Generating Synthetic Data for Data Privacy with R

-Claire McKay Bowen, PhD – Urban Institute (cbowen@urban.org)

Introduction and Motivation

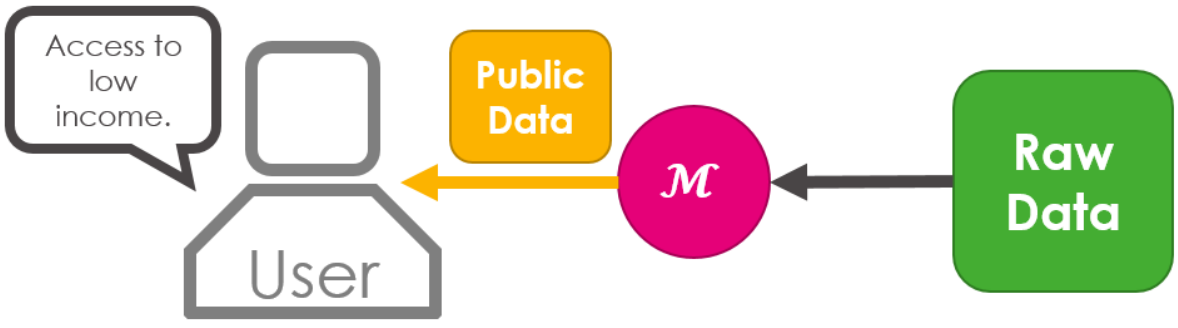
At what point does the sacrifice to our personal information outweigh the public good?

If public policymakers had access to our personal and confidential data, they could make more evidence-based, data-informed decisions that could accelerate economic recovery and improve COVID-19 vaccine distribution. However, access to personal data comes at a steep privacy cost for contributors, especially underrepresented groups. Revealing too much location information places people at risk such as empowering stalkers to track people more easily, but too little personal, location information will severely hinder the effectiveness of contact tracing.

Why Use Synthetic Data Generation?

Statistical disclosure control (SDC), or limitation (SDL), is a field of study that aims to develop methods for releasing high-quality data products while preserving the confidentiality of sensitive data. These techniques have existed within statistics and the social sciences since the mid-twentieth century, and they seek to balance risk against the benefit to society, also known as the utility of the data. While this field has existed for some time, over the past two decades the data landscape has dramatically changed. Data adversaries (also referred to as intruders or attackers) can more easily reconstruct data sets and identify individuals from supposedly anonymized data with the advances in modern information infrastructure and computational power. One of the more recent innovations in SDC is a technique known as synthetic data generation, and it has become a leading method for releasing publicly available data that can be used for numerous different analyses

What is Synthetic Data?

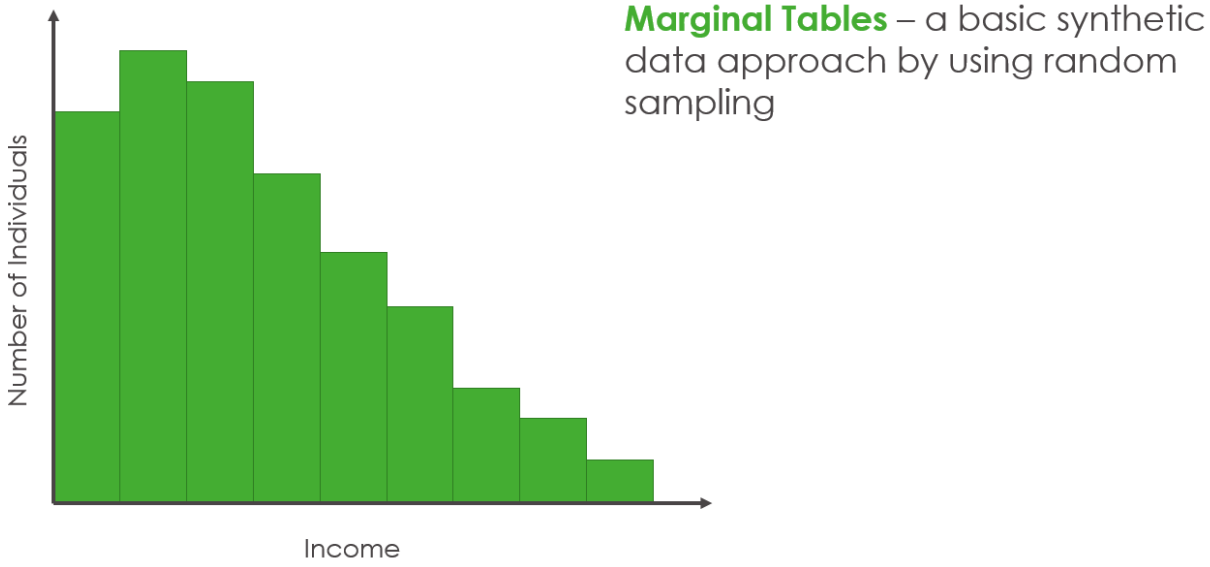


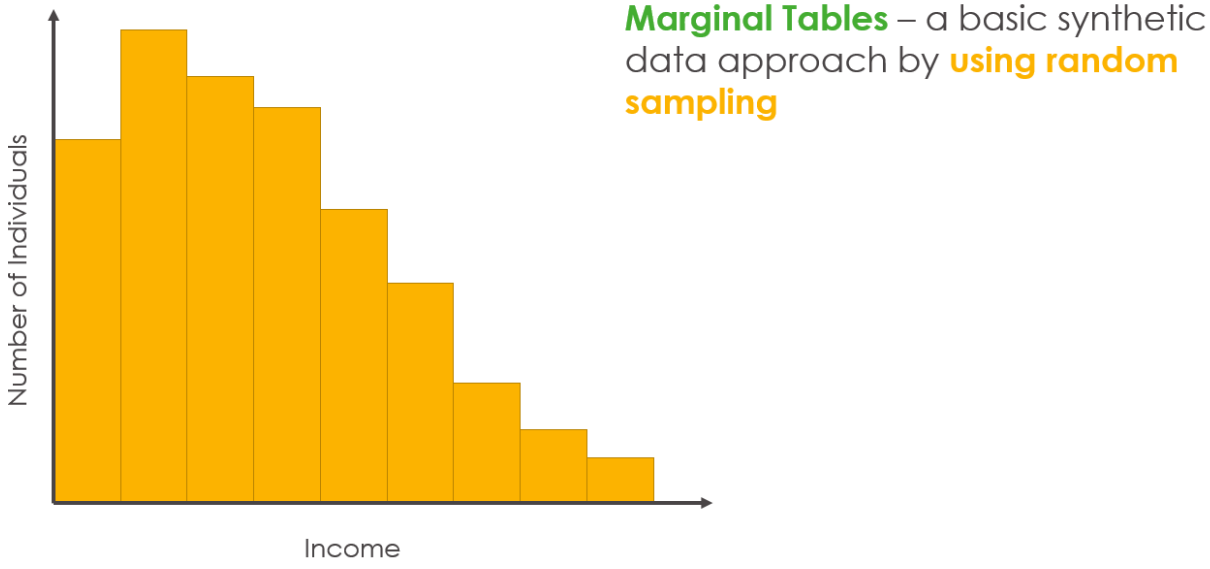
Synthetic data is typically generated from probability distributions that are identified as being representative of the original data in order to preserve as many of the original data

relationships as possible. There are many ways to generate synthetic data, which can be roughly grouped as non-parametric (synthetic data generation based from an empirical distribution) or parametric (synthetic data generation based from a parametric distribution or generative model). We will cover the most basic non-parametric and parametric approaches.

Marginal Tables

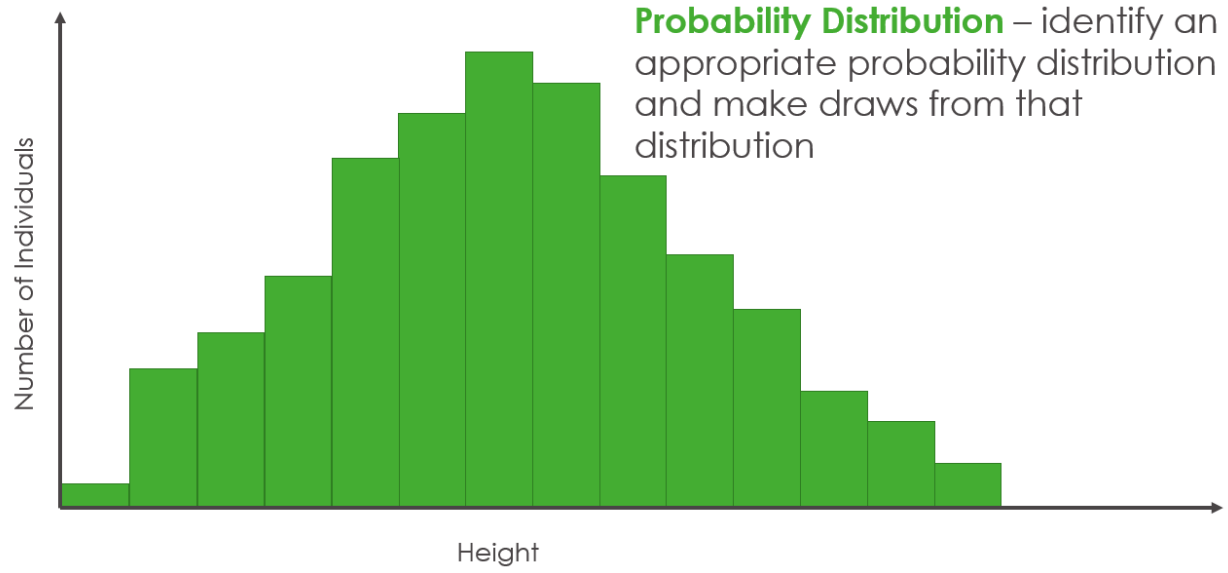
The most basic way to generate synthetic data is from marginal tables of the data, where the data is changed based on the random sampling. For instance, you can categorize your data into groups or bins, calculate the proportion, and randomly sample new values (synthetic values) based on the proportion.

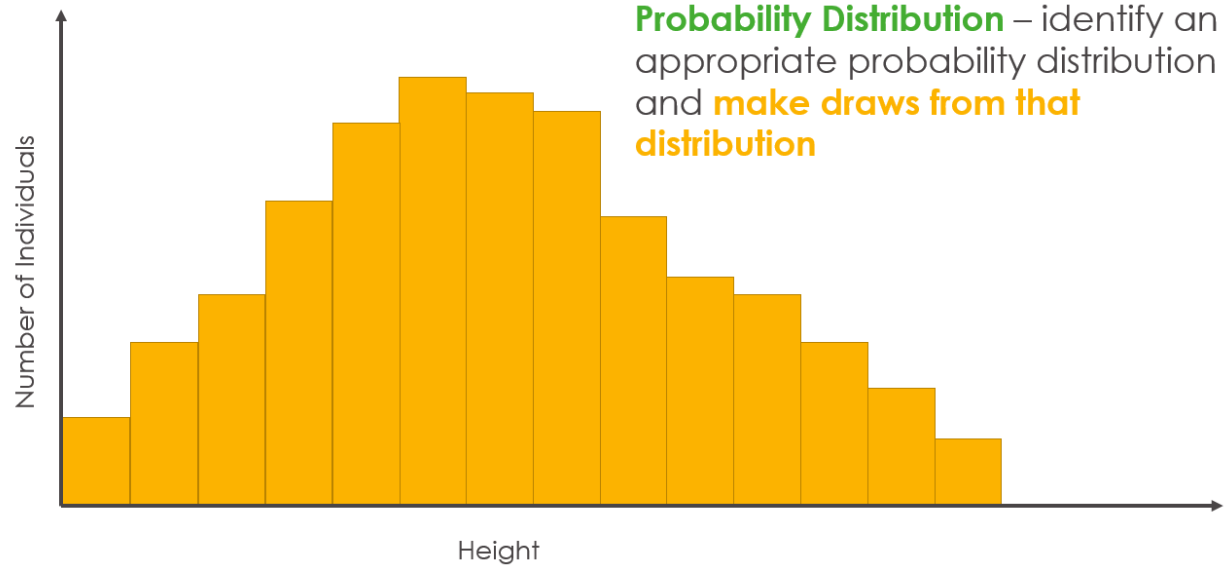




Probability Distributions

Another simple way to generate synthetic data is from an appropriate probability distribution. First, identify an appropriate distribution, then make draws from that distribution based on the parameters or sufficient statistics on the original data. E.g. a normal distribution requires knowing the mean and variance of the data. Once you identify those values, make random draws from a normal distribution using the original data's mean and variance.





Generating Synthetic Data in R

There are many ways to generate synthetic data, but we will focus on the most basic approaches:

- marginal tables
- probability distributions
- regression models

As an example, we'll be using a subset of the Star Wars data in the R package tidyverse to demonstrate each method. The rows of the data represent a Star Wars character with various information such as *mass* and *species*. We'll focus on these variables for our examples.

Marginal Tables

The most basic way to generate synthetic data is from marginal tables of the data. Suppose we want to generate a synthetic data on the number of humans, droids, and aliens. Below is the marginal table showing the true proportion of humans, droids, and aliens in the data.

Species (n = 57)		
Human	**Droid**	**Alien**
22	4	31
38.6%	7.0%	54.4%

Instead of releasing the original values, we can make 57 weighted samples based on the true data proportions. i.e. for each new synthetic observation we create, there is a 38.6%, 7.0%, and 54.4% that the sample will be human, droid, and alien, respectively.

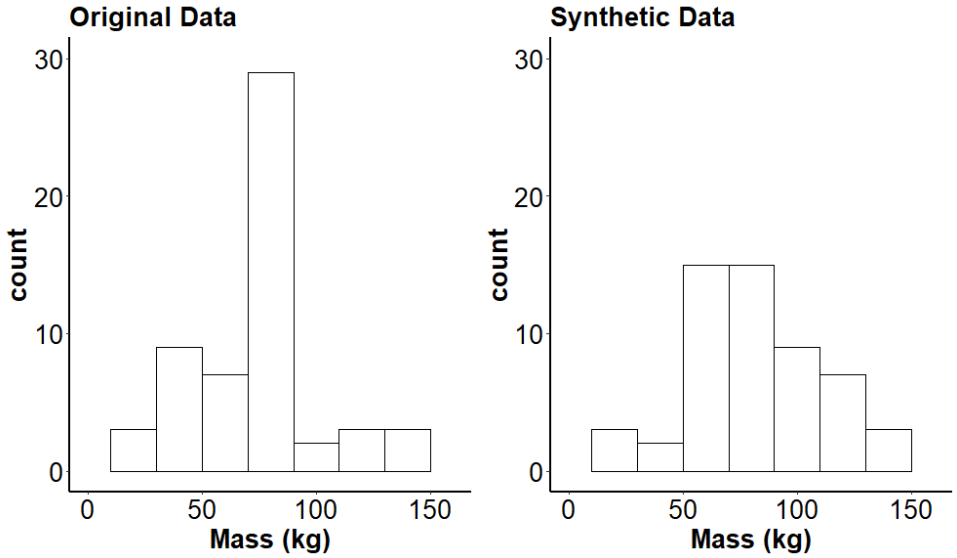
The following table is the synthetic data generated, which used R function *sample()* and a seed of 42.

Species (n = 57)		
Human	**Droid**	**Alien**
23	9	25
43.9%	15.8%	40.3%

We see the values are close, but not exactly the same as the original data. Keep in mind that data privacy and data utility (the usefulness of the data) are opposed to one another. If we want more privacy, we have to lower the utility and vice versa. With our example, there is a layer of randomness when drawing our new values. In addition, we can create synthetic data based on higher dimensional marginal tables (e.g. 2-way or 3-way marginal).

Probability Distributions

Another simple way to generate synthetic data is from a specific probability distribution. From our Star Wars data set, we want to create synthetic data of the characters' mass.



Histogram on the left is the original data and the histogram on the right is the synthetic data. The data is the Star Wars characters' mass.

Based on the figure above, the data resembles roughly a normal distribution. We then gather the sufficient statistics (the required statistic or parameter for the particular model) for a normal distribution, which are the sample mean and standard deviation. For our data, the sample mean and standard deviation are 76.1 and 29.3, respectively. We draw new observations from a normal distribution with mean 76.1 and standard deviation of 29.3, using the R function *rnorm()* and a seed of 42. The figure above shows our a synthetic data set, which has a sample mean and standard deviation 76.6 and 32.5, respectively. Normal distribution is one of many distributions we can use to create synthetic data such as multinomial, Poisson, and multivariate distributions.

Note: Sometimes values you draw from your probability distribution will not make any sense for the application. For instance, when generating the synthetic data of Star Wars characters' mass, one of the values came out negative. When this occurs, we need to perform post-processing techniques such as bounding the values to ensure the synthetic data makes sense for the particular application.

Regression Model

We can also generate synthetic data from a specified model. Continuing our Star Wars example, suppose we want to use a linear regression model to generate synthetic data to preserve the relationship between the mass and species. The original data has the linear regression model as

$$mass = \beta_0 + \beta_1 Droid + \beta_2 Human + \eta$$

where $\beta_0 = 72.1,\ \beta_1 = -2.353, \beta_2 = 10.7$ and η is some random noise. This random noise is typically generated from a normal distribution with a mean of 0. The standard deviation can be adjusted to how much noise we think either the original data might have or how much privacy (more noise leads to more privacy, but less accuracy) we want to add. For our example, let's use a standard deviation of 29.3 given this value is the sample standard deviation of mass.

To generate the synthetic data, we first need to create the synthetic values of the species, which is the dependent variable in our linear regression model. We will use the synthetic data we created from the *Marginal Tables* section. We then input those values to generate the synthetic values of the mass, using linear regression model as described above. The R function *rnorm()* and a seed of 42 was used for calculating η. The synthetic data has a sample mean and standard deviation of 76.9 and 32.6, respectively, for the mass.