Crime Count, Weather Condition, and Day of Week:
Are crime activities and weather conditions good excuses to cancel weekend outings with friends?

Chun-Li Chuang, Zechen Jin, Yiwen Liu, Nehar Poddar, Zheying Yu
April 13th 2020

***Abstract*—We all have times when priorities come up after making outdoor plans with friends, even if the priorities are about ourselves. However, it is always hard to come up with a convincing excuse to cancel on friends. In this paper, we conduct statistical analysis to investigate the association between the two factors of interest – crime count and weather condition – with weekends. We set up objectives and conduct various hypothesis testings to form conclusion respectively. Finally, we gather the conclusion of our analysis and show whether claiming high crime activities or bad weather conditions can be a statistically significant reason to cancel outdoor plans on others.**

I. Introduction

How many times has it happened that after a stressful, assignment full week that you do not feel like going out and having an eventful weekend? You just want to lay in bed and watch your favorite TV show but you do not know what to tell your friends. To come up with the perfect excuse without making your friends feel bad is tough and exhausting. Hence to ease your concern, the report aims to support the proposed two possible excuses, criminal activities and weather conditions, statistically.

Nobody wants to be the victim of a criminal case, especially within the city of Boston where more than 152,000 students attend college here. Still, there are about more than 15,000 major crimes that occurred within the past 12 months in Boston. Furthermore, rumor has it that while we are busy exploring Boston commons and the different party places during our long-awaited weekends, there is a higher chance of becoming a victim of a criminal case. It certainly seems like a probable issue to be concerned about during the outings.

It is never a great feeling when the weather ruins an outdoor plan. In fact, rumor has it that rain occurs more often during the weekends than weekdays, which causes many cancellations, postponements, and disappointments. However, it consequently becomes a seemingly probable cause to cancel plans on others. Even though we do have weather forecasts on the news reporting the weather prediction for the upcoming days, the prediction could vary before the occurrence and inaccurate.

With the two excuses in mind, we would like to conduct a series of statistical analyses to identify any potential associations between each excuse and weekends as well as associations between criminal activities and weather themselves. From the test results and the conclusion we established within each analysis, we will then compile our final conclusion to answer the question of whether crime activities, weather conditions, or both can be statistically significant excuses to cancel outdoor plans with friends.

II. Objective

The main objective of the report is to identify potential associations between criminal activities, weather conditions, and the day of the week. Specifically, our analyses can be broken down into three parts.

1) Crime - Weekends
   - Proportion equivalence between the crime counts on weekdays and weekends
2) Weather - Weekends
   - Proportion equivalence between the frequency of precipitation on weekdays and weekends
3) Crime - Weather
   - Existence of linear correlation between daily crime count and daily temperature
   - Association between crime count and amount of precipitation

III. Study Design

In order to conduct the analysis of our interest, we collected our datasets from the Internet. Since both datasets are collected prior to the testings, all analysis conducted throughout the report will be considered observational studies, which means that we can at most conclude our analysis with associations but not causations.

We will be using the crime incident report provided by the Boston Police Department [1]. It documented the initial details of the reported incident to which BPD officers respond. This includes the date, time, location, and type of incident for each crime. Since the dataset includes both major and minor incidents, we decided to only consider crime entries categorized as "Part 1" within the Uniform Crime Reporting Program (UCR). The category includes the more serious offense such as murder, larceny, and rape, etc. After the filtration, a sample of 72,743 entries was recorded between 2016 and 2019.

To account for the weather of Boston, we collected our data from the website "*Wunderground*"[2]. It provides detailed historical weather data for Boston, specifically Boston Logan

Airport. This indicates that throughout the report, we will be assuming the measurement of the entire Boston area is unison along with that of Boston Logan Airport. However, in reality, measurements of a region in Boston does not represent that of the entire Boston Area. The weather dataset includes maximum, minimum, and average weather measurements for temperature, dew point, humidity, wind speed, pressure, and precipitation. Since the measurements are recorded daily, there are a total of 1,461 records for each corresponding day between 2016 and 2019.

Lastly, the definition of "Weekends" could be ambiguous and varies depending on the person. Therefore, throughout the report, we define the term "Weekends" to consist of Friday, Saturday, and Sunday within a week and "Weekdays" consist of Monday, Tuesday, Wednesday, and Thursday within a week.

IV. Data Analysis / Result

A. Crime - Weekends

We would like to investigate whether daily crime counts have any potential association with the day of the week. Specifically, we would like to approach it by testing whether the proportion of a certain amount of crime is different among weekdays and weekends.

To begin our analysis and find the appropriate benchmark, we first take a look at the average count by the day of the week as well as by the category "Weekdays" and "Weekends".

	Weekdays				Weekends		
$\bar{x}$	49.71103				49.89474		
	Mon	Tue	Wed	Thr	Fri	Sat	Sun
$\bar{x}$	49.636	49.756	49.692	49.760	52.258	50.708	46.718

Based on the table above, we notice that both Weekdays and Weekends shares a similar mean daily crime count. Also, further investigation showed us that the median and mode are both approximately 50. Therefore, we decided to set the benchmark as 50 crime cases. The hypothesis can be shown as the following

$$p = \text{Proportion of days with more than 50 crimes}$$
$$H_0 : p_{\text{Weekdays}} = p_{\text{Weekends}}$$
$$H_A : p_{\text{Weekdays}} \neq p_{\text{Weekends}}$$

We will then use this benchmark to construct a 2×2 contingency table to conduct the χ^2 test. To avoid confusion, we categorized the data to as "Below 50" if it is exactly 50.

```
table(Benchmark,DoWCategory)
```

```
            Weekday Weekend
Above 50        428     327
Below 50        406     300
```

We assumed that the cells within the tables are *i.i.d*. Therefore it satisfies the assumptions to consider the table to be valid. We conduct the χ^2 test using R

Assuming $\alpha = 0.05$

```
chisq.test(Benchmark,DoWCat)
```

```
Pearson's X^2 test w/ Yates' cont' correct'
X^2 = 0.069125, df = 1, p-value = 0.7926
```

Based on the output presented above, we failed to reject the null hypothesis since $0.7926 > 0.05$. We failed to conclude that the proportion of days with more than 50 crimes are not the same between weekdays and weekends. In addition, since 50 is approximately close towards the mean of the distribution, we can extend the analogy can conclude that daily crime counts among weekdays and weekends are similar within the significant level of 0.05.

B. Weather - Weekends

In this section, we will be investigating the association between weather conditions and Weekends. To formulate into a more formal question, we would like to find out whether precipitation happens more often during weekdays than weekends?

To study this question, we first realize that since the frequency is the area of interest, it would be most appropriate to examine the data by proportion and conduct a proportion test. Furthermore, we notice that we can categorize each parameter into two groups. Specifically, we categorize the precipitation data into the days with precipitation of 0 in. and the days with precipitation of over 0 in.. We then construct the following 2 $\times$ 2 contingency table.

Observation	Weekends or Weekdays?		
Precipitation?	Weekday	Weekend	Total
Yes	307	225	532
No	527	402	929
Total	834	627	1461

The contingency table requires entries to be *i.i.d* across all four cells. We can clearly see that the categories across each parameter are independent of each other. The distribution within each cell is also approximately the same distribution shape. Therefore it fulfills all assumptions to consider the table to be a valid and reliable contingency table. Hence, we begin the proportion test by setting the following hypothesis:

$$p = \text{Proportion of days with precipitation}$$
$$H_0 : p_{\text{Weekdays}} = p_{\text{Weekends}}$$
$$H_A : p_{\text{Weekdays}} \neq p_{\text{Weekends}}$$

First create an expected contingency table assuming that H_0 is true.

Expected	Weekends or Weekdays?		
Precipitation?	Weekday	Weekend	Total
Yes	303.69	228.31	532
No	530.31	398.69	929
Total	834	627	1461

We can then use the two tables to conduct a χ^2 test
Calculate the χ^2 test statistics with continuity correction

$$\chi^2_{\text{calc}} = \sum_{i=1}^{4} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

$$= \frac{7.8961}{303.69} + \frac{7.8961}{228.31} + \frac{7.8961}{530.31} + \frac{7.8961}{398.69} = 0.0953$$

Assuming $\alpha = 0.05$
Calculate the critical value using R
df: Number of cells $-$ Number of estimated parameters -1
$= 4 - 2 - 1 = 1$

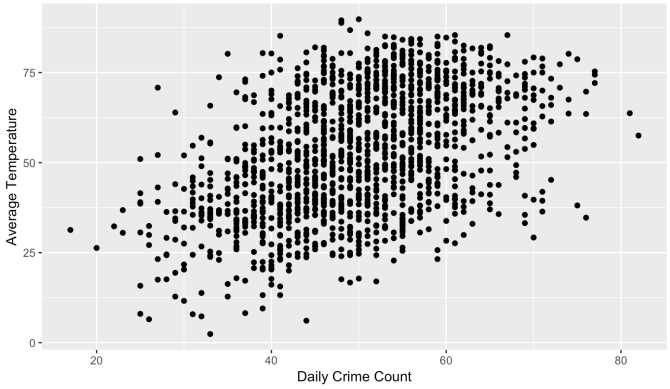
```
pchisq(0.0953, df=1, lower.tail = F)
[1] 0.7575444
```

Since the p-value is less than α, we failed to reject the null hypothesis. We, therefore, conclude that the dataset failed to show the proportion of days with precipitation among the weekdays is not equal to the proportion of days with precipitation among the weekends.

C. Crime - Weather

1) Linear Correlation: We would like to examine if there's any possible relationship between the daily crime count and the daily average temperature in Boston. We first begin by plotting a scatter plot between the two parameters.

Figure 1: Daily Crime Count vs. Average Temperature



It seems like there might be a slight linear trend within the two parameters. Indeed, the point estimate of the correlation is calculated to be 0.46. Therefore we would like to conduct a correlation test to confirm whether a linear correlation exists between them or not.

$$\mathrm{H}_0 : \rho = 0$$
$$\mathrm{H}_\mathrm{A} : \rho \neq 0$$

Before conducting the test, we need to have a look at the nature of each parameter. Referring to the Appendix, we can see that although the daily crime count represents a normal distribution, the daily average temperature shows a bimodal distribution. This indicates that we would not be able to conduct the test using Pearson's correlation. Instead, we will be using Spearman's rank correlation, which is a non-parametric test.

Assuming $\alpha = 0.05$

```
cor.test(daily_avg_temp,daily_crime_count,
                          method="spearman")
```

```
     Spearman's rank correlation rho
data: daily_avg_temp and daily_crime_count
S = 281820000, p-value < 2.2e-16
alt hypothesis: true rho is not equal to 0
```

According to the test result, the p-value of a two-tailed test is 2.2×10^{-16}. Since the p-value is less than α, we reject the null hypothesis at 0.05 level. Therefore, we can conclude that the linear correlation coefficient between the average temperature and the daily crime count is not 0.

Since we have concluded that a linear correlation exists between the two parameters, we can go ahead and find the least-square line that quantifies the relationship.

```
lm(daily_avg_temp~daily_crime_count)
```

```
Coefficients:
    (Intercept) daily_crime_count
       13.5776             0.7882
```

This means that the linear relationship between the two can be best shown as the following equation

$$\text{Daily Avg Temp} = 0.7882 \times \text{Daily Crime Count} + 13.5776$$

2) Association Between Crime and Participation: We now would like to investigate the possible association between daily precipitation amount and the daily crime count. First, we would like to test whether the mean daily crime count among the days without precipitation is different from the mean daily crime count among the days with precipitation.

$$\mu = \text{mean daily crime count}$$
$$\mathrm{H}_0 : \mu_{\text{No Precip}} = \mu_{\text{Precip}}$$
$$\mathrm{H}_\mathrm{A} : \mu_{\text{No Precip}} \neq \mu_{\text{Precip}}$$

Looking at the basic statistics between the two groups within the Appendix, both groups have a similar size of variance. In addition, the distribution for both groups is normally distributed whose graphs are also shown in the Appendix.

With both characteristics in mind, we can now choose to conduct t-test with confidence. Since the hypothesis involves two groups, we should be using the two-sample t-test.

Assuming $\alpha = 0.05$

```
t.test(daily_crime_count,daily_Precip,
                  paired = F, var.equal = T)
```

```
             Two Sample t-test
t = 188.43, df = 2920, p-value < 2.2e-16
95 percent confidence interval:
  49.15756 50.19135
```

Since the p-value is less than α, the null hypothesis is rejected, which implies that the mean values of daily crime

count among no precipitation days and precipitation days are not identical.

After conducting the test above, we would like to dig deeper and try to figure out what degree of precipitation amount most significantly affected the daily crime counts. Defined by the American Meteorological Society[3], the precipitation can be separated into 5 levels:

- No Precipitation: 0 in
- Light Precipitation: 0.001 in ~ 0.099 in
- Moderate Precipitation: 0.1 in ~ 0.399 in
- Significant Precipitation: 0.4 in ~ 0.999 in
- Heavy Precipitation: 1 in$^+$

We would like to figure out the most significant group among these 5 groups, but first, we need to check whether they share the same mean daily crime count.

$$\mathrm{H}_0 : \mu_{\text{None}} = \mu_{\text{Light}} = \mu_{\text{Mod}} = \mu_{\text{Sig}} = \mu_{\text{Heavy}}$$
$$\mathrm{H}_\mathrm{A} : \text{At least one } \mu_i \text{ is different from others}$$

In order to test this, one-way ANOVA would be appropriate. Before conducting the test, we examine the distribution of each group and their variances. As we can see within the Appendix, we conclude that each group of data is approximately normally distributed. Next, we examine the variances.

	None	Light	Moderate	Significant	Heavy
Variance	99.83141	107.6993	91.9309	119.7259	85.63946

Although the variance is slightly different, we can assume that they have a approximately the same variance. We begin our ANOVA test.

Assuming $\alpha = 0.05$

```
aov(daily_crime_count~Category, data=data)
summary()
             Df Sum Sq Mean Sq F value Pr(>F)
Category      4   1264   315.9   3.133 0.0141*
Residual   1456 146849   100.9
```

The p-value is less than 0.05 which we rejected the null hypothesis. We conclude that at least one group of mean daily crime count is different from others.

With that in mind, we want to know which category has the highest mean daily crime count by conducting multiple pairwise comparison tests. As for correction, we decided to use Tukey's HSD for these tests because the hypothesis is set after the data is collected. Bonferroni correction is valid only if hypotheses were set before the data collection.

Based on the Tukey's test results within the Appendix, we can see that only the pair `Zero-Significant` has a p-value of less than 0.05. this indicates that either the mean daily value of Zero or that of Significant precipitation is different from the rest of others. Furthermore, when examining the p-value among the other group, we notice that all pairs with the group "Significant Precipitation" have a lower value than others. Therefore, we can assume that this group should be the one that's different from the others. We now conduct a one-sided t test between "None" and "Significant" to verify which group has a higher daily mean crime count.

$$\mathrm{H}_0 : \mu_{\text{None}} \leq \mu_{\text{Sig}}$$
$$\mathrm{H}_\mathrm{A} : \mu_{\text{None}} > \mu_{\text{Sig}}$$

Assuming $\alpha = 0.05$

```
t.test(Zero_p,Sig_p,alternative = "greater")
```

```
t = 2.9868, df = 111.99, p-value = 0.001732
H_A: true diff in means is greater than 0
95 percent confidence interval:
 1.545923     Inf
```

From the result of the test, we reject the null hypothesis ($0.0017 < 0.05$) and conclude that the mean daily crime count of significant precipitation is lower than that of no precipitation. Combining the conclusion we get from the Tukey's Test, we can now conclude that the mean daily crime count does not affect by the amount of precipitation, with an exception of a lower mean daily crime count when the precipitation is amounting between 0.4 in. and 0.999 in..

V. Discussion

Now that the analyses are completed, we can revisit the conclusion of them and answer our main question. Are crime activities and weather conditions good excuses to cancel weekend outings with friends?

By referencing the test conducted in *Sec. IV Pt. A*, daily crime counts among weekdays and weekends are has no difference within the significant level. In other words, crime count is not a statistically sufficient reason to cancel weekend plans with your friends. A similar conclusion can be made for weather conditions. Referring back to *Sec. IV Pt. B*, the four-year dataset showed the proportion of days with precipitation among the weekdays is approximately equal to the proportion of days with precipitation among the weekends, meaning precipitation does not have a bias towards the weekends. This helped us answer our question - weather conditions are not statistically sufficient enough to be an excuse to cancel weekend outdoor plans with friends.

Although the two excuses we proposed are not strong enough to become valid reasons, there is a silver lining within the analysis we conduct. Through analyzing between daily crime count and daily average temperature, we discovered sufficient evidence showing that a linear correlation coefficient between the average temperature exists and, in fact, shows a moderate positive trend. Therefore, we can claim that the hotter the day is, the stronger the "criminal activity" excuse holds up when using it as the reason to cancel plans with friends.

VI. Potential Improvements / Future Work

Had the datasets not contained unavoidable defects, there could have been many improvements and analysis to look

into. Here we will list out two deficiencies within the data that limited us on our analysis.

While doing basic point estimate analysis using the parameters within the crime incident report, we discovered cases that happened at exactly at 12 a.m. are much higher than any other hours. The probable reason is that midnight (0:00) is the default time when the policemen recorded the crime cases without recording the hours. This causes the data to be accurate up until the day of the incident, instead of the hour of the incident. With the reported hour not accurate, we are unable to conduct more in-depth analyses by the hour without manipulating or removing specific records.

As in the study design section mentioned, our weather data only considered weather status in Boston Logan Airport. A potential improvement can be made here if we were able to access weather statistics throughout each district within Boston. We could pair the weather data according to the location of the incident. This would make the analysis more accurate and, consequently, form a stronger conclusion.

Due to our topic fixating towards the binary categories (Weekday and Weekend), many potential analyses were limited. When focusing on the crime incident report, there exist many questions that can be investigated statistically. To name a few, we could analyze the crime count based on the reported area to find out which part of Boston has a higher rate of criminal activity. We could also investigate by crime types. Has UCR Part 1 crime rate gotten lower over the four years? What type of crime happens the most within each district of Boston? These are all possible topics for future analysis.

REFERENCES

[1] Crime Incident Reports, Analyze Boston <https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system>

[2] Boston, MA Monthly Weather History, Weather Underground <https://www.wunderground.com/history/monthly/us/ma/boston/KBOS>

[3] Rain, Glossary of Meteorology <https://web.archive.org/web/20121205015130/http://amsglossary.allenpress.com/glossary/search?id=rain1>