Statistical Analysis of the Chicago Crime Database

Sophia Milask

Stockton University

CSCI-3327-001 – Probability and Applied Statistics

Byron Hoy

May 5, 2025

Abstract

This report provides statistical analysis of crime data from the Chicago Data Portal, focusing on probability and discrete and continuous probability distributions. The data considered for this report uses all crimes reported in the Chicago Data Portal from 2001 to present. Patterns in crime occurrence and frequency, arrest rates, and crime location and timing are investigated through set theory, conditional probabilities, and several probability distributions. Findings suggest discrepancies with arrest rates, as well as a possible call for reallocation and implementation of resources.

Introduction

Crime in cities like Chicago is influenced by a multitude of social and economic factors. Viewing crime trends through a statistical lens offers key insight into the effectiveness of law enforcement and resources. Principles of probability and statistics applied to the Chicago database can be used to analyze several trends and factors of crime.

Basic Probabilities in Crime Patterns

Many crimes committed in Chicago are volatile and contain complex events. We can apply fundamental probability principles to better interpret this data. The likelihood of certain events occurring in each crime heavily depends on other key factors and events.

Defining Crime Categories with Sets

To better understand the relationship between categories of crimes, each event can be organized into a set.

- Let A be the set of all crimes marked primarily as “theft”.
- Let B be the set of all crimes which are primarily drug-related offenses.
- Let C be the set of all crimes which resulted in an arrest.

By applying set notation, important questions can be answered such as

- $P(A \cap C)$: What is the probability that a theft resulted in an arrest
- $P(A \cup B)$: What is the probability any given crime is either theft or drug related?
- $P(C|A)$: What is the probability of a crime resulting in an arrest given it was theft?

Calculating Probabilities

Out of 1,048,567 crimes reported, 319,556 were reported as theft. This means that

$$P(A) = .\mathbf{3048}.$$

Further, 24,639 of those total crimes were reported to be drug-related offences. So,

$$P(B) = .\mathbf{0235}.$$

Finally, 135,452 of those crimes resulted in an arrest, which means

$$P(C) = .\mathbf{1292}.$$

Using these results, we can use set theory and conditional probability to answer our questions.

First, we want to find the probability that a crime marked as theft resulted in an arrest.

$$P(A \cap C) = \frac{12{,}278}{1{,}048{,}567} =. \mathbf{0117}.$$

Next, we need to find the probability that all the reported crimes are either theft, drug-related, or both. A crime cannot be marked primarily as both theft and drug-related, so the overall does not need to be subtracted out. Therefore,

$$P(A \cup B) = P(A) + P(B) = \ .3048 + \ .0235 = .\mathbf{3283}.$$

Finally, we need to find the probability that a crime resulted in an arrest given the crime was marked primarily as theft. Since the overlap of theft crimes and arrests were previously found, that information can be used to show

$$P(C|A) = \frac{P(A \cap C)}{P(A)} = \frac{.0117}{.3048} =. \mathbf{0384}.$$

Real-World Implications

The probability estimates demonstrate important patterns of crime in Chicago. Approximately 30% of reported crimes were primarily theft, and only around 2.3% were marked as drug-related. This shows that the residents of Chicago tend to commit property crimes more frequently than other crimes. However, this may also reflect the behavior of Chicago police officers. The data suggests that officers may be inclined to prioritize property crimes over crimes involving narcotics.

Surprisingly, the data analysis shows that only 3.84% of theft results in an arrest. This demonstrates that theft is so prevalent in Chicago that most cases go unresolved. This could be the result of owners deciding not to press charges, limited resources, petty and nonviolent nature of most theft cases, or a prioritization of violent crimes over misdemeanors.

These findings suggest that Chicago needs to take further precautions to both reduce and prevent the influx of theft occurring in the city.

Discrete Probability Distributions

Crimes reported in Chicago are discrete events. Each report can be classified in several series of events. Many characteristics of a crime are countable and follow discrete probability distributions. These distributions are important because they can reveal patterns and predictions of crimes depending on time, category, location, and more.

Arrest Probability Distribution by Crime Type

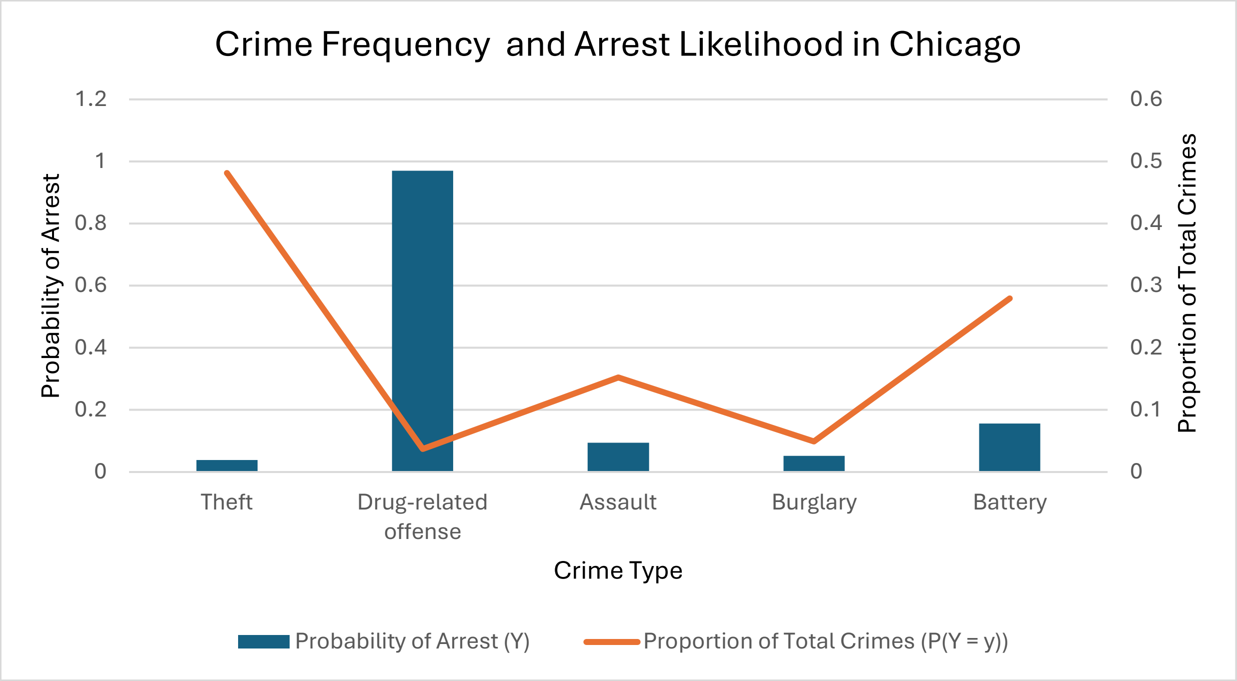
Let discrete random variable Y represent the number of arrests associated with a given mark of primary crime, and let there be five main crimes: Theft, Drug Offense, Assault, Burglary, and Battery. Let each value of Y correspond to a different crime type and Y map to the probability of that crime resulting in an arrest.

According to the Chicago Data Portal, since 2001 there has been

- 319,556 thefts, and 12,278 resulting in an arrest
- 24,639 drug-related offenses, 23,901 resulting in an arrest
- 101,060 assault cases, 9,551 resulting in an arrest
- 32,751 burglaries, 1,697 resulting in an arrest
- 185,409 battery cases, 28,979 resulting in an arrest.

We can use this data to create a table and graph for our discrete random variable Y.

Crime Type	**Total Crimes**	**Number of Arrests**	**Y (Probability of Arrest)**	**$P(Y = y)$ (Proportion of Crimes)**
Theft	319,556	12,278	.0384	.4815
Drug-related offense	24,639	23,901	.9700	.0371
Assault	101,060	9,551	.0945	.1523
Burglary	32,751	1,697	.0518	.0494
Battery	185,409	28,979	.1563	.2797



Expected Value of Arrest Likelihood

Let p_i be the proportion of crimes in crime type n, and a_i be the arrest probability for that crime type. Then, the expected value of Y is

$$E[Y] = \sum_{i=1}^{n} p_1 a_i = .4815(.0384) + .0371(.9700) + .1523(.0945) + .0494(.0518) + .2797(.1563) = \mathbf{.1152}$$

This signifies that across all five types of primary crimes, the average probability of an arrest occurring is 11.52%.

Real-World Implications

Across the five main primary categories for arrests chosen above, the average arrest weight is only about 1 in 9. There are many implications we can draw from this data analysis.

First, we can conclude that arrest rates have large discrepancies across the different crime types. Drug-related offences have resulted in an arrest rate of 97%. This may be because police consider narcotics to be a public safety concern and often follow up on these types of reported crimes. On the other hand, theft accounts for significantly more reported crimes than drug-related offenses yet only has an arrest rate of 3.84%. This is further proof that fewer resources are dedicated to these kinds of crime and Chicago needs adapt to better handle theft.

The discrepancies between arrest rates for theft and burglary versus. drug-related offenses demonstrate that the more complex the crime, the less likely for the case to be solved. Drug-related offenses often are easy to arrest, making their arrest rates higher. Since theft and burglary cases do not provide much evidence and are more difficult to solve, not enough time is allocated to these crimes and the arrest rates are much lower.

Exploring Crime Patterns Though Probability Distributions

Statistical distributions can offer insight into more complex crime patterns that occur in Chicago. These distributions help analyze the frequency of events, not just the likelihood. This information can help predict the type of crimes Chicago will experience in the future and provide insight into potential policy changes to combat these crimes.

Analyzing a Fixed Number of Domestic Crimes

All crimes reported in the Chicago Data Portal are marked as either domestic or not domestic. On Ontario Street, there are 1,645 crimes from 2001 reported in the database, and 103 of them are marked as domestic. Let *Y* be the number of domestic crimes occurring on Ontario Street. If we need to know the probability of at least one of the next 20 crimes occurring on Ontario Street being domestic, we can model this scenario with a binomial distribution where

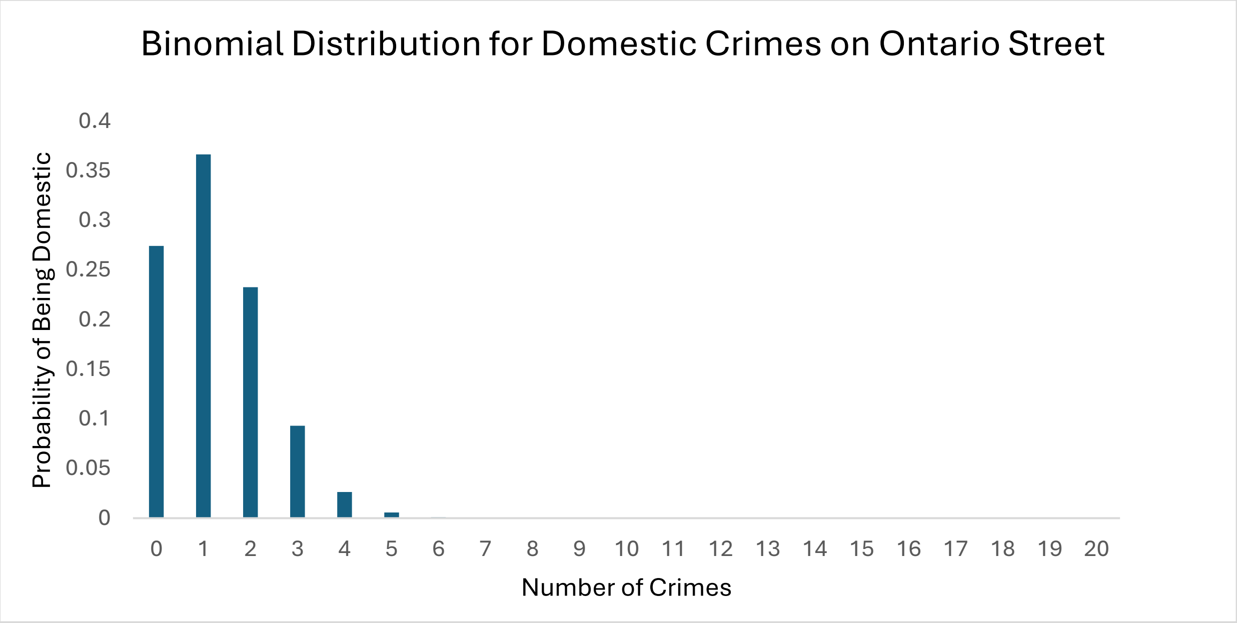
- y = 1, 2, …, 20
- n = 20
- p = .0626
- q = .9374,

then

$$p(y \geq 1) = 1 - \binom{20}{0}.0626^{0}.9374^{20} =\mathbf{.7256.}$$

This means that there is about a 73% chance that at least one of the next 20 crimes occurring on Ontario Street will be domestic.

We can also construct a binomial distribution graph for this scenario to model the likelihood of crimes on Ontario Street being domestic.



We can also find the expected value of the number of domestic crimes in the next 20 by calculating

$$E[Y] = 20(.0626) = \mathbf{1.252}$$

domestic crimes.

Response Timing and Efficiency for Arrests

A “beat” is a geographic area assigned to an officer or team to patrol. Officers are assigned to answer calls for service and perform necessary arrests. Let’s say we are interested in the number of calls for service it takes for different beats to get an arrest. We can model this with a geometric distribution. Let *Y* be the number of calls for service before an arrest is made.

By viewing the number of crimes in a certain beat and the number of those crimes which resulted in an arrest, we can construct a table as such:

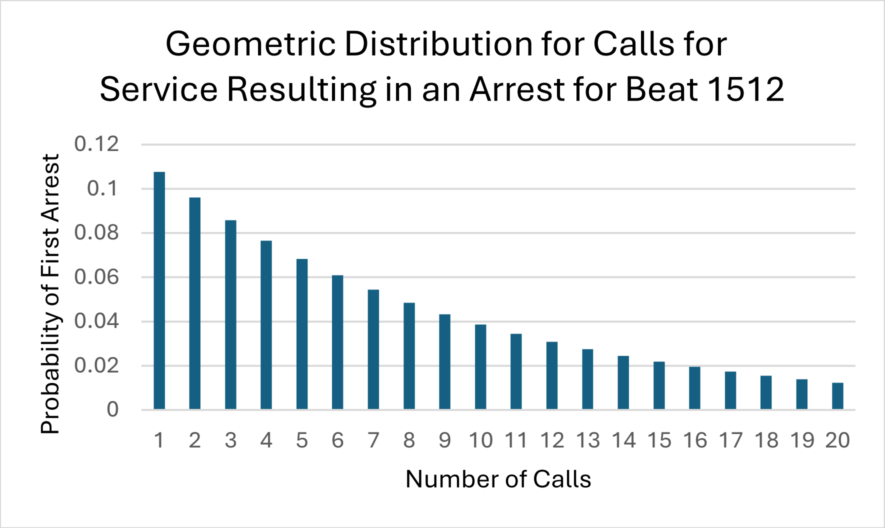
Beat	Number of Crimes	Number of Arrests	Proportion of Crimes Resulting in Arrest
1512	3,092	333	.1077
1834	12,562	2,139	.1703
1215	3,866	183	.0473

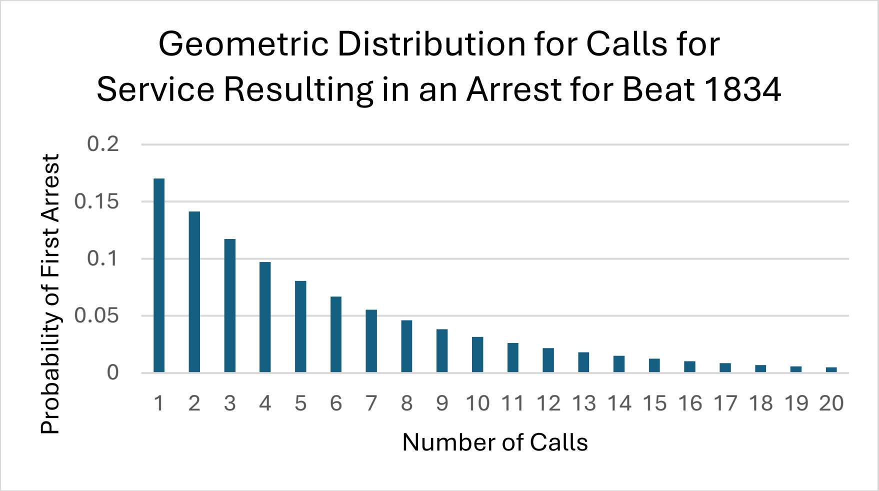
This data allows us to use a geometric distribution to find the probability of the first arrest carried out on different numbers of calls for service.

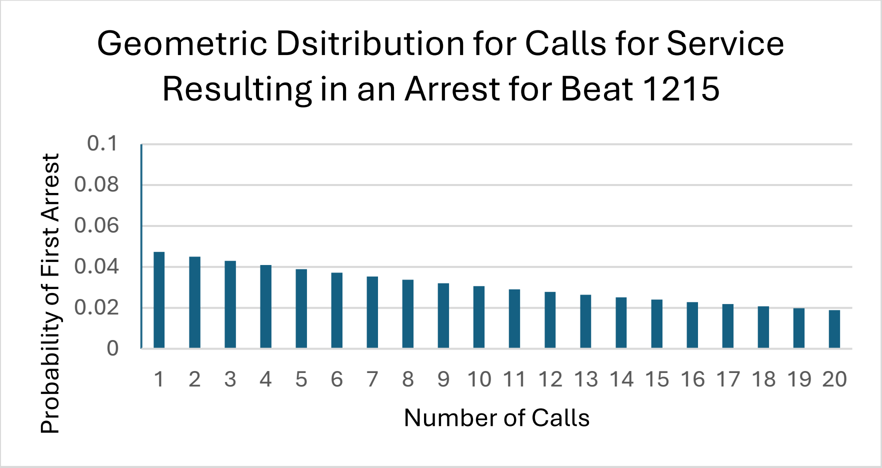
For example, we can ask the probability of an arrest happening on the third call:

- Beat 1512: $P(Y = 3) = (1 - .1077)^2(.1077) = \mathbf{.0858}$
- Beat 1834: $P(Y = 3) = (1 - .1703)^2(.1703) = \mathbf{.1172}$
- Beat 1215: $P(Y = 3) = (1 - .0473)^2(.0473) = \mathbf{.0429.}$

We can also model the geometric distribution for these three beats:







Finally, we can find the expected value for the three different beats to determine the average number of calls for service needed to observe the first arrest.

- Beat 1512: $E[Y] = \frac{1}{.1077} = \mathbf{9.2851}$ calls
- Beat 1834: $E[Y] = \frac{1}{.1703} = \mathbf{5.8720}$ calls
- Beat 1215: $E[Y] = \frac{1}{.0473} = \mathbf{21.1416}$ calls

Here, we can see Beat 1834 is most likely to follow up on arrests and Beat 1215 is least likely. This could be due to several factors such as calls not warranting an arrest, or police officers on duty lacking diligence.

Location-Based Investigations

We can use a hypergeometric model to investigate enforcement and performance trends within the Chicago Transit Authority (CTA) system. This system includes trains, buses, platforms, and more. Suppose we selected a random sample of 50 CTA-located crimes with the goal of assessing arrest rates to see if the rates are up to standards.

From the Chicago crime database, we see

- The total number of CTA-located crimes is 17,057.
- Of the total number of crimes, 3,889 resulted in an arrest.

Let discrete random variable Y represent the number of crimes from the sample which resulted in an arrest. This sampling is done without replacement.

The probability of observing at least 10 arrests from our sample of 50 is

$$P(Y \geq 10) = 1 - P(Y \leq 9) = 1 - \sum_{i=0}^{9} \frac{\binom{3,889}{i}\binom{13,168}{50-i}}{\binom{17,057}{50}} = \mathbf{.7304}$$

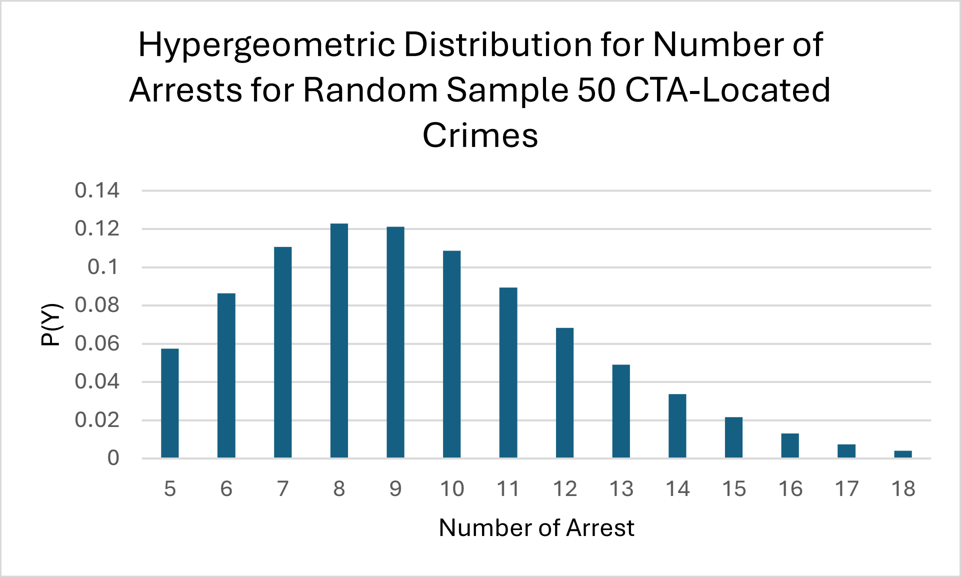
This means that there is a 73% chance of observing 10 or more arrests from a sample of 50 crimes located within the CTA system. It is not unusual to observe at least 10 arrests.

We can also calculate the expected value for the number of arrests within our sample of 50 crimes.

$$E[Y] = \frac{50(3,889)}{17,057} = \mathbf{11.400}$$

So, if we randomly sample 50 crimes located withing the CTA system, we are expecting to see a little more than 11 arrests on average.

Finally, we can model our probability distribution from our sample:



Predicted Repeat Offenses Before an Arrest

Some locations in Chicago are "hot spots" for crime but have low resources patrolling the area to perform arrests. Often, a crime will be repeated until it finally triggers more resources to be able to get the arrest. The negative binomial distribution is useful for determining the number of triggers the department needs before obtaining a fixed number of arrests.

For this example, we will look at State Street, an area with a low arrest rate to a high crime rate ratio. From the database, 17,206 crimes have been reported, but only 1,353 resulted in an arrest.

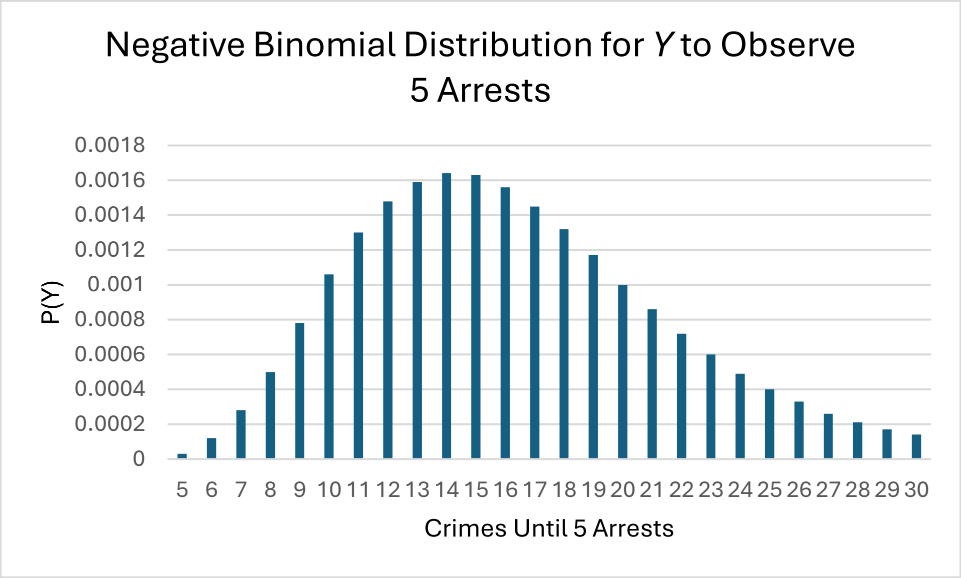
- The probability of arrest is .0787
- We will model a scenario to predict the number of crimes which have to occur to obtain 5 arrests.

Let discrete random variable *Y* represent the number of crimes that occur before a fixed number of crimes happen. We will observe the probability that it will take 20 crimes to observe 5 arrests.

$$P(Y = 20) = \binom{20-1}{5-1}(.0787)^5(.9213)^{20-5} = \mathbf{.00247.}$$

This means that there is a .25% chance it will take 20 crimes to occur on State Street for the police to make five arrests.

We can also graph the negative binomial probability distribution for our random variable *Y* to view the likelihood of different amounts of crimes it will take to observe five arrests.



We can also find the expected value for the number of crimes it will take on average, to make five arrests:

$$E[X] = \frac{5}{.0787} = \mathbf{63.5234}$$

So, on average, it will take around 63 crimes to occur on State Street for the police to make five arrests.

Real-World Implications

These discrete probability distributions offer valuable insight to the Chicago Police Department. Even with a low domestic crime rate, there is still a strong likelihood that these types of crimes will still appear. Additionally, we saw that there are large discrepancies in police efforts and efficiency in different beats. This may call for an evaluation of patrolling officers and teams to make sure they are following up on their reports, making arrests when necessary, and prioritizing important cases.

We also saw that hypergeometric models are very useful when determining if police in different areas are meeting historical arrest quotas, or if they need to be more diligent in their work. Finally, we saw that enforcement intervention often requires a repeated number of offenses in order to be seriously investigated. Chicago may need to be better at responding to criminal acts, even if they only occur a few times in an area.

Daily Crime Events Modeled with Poisson Distribution and Tchebysheff's Theorem

Many types of crimes in Chicago occur throughout the day at an inconsistent rate. In general, crimes are more likely to occur during nighttime, rather than during the day. Specifically, Chicago experienced an influx of vehicle theft in 2023, reaching 29,063 incidents. To analyze this inconsistent pattern, we can use the Poisson distribution to model these events which occur within a fixed interval of time.

Discrepancies in Car Theft Consistency Throughout the Day

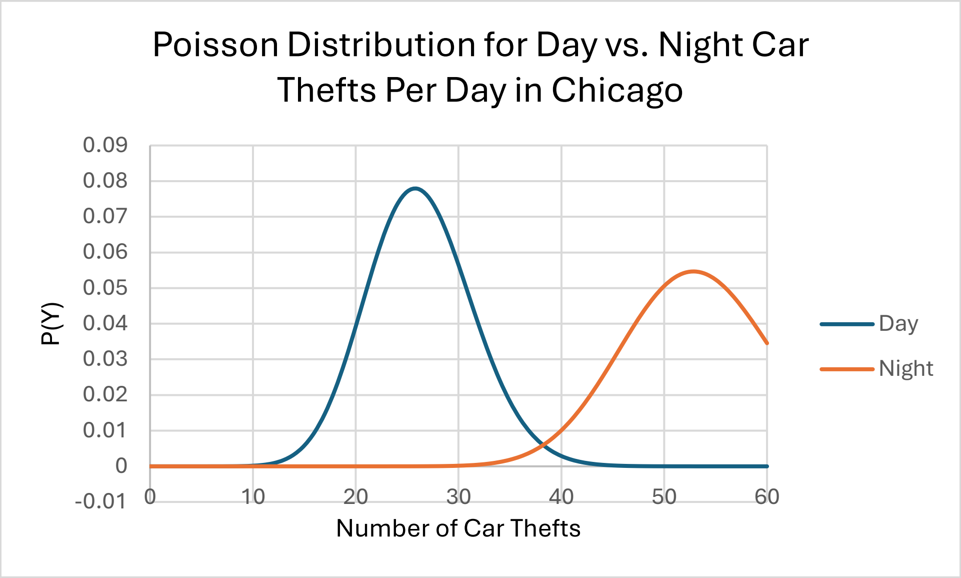
Let discrete random variables Y_d and Y_n represent the number of car thefts occurring from 6AM-6PM (day) and 6PM to 6AM (night), respectively. According to the Chicago database, 67% of car theft crimes occur at night. Given our statistics, we can calculate the expected number of thefts for each time period per day.

- $\lambda_d = .33(29{,}063) = 9{,}590.79$ car thefts per year
- $\lambda_n = .67(29{,}063) = 19{,}472.21$ car thefts per year

These rates apply to a year, but we are interested in the crimes for any given day. So, we must use dimensional analysis to fix our lambda value:

- $\frac{9{,}590.79\ car\ thefts}{1\ year} \times \frac{1\ year}{365\ days} = \frac{26.2761\ car\ thefts}{1\ day} = \lambda_d$
- $\frac{19{,}472.21\ car\ thefts}{1\ year} \times \frac{1\ year}{365\ days} = \frac{53.3485\ car\ thefts}{1\ day} = \lambda_n$

We can graph each distribution:



Intervals of Car Thefts Modeled Day vs. Night

In a Poisson distribution, λ is equal to both the mean and variance. Therefore,

- $\mu_d = 26.2761$
- $\sigma_d = \sqrt{26.2761} = 5.1260$
- $\mu_n = 53.3485$
- $\sigma_n = \sqrt{53.3485} = 7.3040$

Now, we can use Tchebysheff's theorem to find what percent of data lies between a variety of intervals about the mean. We can construct a table to better view the discrepancies between crime frequency during the night and during the day.

Daytime

Lower Limit	Upper Limit	k-Value	Percent of Data
18.0241	38.5281	2	75
15.7611	41.3911	2.5	84
12.8981	43.6541	3	88.8889

Nighttime

Lower Limit	Upper Limit	k-Value	Percent of Data
38.7405	67.9565	2	75
35.0885	71.6085	2.5	84

31.4365	75.2605	4	88.8889

From these tables, we can clearly see the difference between the number of car thefts that are expected to happen during the day vs. at night.

75% of days will have car thefts between 18 and 39, while 75% of nights will have car thefts between 39 and 68. That is a 38.1% increase from day to night.

84% of days will have car thefts between 16 and 41, while 84% of nights will have car thefts between 35 and 72. That is a 48% increase from day to night.

89% of days will have car thefts between 13 and 44, while 89% of nights will have car thefts between 31 and 75. That is a 41.94% increase from day to night.

So, on average, there is a 42.68% increase in car theft at night compared to during the day.

Real-World Implications

There is a clear distinction between the frequency and likelihood of car theft occurring during the day compared to during the night. This shows that the Chicago Police Department should allocate more time and resources to surveil during night to reduce the volume. This shows that crime is more likely to happen with lower visibility and fewer people around.

The Chicago Police Department can consider tactics such as changing surveillance schedules to allot more officer to patrol during night, institute more overnight shifts, and create outreach programs to create a safer community.

Modeling Crime Times with a Continuous Random Variable

In the context of crime analysis, some aspects are continuous and can be modeled through the distribution of a continuous random variable. This information can be useful when considering the time between reported crimes, severity of crimes, and more.

Modeling Times Between Assaults

Let’s assume the reported assault rate in Chicago follows a uniform distribution from 0 to 12 hours. Then, any time interval within that range has the same likelihood of being the time between two consecutive assault reports. The uniform distribution PDF would be as follows:

$$f(y) = \begin{cases} \frac{1}{12}, & 0 \leq y \leq 12 \\ 0, & elsewhere \end{cases}$$

So, we can calculate the expected value for the distribution which is

$$E[Y] = \frac{0+12}{2} = 6$$

This means that on average, there will be a reported assault every 6 hours in Chicago. It is important to note that assaults most likely occur much more frequently than the model suggests. We are only accounting for the reported assaults in the database.

Modeling Times Between Battery

For the next analysis, we will focus on the time between consecutive battery charges. We will model the waiting times with a gamma distribution. Using the database, we can calculate the time differences between battery reports.

Using the data, we find

- Mean of the time differences = $\mu = .15$ days
- Variance of the time differences = $\sigma^2 = .02$ square days.

From mean and variance, we can find our shape and rate parameters:

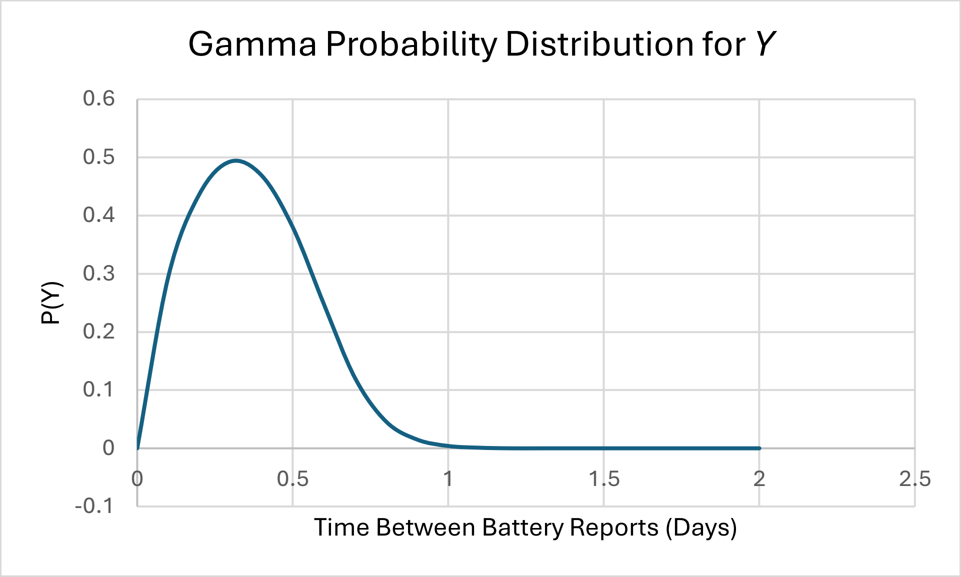
- Shape parameter = $k = \frac{\mu^2}{\sigma^2} = 1.125$
- Rate parameter = $\lambda = \frac{\mu}{\sigma^2} = 7.5$

Let continuous random variable *Y* represent the time between two consecutive battery reports in days. We want to find the probability that two consecutive battery reports take place .2 days apart. Using the formula for gamma distribution, we find that

$$P(Y < .2) =\mathbf{.1192.}$$

This means that there is a 12% chance that two consecutive battery reports will occur within 4.8 hours of each other.

We can graph the gamma probability distribution for our continuous random variable.



Real-World Implications

These distributions provide insight for law enforcement officials regarding resource allocation and time management. If this model suggests that assaults occur every six hours from a database of reported crimes, that means assaults occur way more frequently because many go unreported. Police can use this information to create outreach groups or educational programs to help Chicago residents safely report assaults to keep the community safe. Predictive models can be implemented to help eliminate high frequency windows of crime.

Police can use continuous models to help better understand crime occurrences and how often crimes really happen in their city. By applying these models to a plethora of types of crime, the Chicago Police Department can understand which crimes to prioritize based on frequency.

Multivariable Crime Analysis

Just looking at one continuous random variable in crime data poses limitations. Looking at bivariate distributions allows us to connect different factors together to view underlying trends and patterns.

Relationship Between Crime Type, Location Description, and Arrests

We can use a multivariate probability distribution to understand the relationship between the type of crime committed, what type of area the crime took place in, and the arrest outcome.

First, we will define our random variables. Let

- C represent the type of crime (Theft, Narcotics, Motor Vehicle Theft),
- L represent the location type (Street, Residence), and
- A represent the arrest status (True, False).

Using data from the Chicago database, we can construct a table as follows:

Crime Type (*C*)	**Location (*L*)**	**Total Crimes**	**Arrests**	**P(*C, L*)**	**P(Arrest|*C, L)***
Theft	Street	62,109	3,109	.3097	.0501
Theft	Residence	29,835	1,591	.1487	.0533
Narcotics	Street	6,286	6,087	.0313	.9683
Narcotics	Residence	3,250	3,163	.0162	.9735
Motor Vehicle Theft	Street	25,645	722	.1279	.0282
Motor Vehicle Theft	Residence	11,154	305	.0556	.0273

From this table, we can see that narcotic crimes have the highest arrest rate compared to theft and motor vehicle theft. We see that narcotic arrest rates don't vary by location very much and remain consistent.

Marginal and Conditional Probabilities

We can calculate the marginal probabilities of each crime-location combination. Let's choose Street crimes and Theft crimes as examples.

- P(Theft) = P(Theft, Street) + P(Theft, Residence) = .3097 + .1487 = **.4584**
- P(Street) = P(Theft, Street) + P(Narcotics, Street) + P(Motor Vehicle Theft, Street) = .3097 + .0313 + .1279 = **.4689**
- P(Arrest) = Total Arrests / Total Crimes = 14,977 / 200,279 = **.0748**

This means there is only a 7.5% arrest rate among crimes in Chicago including theft, motor vehicle theft, and narcotics.

We can also calculate conditional probabilities. For example:

- P(Arrest | Narcotics, Residence) = $\frac{3{,}163}{3{,}250} =\mathbf{.9735}$
- P(Arrest | Motor Vehicle Theft, Street) = $\frac{722}{25{,}654} =\mathbf{.0282}$
- P(Theft | Street) = $\frac{62{,}109}{94{,}040} =\mathbf{.0748}$

As we see, these values match the ones from our table above.

Independence Between Crime Type and Arrests

We can test to see if crime types and arrest rates are independent or dependent of each other. As we saw above, the type of crime heavily impacts the rate of arrests. But we can prove this by saying that:

P(Theft)P(Arrest) = .4584(.0748) ≠ P(Theft and Arrest) = .0235.

Real-World Implications

The multivariate distributions between crime types, location types, and arrest status reveal heavy discrepancies between the likelihood the arrests based on crimes. More dangerous crimes that are drug-related are more likely to result in an arrest than crimes such as theft. This most likely reflects public safety concerns as narcotics are much more dangerous than events such as petty theft.

We also saw that the type of crime has more of an impact on arrests rates than the type of location in which the crime took place. This is because no matter the location, the arrest rates for each crime stayed most consistent. Again, this is probably a reflection of public safety concerns.

Conclusion

Statistical analysis of Chicago crime data shows important trends and answers vital questions. We were able to prove discrepancies in law enforcement and show how the Chiago Police Department can better allocate their time and resources to overall provide a safer community for their residents.

We saw that theft cases need more resources devoted to ensure arrest rates rise and meet those of other crimes. Additionally, there needs to be more surveillance during the night to reduce the influx of crimes occurring between the hours of 6PM and 6AM.

By applying these statistical principles to other databases in other cities and to a larger variety of logged information, we will be able to answer more questions and overall provide more effective law enforcement.

References

City of Chicago. (2025). *Crimes - 2001 to Present*. Chicago Data Portal. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

Wackerly, D., Mendenhall, W., & Scheaffer, R. L. (2014). *Mathematical statistics with applications* (7th ed.). Cengage Learning.