Statistical Analysis of Breaking and Enterings in Toronto

The Importance of Location and Time

Yusra Fayyaz, Tanya Thaker, Keya Patel and Andre Graham
Tutorial 0106, Group #4

March 30, 2020

Introduction

We carried out an analysis on the open source data set provided to us from the Toronto Police Service. This data set contains information on break-in and enters for the population of Toronto. We analysed this data set to shed some light on the spatial and temporal significance in regards to break-in and enters. In this process, we looked at specific situations that would be useful for the TPS when handling B&E's. This dataset is a subset of the Major Crime Indicators (MCI) and the data has been collected throughout 2014 to 2019.

Objectives

The questions we are looking to find answers to are:

1. Is the proportion of break-in and enters in commercial buildings 0.3?
2. Is the number of break-ins during the day the same as the number of break-ins during the night?
3. What is the average time it takes to report a break-in in hours after the occurrence?

Why are these questions important?

These questions will be useful in order to...

1. Provide more information as to whether commercial buildings truly are less likely to experience B&Es compared to residential buildings, like apartments and houses. The results to this question would provide the TPS with a better understanding to determine which type of building requires additional precautionary and security measures. For example, the TPS could implement a required number of security systems (alarms, security cameras, etc.) for the buildings that we find are at a higher risk of B&E's.

Why are these questions important? (continued)

2. Provide information in regards to what time of day break-ins are more likely to occur. This would allow the TPS to take neccessary action for when to be more cautious or when to have readily available services. For instance, they can implement additional security such as neighbourhood watch.
3. Use the average time it takes to report a break-in so that when a break-in is reported, there can be an estimated time of occurrence immediately before further research. Also, this can provide more information on the case and approximate times to look back at security camera footage or other factors.

Data Summary for Questions 1

How did we clean the data?

From the original data, we selected the variables needed for the analysis, which were the type of building and month at which the break-in was reported. After calculating the test statistic, we filtered out the break-ins that occurred in houses and apartments, since for our analysis we were only looking at commercial buildings.

What variables did we create?

We created a new variable called "premise" which grouped the houses and apartments from the variable "premisetype" into the residential group, so that we could carry on with calculating our test statistic and hypothesis test more efficiently.

Data Summary for Questions 2

How did we clean the data?

In order to filter and clean the data according to our needs, we selected the month at which the break-in was reported, as well as the break-in time variable. Then, we found the total number of breakins for day and night for each month and saved it into a new data file.

What variables did we create?

We only added two new variables to the dataset. The first additional variable was "day_or_night" that stored whether the break-in occured during the day or night. We considered night as a period that starts around 6 pm and lasts until about 6 am and the remaining time as day. The other variable we created was "total" that contains the total number of breakins per month depending on the time of day.

Data Summary for Questions 3:

How did we clean the data?

To clean the data, first we selected the variables which were of importance to this question, the hour of occurrence and and the hour at which the break-in was reported. Then, we filtered the data so that it would only include observations that were reported on the same day that they occurred. This was to avoid outliers from people who were, for instance, on vacation or absent for a long period of time.

What variables did we create?

We created one additional variable that was important to answer this question, called "time_since_occurrence". This variable is essentially the difference between the reported hour and the occurrence hour in our filtered data in order to see how long each observation took to be reported after it occurred (in hours).

Statistical Methods for Question 1

For question 1, the statistical method we used was a single proportion hypothesis test. This is because we are analyzing whether or not the proportion of B&Es that occur in commercial buildings is about 0.3.

Variables Used

- *premisetype*: the type of building in which the B&E occurred
- *reportedmonth*: the month of year at which the B&E was reported
- *premise*: our new variable that groups houses and apartments as residential buildings, and commercial stays commercial.

Statistical Methods for Question 2

For question 2, the statistical method we used was randomised hypothesis testing since we had two groups of data, i.e. day and night. We used this in order to find evidence to support or reject our null hypothesis that there is no difference in number of breakins during the day and the night.

Variables Used

- *reportedmonth*: the month of year at which the B&E was reported
- *breakin_time*: the time when the B&E occured on a 24-hour clock
- *day_or_night*: whether the B&E occured during the Day or Night

Statistical Methods for Question 3

For question 3, the statistical method we used was bootstrap sampling. We used this in order to get a confidence interval that would provide us with a range for the average time it takes for a break and enter to be reported after occurrence.

Variables Used

- *occurrencedayofyear* : day of the year it occured - in order to filter data
- *reporteddayofyear*: day of year it was reported - set equal to above to filter
- *occurrencehour*: the hour of the day the B&E occurred
- *reportedhour*: the hour of the day the B&E was reported
- *time_since_occurrence*: reportedhour - occurrencehour
- *mean_diff*: the average time it takes to report after occurrence

Results for Question 1: Our Hypotheses

Our hypotheses were as follows:

- H_0: P $=$ 0.3
- H_A: P != 0.3

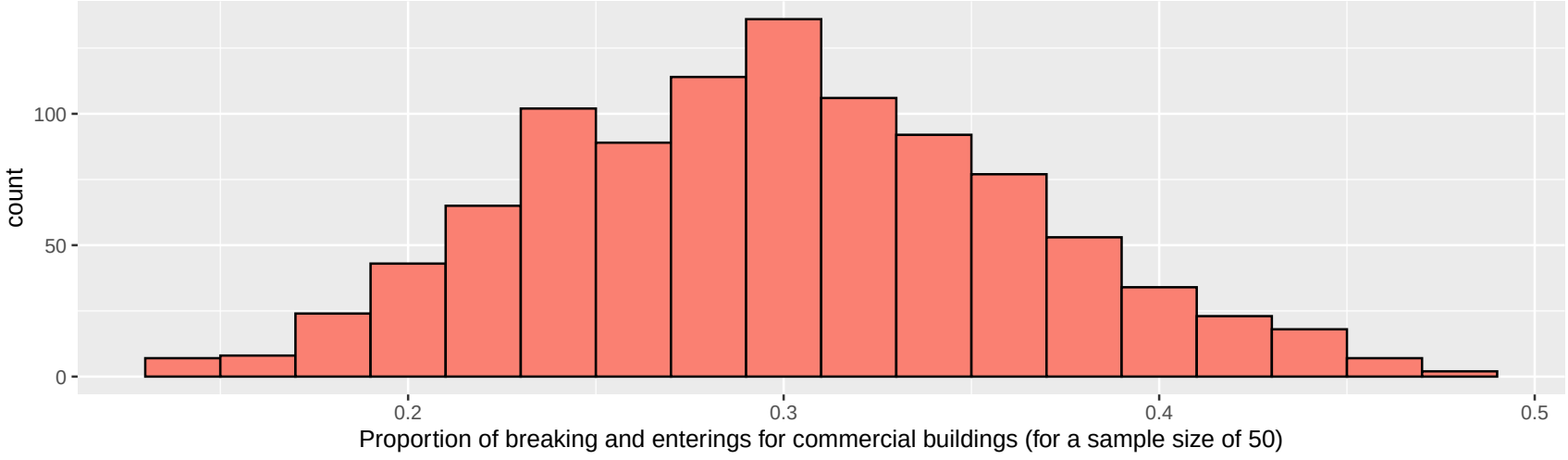
where P is the proportion of break-ins in commercial buildings

Results for Question 1: Test Statistic and Sample Simulation

Test Statistic Being Used

```
## [1] 0.336564
```

Histogram of Simulated Values



Results for Question 1: Our p-value

Based on this p-value, we are not able to reject the null hypothesis because the value is greater than 0.05 which is the significance level.

```
## # A tibble: 1 x 1
##   pvalue
##    <dbl>
## 1  0.694
```

Results for Question 2: Our Hypotheses

Our hypotheses were as follows:

- H_0: D = N
- H_A: D != N

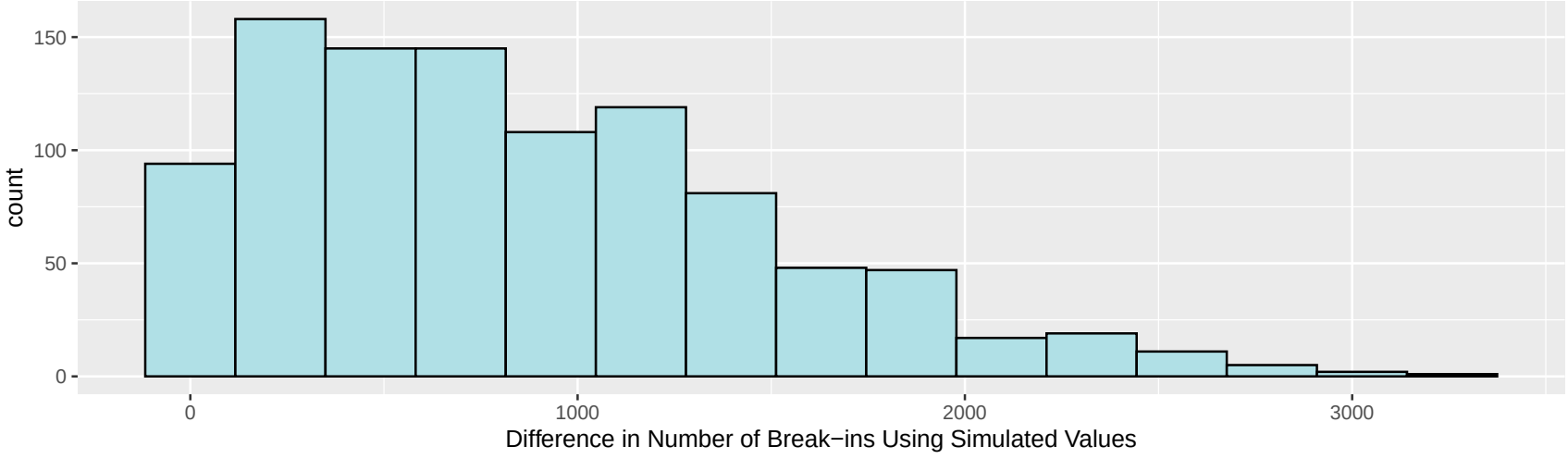
where D is the number of break-ins during the day and N is the number of break-ins during the night.

Results for Question 2: Test Statistic

Test Statistic Being Used

```
## [1] 3190
```

Histogram of Simulated Values: Differences Between Number of Break Ins for Day vs Night

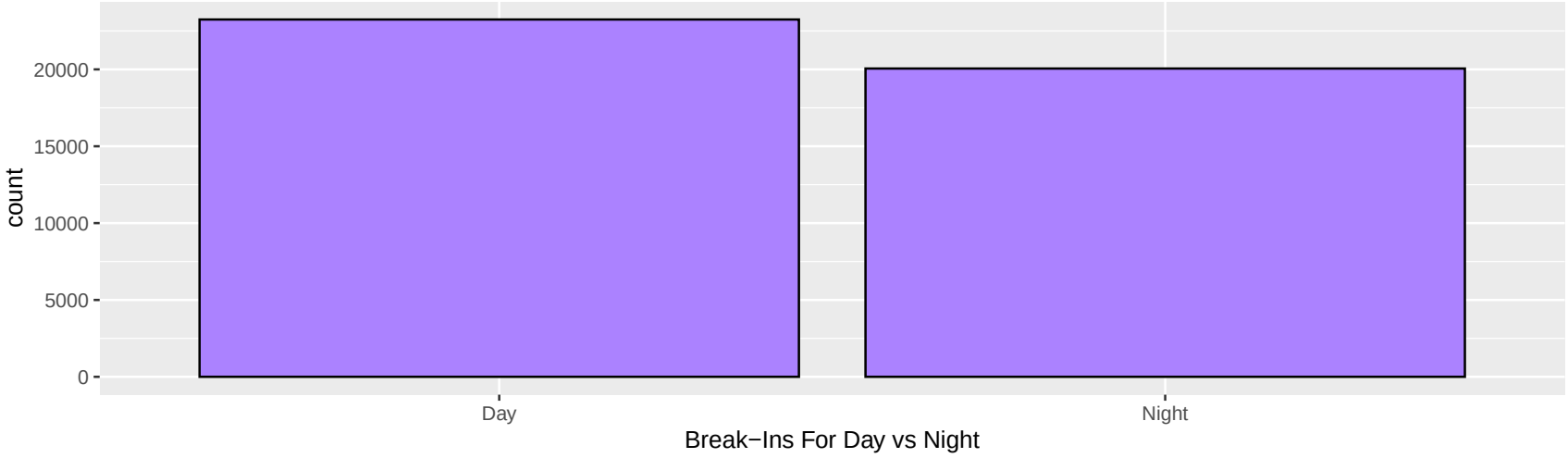


Results to Question 2: Finding the P-Value and B&Es during the Day and Night

Our P-Value: We can reject the null hypothesis because the p-value is less than our significance level of 0.05

```
## # A tibble: 1 x 1
##   pvalue
##    <dbl>
## 1  0.001
```

Comparison of Number of B&Es During the Day vs Night



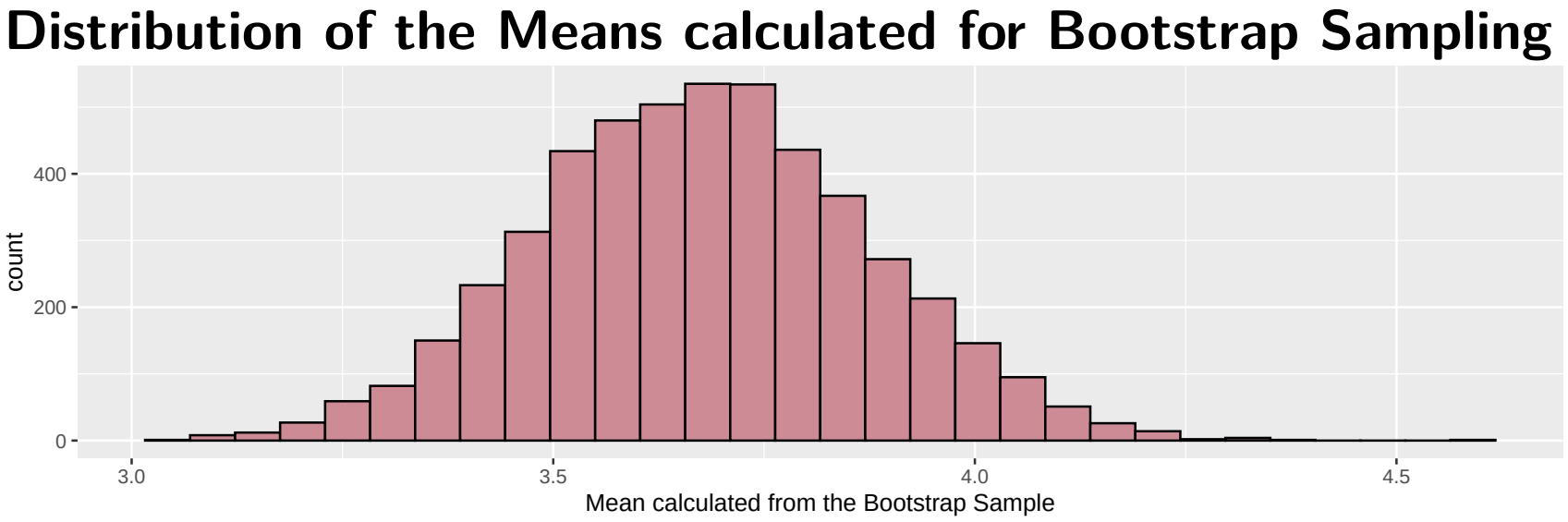
Results for Question 3: Summary Values for Initial Data

Summary Values (Mean, Median and Max)

```
## # A tibble: 1 x 3
##   mean_diff median_diff max_diff
##       <dbl>       <dbl>    <dbl>
## 1      3.68           2       23
```

Therefore, the average time it takes to report a B&E is about 3 hours and 40 minutes (3.68), the median time is 2 hours and the maximum time from our data is 23 hours.

Results for Question 3: Histogram of Means Calculated from Bootstrap Samples



Results for Question 3: Quantiles for Confidence Interval

The Values for the 2.5th and 97.5th Percentiles

```
##  2.5% 97.5%
## 3.292 4.070
```

Therefore, the interval that is the middle 95% of our bootstrap distribution is (3.3, 4.07).

Conclusion

The answers to our questions according to our results are as follows:

1. We had set a signficance level of 5%, and the p-value that resulted from the calculations is approximately 0.69. This lets us conclude that we do not have enough evidence to reject the null hypothesis. Thus, we can say that commercial buildings are indeed less likely to experience B&Es compared to residential buildings. In other words, the proportion of B&E's in commercial buildings is in fact about 0.3, and the proportion of B&E's in residential buildings is approximately 0.7.

Conclusion (continued)

2. Given that we had set our significance level at 5%, and our resulting p-value was 0.001, we can say we have strong evidence to reject the null hypothesis. Thus we can conclude the number of breakins during the day and night are not the same. Furthermore, we conclude that, based on our dataset, the number of break-ins during the day is more than the number of break-ins at night.

3. The overall results for this analysis were that our 95% confidence interval is about 3.3 - 4.07, which includes our population mean for the dataset (3.67). This suggests that we can be 95% confident that the average time it takes between occurence of a break-in and reporting a break-in is between (about) 3 hours and 18 minutes, and 4 hours and 4 minutes.

Possible Limitations

What are possible limitations of our analyses?

A possible limitation to our analyses is under recording and under reporting. It is possible that our dataset is too small compared to the population of Toronto which is about 2.93 million people, so this could lead to inaccuracy. Furthermore, not all neighbourhoods in Toronto where included in this dataset, which may result in a misrepresentation of the entirety of B&E's in Toronto.