# CS 208 - Applied Privacy for Data Science
# Homework 1

Jason Huang
Spring 2019 - Harvard University

## Problem 1

The dataset was loaded into `R`, where preliminary data exploration took place. Most notably, there are 25766 entries and 18 variables in the dataset, with the variables being:

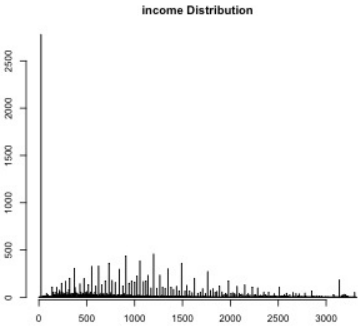
`state, puma, sex, age, educ, income, latino, black, asian, married, divorced, uscitizen, children, disability, militaryservice, employed, englishability, fips`

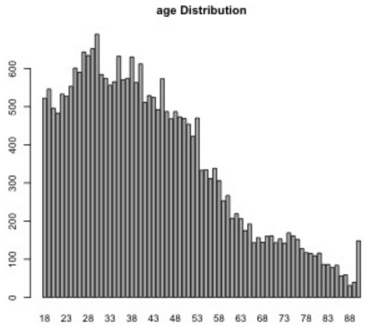
The naive and most effective starting point is tallying the unique values for each of these variables. Two of the variables, `state` and `fips`, have the same value for all rows and therefore will not be considered at all.

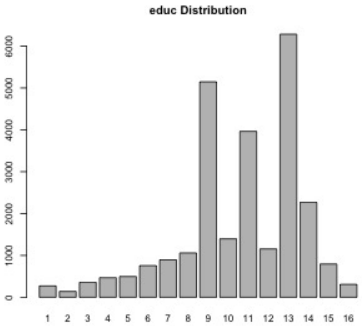
Now examine the most identifying variables (i.e. the variables with most unique values in the dataset, as determined by taking the length of its count table in `R`).

| `income` | `age` | `educ` | `puma` |
|---|---|---|---|
| 2763 | 73 | 16 | 7 |

All the other variables are binary variables (can only have two distinct values). `puma` is implicitly included (since the goal is to uniquely identify individuals *within a PUMA region*) and will be accounted for in the last step. However, these other three variables should be explored, particularly by examining their distributions, as plotted below.







The goal now is to quantify exactly how identifiable each variable is. Estimates for the probability of collision (of another individual having the exact same value) for each variable based on the largest bin (i.e. worst case) are made below. Actual probabilities for collision can also be calculated by the following formula:

$$p_{collision} = p(X_1 = X_2) = \sum_x p(X_1 = x)p(X_2 = x) = \sum_x \left(\frac{\text{count}(x)}{25766}\right)^2$$

where $x$ is to take on all possible values for that variable.

`income` is the most identifiable variable but also raises an issue when examining the distribution histogram: a large number of individuals have zero income, and are therefore substantially more difficult to uniquely identify. Apart from this unique case, though, no bin has more than 500 individuals, which corresponds to approximately $500/25766 \approx 2\%$ of collision with another individual (actual: 0.01555964). `age` is distributed a lot better, with the worst case bin having no more than 800 individuals with the same age. When considering the overall size of 25766, that means that there is *at most* a $800/25766 \approx 3\%$ chance of having the same age as a randomly chosen person in the dataset (actual: 0.01831928). Lastly, for `educ`, with up to 6500 in the same bin, the probability of having the same education level as someone else would be $6500/25766 \approx 25\%$ (actual: 0.1416453).

Percentage estimates where given above to provide intuition and also as a sanity check, but the actual values calculated will be the ones used from here on.

When taking these three variables together, the probability of the collision is the intersection of all three variables colliding, which is the product of the three probabilities. This gives an overall probability of colliding to be $0.01555964 \times 0.01831928 \times 0.1416453 \approx \mathbf{0.00004} = p_{collision}$.

Now, the geographic identifier of PUMA regions needs to be taken into account. Here is a summary of counts for each PUMA region.

| **PUMA** | 1101 | 1102 | 1103 | 1104 | 1105 | 1106 | 1107 |
|---|---|---|---|---|---|---|---|
| **Count** | 3215 | 5736 | 3728 | 3740 | 3128 | 3236 | 2983 |

These are very roughly equal (i.e. on the same order of magnitude), and these do appear to be 5% samples (with full populations of 50000-120000 in each PUMA, which is appropriate). Assume the average PUMA region to therefore be 1/7 of the total dataset: $(1/7) \times 20 \times 25766 = 73617$.

To finally determine the percentage of individuals $p(U)$ within a PUMA that can be uniquely identified by the aforementioned variables, this calculation simply involves the probability of no collision with anyone in that region of size $n$, or:

$$p_{unique} = (1 - p_{collision})^n$$

In this situation, $p_{collision} = 0.00004, n = 73617$ as calculated above, so:

$$p_{unique} = 0.051183922689063$$

$\therefore$ With the three variables of `income`, `age`, `educ`, approximately $\boxed{5\%}$ of the population within a single PUMA can be identified.

Better reconstruction results can be achieved using more variables. The only remaining variables are binary, which should have roughly 50% chance of collision (though in actuality it will be higher due to uneven distribution). The lowest probabilities of collision are for `sex` (0.5014973) and `married` (0.5046546), two common and relatively evenly distributed binary indicators. When factoring these in, then:

$$p_{collision} = 0.00001 \implies p_{unique} = 0.471312198919265$$

$\therefore$ With the five variables of `income`, `age`, `educ`, `sex`, `married`, approximately $\boxed{47\%}$ of the population within a single PUMA can be identified.

While it is more difficult and less likely to orchestrate a reconstruction attack with large numbers of variables, out of interest and completeness, here are some further results. With a sixth variable of `black`, 68% of the population can be uniquely identified. With a seventh added variable of `employed`, 81% of the population can be uniquely identified. The remaining binary variables have sharply decreasing utility due to their uneven distribution (i.e. almost all individuals have the same value and therefore it is not very helpful in identifying someone).

Finally, some small disclaimers. The results above are all approximate and derived from rough back-of-the-envelope calculations. They are also assuming that these variables are available in external sources for cross-referencing, which may present practical obstacles that may or may not be easy to handle. For example, a specific education encoding is used here. Other government databases may use the same schema, making cross-referencing trivial. Furthermore, knowing an individual's exact education level may also be sufficient to map it to one of the factors in this dataset's `educ` schema. However, if a different, less granular schema was used (i.e. only 5 levels instead of 16), then cross-referencing may not really be possible. Different levels of granularity would be a particularly common issue for `income`, with added concerns such as rounding. It may also affect `age`, though probably to a much lesser extent. Binary variables should otherwise be less problematic in this regard.

## Problem 2

## Problem 3

## Problem 4