : 14.45   Mean   : 34.4   
 1/10/14:    1                 3rd Qu.: 18.00   3rd Qu.: 42.0   
 1/10/15:    1                 Max.   :267.00   Max.   :294.0   
 (Other):1428                                                   
```

```
victim              178950 obs. of 7 variables
Victim_10_24        48471 obs. of 7 variables
Victim_10_24_2014   11358 obs. of 7 variables
  Date : Factor w/ 1442 levels "1/1/14","1/1/15",..: 1 1 1 1 1 1 1 1 1 1 ...
  State : Factor w/ 51 levels "Alabama","Alaska",..: 33 33 11 39 47 1 33 33
  city.or.county: Factor w/ 9831 levels "Abbeville","Abbott",..: 7226 8643 7
  no_killed : int 1 0 0 0 0 1 0 0 0 0 ...
  no_injured : int 0 1 1 1 0 0 1 1 2 3 ...
  Type : Factor w/ 1 level "Victim": 1 1 1 1 1 1 1 1 1 1 ...
  Age : int 22 20 24 18 19 21 23 19 24 17 ...
Victim_10_24_2015   11875 obs. of 7 variables
Victim_10_24_2016   13087 obs. of 7 variables
Victim_10_24_2017   12151 obs. of 7 variables
victim_final        1434 obs. of 4 variables
```

	Name	Description
User Library		
☑	BSDA	Basic Statistics and Data Analysis
□	e1071	Misc Functions of the Department of Statistics, Probability Th Group (Formerly: E1071), TU Wien
☑	hypothesestest	Confidence Intervals and Tests of Statistical Hypotheses
□	lubridate	Make Dealing with Dates a Little Easier
☑	openxlsx	Read, Write and Edit XLSX Files
□	rJava	Low-Level R to Java Interface
☑	stringr	Simple, Consistent Wrappers for Common String Operations
□	xlsx	Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003
□	xlsxjars	Package required POI jars for the xlsx package
System Library		
□	assertthat	Easy Pre and Post Assertions
□	backports	Reimplementations of Functions Introduced Since R-3.0.0
□	base64enc	Tools for base64 encoding
□	BH	Boost C++ Header Files
□	bindr	Parametrized Active Bindings
□	bindrcpp	An 'Rcpp' Interface to Active Bindings
□	bitops	Bitwise Operations
□	boot	Bootstrap Functions (Originally by Angelo Canty for S)
□	broom	Convert Statistical Analysis Objects into Tidy Tibbles
□	callr	Call R from R
□	cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns
□	class	Functions for Classification
□	cli	Helpers for Developing Command Line Interfaces
□	clipr	Read and Write from the System Clipboard
□	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeu al.
□	codetools	Code Analysis Tools for R

The estimated sample mean is 14 and the sample standard deviation is 9.560137. The population standard deviation is unknown.

From the below-mentioned concept in the book (Probability and Statistics for Engineers & Scientists), the hypothesis test was done in R using sample standard deviation and sample mean, in the case of n>30 and population standard deviation being unknown with a 95% confidence interval.

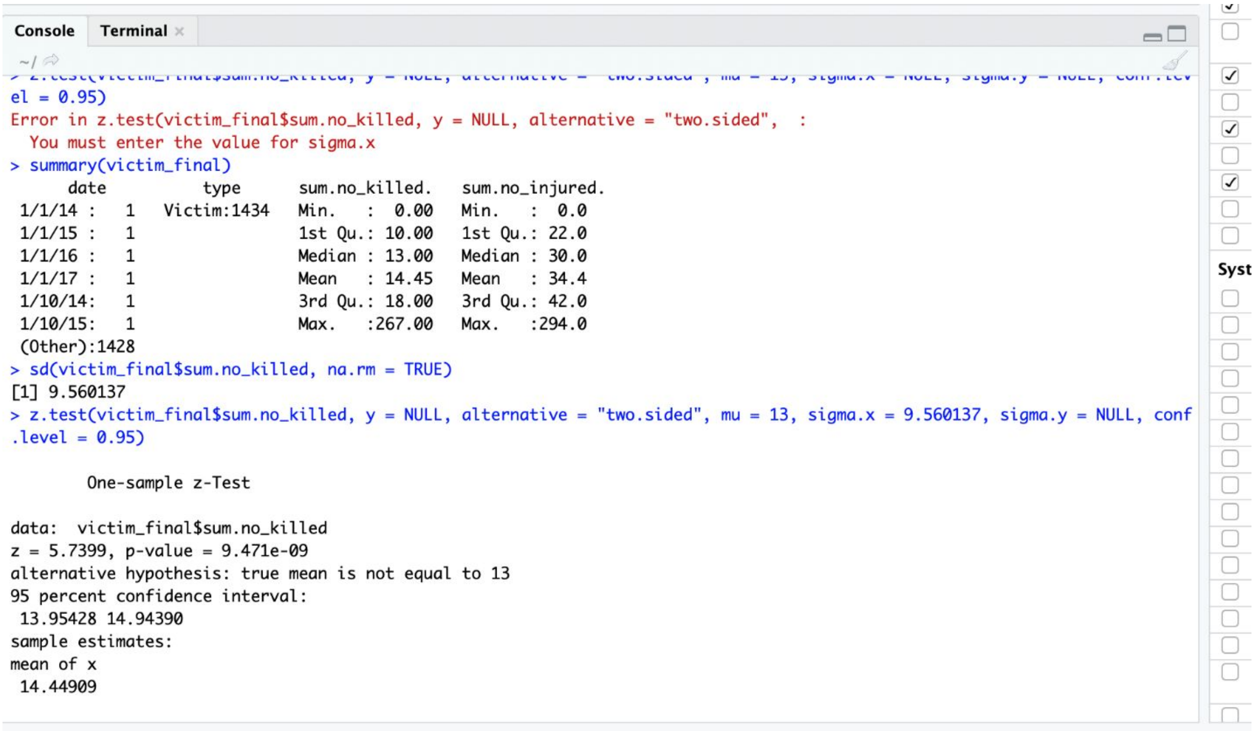


Concept of a Large-Sample Confidence Interval

Often statisticians recommend that even when normality cannot be assumed, σ is unknown, and $n \geq 30$, s can replace σ and the confidence interval

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

may be used. This is often referred to as a *large-sample confidence interval*. The justification lies only in the presumption that with a sample as large as 30 and the population distribution not too skewed, s will be very close to the true σ and thus the Central Limit Theorem prevails. It should be emphasized that this is only an approximation and the quality of the result becomes better as the sample size grows larger.



```
> z.test(victim_final$sum.no_killed, y = NULL, alternative = "two.sided", mu = 13, sigma.x = NULL, sigma.y = NULL, conf.lev
el = 0.95)
Error in z.test(victim_final$sum.no_killed, y = NULL, alternative = "two.sided",  :
  You must enter the value for sigma.x
> summary(victim_final)
      date           type      sum.no_killed.   sum.no_injured.
 1/1/14 :   1   Victim:1434   Min.   :  0.00   Min.   :  0.0
 1/1/15 :   1                 1st Qu.: 10.00   1st Qu.: 22.0
 1/1/16 :   1                 Median : 13.00   Median : 30.0
 1/1/17 :   1                 Mean   : 14.45   Mean   : 34.4
 1/10/14:   1                 3rd Qu.: 18.00   3rd Qu.: 42.0
 1/10/15:   1                 Max.   :267.00   Max.   :294.0
 (Other):1428
> sd(victim_final$sum.no_killed, na.rm = TRUE)
[1] 9.560137
> z.test(victim_final$sum.no_killed, y = NULL, alternative = "two.sided", mu = 13, sigma.x = 9.560137, sigma.y = NULL, conf
.level = 0.95)

	One-sample z-Test

data:  victim_final$sum.no_killed
z = 5.7399, p-value = 9.471e-09
alternative hypothesis: true mean is not equal to 13
95 percent confidence interval:
 13.95428 14.94390
sample estimates:
mean of x 
 14.44909 
```

<u>CONCLUSION:</u>

Since P-value is very low, the null hypothesis that the average number of victims in the age group of 10-24 years old, who are either killed or injured due to gun violence is equal to 13 per day is rejected. The average number of victims in the age group of 10-24 years old, who are either killed or injured due to gun violence is varies from 13 per day, according to the source data.