311 statistics – methodology

Joel Miller

July 6, 2020

The Chicago 311 dataset lists all 311 requests made in the past few years, along with the request type, location (expressed in coordinates), and open/close time of each request. Meanwhile, ACS/census data gives demographic information for each block group in Chicago. We're interested in using these datasets to see whether different demographic groups enjoy different average response times for various types of 311 requests.

1 Notation

- Let B denote the set of block groups.
- Let A be a table listing average response times for different requests by block group, such that $A[b, r]$ gives the average response time for requests of type r sent out from block group b.
- Let D be a table listing demographics by block group, such that $D[b, d]$ gives the number of people in demographic d who live in block group b.
- Let $T(d)$ denote the total number of people in demographic d across the whole city.

Among other things, we're interested in calculating $R(d, r)$ – the average response time for 311 requests of type r for people in demographic d.

2 Calculating $R(d, r)$

One idea is to calculate $R(d, r)$ as

$$\begin{aligned}&\sum_{b\in B} P(\text{I living in } b \mid \text{I belong to demographic } d) \times A[b, r]\\ =&\sum_{b\in B} \frac{D[b, d]}{T(d)} \times A[b, r] = \frac{1}{T(d)} \sum_{b\in B} D[b, d] \times A[b, r]\end{aligned}$$

But there's an issue – not all block groups have made 311 requests for all request types, so $A[b, r]$ is undefined in some cases. To get around this, we could let $B_r = \{b \in B \mid A[b, r] \text{ is defined}\}$ and compute $R(d, r)$ as

$$\frac{1}{\sum\limits_{b\in B_r} D[b, d]} \times \sum_{b\in B_r} D[b, d] \times A[b, r]$$

and this is what I do in my code. However, I am *not* necessarily convinced that this is the best way to deal with missing data.

3 Calculating response times for larger demographic groups

The ACS data is pretty specific with some demographics, especially demographics related age and sex. I had the notion that it might be more informative to look at average response times for, say, all people over 60, rather than just people aged 60-61, or just people aged 62-64, etc.

Accordingly I was interested in calculating $R(\{d_1, ..., d_n\}, r)$, the average response time for request type r for by people in a set of disjoint demographics $\{d_1, ..., d_n\}$ (where, say, d_1 = people age 60-61, d_2 = people age 62-64, etc). One idea would be to calculate this as

$$\begin{aligned}
&\sum_{d_i} P(\text{I belong to demographic } d_i \mid \text{I belong to one of } d_1, ..., d_n\} \times R(d_i, r) \\
=&\sum_{d_i} \frac{T(d_i)}{\sum\limits_{d_x} T(d_x)} \times R(d_i, r) = \frac{1}{\sum\limits_{d_x} T(d_x)} \times \sum_{d_i} T(d_i) \times R(d_i, r)
\end{aligned}$$

But again, undefined values lurk – $R(d_i, r)$ can be undefined if no members of d_i live in *any* of the block groups that have ever produced 311 calls of type r. I found that this was occasionally true when d_i was a small population, like Pacific Islanders (code B03002007), and r was an uncommon 311 call, like a severe weather call (code JNS).

To deal with this I did something similar to what's described in the last section: I only summed over the demographics d_i for which $R(d_i, r)$ was defined, and only divided that sum by the sum of $T(d_i)$ for which $R(d_i, r)$ was defined. Again, I'm by no means convinced that this is the best way to do things – it just seemed like the most straightforward approach.