Probabilistic Population Synthesis for Decision-Making

Christopher Grubb Dr. David Higdon Dr. Leanna House

Statistics Department
Virginia Tech

12 August, 2021

Outline

1. Introduction
2. Simple Example
3. Theory
4. Less Simple Example
5. Future Work

Section 1

Introduction

Introduction

What do we mean by population synthesis?

- Create a posterior over finite (but potentially large) populations, and sample from it
- Propagate uncertainty from sample(s) into populations

Why would we want to do this?

- Decision makers are typically not statisticians
- With complex data, it is very difficult to create a population that satisfies certain criteria

Section 2

Simple Example

Simple Example (Introduction)

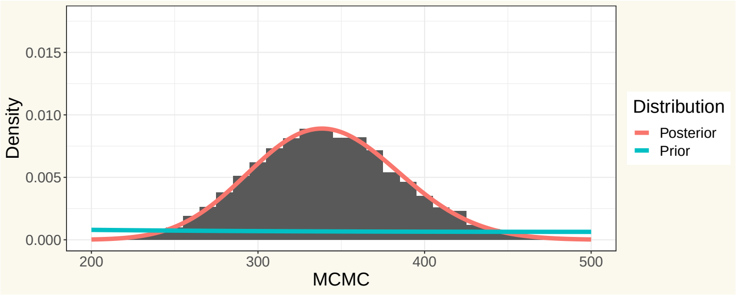
Suppose our variable of interest is categorical with only 2 categories. We observe a sample of size $n = 100$, with category tallies of 34 and 66, respectively.

Without any population synthesis, what would we do?

- Choice of distribution
    - Binomial
    - Hypergeometric
- Approach
    - Classical – Exact confidence interval or Normal approximation
    - Bayesian – Construct a posterior

Simple Example (Comparison)

If we compare a synthetic population approach to a traditional Bayesian approach – using the same distributional assumptions and priors – inference on parameters should be identical.



Comparison of Analytic Prior/Posterior to Synthetic Populations for First Category

Simple Example (Process)

How do we actually create the populations? For simple examples like this, importance sampling works very well; however, as data complexity increases, the effectiveness decreases rapidly. Our approach is to jitter populations and accept or reject using the posterior.

- Basically any variant of MCMC will work
- Need to be careful – depending on how you jitter, a correction factor may be needed

Section 3

Theory

Posterior Form

Treating the population **Y** as a random variable, we can use Bayes' theorem to derive the posterior distribution for populations, in order to sample many possible populations.

- Sample distribution – How likely is our sample given a specific population?
- Population distribution – How likely is a specific population given the parameters that control the distribution of populations?
- Prior distribution – What do we think are reasonable values for said parameters?

$$\begin{aligned} f(\theta, \mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}) &= \frac{f(\mathbf{X} = \mathbf{x} | \theta, \mathbf{Y} = \mathbf{y}) f(\theta, \mathbf{Y} = \mathbf{y})}{f(\mathbf{X} = \mathbf{x})} \\ &\propto f(\mathbf{X} = \mathbf{x} | \mathbf{Y} = \mathbf{y}, \theta) f(\mathbf{Y} = \mathbf{y} | \theta) \pi(\theta) \end{aligned}$$

Depending on the form of **X**, the choices we make for these three distributions can have a very large impact on what the resulting populations look like.

Choosing a sample distribution

This may *seem* like the easiest choice out of the three, but some unintended complications can occur depending what choice we make.

For example, suppose the variable of interest is numeric and we observe $\mathbf{x} = \{0.2, 0.5, 0.9\}$. Consider the following populations,

$$\mathbf{y}_1 = \{0.2, 0.4, 0.5, 0.7, 0.9\} \qquad \mathbf{y}_2 = \{0.2, 0.3, 0.5, 0.8, 0.9\}$$

For these two populations, $f(\mathbf{x}|\mathbf{y}_1) = f(\mathbf{x}|\mathbf{y}_2) = \binom{5}{3}^{-1}$, if we use a discrete sample distribution that assigns equal probability to all values in the population. In fact, this will always be the case if our resolution on $\mathbf{x}$ is high enough to prevent the possibility of duplicates.

Luckily, we typically don't have high resolution data. Many variables that are continuous (e.g., age, salary) are often observed on discrete scales (e.g., years, thousands of dollars), and are sometimes even binned into categories.

Choosing a sample distribution (cont'd)

For categorical data, this is a much easier choice. In theory, we would like to use a (potentially multivariate) Hypergeometric distribution, as this correctly explains the process of sampling that was used, assuming $\mathbf{x}$ came from sampling without replacement.

- $\mathbf{x}$ sampled with replacement → use Binomial or Multinomial distribution instead
- Might have to use Binomial or Multinomial anyway, because of computational difficulties

Choosing a population distribution

Likely the most difficult and potentially most influential choice, it is critical that whatever distribution is picked can actually create populations similar to the one being examined.

- This is not as easy as it sounds
- Similar can mean many things
- May need to be rather complex

Choosing a prior distribution

Depending on the type of variable being created, as well as the previous two choices, this can range from relatively straightforward to very confusing.

- Probably a good idea to use a range of values via a hyperprior
- Imagine **x** is categorical and we use a Hypergeometric distribution. What is our parameter?

Section 4

Less Simple Example

Less Simple Example (Introduction)

This time, suppose the variable of interest is numeric with a range of $(0, 1)$, however we only observe counts within bins of $(0, 0.5)$ and $(0.5, 1)$, totaling 34 and 66 respectively.

We want to create populations with numeric values, while respecting the likelihood of sampling our observed binned data.

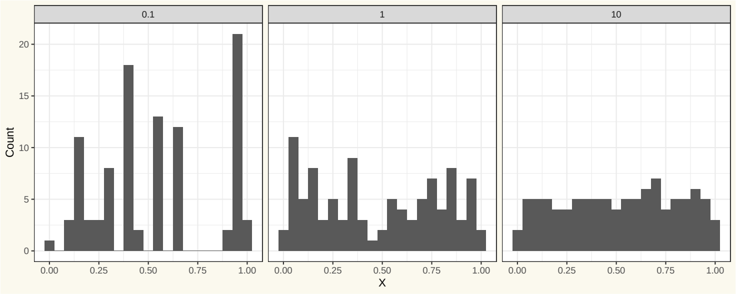
A few caveats:

- We want to choose a population distribution that make sense (probably requires expert knowledge or historical data)
- Have to make sure our sample is actually plausible, given whatever population distribution we pick, or samplers will get stuck

Less Simple Example (Choosing Population Distribution)

Since our data is in the region (0, 1), we can exploit the fact that the *spacings* of $\{0, x, 1\}$, where x is a sorted vector (i.e., a sorted population), sum to 1.

Thus, we can use a *Dirichlet* distribution, which lets us control how clustered our population is.



Sample draws with alpha of 0.1 (left), 1 (center), and 10 (right)

Section 5

Future Work

Future Work and Goals

Much work still needs to be done before we can tackle our initial chosen application.

- Extending to other types of data sources
- Combining information across various data sources
  - Different types of data; simple random samples, marginal tables, histograms, etc.
  - Different resolutions and scales

Once finished, the goal is for synthetic populations to aid decision-making for public policy.

- Potential infrastructure changes
- Public policy