Jess Moslander
10 December 2012

You Want to Live Where?
Crime versus Population
MATH 3220 Report – Final

Executive Summary

Since we are looking for the correlation between crime and population, one can use percentages in order to find the safest cities in the United States. Off of the UNI Machine Learning Repository: Communities and Crime Unnormalized Data Set, one can retrieve the population count for specific cities for each state, recorded violent incidents, and non-violent circumstances. (Redmond) By taking the total recorded circumstances for each violence type and dividing them by the population size, one can find percentages of violence within specific United States' cities. In addition to further the investigation, one can separate the violent circumstances into their respectable groups and divide them by the population to get where certain violent and non-violent crimes are more popular. We have found that there is extreme little correlation between population and crimes.

Problem Description

In the Communities and Crime Unnormalized Data Set, there are given records of violent and non-violent crimes in specific United States' cities. There are multiple cities per state in order to represent a wide variety of portions in each state. The difficulty arises when one looks at the data and notices there is missing values. Since some cities have different definitions of certain sub-violent categories, such as rape, there may not be a value given for that category. Then one needs to find a way to calculate in order to fill in the missing data or a way to maneuver around the blank.

Many variables contribute to discovering which locations have the highest crime for the population. The variables included in the dataset involve the community, the median family income, the involving of law enforcement, and the percent of officers assigned to drug units. The crimes contributing to the statistics could be predicted are the 8 crimes considered "index crimes" by the FBI (Murder, Rape, Robbery, Assault…), per capita versions of each, and the violent crimes and nonviolent crimes rates. The violent sub-categories we will be working with are murder, rape, robbery, and assault. For the non-violent sub-categories, we will be working with burglary, larceny, auto theft, and arson.

The data set was taken of different sized police departments members. With the departments with at least 100 officers (give or take for the smaller divisions), some of the communities were not found in the census or crime data sets were omitted for major lack of informational purposes. However, there is missing information in the data set still. The calculations of the used data set are using population values included in the 1995 FBI data. This differs from the 1990 Census values.

Since there is a controversy in some areas involving rape, the values for rape may be less than recorded. Many of the communities omitting their values altogether concerning rapes were from the Midwestern states. For non-violent crimes in the United States (burglaries, larcenies, auto thefts, arsons…) are calculated by the sum of crime variables versus the population numbers.

Analysis Technique

Since one is trying to find the percentages of crimes by populations, we are going to run tests through a comparison methodology. Each of the statistics was calculated using population and the sum of crime variables considered violent crimes (murder, rape, robbery, assaults…) in the United States.

Equations used to find statistics for crimes by populations for communities in the United States

For Considered Violent Crimes:

Population of the United States' Community ÷ considered violent crimes (murder, rape, robbery, assaults…)

-OR-

Population of the United States' Community ÷ [number of murders + number of rapes + number of robberies + number of assaults + …]

For Considered Non-Violent Crimes:

Population of the United States' Community ÷ considered non-violent crimes (burglaries, larcenies, auto thefts, arsons…)

-OR-

Population of the United States' Community ÷ [number of burglaries + number of larcenies + number of auto thefts + number of arsons + …]

However, some of the values are missing from the data set. In this case, in order to fill in the values and calculate the percentages of the total crime, violent crimes, and the non-violent crimes, we would work to find the averages of the given values and find the totals without using that specific missing piece. For instance, if one was given the murder, robbery, and assault amount but not the rape records, we could add the amounts of murder, robbery, and assault and divide by three (instead of four).

For instance, Chicago is missing its rape records:

City	State	Population of Community	Murders	Rapes	Robberies	Assaults	Total Violent Crimes	Total Violent Crimes ÷ population
Chicago City	IL	2783726	845	**?**	35189	39753	**25262.33**	0.91%

Number of Murders, Robberies, and Assaults ÷ 3 = Total Violent Crimes ÷ Population = Total

845 + 35189 + 39753 ÷ 3 = 25262.33 ÷ 2783726 = .91%

(Took values for violent or non-violent data and averaged in order to fill in missing totals)

This would allow us to fill in the total percentage for the violent crimes without factoring the rape category. This does screw the data; however one is given an amount to work with so the city is not thrown to the side and not added to the results. In addition, finding the average of the given amounts is more efficient than plugging zeros in for the missing data. That would screw the data massively leaving not sufficient results.

Depend on what specific data one is looking for, the results charts differ. For the top ten most populous cities, we would just be observing the population with the crime amount. However if one is trying to find the most dangerous cities based off the population in the United States, we would need to sort by "total violent + non-violent crimes ÷ population". This gives a different set of cities – supposedly the most dangerous cities. *See attached charts for details.

Assumptions

We can assume the statistical information provided has assumptions made concerning the number of crimes represented for the United States communities are correct. However, the amount of rapes could be incorrect. Depending on which communities the incident happened, there is a major controversy to what entitles rape. Some of the communities have voided the

count in the data set. In addition, we assume that the number of violent crimes is not counted twice for multiple violent crime incidents. For crimes which cannot be identified, we can assume they are placed into the most logical areas.

We also assume that by dividing the given data by the number of values we have that it will present an acceptable value in order to maneuver around missing data. Plugging in zeros would screw the data massively so going around the missing data screws the data less. In addition, it is giving us a value for the total instead of eliminating the city's records completely from the data set.

Results

Once all the mathematic work is complete, one can read that there is little correlation between population and crimes. True, with larger populations, there is less crime. However, there are more people to commit crimes over a higher population. On the flip side, smaller cities have a higher rate of crime since there are less people. It is more noticeable but does not mean there is more crime. All of the percentages are based off of the population of each individual city. There is a margin of error. We have concluded that many of the top ten largest populated cities are on the 10 smallest total non-violent plus violent percentages chart. Basically, this proves that the higher the population, the lower the amount of crime percentage there is. Little correlation is made for crime versus population showing that it does not matter where you live. There is crime everywhere; however the "safest" city in the United States, based off of Communities and Crime Unnormalized Data Set (and the 1995 records), would be New York City (based off of the total violent plus total non-violent crimes divided by the population).

Issues

Issues I came across working with the UNI Machine Learning Repository: Communities and Crime Unnormalized Data Set was how to test my hypothesis – a larger population would result in a higher crime rate since there are more people to commit offenses. I had no clue where to start with the data. The only thing I knew was that I needed to somehow get it into an excel document and eliminate unless data. I needed what was important in order to get the total amount of violent and non-violent crimes for each city. In addition, I needed to figure out what to do with the missing data. I just assumed that an average was acceptable in order to find the

total amounts. Since there were a number of cities with missing data, I could not throw them aside.

My original thoughts when I looked at the topic I chose and the data set was to run it through a computerized system since we have been doing that all semester. However, with some help, I realized it was not necessary. My main goal was to find if there was a correlation between the population and crime. Honestly, I do not know if there is a computerized system that would give me what I need.

Appendices and Works Referenced

After reading Levitt's and Dubner's article on Freakonomics, the data on crime versus population made sense. However, the reading was perplexing. Stating there is a correlation between a baby boom's education and abortion banned was confusing. Even with the banning of abortion, what does it have in common with a child's education? From what I gathered, Freakonomics is the data collected but does not correlate with the original expectations. In the article, it stated that the infants born after the abortion ban would suffer, have low grades, and a low track life but it resulted in higher test scores, a struggle but not suffering, and an average life.

Would this information relate to my data in the sense that what is expect is not what resulted? It would make sense more populous cities to have a higher crime rate. However, there seems to be no correlation between the population size and crime rate. True, there is evidence that the higher the population, the lower the crime but that could be due to the amount of crime divided by the amount of people. It balances out and makes it seem like there is less crime. In smaller cities, the crime is noticeable since the population is lower.

DiNardo, J. (n.d.). Freaknomics: Scholarship in the service of storytelling. American law and economics review, 8(3), 615-626. doi: Oxford Journals

Levitt, S., & Dubner, S. (n.d.). Freakonomics: A rogue economist explores the hidden side of everything. Retrieved from Http://contemporarylit.about.com/od/socialsciences/fr/freakonomics.htm

Redmond, M. (2011, March). Communities and crime unnormalized data set. Retrieved from http://archive.ics.uci.edu/ml/datasets/Communities and Crime Unnormalized

U. S. Department of Commerce, Bureau of the Census, Census of Population and Housing 1990

United States: Summary Tape File 1a & 3a (Computer Files)

U.S. Department Of Commerce, Bureau of the Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan. (1992)

U.S. Department of Justice, Bureau of Justice Statistics, Law Enforcement Management And Administrative Statistics (Computer File) U.S. Department Of Commerce, Bureau Of The Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan. (1992)

U.S. Department of Justice, Federal Bureau of Investigation, Crime in the United States (Computer File) (1995)