A STATISTICAL ANALYSIS OF POLICE BRUTALITY

STAT476 Project I

Abstract

An analysis of police-caused deaths in the U.S. in 2015 was performed using NBD probability models. Questions considered were 1. whether different months are more prone to violence than others, and 2. whether claims can be made regarding the death-causing propensity of the police of different states. The resulting models are surprisingly powerful and provide a strong framework for further, similar analyses.

Word count: 2498, excl. cover

STAT476 Spring 2016

1. Introduction

Police brutality, an ugly issue brought to the limelight by several high-profile cases in the past few years, is an example of a social issue that people were unable to quantify and examine meaningfully until recently. New centralized efforts for monitoring police shootings and other violent incidents have produced a wealth of data, with which we can extract information about crime and violent police incidents.

In this study, data on deaths caused by law enforcement in the U.S. in 2015-2016 is analyzed. Using probabilistic models, we will formulate conclusions about particularly violent months of the year, and differences in violent tendencies among police forces in different states. Hopefully, through this exercise, actionable insights can be produced to tackle one of the pricklier issues in the public consciousness.

2. Dataset

2.1 Data Collection

The data is provided by "The Counted," a project by the Guardian that counts the number of people killed by law enforcement in the U.S. since 2015. Causes of reported deaths include shootings, Tasers, police vehicle collisions, and deaths in custody; self-inflicted deaths and suicides are excluded. The database is built by Guardian reporters as well as verified crowdsourced accounts. It can be found at the following link: http://www.theguardian.com/us-news/ng-interactive/2015/jun/01/the-counted-police-killings-us-database

The project was motivated by the stunning fact that the federal government keeps no record of those killed by police. The extant system, run by the FBI, calls on law enforcement agencies to *voluntarily* report the number of "justifiable homicides." The problems with this are blatant. Agencies are incentivized not only to abstain from reporting, but to undercount the figures if they do report.

Notably, the data was collected both from journalists, who we can generally trust as a source, as well as the public, who are much less trustworthy. However, there is a process involved in verifying user reports, so it is probably safe to assume that the data is correct. What is likely the greater danger is underreporting of violent incidents, rather than overreporting caused by falsified or incorrect entries in the data. While there is nothing we can do to check or correct for this, it is good to keep in mind.

2.2 Data Description

The dataset is split into incidents in 2015, and incidents in 2016. The entries include the following variables:

1. Name
2. Demographic data – age, gender, and ethnicity
3. Date – month and day
4. Location – street address, city, and state
5. Misc. details – cause of death, the police dept. involved, and whether the deceased was armed

For the purposes of our study, only the month and state are considered.

2.3 Data Processing

The data was broken down into the number of incidents in each month, for every state. Note that all 50 states as well as Washington D.C. are included. A snapshot of the processed data is shown, for the first 5 states in alphabetical order (Figure 2.1).

	AL	AK	AZ	AR	CA	Total
January	0	1	5	2	11	91
February	3	1	6	0	11	83
March	4	0	7	0	24	114
April	0	0	4	1	20	102
May	3	0	4	1	8	86
June	1	0	1	0	12	79
July	1	0	3	1	24	123
August	2	0	5	0	20	102
September	2	2	2	0	20	95
October	3	1	1	0	22	90
November	0	0	2	0	16	81
December	0	0	4	0	22	94

Figure 2.1 – Processed data. Note that the total number of incidents is 1140.

From here, binned data on the number of incidents in each month was built. For instance, the January data is shown (Figure 2.2). The interpretation of the first row is that in January, there were 16 states with 0 police killings. This step allows us to proceed with the analyses.

x	n_x
0	16
1	15
2	12
3	2
4	3
5	1
6	0
7	0
8	0
9	0
10+	2

Figure 2.2 – Re-binned data for analysis.

3. Data Analysis

3.1 Summary Statistics

Average, variance, and select percentiles are displayed for each month (Figure 3.1). Note the high positive skew. The maximums are extremely large in comparison to the rest, indicating some states have substantially more incidents than others. They drag the averages and variance up by a huge amount. Central tendency is thus more accurately reflected by the median.

Note the differences in averages among the months. We may expect, intuitively, that summer months have more incidents than others. There are various explanations, such as the academic summer break, the weather making it more likely for people to go outside as compared to winter months, mental issues caused by the heat, etc. This is generally shown here, although May and June are noticeably low.

	Avg #	Variance	25th Perc	Median	75th Perc	Max
January	1.8	8.6	0	1	2	18
February	1.6	5.7	0	1	2	11
March	2.2	12.8	0	1	3	24
April	2.0	11.4	0	1	3	20
May	1.7	3.6	0	1	3	8
June	1.5	6.2	0	1	2	12
July	2.4	14.0	0	2	3	24
August	2.0	11.0	0	1	2.5	20
September	1.9	9.2	0	1	2	20
October	1.8	11.6	0	1	2	22
November	1.6	7.3	0	1	2	16
December	1.8	10.2	0	1	2	22

Figure 3.1 - Summary statistics for the data. Note the minimum is 0 for every month.

3.2 Questions, Motivations, and Approaches

Our main two questions are:

1. Are certain months more prone to police incidents than others?
2. What can we say about the propensities for violence in different states?

We can answer the first question by running NBD models for each month and then comparing the gamma distributions and parameters (r, α). In doing so, we are implicitly assuming that the number of violent incidents in each state is distributed according to a Poisson distribution. This assumption is not intuitive. It is easy to assume that individuals deciding on the number of products to buy can be modeled by a Poisson process, but the logical step is considerably larger in our case.

The distribution of the data for each state, however, does lend itself nicely to a Poisson distribution. And conceptually, individuals' purchasing behaviors are also quite complicated, yet can still be treated as Poisson distributed. It is not an outrageous claim that the number of police-caused deaths in each state can be similarly modeled.

If we can accept that the complex factors involved in determining the number of police-caused deaths can all be wrapped into a single λ parameter of the Poisson distribution, the rest falls into place. Using Bayes' Theorem on our NBD models, we can produce estimates of λ for a given state in a given month. Comparing λ's allows us to make claims about the violent tendencies of police forces in each state.

Finally, we discuss some hypotheses regarding the values of r for each month. There is clearly significant heterogeneity amongst the states, since many have 0 incidents per month, while some have many incidents. Thus we expect r values that are < 1. This implies that months that are more violent would have higher r values, since more states having incidents would decrease heterogeneity. They wouldn't be much higher, however, due to the effects of the extreme outliers – probably ≥ 1.

3.3 Models

Typical NBD models were fitted to each month. The spike-at-zero model was considered but rejected. While the idea that some states' police forces are "hard-core-never-killers" is appealing in several senses, when running the models, some months fit the spike model while most did not. The narrative that some months, a state's police force is a "hard-core-never killer" while in other months, it isn't, seems ridiculous. Therefore, while some months' models had improved goodness of fit under the spike model, all spike models were rejected.

The histograms of the model fits are included but shown in the appendix, in the interest of conserving space. Relevant information about the fits is included in Figure 3.2.

	r	α	LL	chisq	GoF p-val
January	1.18	0.71	-86.51	15.15	0.03
February	0.69	0.40	-86.00	9.19	0.24
March	1.33	0.67	-95.17	3.76	0.81
April	0.61	0.32	-89.98	4.01	0.78
May	1.27	0.75	-90.26	3.28	0.86
June	0.70	0.45	-83.20	10.16	0.18
July	0.96	0.43	-97.83	16.89	0.02
August	1.02	0.56	-89.77	6.61	0.47
September	1.28	0.75	-88.78	5.72	0.57
October	0.79	0.50	-85.41	12.45	0.09
November	0.53	0.35	-82.66	6.51	0.48
December	1.30	0.80	-87.37	5.20	0.64

Figure 3.2 - Model parameters and test statistics.

r, α are the model parameters. LL is the log likelihood that is maximized to fit the models. Chisq is the chi-square statistic for goodness-of-fit, and GoF p-val is the corresponding p-value.

Considering the goodness-of-fit tests, we gain confidence that the NBD model assumption – the Poisson nature of police-caused deaths – is reasonable. All months have high p-values > .20, indicating good fit, except January, July, and October (June, at .18, is still considered good fit). Referring to the appendix, in January, there were more occurrences of 1's and 2's than expected. For July, there were a surprisingly large number of 3's. In October, there were more 1's and 3's than expected.

July, as the apex of the summer season, corresponds to the phenomenon that there are more crimes in summer months. The other two months are more difficult to explain. January could be due to the Eric Garner case, a widely publicized incident that was held up as an example of police brutality. In Dec. 2014, the courts ruled not to indict the officer involved, which could have led to increased violent incidents in January. Another explanation is simply chance.

4. Results

4.1 Monthly Differences

We can now tackle question 1 in detail, by examining the model parameters. Specifically, we are concerned with shape parameter r and how it varies.

A visualization of the gamma distributions implied by the fitted r, α parameters is shown in Figure 4.1. We note that the r values match up to our earlier hypothesis, being clustered around 1. More interestingly, we can divide the months into two groups – those with $r < 1$ and those with $r \geq 1$. January, March, May, August, September, and December appear to belong to the latter. These are months where the distribution of λ is less positively skewed and more homogeneous. In other words, these are months with more deaths.

Why these months? Interpreting the results is challenging. There is nothing obvious that relates them. How much of this is due to abnormal events occurring in those months, such as the Eric Garner case, and how much of this can be attributed to real differences in crime rate between months? Without additional data for more years, it is difficult to tell.

What we do have is the complete data for incidents in January 2016. This can serve as a small test set for the fit. The relevant summary statistics and model fit for January 2016 are displayed (Figure 4.2, 4.3).

The parameters match up astonishingly well, which supports our implicit hypothesis that every month has some innate characteristics regarding occurrences of deaths caused by police. Furthermore, the goodness of fit is good for 2016 but not 2015, which lends some credence to the argument that January 2015 was affected by some external event. The lower variance and maximum in 2016 further assists this theory.

Therefore, while we cannot offer a compelling argument for why some months are more problematic, we can state with some confidence that the models possess some predictive power and that those months are indeed more prone to police-caused deaths.

4.2 State Differences

Now we arrive at perhaps the more interesting question of the two. Can we make statements about the violence propensities of different states? Arriving at λ values for each state is simple thru Bayes' Theorem. However, there is an elephant in the room: different states are not equal. California has ~53 times the population of Alaska. So while there were 11 deaths in January 2015 in California and 0 in Alaska, Alaska could still have a higher propensity, which is not reflected in the Bayes λ calculation.

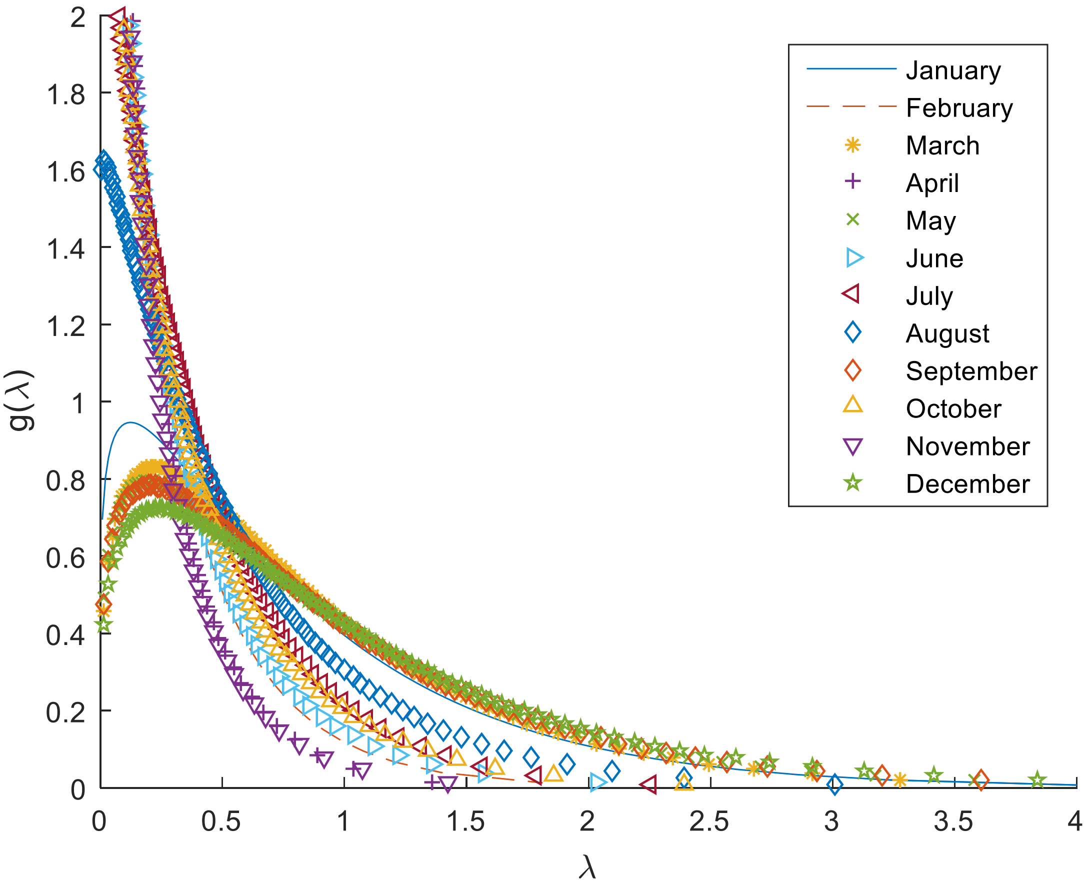


Figure 4.1 - Gamma distributions implied.

	Avg #	Variance	25th Perc	Median	75th Perc	Max
Jan-16	1.627451	4.508266	0	1	2	12

Figure 4.2 - Summary statistics for January 2016.

	r	α	LL	chisq	GoF p-val
Jan-15	1.18	0.71	-86.51	15.15	0.03
Jan-16	1.25	0.77	-86.94	5.08	0.65

Figure 4.3 - Model parameters and test statistics for January 2016.

A simple normalizing factor is employed, by dividing each λ by state population. Results for January 2015 and 2016 are shown, for select states (Figure 4.4, 4.5). California, Texas, and New York, as the most populated states, are chosen to test our normalization. The most dangerous states in the US – Alaska, New Mexico, and Nevada, in order, are also chosen. Lastly, Pennsylvania is included.

	# of deaths	λ	Population (millions)	Normalized λ
PA	2	1.86	12.8	0.15

CA	11	7.14	38.8	0.18
TX	18	11.24	27.0	0.42
NY	1	1.28	19.7	0.06
AL	0	0.69	0.73	0.94
NM	1	1.28	2.1	0.61
NV	1	1.28	2.8	0.46

Figure 4.4 - λ estimates for select states in January 2015.

	# of deaths	λ	Population (millions)	Normalized λ
PA	3	2.40	12.8	0.19
CA	12	7.48	38.8	0.19
TX	7	4.65	27.0	0.17
NY	0	0.70	19.7	0.036
AL	4	2.96	0.73	4.06
NM	2	1.83	2.1	0.87
NV	2	1.83	2.8	0.65

Figure 4.5 - λ estimates for select states in January 2016.

The normalized λ's work incredibly well. Not only do they reflect that AL, NM, and NV are the most dangerous states, but they even capture their relative ranking. The normalization also manages to adjust down the large population states adequately.

One quick observation we can make is that the police in New York are surprisingly efficient and safe.

Extending this analysis to all other states for all the months is simple, and would produce more interesting results.

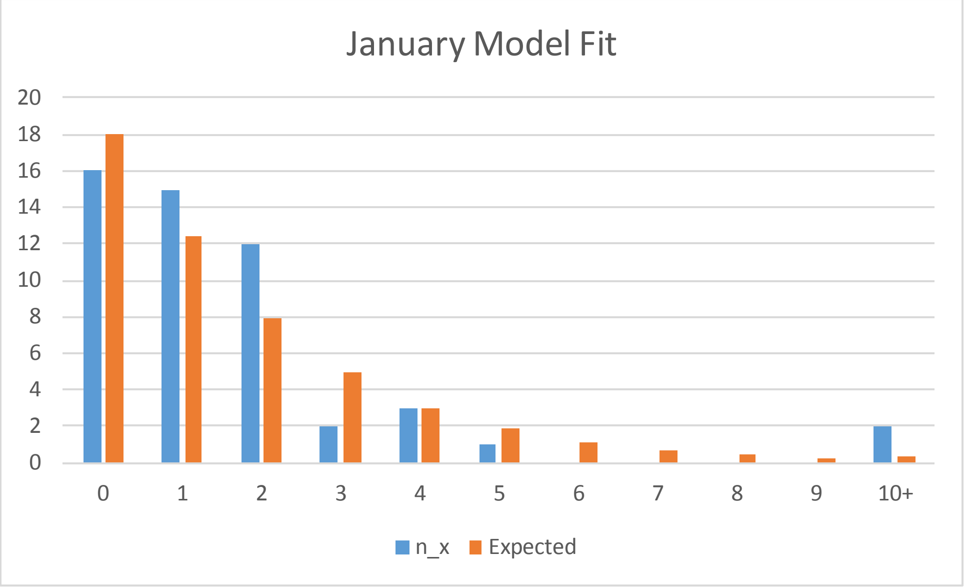
5. Conclusion

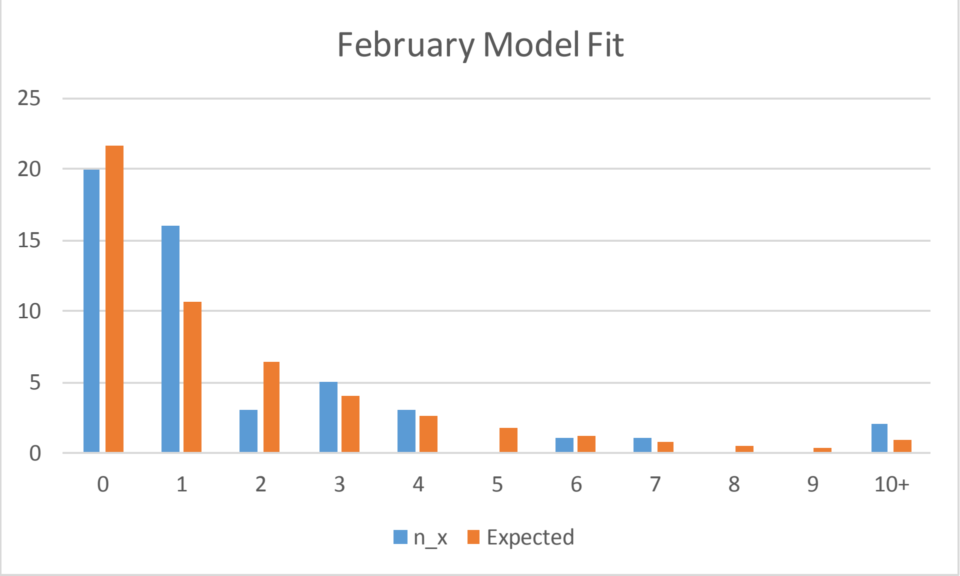
Ultimately, this paper showed that NBD models for the number of police-caused deaths in a given month are incredibly powerful, despite their simplicity. With these models, we were able to highlight more violent months of the year using model parameters, as well as identify which states had greater rate of violent incidents. These are real, actionable insights – we have pinpointed months police should be more careful in, as well as which police forces need reform.

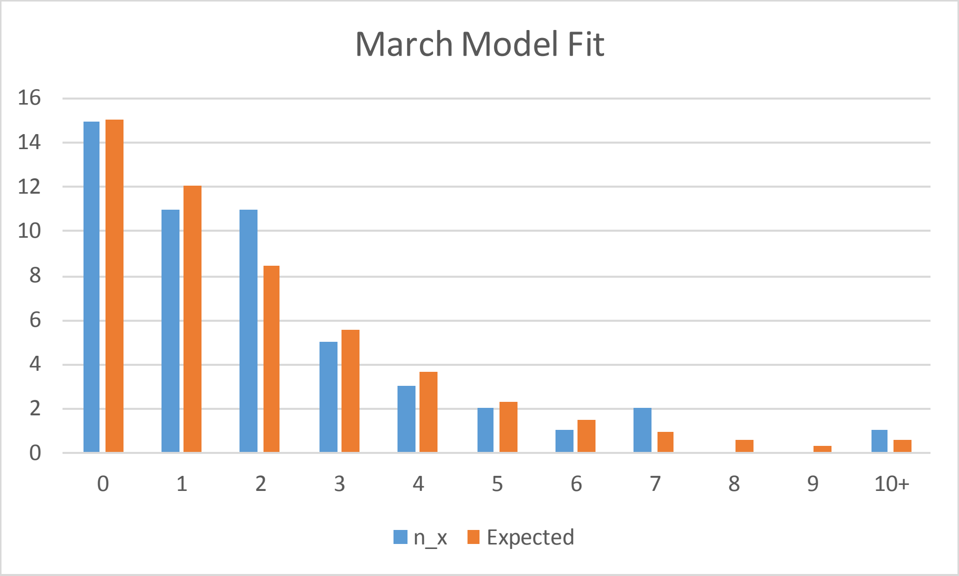
Our limited data is the main caveat. Though the models have proven to be quite useful, further confirmation of their accuracy is the logical next step. With more time, and thus more data rolling in for 2016, we can further test the accuracy of these models.

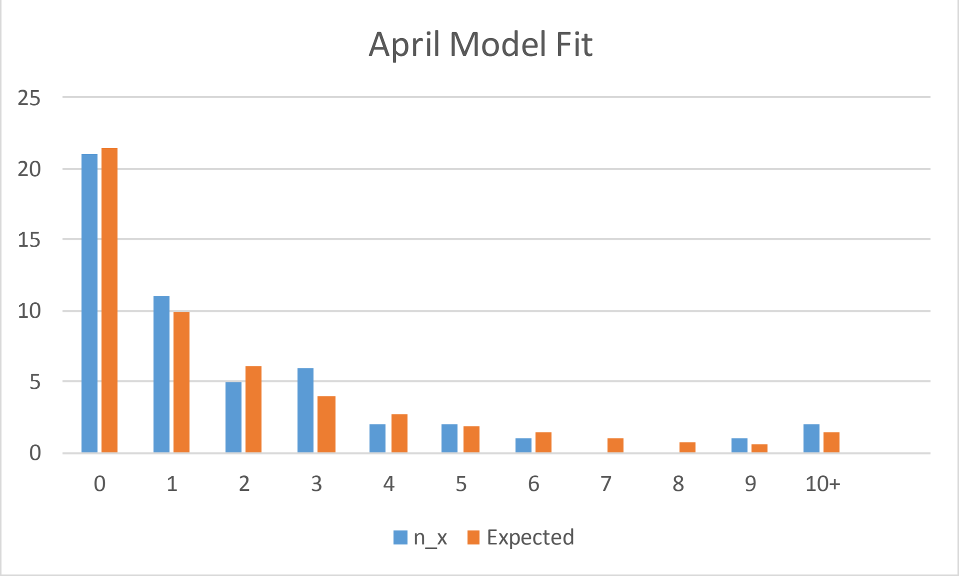
Further analyses could include the covariate data excluded from this study, such as which states had more deaths of a certain race/gender. Such analyses would clearly have their uses. There are many, many steps we could take next with the framework established here, each likely producing compelling insights.

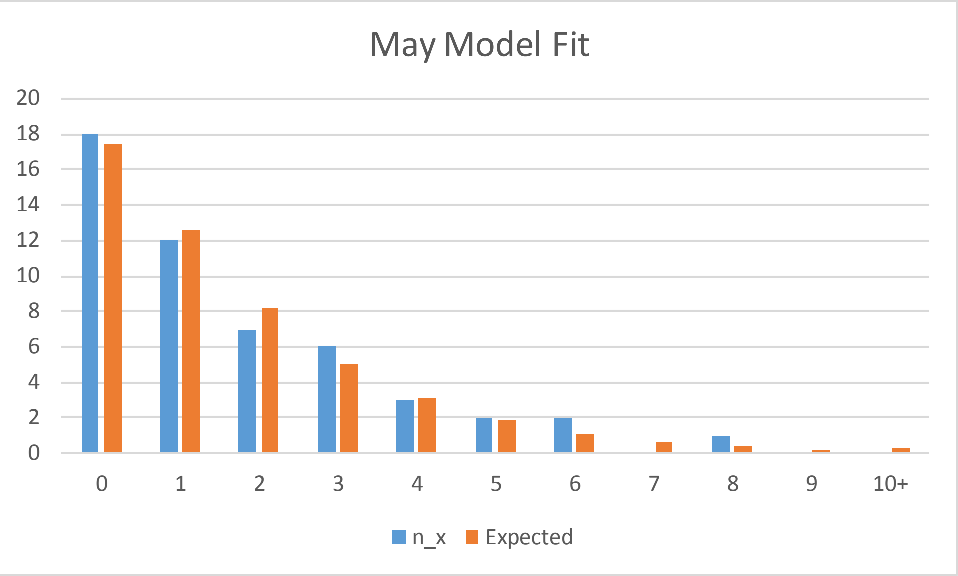
6. Appendix

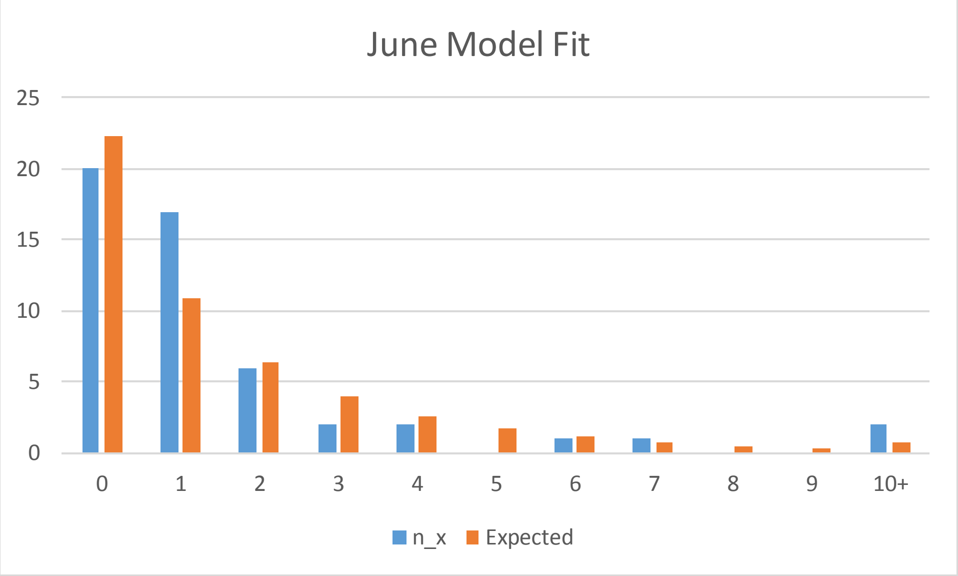

January Model Fit
20
18
16
14
12
10
8
6
4
2
0
0
1
2
3
4
5
6
7
8
9
10+
n_x
Expected

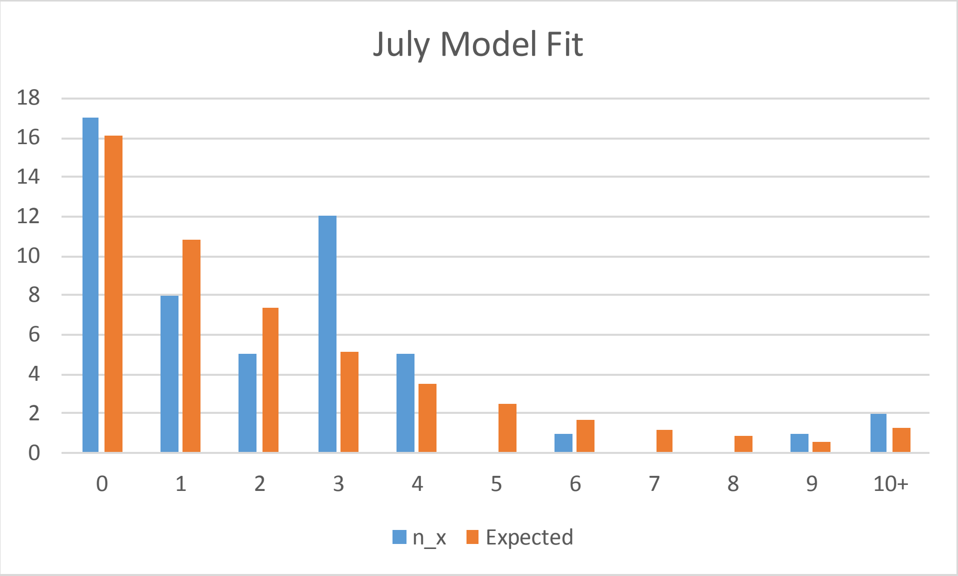

February Model Fit
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10+
n_x
Expected

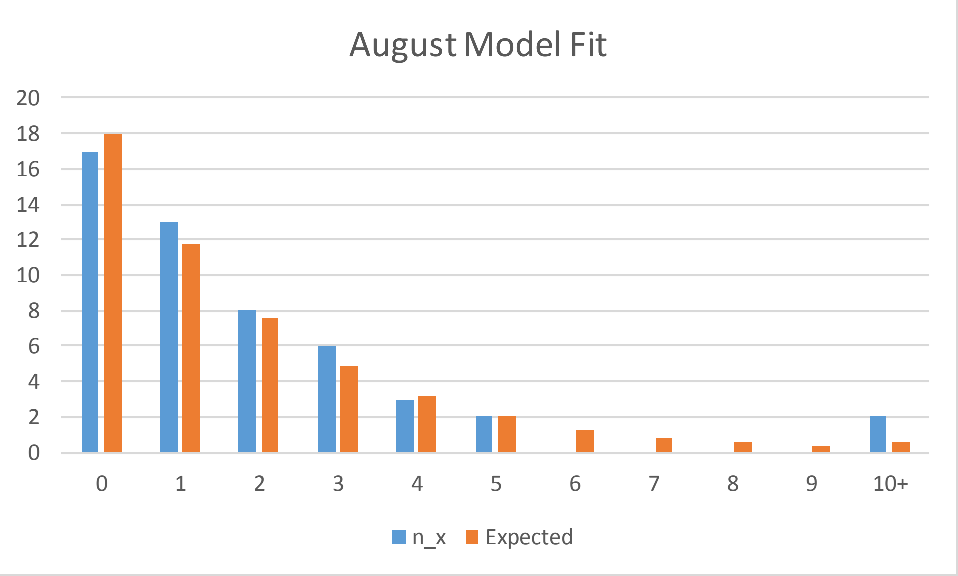

March Model Fit
16
14
12
10
8
6
4
2
0
0
1
2
3
4
5
6
7
8
9
10+
n_x
Expected

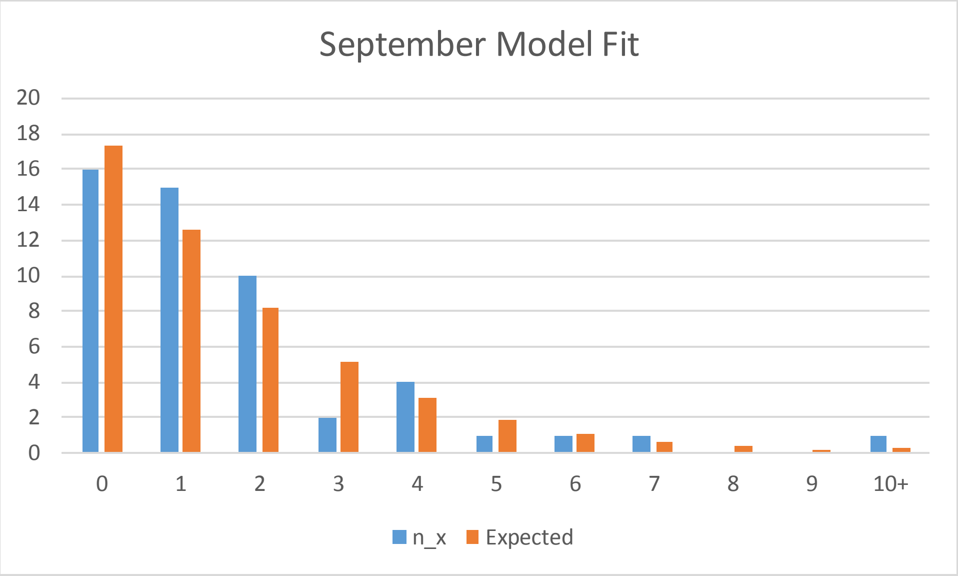

April Model Fit
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10+
n_x
Expected

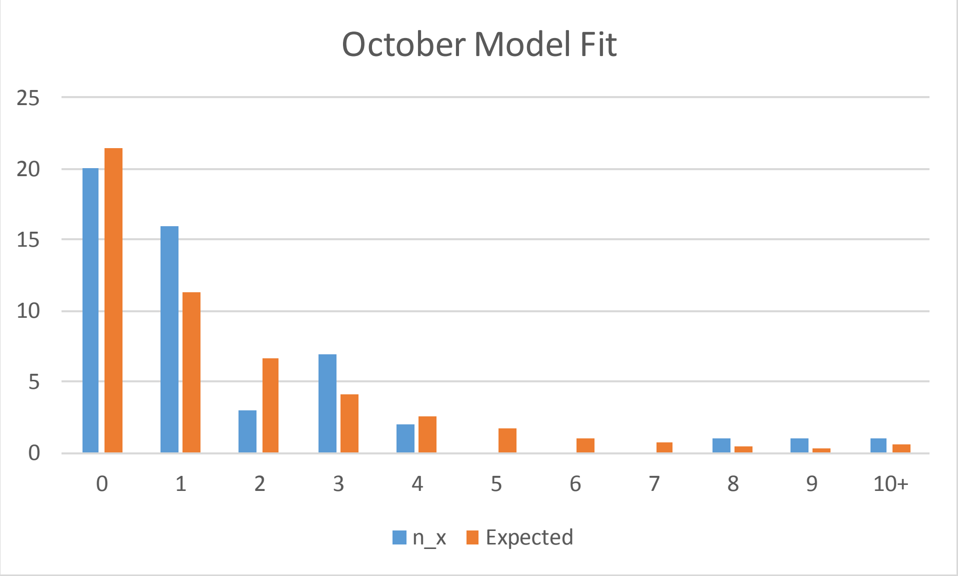

May Model Fit
20
18
16
14
12
10
8
6
4
2
0
0
1
2
3
4
5
6
7
8
9
10+
n_x
Expected

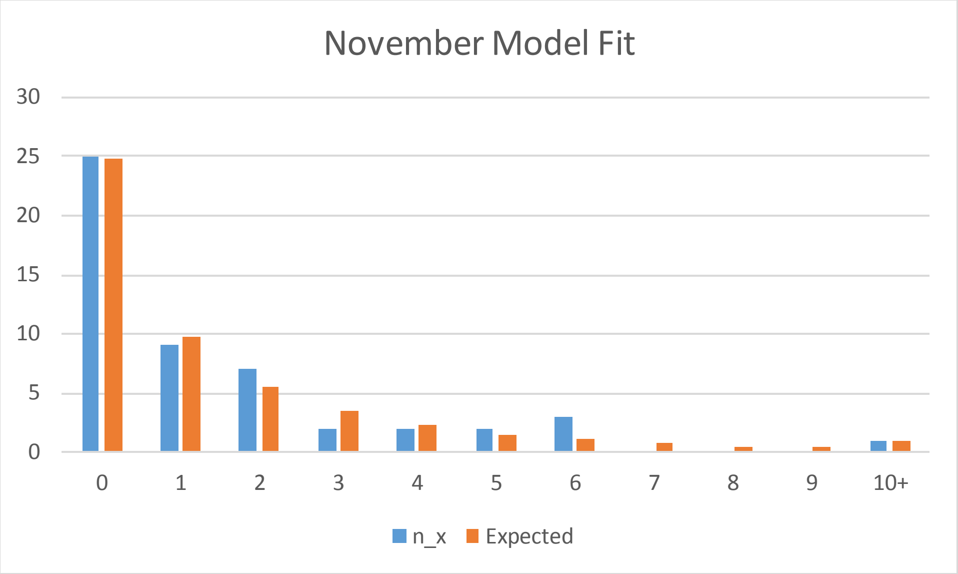

June Model Fit
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10+
n_x
Expected

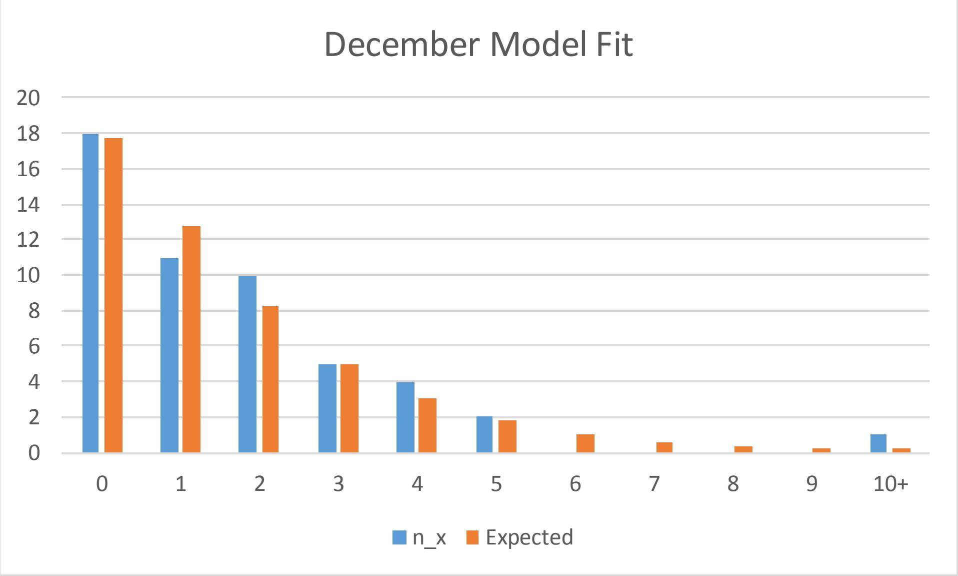

July Model Fit
18
16
14
12
10
8
6
4
2
0
0
1
2
3
4
5
6
7
8
9
10+
🟦 n_x
🟧 Expected


August Model Fit
20
18
16
14
12
10
8
6
4
2
0
0
1
2
3
4
5
6
7
8
9
10+
🟦 n_x
🟧 Expected


September Model Fit


October Model Fit


November Model Fit
30
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10+
n_x
Expected


December Model Fit
20
18
16
14
12
10
8
6
4
2
0
0
1
2
3
4
5
6
7
8
9
10+
n_x
Expected