Statistical Analysis of the 2016 San Francisco Crime Data

MATH1324 - Applied Analytics

Rinaldo Gagiano S3870806

Last updated: 17 October, 2020

RPubs link information

- Rpubs link: https://rpubs.com/Od-Lanir/Applied_Analytics
- This assignment looks better through this link ^

Introduction

- Police forces around the world have been logging data since the beginning of their respective formations.
- It is with this data that people have been able to gather many insights, and now with the recent change to log data electronically, even greater insights can be made with mathematical assurances.
- "In a 2016 instance where a woman named Karina Vetrano got murdered while running in Queens, N.Y., prosecutors used cell phone records to put the suspect — who was eventually found guilty and convicted — near the scene of the crime. Moreover, the phone's web browser contained 137 links related to the crime." - (Matthews How Data Analytics Are Solving Murders)
- "…in a New Hampshire double murder case, a judge ordered Amazon to release Amazon Echo records, saying that the speaker may have picked up parts of the attack." - (Matthews How Data Analytics Are Solving Murders)

Introduction Cont.

- Have you ever wondered if there is a particular day that more crime is reported within the week?
- Is the number of crimes reported every day equal?
- Could we see an increase in crimes reported over weekends or weekdays?
- Using the data set from San Francisco's 2016 crime log, I will venture down this insight rabbit hole and attempt to see if there is a statistically significant difference between the days of the week and crime reported all within San Fran in the year 2016.

San Francisco Golden Gate Bridge ~ https://www.mercurynews.com/wp-content/uploads/2018/10/San-Francisco-2-2.jpg?w=867

Problem Statement

Is there an unequal distribution of crimes reported daily, for each day within the week, in San Francisco?

- To answer this question, a Chi-square Goodness of Fit Test will be conducted.
- I will be assuming that each day as an equal population proportion of 0.143. (0.143 = 1/7 rounded to 3 decimal places)

Crime Tape ~ https://s3.envato.com/files/296610.jpg

Data

- The data set 'Police_Department_Incidents_-_Previous_Year__2016_.csv', I will be using is an open-source data set found on Kaggle - https://www.kaggle.com/roshansharma/sanfranciso-crime-dataset
- This particular data set is a log of every police report made within the San Fran police department in the year 2016, making it a sample data set.

Kaggle Logo ~ https://miro.medium.com/max/640/0*ftOal7fKVCNtJr4N.png

Data Cont.

- Here we can see the import of the data set into a variable named 'crime':

```
crime <- read_csv("SF_Crime.csv")
```

- There are 13 variables as shown:

```
colnames(crime)
```

```
##  [1] "IncidntNum" "Category"   "Descript"   "DayOfWeek"  "Date"
##  [6] "Time"       "PdDistrict" "Resolution" "Address"    "X"
## [11] "Y"          "Location"   "PdId"
```

- To determine a conclusion to the problem statement, I will only require the variable "DayOfWeek" as well as a count of each observation in the variable.
- Seen here is a subset of this variable:

```
crimeDay <- data.frame(table(crime$DayOfWeek))
crimeDay %>% kbl()
```

Var1	Freq
Friday	23371
Monday	20783
Saturday	22172
Sunday	20205
Thursday	21395
Tuesday	21242
Wednesday	21332

- This data frame doesn't look the best… Let's give it a clean.

Data Frame Clean

- As we could see, the variable name in our new data frame wasn't correct so let's change this:

```
colnames(crimeDay)<- c("Day","Observed")
```

- Now let's check the levels of our variable 'Day':

```
levels(crimeDay$Day)
```

```
## [1] "Friday"    "Monday"    "Saturday"  "Sunday"    "Thursday"  "Tuesday"
## [7] "Wednesday"
```

- We see this is not in order, let's fix that:

```
crimeDay$Day <- factor(crimeDay$Day, levels = c("Monday","Tuesday","Wednesday","Thursday",
                                                "Friday", "Saturday", "Sunday"),ordered = TRUE)
levels(crimeDay$Day)
```

```
## [1] "Monday"    "Tuesday"   "Wednesday" "Thursday"  "Friday"    "Saturday"
## [7] "Sunday"
```

- We can use the 'kableExtra' package to produce a cleaner table and have one final Check:

```
crimeDay %>% kbl() %>% kable_styling()
```

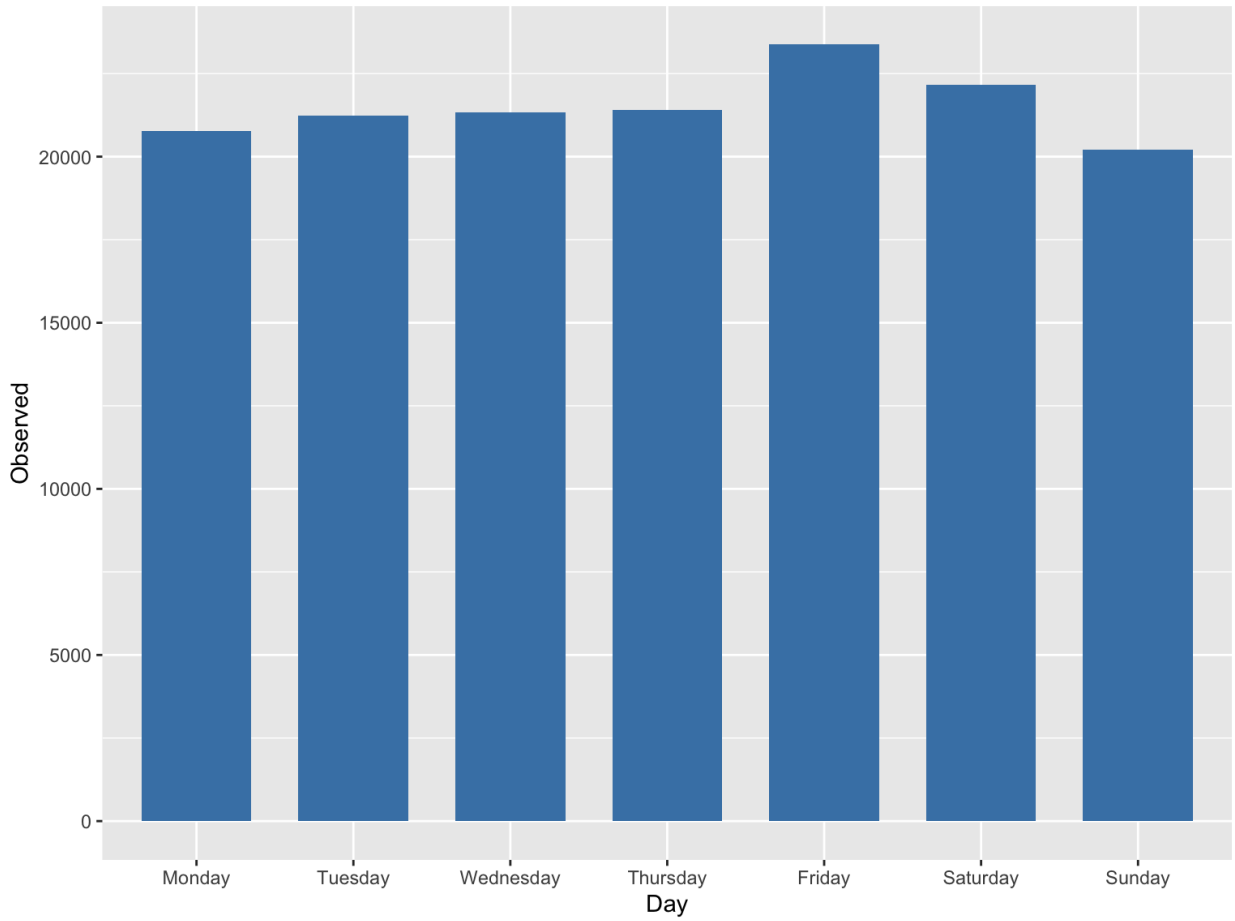
Day	Observed
Friday	23371
Monday	20783
Saturday	22172
Sunday	20205
Thursday	21395
Tuesday	21242
Wednesday	21332

- Here we can see each day and the corresponding count of reported crimes, for that particular day.

Descriptive Statistics and Visualisation

- Let's have a general look at our data frame through a bar plot and see if we can spot any trends:

```
ggplot(data = crimeDay, aes(x=Day,y=Observed)) +
  geom_bar(stat="identity",width=0.7,fill="steelblue")
```



- We can begin to see that the day Friday has the most reported crimes and that each day varies in the number of total crimes reported, but is this enough to be statistically significant?

Observed vs Expected Table

- Before we begin our Chi-square Goodness of Fit Test, let's construct an observed vs expected table including the respective proportions:

```
Obs_Proportion <- round(prop.table(crimeDay$Observed),3) #Proportion table of our observed count
Expected <- rep(150500/7,7) #Expected number of crimes per day if everyday was equal
Exp_Proportion <- rep(1/7,7) #Expected proportion of crimes per day if everyday was equal
Day <- c("Friday","Monday","Saturday","Sunday","Thursday","Tuesday","Wednesday")
Observed <- c(crimeDay$Observed)
OvsE_Table <- data.frame(Day,Observed,Expected,Obs_Proportion,Exp_Proportion)
OvsE_Table <- t(OvsE_Table)
OvsE_Table %>% kbl() %>% kable_material(c("striped", "hover")) %>%
  kable_styling(font_size = 15, full_width = F)
```

Day	Friday	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday
Observed	23371	20783	22172	20205	21395	21242	21332
Expected	21500	21500	21500	21500	21500	21500	21500
Obs_Proportion	0.155	0.138	0.147	0.134	0.142	0.141	0.142
Exp_Proportion	0.1428571	0.1428571	0.1428571	0.1428571	0.1428571	0.1428571	0.1428571

- From this table, we can see that the expected proportions and observed proportions do not equal one another. We can also see that again, Friday has the most crimes reported, while Sunday has the least… Is this sample statistically significant though?

Hypothesis Testing

- Let's define our Null Hypothesis:

$$H_0 : \ The\ population\ distribution\ of\ Crime\ Days\ are\ equal$$

- And our Alternative hypothesis:

$$H_A : \ The\ population\ distribution\ of\ Crime\ Days\ are\ not\ equal$$

- In order to reject or fail-to-reject the Null Hypothesis we need two particular values:
- P-Value: the probability of obtaining results assuming the null hypothesis is true.
- Chi-square statistic, χ2:

$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$

Chi-square Goodness of Fit Test

- A Chi-square goodness of fit test is a hypothesis where we see if the observed values are consistent with the hypothesized distribution.
- As noted previously, the Chi-square statistic is the sum of each observed value minus the expected value squared, divided by the original expected value.
- The 'chisq.test' function allows us to easily calculate both required values:

```
chisq.test(table(crime$DayOfWeek), p = Exp_Proportion)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(crime$DayOfWeek)
## X-squared = 290.66, df = 6, p-value < 2.2e-16
```

- Here we can see that X-squared = 290.66 and p-value < 2.2e-16

But What Does This Mean?

Chi-square statistic

- The X-squared value represents the discrepancies between what values were expected and what values were observed.
- In order to decide on the Null Hypothesis, we need to calculate the Critical Value.
- In our instance, if our X-squared value is greater than our critical value, we must reject the Null Hypothesis, if it is less, then we may fail to reject.
- In R we can calculate the critical value by using the 'qchisq' function, where p = confidence level, df = degrees of freedom, and lower.tail = FALSE

```
qchisq(p = 0.05, df = 6, lower.tail = FALSE)
```

```
## [1] 12.59159
```

- Our critical value is equal to 12.59159, which puts our X-Squared value above this range.
- Since this is the case we can form a decision to reject the Null Hypothesis… but first, let's perform our p-value tests.

P-Value

- Another way to test the Null Hypothesis would be to look at the p-value.
- If our p-value is less than the standard significance level (0.05), we should reject the Null Hypothesis.
- In our case, the p-value (< 2.2e-16) is less than our significance level, and therefore we should reject the Null Hypothesis.
- This is good because both our critical value and p-value test come to the same conclusion.

Reject the Null Hypothesis ~ https://i.ytimg.com/vi/lKtfJYR0cUo/maxresdefault.jpg

Discussion

- A Chi-square goodness of fit test was used to determine whether the distribution of crimes reported was equal among every day of the week.
- The test was statistically significant, X-Squared = 290.66, df = 6, $p < 0.001$.
- This suggests that the distribution of reported crime does not follow an equal distribution between each day within the week.

Discussion Cont.

- In layman's terms, the number of reported crime varies depending on the day.
- This data is a good source of data yet it has its problems as the sample may not accurately reflect on the other years of crime reported in San Fran.
- This is the issue with sample data, as there is always sample error.
- A proposed fix to this issue would be to take multiple years worth of data and perform the same experiment, especially data that is closer to the current date of the replicated investigation.

References

- Data Source: Sharma, Roshan. "Sanfranciso Crime Dataset." Kaggle, 29 May 2019, www.kaggle.com/roshansharma/sanfranciso-crime-dataset
- Matthews, Kayla. "How Data Analytics Are Solving Murders." Medium, Towards Data Science, 23 July 2019, www.towardsdatascience.com/how-data-analytics-are-solving-murders-1cdac5432d6e
- Baglin, James. "Module 8 Categorical Associations." Applied Analytics, 2016, www.astral-theory-157510.appspot.com/secured/MATH1324_Module_08.html#example_write-up
- Science, ODSC - Open Data. "10 Tips to Get Started with Kaggle." Medium, Medium, 14 Jan. 2019, www.medium.com/@ODSC/10-tips-to-get-started-with-kaggle-fc7cb9316d27
- News, Mercury. "San Francisco Golden Gate Bridge." Https://Www.mercurynews.com/Wp-Content/Uploads/2018/10/San-Francisco-2-2.Jpg, www.mercurynews.com/wp-content/uploads/2018/10/San-Francisco-2-2.jpg
- Zhu, Hao. Create Awesome HTML Table with Knitr::Kable and KableExtra, 6 Oct. 2020, www.haozhu233.github.io/kableExtra/awesome_table_in_html.html
- Market, Envato. Police Crime Scene Tape HD. www.videohive.net/item/police-crime-scene-tape-hd/93420
- Jawlik, Andrew. "Reject the Null Hypothesis." Statistics from A to Z – Confusing Concepts Clarified, 15 Nov. 2016, www.i.ytimg.com/vi/IKtfJYR0cUo/maxresdefault.jpg