1. Introduction

An exhaustive collection of all reported crimes committed in the city of Vancouver between 2003 and 2016 was examined for this study. The data used in this analysis was provided by the Vancouver Police Department (VPD) from the VPD Records Management. It was refined to exclude data from crimes that may contain sensitive information pertaining to current investigations. The crime records contain the type of the crime, as well as the year and month in which they were committed. The locations of these crimes are described by the neighbourhood where the incidents occurred and by their corresponding longitudinal and latitudinal coordinates. The data does not reflect the total number of calls or complaints made to the VPD; only crimes that fit the attributes specified are included. This study focused on the total number of violent crimes committed in the city of Vancouver between 2003 and 2016.

The **objective** of this report is to compare estimates obtained using three sampling strategies for the crime data population. With our results, we hope to better understand the accuracy of sampling strategies and their uses in the real world. Due to the large size of our data, the purpose of finding an effective sampling strategy is to reduce the hours of research, investigation, and data entry required to obtain specific crime proportions. This can also aid us in understanding the dynamic status of crime prevalence in Vancouver. If a certain statistic is desired, we wish to find out which one of the sampling strategies to employ, and only crime types from the sample will need to be investigated to obtain an estimate for that proportion.

This data is to be viewed as population data (census) for our purposes; thus, statistics from the total data set are to be interpreted as true population data. We classified non-violent crimes as commercial and residential breaking and entering, mischief, theft of and from a vehicle, and other miscellaneous thefts. The violent crimes were specific to homicide and offence against a person. A more thorough description of the crimes is included in Appendix B. The year and month are numeric data that range from 2003-2016 and 1-12 respectively. The location data are character strings and are censored for violent crimes to protect the investigation and individuals involved.

2. Sampling Methodology

The sampling Strategies employed in the analysis were simple random sampling (SRS), stratified sampling, and systematic sampling. These sampling methods were repeated 10,000 times using bootstrapping methods to obtain the distributions of our estimates as well as the point estimates and standard errors. The point estimate is the expected value of our bootstrap estimates, and the standard error is the standard deviation of our bootstrap

estimates. The point estimates and standard error were used to calculate the mean square error. The mean square error is a measure of the effectiveness of our estimator in predicting the true value of the population statistic, and will be used to decide which of our sampling strategies is most effective in predicting the proportion and total number of violent crimes committed in Vancouver during the last thirteen years.

The true total number of violent crimes in our data set, τ, is 50,137, yielding a proportion p of 0.093. We also found a value for our sample size n using the strategy defined by Thompson (2012), which requires us to specify an allowable threshold difference d of 0.02. The worst case estimate for our population standard deviation was used with a p of 0.5. We used the proportion sample size equation, as seen in Equation (1) to find the sample size. Due to the large size of our population in comparison to our sample size, the correction factor was ignored. The resulting sample size is 2401 crime incidents.

$$n_0 = \frac{z^2 p(1-p)}{d^2} \tag{1}$$

A confidence interval for this estimate was also calculated using Equation (2). Due to the large sample size, we can use the Z-statistic obtained from the Normal distribution. The mean of our bootstrap estimates is our point estimate with the standard error being the square root of the variance of our bootstrap distribution

$$\tau = \hat{\tau} \pm Z_{\alpha/2} SE(\hat{\tau}) \tag{2}$$

2.1 Simple Random Sampling

The first sampling method that was used was the simple random sampling design, otherwise known as random sampling without replacement. In this design we selected n distinct units from our population N. Each unit selected had equal probability of being selected during each step of selection. The probability of selection is $\frac{n}{N}$ with marginal unit decreases from n and N as each unit is selected. Our unbiased estimator is defined in Equation (3).

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^{n} y_i \tag{3}$$

where y_i is the value of the i-th unit that is selected.

2.2 Stratified Sampling

In stratified sampling, we partitioned our population data by season, where each season strata contains 3 months. An estimate for each of the four strata was obtained using simple random sampling methods. We used proportional allocation to assign sample sizes for each strata accordingly. An unbiased estimator of our population total is obtained by summing the totals from all strata as can be seen in Equation (4).

$$\hat{\tau}_{st} = \sum_{h=1}^{L} \hat{\tau}_h \tag{4}$$

where $\hat{\tau}_h$ is the estimate obtained for each h-th stratum

2.3 Systematic Sampling

The third sampling method that was used was systematic sampling. Under this method, we randomly selected without replacement, a sample unit from the first 223 units that is used as our starting point; where 223 is obtained from diving the population size by the sample size desired. Our sample units were then systematically obtained by selecting every 223rd unit thereafter.

The unbiased estimator of the population total, τ can be seen in Equation (5):

$$\hat{\tau} = \frac{N}{n} \sum_{j=1}^{S} y_j \tag{5}$$

Where S =2401 is the sample size and N=223 is the number of possible primary units (systematic samples), of which n=1 will be selected. Additionally, y_j is the value of the j-th secondary unit that we would select.

3. Results

The total number of violent crimes in Vancouver from the population is found to be τ= 50,137 The estimators, the variances and corresponding 95% confidence limits that were obtained from the three sampling designs are summarized in Table 1 (below). See Appendix A for the histograms of the sampling distributions for all three sampling designs.

Table 1: Summary of Estimates Obtained by Sampling

Sampling Method	Total Number of Violent Crimes	Variance of Estimator	Lower Confidence Limit	Upper Confidence Limit	Bias	Mean Squared Error
SRS	50, 158	10,136,890	43,917	56,398	1.48	10,137,317
Stratified	50, 104	10,128,429	43,866	56,342	33.24	10,129,534
Systematic	50, 222	8,887,352	44,378	56,065	84.93	8, 894, 564

We expected the estimates of τ for the population to be unbiased, regardless of the sampling method used. This is confirmed by the biases that are obtained for the three sampling methods. The biases are extremely small, and can be considered unbiased. As the estimators are unbiased, all three sampling methods produce values of τ that are close to the true population.

Our method of choice is the systematic sampling method, because it generates the lowest Mean Squared Error (MSE).

4. Discussion:

Our results provided us with the conclusion that systematic sampling produces our best estimate due to its low MSE value compared to the other sampling methods. This result is attributed to the fact that systematic sampling accounts for the seasonality in our data from year to year, and this induces bias in our estimator as a result. However, even as a result of an increase in bias, it still retains the lowest MSE.

In the wake of multiple warnings of sexual assaults and random attacks, it is interesting that violence against other persons only accounts for approximately 9% of crimes committed in Vancouver. It appears in today's multimedia world, the majority of crimes reported in the news revolve around crimes against other persons. Of course, the protection of people is more important than the protection of property, but there is a significant amount of non-violent crime occurring that the public may not be as aware of as the violent crimes. Effective sampling strategies, such as stratifying by neighbourhood or by month, may allow the VPD to better react to or even prevent crimes based on sampling estimates.

Stat 410 Project

Abe Adeeb, Derek Qiu, Sumayya Anwer, Vinnie Liu

April 5, 2016

1 Preliminaries

Set the seed for reproducibility.

```
set.seed(123456789)
```

Load in the csv file.

```
crime = read.csv(file.choose())
```

Find the unique types of crime

```
unique(crime$TYPE)

## [1] Mischief                           Theft from Vehicle
## [3] Break and Enter Residential/Other Other Theft
## [5] Break and Enter Commercial        Offence Against a Person
## [7] Theft of Vehicle                  Homicide
## 8 Levels: Break and Enter Commercial ...
```

The violent crimes would be the following 2 types. The remaining crimes would be non-violent.

```
violent_types = list("Offence Against a Person","Homicide")
```

Now, let's find the number of violent crimes.

```
offence  = length(which(crime$TYPE==violent_types[[1]]))
homicide = length(which(crime$TYPE==violent_types[[2]]))
```

The total number of violent crimes.

```
(total = violent=offence+homicide)

## [1] 50137
```

The proportion of violent crimes.

```
(prop = violent/length(crime$TYPE))

## [1] 0.09334552
```

2 Simulation

For the acutal simulation study, I will be considering:

1. SRS
2. Stratified
3. Systematic

2.1 Simple Random Sampling Design

Total number of observations

```
L = length(crime$TYPE)
```

Sample size desired.

```
n = 2401
```

Number of bootstrap iterations

```
B = 10000
```

Performing the bootstrap.

```
boot_prop = numeric()
for (i in 1:B){
  bootsample   = sample(crime$TYPE, n, replace=TRUE)
  boot_prop[i] = (length(which(bootsample==violent_types[[1]]))+
                  length(which(bootsample==violent_types[[2]])))/n
}
```

Get our population total, that is the total number of violent crimes in our pop.

```
boot_tauhat = length(crime$TYPE)*boot_prop

#the expected value of our bootstrap estimator
tauhat_est=mean(boot_tauhat)
tauhat_est

## [1] 50157.67

#the variance of our estimator
tauhat_var=var(boot_tauhat)
tauhat_var

## [1] 10136890

tauhat_se=sqrt(tauhat_var)
tauhat_se

## [1] 3183.848
```

```r
#
UCL_tauhat=mean(boot_tauhat)+1.96*sqrt(var(boot_tauhat))
UCL_tauhat
```

```
## [1] 56398.01
```

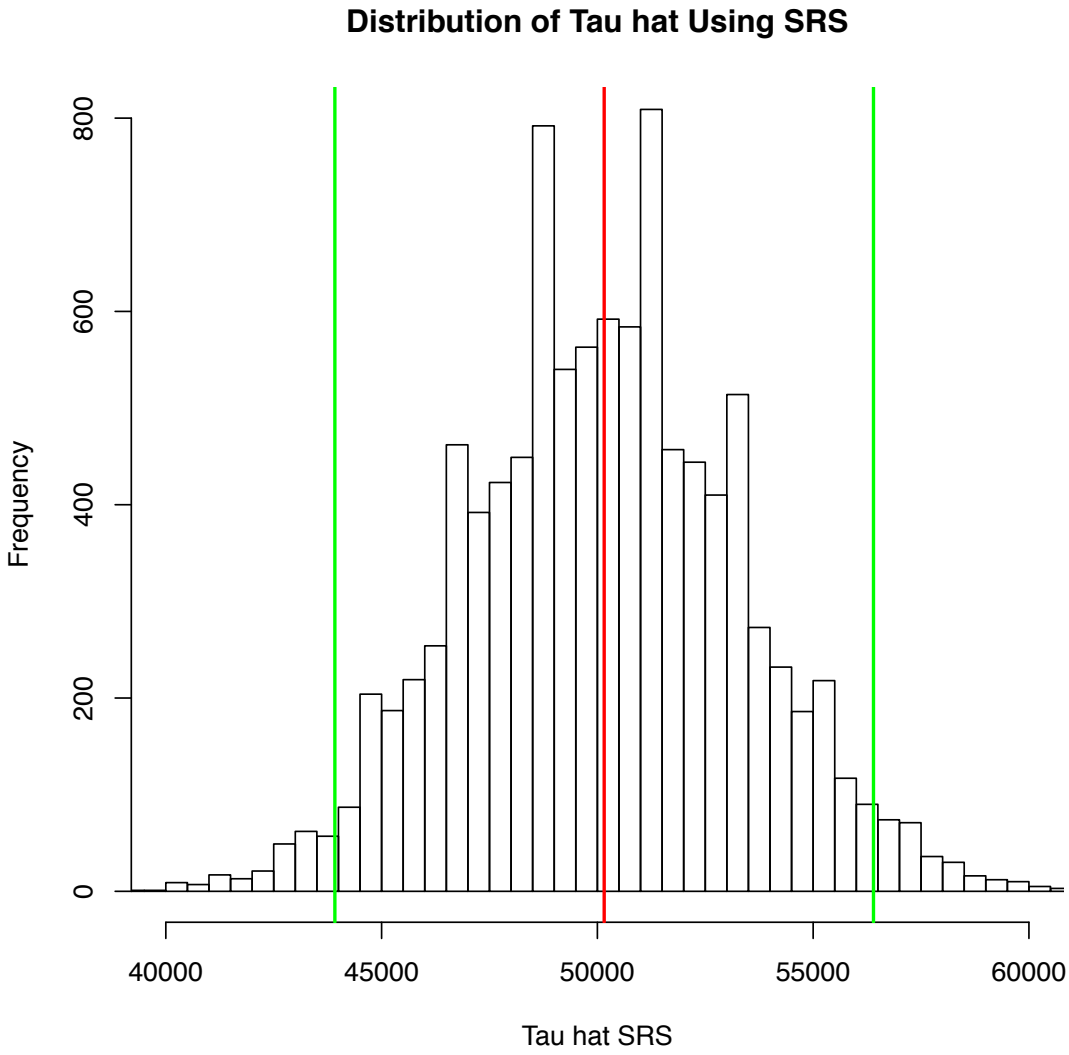
```r
LCL_tauhat=mean(boot_tauhat)-1.96*sqrt(var(boot_tauhat))
LCL_tauhat
```

```
## [1] 43917.33
```

```r
# MSE calculations
bias_srs=tauhat_est-total
MSE_srs= bias_srs^2+tauhat_var
MSE_srs
```

```
## [1] 10137317
```

```r
#histogram of distribution with mean and confidence intervals included
hist(boot_tauhat,nclass=50, xlab="Tau hat SRS",
     main ="Distribution of Tau hat Using SRS",
     xlim=c(40000,60000),ylim=(c(0,800)))
abline(v=mean(boot_tauhat),col="red",lwd=2)
abline(v=UCL_tauhat,col="green",lwd=2)
abline(v=LCL_tauhat,col="green",lwd=2)
```



2.2 Stratified Sampling (Proportional Allocation)

Find the unique months, and sort them in numerical order.

```
(U = sort(unique(crime$MONTH)))

##  [1]  1  2  3  4  5  6  7  8  9 10 11 12
```

Since we now have 12 unique months, I just split them into 4 strata based on season.

```
# winter
S1 = U[c(1,2,12)]
# spring
S2 = U[c(3,4,5)]
# summer
S3 = U[c(6,7,8)]
# fall
S4 = U[c(9,10,11)]
```

Given that we have our stratum assignments, we can now perform bootstrap.

```
#Get all the neighbourhood names.
NH = crime$MONTH

#Find the population strata sizes.
N1 = sum(NH %in% S1 * 1)
N2 = sum(NH %in% S2 * 1)
N3 = sum(NH %in% S3 * 1)
N4 = sum(NH %in% S4 * 1)
N  = N1+N2+N3+N4

#the sample size we are using. I believe it's 2401?
n=2401

#Use proportional allocation to assign sample sizes.
n1 = (n*N1)/N
n2 = (n*N2)/N
n3 = (n*N3)/N
n4 = (n*N4)/N


tauhat1 = numeric()
tauhat2 = numeric()
tauhat3 = numeric()
tauhat4 = numeric()
tau_hat = numeric()


B = 10000

#This takes a long time to run, so I already ran the simulatio and saved the result in a RData file
```

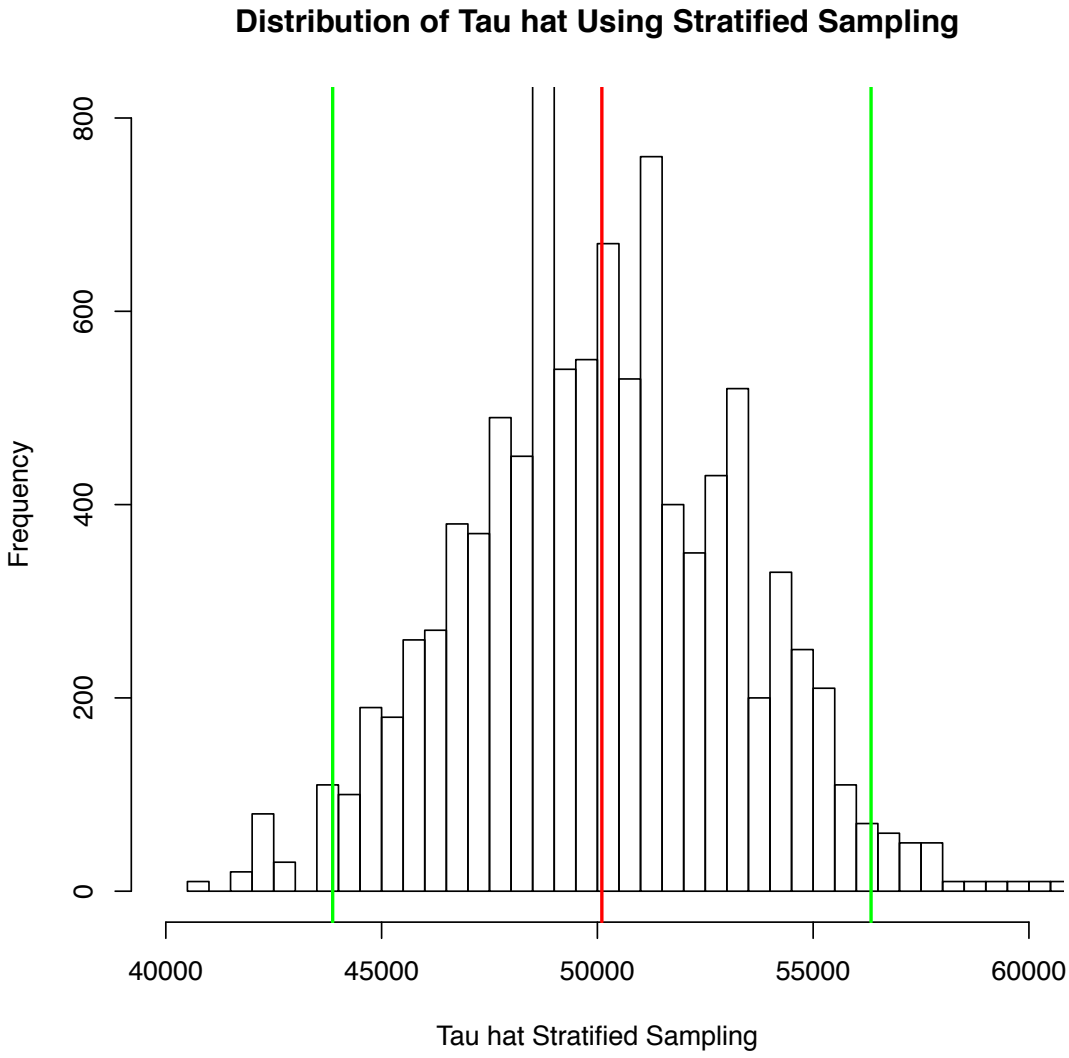
```
#if (!file.exists("stratified.RData")){
#  for(k in 1:B){
#  index_1       = sample(N1,n1)
#  bootsample_1  = ((crime[crime$MONTH %in% S1,])$TYPE)[index_1]
#  tauhat1[k]    = N1*(length(which(bootsample_1==violent_types[[1]]))+
#                      length(which(bootsample_1==violent_types[[2]])))/n1
#
#  index_2       = sample(N2,n2)
#  bootsample_2  = ((crime[crime$MONTH %in% S2,])$TYPE)[index_2]
#  tauhat2[k]    = N2*(length(which(bootsample_2==violent_types[[1]]))+
#                      length(which(bootsample_2==violent_types[[2]])))/n2
#
#  index_3       = sample(N3,n3)
#  bootsample_3  = ((crime[crime$MONTH %in% S3,])$TYPE)[index_3]
#  tauhat3[k]    = N3*(length(which(bootsample_3==violent_types[[1]]))+
#                      length(which(bootsample_3==violent_types[[2]])))/n3
#
#  index_4       = sample(N4,n4)
#  bootsample_4  = ((crime[crime$MONTH %in% S4,])$TYPE)[index_4]
#  tauhat4[k]    = N4*(length(which(bootsample_4==violent_types[[1]]))+
#                      length(which(bootsample_4==violent_types[[2]])))/n4
#  }}else{
#    load("stratified.RData")
#    tau_hat=strat
#  }

load("C:/Users/admin123/Documents/R/stratified.RData")
boot_tauhat_str=strat
```

Its mean and variance.

```
#the expected value of our bootstrap estimator
boot_tauhat_str_est= mean(boot_tauhat_str)
boot_tauhat_str_est

## [1] 50103.76

#the variance of our estimator
boot_tauhat_str_var=var(boot_tauhat_str)
boot_tauhat_str_var

## [1] 10128429

boot_tauhat_str_se=sqrt(boot_tauhat_str_var)
boot_tauhat_str_se

## [1] 3182.519

# 95% CI of our estimate
UCL_tauhat_str=mean(boot_tauhat_str)+1.96*sqrt(var(boot_tauhat_str))
UCL_tauhat_str
```

```
## [1] 56341.5

LCL_tauhat_str=mean(boot_tauhat_str)-1.96*sqrt(var(boot_tauhat_str))
LCL_tauhat_str

## [1] 43866.02

# MSE calculations
bias_str=boot_tauhat_str_est-total
bias_str

## [1] -33.24195

MSE_str= bias_str^2+boot_tauhat_str_var
MSE_str

## [1] 10129534
```

```r
#histogram of distribution with mean and confidence intervals included
#histogram of distribution with mean and confidence intervals included
hist(boot_tauhat_str,nclass=50, xlab="Tau hat Stratified Sampling",
     main ="Distribution of Tau hat Using Stratified Sampling",
     xlim=c(40000,60000),ylim=(c(0,800)))
abline(v=mean(boot_tauhat_str),col="red",lwd=2)
abline(v=UCL_tauhat_str,col="green",lwd=2)
abline(v=LCL_tauhat_str,col="green",lwd=2)
```



2.3 Systematic Sampling

Total number of observations.

```
L = length(crime$TYPE)
```

Sample size desired.

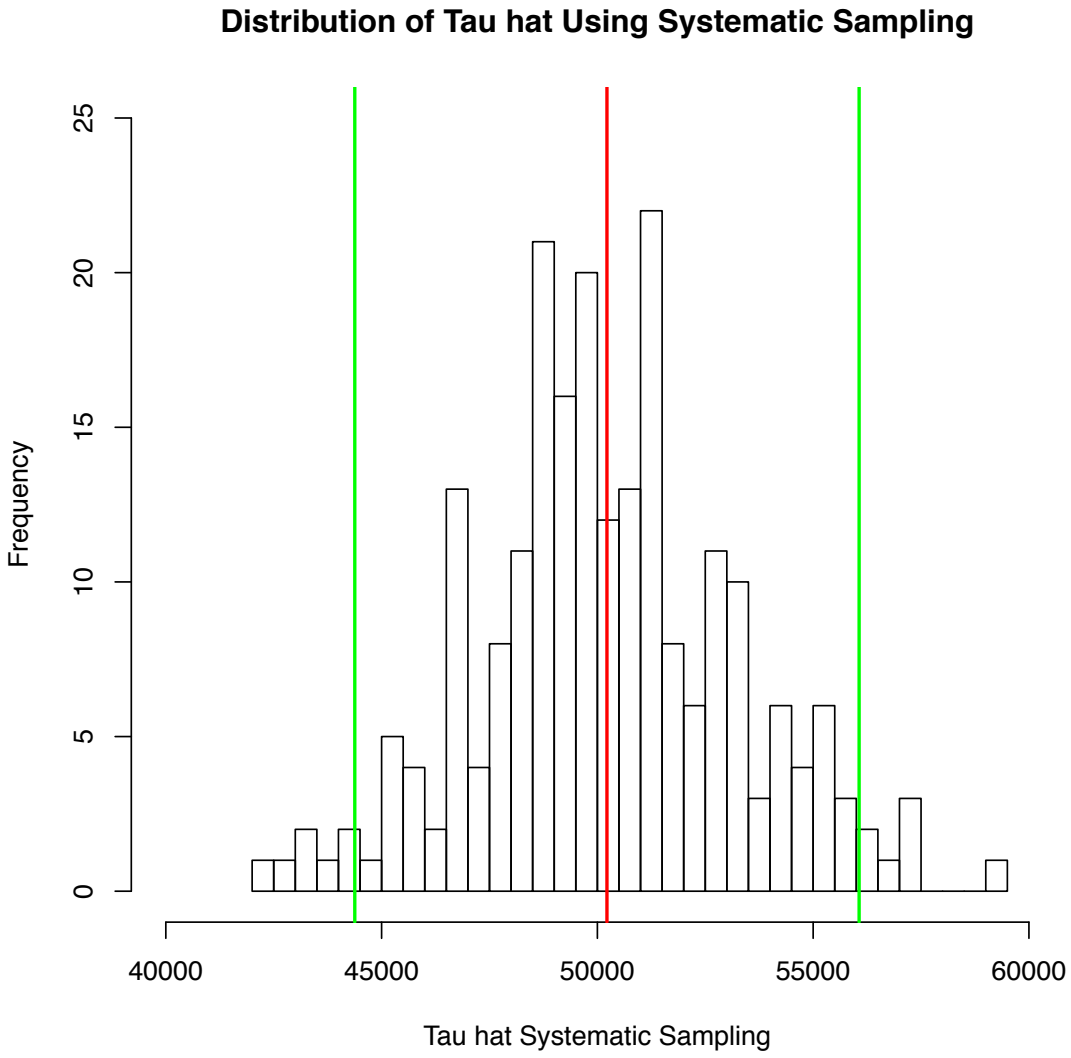
```
S = 2401
n = floor(L/S)


# Number of bootstrap iterations
B = 10000


#Performing the bootstrap sampling. I'm using Sys.time to keep track
#of how long it took for my code to run.

boot_prop_sys = numeric()

for (i in 1:B){
  i = sample(1:n,1)
  boot_index       = seq(from = i, by = n, length=S)
  bootsample       = as.character(crime$TYPE[boot_index])
  boot_prop_sys[i] = (length(which(bootsample==violent_types[[1]]))+
                        length(which(bootsample==violent_types[[2]])))/S
  }

boot_tauhat_sys = L*boot_prop_sys
```

Get our population total, that is the total number of violent crimes in our pop.

```
#the expected value of our bootstrap estimator
boot_tauhat_sys_est = mean(boot_tauhat_sys)
boot_tauhat_sys_est

## [1] 50221.93

#the variance of our estimator
boot_tauhat_sys_var=var(boot_tauhat_sys)
boot_tauhat_sys_var

## [1] 8887352

boot_tauhat_sys_se=sqrt(boot_tauhat_sys_var)
boot_tauhat_sys_se

## [1] 2981.166

# 95% CI of our estimate
UCL_tauhat_sys=mean(boot_tauhat_sys)+1.96*sqrt(var(boot_tauhat_sys))
UCL_tauhat_sys
```

```
## [1] 56065.01

LCL_tauhat_sys=mean(boot_tauhat_sys)-1.96*sqrt(var(boot_tauhat_sys))
LCL_tauhat_sys

## [1] 44378.84

# MSE calculations
bias_sys=boot_tauhat_sys_est-total
bias_sys

## [1] 84.92765

MSE_sys= bias_sys^2+boot_tauhat_sys_var
MSE_sys

## [1] 8894564
```

```r
#histogram of distribution with mean and confidence intervals included
#histogram of distribution with mean and confidence intervals included
hist(boot_tauhat_sys,nclass=50, xlab="Tau hat Systematic Sampling",
     main ="Distribution of Tau hat Using Systematic Sampling",
     xlim=c(40000,60000),ylim=(c(0,25)))
abline(v=mean(boot_tauhat_sys),col="red",lwd=2)
abline(v=UCL_tauhat_sys,col="green",lwd=2)
abline(v=LCL_tauhat_sys,col="green",lwd=2)
```



Open Data Catalogue

Crime

Data custodian	Vancouver Police Department
Data currency comments	The data on this site is scheduled to be updated every Sunday morning. Note: there can be a delay of up to a week between when data is updated in its home system and the publication to the Open Data feed.
Data set description	This is a dataset of crime data on a year-by-year basis beginning in 2003. Legal Disclaimer from Vancouver Police Department The release of Vancouver Police Department (VPD) crime data is intended to enhance community awareness of policing activity in Vancouver. Users are cautioned not to rely on the information provided to make decisions about the specific safety level of a specific location or area. By using this data the user agrees and understands that neither the Vancouver Police Department, Vancouver Police Board nor the City of Vancouver assumes liability for any decisions made or actions taken or not taken by the user in reliance upon any information or data provided. While every effort has been made to be transparent in this process, users should be aware that this data is designed to provide individuals with a general overview of incidents falling into several crime categories. The information provided therefore does not reflect the total number of calls or complaints made to the VPD. Please refer to the FAQ for further details. The data provided is based upon information contained in the VPD Records Management System. The crime classification and file status may change at any time based on the dynamic nature of police investigations. The VPD has taken great care to protect the privacy of all parties involved in the incidents reported. No personal or identifying information has been provided in the data. Locations for reported incidents involving Offences Against a Person have been deliberately randomized to several blocks and offset to an intersection. No time or street location name will be provided for these offences. For property related offences, the VPD has provided the location to the hundred block of these incidents within the general area of the block. All data must be considered offset and users should not interpret any locations as related to a specific person or specific property.

Data accuracy comments	The Vancouver Police Department's GeoDASH Crime Map remains the authoritative source. Note: GeoDASH stands for Geographic Data Analysis and Statistics Hub. It is a crime mapping tool used by Vancouver Police Department (VPD) to inform residents on the crime activities happening in Vancouver.
Attributes	▪ TYPE ▪ YEAR ▪ MONTH ▪ HUNDRED_BLOCK ▪ NEIGHBOURHOOD ▪ X ▪ Y
Websites for further information	▪ GeoDASH Crime Map ▪ GeoDASH FAQ
Coordinate system	▪ All coordinates data in this data set must be considered offset and is projected in latitude and longitude. Users should not interpret any locations as related to a specific person or specific property. ▪ Coordinates data for records with "Offset to Protect Privacy" is not disclosed to provide privacy protection.
Data set details	I have read and understood the disclaimer above. Please click the "I Agree" button below to access crime data files. I Agree
Note	Data provided does not reflect the total number of calls or complaints made to the VPD. Only the categories described in the attributes and that occurred from January 1, 2003 are included. Certain crimes are excluded for privacy and investigative reasons. Please refer to the above FAQ for further details. If you have a questions that was not answered on this page or the above FAQ page, please email VPD.