Weighted likelihood to recover unbiased population estimates from a weighted sample.

WorldPop, University of Southampton

11 December 2020

Contents

1 Introduction

Population estimates that are accurate and up-to-date are critical for government planning and development projects in low and middle income countries where recent census results may not be available. Bottom-up population models can be used to produce gridded population estimates from population enumeration data from a representative sample of locations across the study area (i.e. nationally). A stratified random sample of locations would be appropriate to recover unbiased estimates of average population density (e.g. people per hectare). Household surveys are routinely conducted with national coverage. These surveys often use PPS sampling designs (i.e. probability proportional to size) in which locations are selected for the sample based on the number of people or households in that location. The PPS sampling design is intended to selected a sample of locations that represent a random sample of individual people. This is ideal when using the data to estimate population

characteristics such as fertility, mortality, poverty, etc. But, because the weighted sample is based on population size, it will produce a biased result if used to estimate population size.

Our objectives are to:

1. Demonstrate that a population-weighted sampling design results in biased estimates of population density,
2. Demonstrate that population totals are sensitive to this bias when estimated population densities are applied across large areas,
3. Propose a Bayesian weighted-likelihood approach to account for the bias in weighted samples, and
4. Demonstrate that the weighted-likelihood model recovers unbiased estimates of population densities and population totals across large areas.

In this paper, we will explore the weighted likelihood approach using simulated data to represent samples that are random, weighted, and mixed. We will stratify the samples based on urban (higher density) and rural (lower density) settlement types. For simplicity, we will ignore spatial effects (i.e. relative locations of samples) and effects of geospatial covariates.

This analysis is intended as a theoretical foundation to support future empirical studies.

2 Methods

2.1 Data

Populations were simulated using 10,000 random draws from log-normal distributions to represent a census of a population with people in 10,000 different locations (i.e. think of each location as a one hectare grid square). Population densities in urban areas were simulated using a median population density of 750 people per hectare with a standard deviation of 500 to parameterise the log-normal distribution. Populations in rural areas were simulated using a median of 100 and a standard deviation of 250. Half of the locations were urban and half were rural.

These simulated populations were sampled using random sampling, population-weighted sampling, or both. Samples comprised a total of 1,000 locations that were enumerated in each simulated population (i.e. a 10% sample from the total population of 10,000 locations). In simulations that included both types of sampling, half of the samples were random and half were population-weighted. Stratified sampling was implemented so that an equal number of samples were collected from urban and rural areas.

Population-weighted sampling used sample weights w_j to define the probabilities for each location j being sampled:

$$w_j = \frac{N_j}{\sum_{j=1}^{J} N_j} \tag{1}$$

where N_j is the population size at location j, and J is the total number of simulated locations (*i.e.* $J = 10,000$).

Four scenarios were simulated:

1. Stratified random sampling with an unweighted model,
2. Stratified weighted sampling with an unweighted model,
3. Stratified weighted sampling with a weighted-likelihood model, and
4. Stratified random and weighted sampling with a weighted-likelihood model.

2.2 Analysis

We used a log-normal weighted-likelihood model to represent the distribution of population densities among locations:

$$y_i \sim LogNormal(log(\mu_t), \tau_{t,i}) \tag{2}$$

where y_i is the observed number of people at sampled location i, and mu_t is the median population size for the settlement type t to which location i belongs. Sample weights w_i were used to calculate model weights v_i to account for the sampling bias:

$$v_i = \frac{w_i^{-1}}{\sum_{i=1}^{I} w_i^{-1}} \tag{3}$$

where i is a sampled location and I is the total number of sampled locations (i.e. $I = 1,000$). Model weights are the inverse of sample weights rescaled to sum to one among all sampled locations.

τ_i is a location-specific estimate of precision (i.e. the inverse of variance; on the log scale) that is dependent on the model weights v_i and a global estimate of precision $\bar{\tau}_t$:

$$\tau_{t,i} = \bar{\tau}_t v_i \tag{4}$$

We defined the prior distributions of μ_t and $\bar{\tau}_t$ using uninformative uniform distributions:

$$\mu_t \sim Uniform(0, 5000) \bar{\tau}_t \sim Uniform(0, 3) \tag{5}$$

$\bar{\tau}_t$ cannot be used to predict y in new locations where model weights are not available. So, we used a weighted average of $\tau_{t,i}$ to derive an estimate of precision for each settlement type that can be used for predictions in new locations:

$$\theta_t = (\frac{\sum_{i=1}^{I_t} v_i \sqrt{\tau_{t,i}^{-1}}}{\sum_{i=1}^{I_t} v_i})^{-2} \tag{6}$$

where I_t is the total number of samples from settlement type t. The weighted average was calculated based on the standard deviation $\sqrt{\tau_{t,i}^{-1}}$ rather than precision $\tau_{t,i}$ because ?????. This derived estimate of precision was used for making posterior predictions without the need for model weights:

$$\hat{y}_t \sim LogNormal(log(\mu_t), \theta_t) \tag{7}$$

3 Results

This section should describe the results of the analysis.

Links to the full results should be provided when possible (e.g. WOPR, woprVision, DOI link, etc).

Sufficient diagnostics should be included to demonstrate that the results are fit for the intended uses.

4 Discussion

This section should describe the known limitations and assumptions associated with the results.

This section should recommend next steps that the authors see as most promising to build on the work.

Contributing

This section should describe funding sources, authors, contributers, external collaborators, etc. and their contributions.

Suggested Citation

Leasure DR, Dooley CA. 2020. Weighted likelihood to recover unbiased population estimates from a weighted sample. WorldPop, University of Southampton. doi:10.5258/SOTON/WP00XXX

License

References

{APPENDIX} Appendix A: Supplementary Files

Model Data

This sub-section describes all input data published as supplementary information.

Model Code

This sub-section describes all code published as supplementary information.