# Minneapolis police use of force

Data set analysis: mn_police_use_of_force.csv

This data set provides information about police interventions in Minneapolis from 2016 to 2021 presenting the following 13 variables:
*response_datetime*: date and time of intervention, *problem*, *is_911_call*, *primary_offense*: the primary offense charged to the subject, *subject_injury*: whether someone was injured or not by the time the police arrived, *force_type* and *force_type_action*: type of force applied by the police, *race*, *sex*, *age*, *type_resistance*: how the subject reacted, *precinct*: id of the police district, *neighborhood*: where the problem occurred.
There are different interesting observations we shall capture from such data, so let's get started.

## 1 Variation among neighborhoods

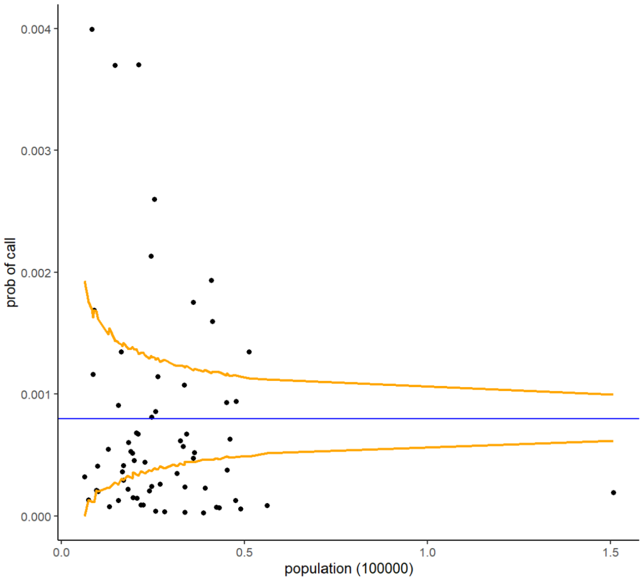
The first figures we might be interested in are the police calls per neighborhood: in fact, by assessing their distribution we can try to guess whether there is a variation or not in the number of crimes committed and this could be insightful for countless reasons. For this purpose let's just focus on last two years and consider the calls regarding serious crimes, like burglary, domestic abuse, fights, shootings, stabbings and so on.
Our hypothesis is that there is no meaningful difference among neighborhoods and we should proceed in testing it either to accept or reject it: to do so, it seems reasonable to compare the specific probabilities of call per district and assess their distribution around the mean value, which should be, according to our hypothesis, the true value for any district. The problem is that the data set does not provide any information about the population per district, so we must find a way to retrieve these values: actually, the most feasible and smart thing that came to my mind, even though it is not smart at all, was to measure the areas of all districts and approximate the relative population by multiplying the total population of Minneapolis times the fraction of the total area occupied by each neighborhood (such a painful job on google earth is worth alone the whole reading).
A clever move now is to think that for each person in a district the event of coming across a single criminal is nothing but a Bernoulli random variable which takes on a positive value with a probability equal to the ratio we have previously computed: therefore, the total number of committed crimes per neighborhood is approximately equal to a binomial random variable with the same probability and size of the sample equal to the size of the relative population.
Well, our aim is to test the *"null"* hypothesis, so we have to figure out whether the true value is reasonable or not under our approximation and the best way to do so here is by using confidence intervals: specifically, we'd better set a large confidence interval, say 99% since our previous measurements could have been somehow a little incorrect and we must take it into account. As a result we get:

$$CI = (qbinom(.005, population, mean[prob\_call]), qbinom(.995, population, mean[prob\_call]))$$

so for example for Armatage it is:

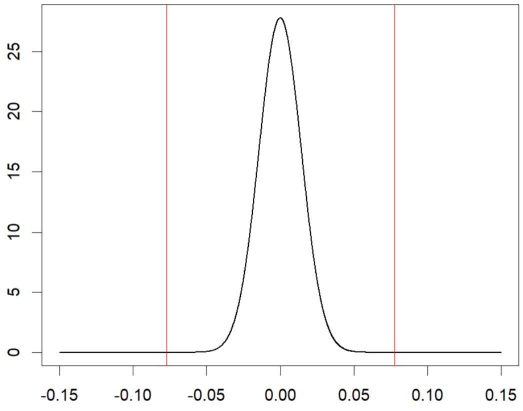
$$\begin{aligned} CI &= (qbinom(.005, 38816, 8e-4), qbinom(.995, 38816, 8e-4)) \\ &= (4.6e-4, 1.2e-3) \end{aligned}$$

and here it is a graphic representation of the overall situation in a funnel plot, where the dots stand for the true rates and the orange lines represent the 99% confidence interval around the mean value in blue.



Clearly, it is neither a matter of bad measurements nor a matter of outliers influence, even though I had to cut some of them from the image for visual purposes: the truth is that our hypothesis is false and that crime rates actually vary among different districts according to many factors such as position and people.
For the record, a quick calculus of the p-value is going to confirm our findings for Armatage and eventually for many other neighborhoods:

$$
\begin{aligned}
sigma &= \sqrt{\frac{mean[prob\_call] \cdot (1 - mean[prob\_call])}{population}} \\
&= \sqrt{\frac{8e-4 \cdot .99}{38816}} = 1.4e-4 \\
&\rightarrow 2 \cdot pnorm(prob\_call[Armatage], mean[prob\_call], sigma) \\
&= 2 \cdot pnorm(2.57e-5, 8e-4, .1.4e-4) = 6.82e-8
\end{aligned}
$$



## 2 Changes throughout the years

Even though in the previous analysis we did not confirm our initial hypothesis, we can't say that it was totally inconclusive: actually, we learnt something important about the distribution of serious crimes in Minneapolis, and yet there are plenty other pieces of information waiting to be processed, for example whether over the last years there were some sort of changes or if the number of serious committed crimes has not varied that much.
There are different paths one may follow to answer such question, but maybe the quickest and still insightful one is probably via construction of confidence intervals: they can be highly helpful to assess the possible variability of

a measurement indeed.
To compute an adequate confidence interval we shall consider that in a measurement more or less 95% of the obervations are included in the interval $(\bar{x} \pm 1.96\sqrt{var})$ and that since we are investigating an event with very small probability to happen within a large population the distribution of serious crimes is approximatively poissonian with both mean and variance equal to the mean value.
Specifically, let's focus on the last 3 years, when the number of serious committed crimes and relative confidence intervals were:

$$\begin{aligned} 2019)\ 639 &\rightarrow CI = (589, 689) \\ 2020)\ 763 &\rightarrow CI = (709, 817) \\ 2021)\ 825 &\rightarrow CI = (769, 881) \end{aligned}$$

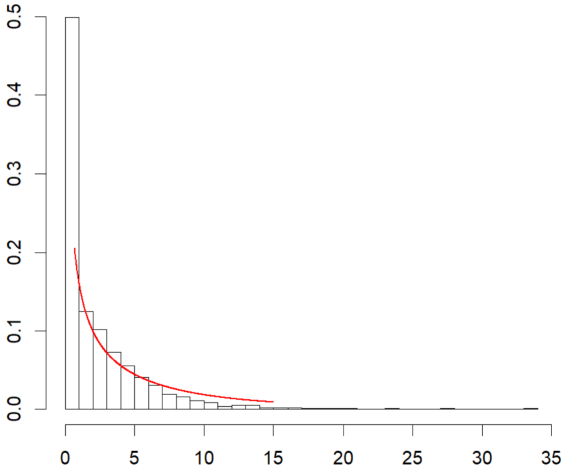
Evidently, none of the intervals contain the true value for another year, so the numbers don't appear strongly linked, but surely the intervals for the last two years overlap much more with respect to the ones for 2019 and 2020, hence they could be somehow related. Yet, to correctly understand whether there was a real variation or not, it is important to compute also the confidence intervals for the difference between adjacent years:

$$\begin{aligned} 2019/2020)\ CI &= (763 - 639 \pm \sqrt{763 + 639}) = (50.6, 197.4) \\ 2020/2021)\ CI &= (825 - 763 \pm \sqrt{825 + 763}) = (-16.1, 140.1) \end{aligned}$$

The golden rule to apply here is that if such intervals contain 0 it means that the variability has too influence on the true value, so we can't conclude anything; otherwise it is likely that there has been a change. In this case, the left extremum of the 2019/2020 interval is pretty far from 0, while the 2020/2021 includes it: as a consequence, we can affirm that there was a significant change between 2019 and 2020, but not between 2020 and 2021.
Since we are already dealing with numbers describing the distribution of the crimes it is also interesting to model a common pattern in order to describe their behavior and predict future observations: for example, imagine we want to know how many days in a year we should expect to record more than 10 serious crimes. Well, even though we are provided with a discrete random variable, it actually takes on integer values within a wide range, so we can approximately consider it as continuous; furthermore, it is evident that the distribution follows an exponential function, hence after a couple of trials and adjustments we can come to the conclusion that

$$\Gamma(\sqrt{\frac{mean[daily\_crimes]}{10}}, .1) = \Gamma(.438, .1)$$

seems a good upper bounding alternative.



The probability associated to the event of more than 10 serious crimes happening is:

$$pgamma(11, .438, .1, lower.tail = FALSE) = .117$$

which results in about 43 days per year against the true 50.
P.S. What we have done here is just a trivial approximation: it is not the best one!

## 3 Different reactions

Another interesting analysis might be carried out on people's tendency to react violently or not in presence of the police: in particular, we want to investigate whether in such situations males and females behave the same way or not. To do so, let's create a table distinguishing between males and females who committed or did not commit a crime after the police had arrived.

| | crime | no crime | tot |
|---|---|---|---|
| male | 2056 | 2288 | 4344 |
| female | 548 | 311 | 859 |
| tot | 2604 | 2599 | 5203 |

The relevant values for us are the probability for men and women to commit a crime and the average probability:

$$prob\_male = \frac{2056}{4344} = .473$$
$$prob\_female = \frac{548}{859} = .638$$
$$mean\_prob = \frac{2604}{5203} = .5$$

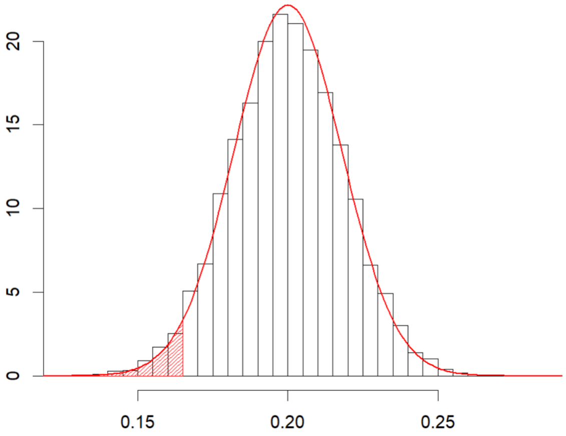
It is clear that in these situations women tend to be more aggressive than men, but what about the following question: do females have violent reactions 20% times more than males?
To provide an answer, let's just carry out a proper one-sided hypothesis testing.

- **STEP 0**: set a null hypothesis $H_0$.
  Suppose that females have violent reactions 20% times more than males.

- **STEP 1**: compute a test statistic $t_0$ of the data.
  Since we are comparing different behaviors it seems reasonable to set the test statistic equal to the difference between the two probabilities of committing crimes for males and females.

$$t_0 = prob\_female - prob\_male = .638 - .473 = .165$$

- **STEP 2**: assess the distribution $T$ of $t_0$ under $H_0$.
  Given the data and the scope we have, it is logic to simulate the distribution of our difference by randomly generating data with similar characteristics to the ones picked from the table. In fact, we can assume that the event of committing a crime or not before an officer is well represented by a Bernoulli random variable which takes on a positive value with the probability we have calculated before: accordingly, the total numbers of such crimes is nothing but a binomial random variable with the same probability and size equal to the total number of men or women.
  Notice that in order to have meaningful results it is fundamental to repeat the generation of random samples multiple times, as with bootstrap: the distribution of the differences simulated in the experiment are represented below.



  Naturally, the differences are normally distributed around the mean value, which is 20% under $H_0$ and the related standard deviation is equal to the square root of the sum of the variances for the two groups.

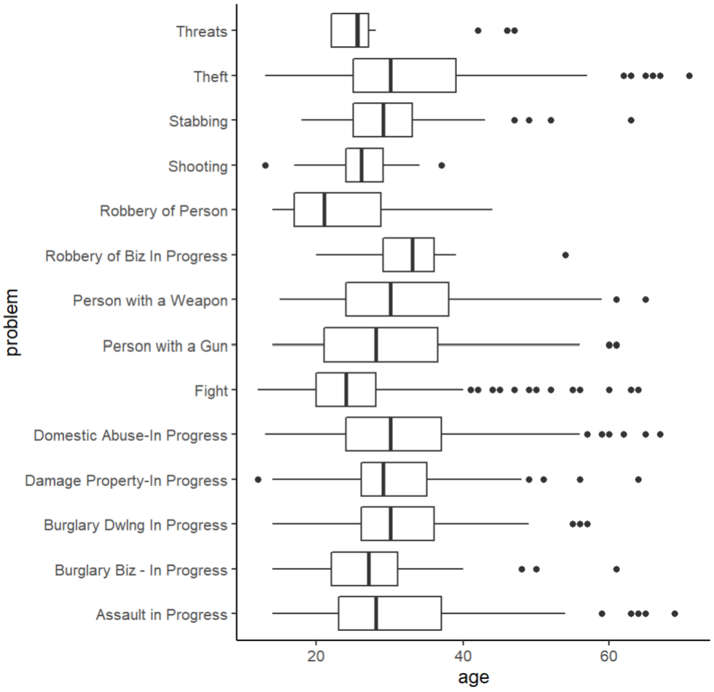
- **STEP 3**: computation of the p-value.
We are left with the last step of the process, that is the simple calculus of the p-value for a one-sided test:

$$
\begin{aligned}
sigma &= \sqrt{\frac{prob\_male \cdot (1 - prob\_male)}{tot\_males} + \frac{prob\_female \cdot (1 - prob\_female)}{tot_f emales}} \\
&= \sqrt{5.74e-5 + 2.69e-4} = .018 \\
&\rightarrow pnorm(.165, .2, .018) = .025
\end{aligned}
$$

and it leads to reject the null hypothesis $H_0$ in favor of a new hypothesis $H_1$: females have violent reactions less than 20% times more than males.

# 4 Criminals' age

Among the numerous insights that this data set offer, the last correlation I want to report is between the type of crime committed and the related criminal's age: in fact, having a look to such type of data could be very helpful in order to develop and implement specific strategies to reduce the probability of recording high numbers of crimes committed. For this reason, let's consider again just serious crimes, leaving the others apart.



Naturally, the majority of people committing such crimes are between 20 and 40 years old and the median values seem to be quite similar, but what surprises me the most is the width of the interquartile range and the presence of so many upper outliers: more than 1% of the crimes are committed by people over 60 years old, where such percentage represents more or less 120 observations, that is by the way also the number related to minors under the age of 14.

Not only Minneapolis has an high crime rate, but also people of any age contribute in making such index arise and clearly this is not something positive. What would be interesting now is investigating whether the conclusions we have inferred here would be the same for other cities, both in US and around the whole world.