JCPSS Anonymization and Derivative Tables

June 30, 2017

Table Overview

This document produces three tables from an anonymized version of the jcpss and will be further amended to output a public-use file. Those tables are:

- `jcpss_statewide.csv` offense level and sustained offense level, group offense level and sustained group offense level, and demographics by year statewide (no counties).
- `jcpss_county.csv` offense level and sustained offense level, offense category and sustained offense category, and disposition totals by year, county, and demographics

Sustained data only provided when disposition=wardship. Tables are derived from the same source so that what has been suppressed in one is also suppressed in another.

Disclosure Overview

From the jcpss, one can learn:

- whether or not a person has been arrested/referred
- knowing that a person has been arrested/referred, the arrest offense
- knowing that a person has been arrested/referred, the severity of the arrest offense
- knowing that a person has been found to have commited a crime, the sustained offense level and information about the sustained offense.
- knowing that a person has been found to have commited a crime and placed under wardship, the type of wardship
- knowing that an arrest has been made for a specific offense, the demographics of the person arrested
- knowing that someone was found to have committed a crime, the demographics of that person

In general we hold that learning the outcome of a crime when nearly all other information is available does not constitute disclosure - to know a person's complete demographics and disposition or demographics and referral offense is a rare and highly specific situation.

Anonymization Techniques

We aggregate race and age into larger categories. When the groups corresponding to demographic key variables are deemed at risk, information is selectively suppressed. This act of suppression allows individuals from that category and other categories to blend together, so that one arrest instance can stand in for two or more demographic keys. When suppression of information is necessary, we enforce the order:

1. gender
2. age group
3. race

This means that, if for a group it is determined that their presence in the dataset could lead to disclosure, we place a priority on discarding the gender of individuals in it and in similar categories. Considering the suppression of subsequent variables comes at a higher cost, and thus occurs less frequently.

Risk Measures

The possible disclosures that can occur lead to the following measures of risk:

- k-anonymity - the total number of people sharing a given demographic key (e.g., “female”/“10 to 17”/“Asian or Pac Islander”)
- l-diversity - for a given demographic key, the number of different associated outcomes (e.g., for the class of personal above, how many different arrest codes were used)
- p-diversity - for a given demographic key, the percentage of outcomes in a given class (e.g., % of arrests that are for misdemeanors) (note, the term p-diversity is not standard)

At-Risk Subpopulations

Due to external sources of information, we provide additional protection to people in these classes:

- Non-wardship dispositions (including the equivalent of released at any stage)

Protections Afforded

In light of the above, we make the anonymize in the following fashion:

1. protect that referred ($k = 5$) in the total population (county)
2. protect that referred ($k = 5$) in the non-wardship population (county)
3. diversity ($l = 3$) in offense codes in the total population (county)
4. diversity ($l = 3$) in offense codes in the non-wardship population (county)
5. protect offense severity (misd % of total $p = 0.2$) in total population (county)
6. protect offense severity (misd % of total $p = 0.2$) in non-wardship population (county)
7. violent felony diversity ($l = 3$) in demo-key at in total pop (county)

Replicability

As the anonymization uses random suppressions, making the procedure replicable in the future requires

- no changing the results of the past, or when doing so in a way that does not change which data gets suppressed
- setting a random seed

Because the data comes in sequentially it would be ideal to run on individual years at one time, possibly with random seeds for each.

Protecting That Referred or Found to have Committed a Crime

It is possible to determine that a person has been arrested, referred, or found to have committed a crime if a count/row is shown matching his or her demographics, and those demographics are rare in the population. When the total for any combination of year/jurisdiction/gender/race/age group is less than $k = 5$, we enact suppression.

Protecting Offense Diversity

If ever it is the case that for a demographic key all of the referrals are of a similar nature (or have the same offense), then one can infer the offense for a referral without knowing the specific instance. Consequently, we guarantee that there are at least $l = 3$ different kinds of offenses for every demographic key.