Singleton Error Experiments

Andee Kaplan

This is a set of small experiments with the goal of determining how errors in the contingency table result in errors in the population size estimate. We are interested ultimately in how errors from record linkage are propagated through a capture-recapture method and result in biased results. Specifically we will explore in two questions.

1. Which type of errors are worse–errors in the number of records counted in only one database or errors in the number of records counted in multiple databases?
2. In a record linkage setting, at what point to singleton errors result in capture-recapture estimations of the population size that are no longer useful (coverage $< 95\%$)

1 Experiment Setup and Data

To set up our experiments, we start with a population of size $M = 1000$ and select with replacement to create $D = 5$ databases according to the inclusion probabilities in Table 1. The inclusion probabilities were randomly generated from a $Beta(a_0, b_0)$ distribution. This results in $n(0) = 262$ records never having been recorded in any database, which is the value that we will estimate using a capture-recapture procedure.

Table 1: Inclusion probabilities for each of the databases.

	1	2	3	4	5
inclusion	0.1	0.2711	0.4202	0.0307	0.2953

To perform capture-recapture, I used the nonparametric latent class model (NPLCM) of Manrique-Vallier (2016) with default priors and hyperparameters. The posterior distribution of M is shown in Figure 1, with the true value of $M = 1000$ shown as a vertical line. The posterior contains the true population size nicely, indicating that in the absence of errors, the NPLCM model with default priors and hyperparameters is working adequately.

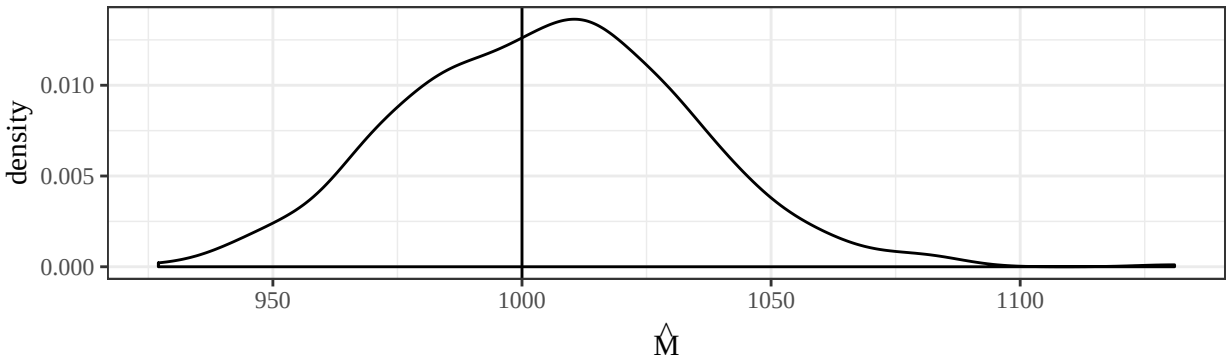

Figure 1: The posterior distribution of M as generated by the NPLCM, with the true value of $M = 1000$ shown as a vertical line. The posterior contains the true population size nicely, indicating that in the absense of errors, the NPLCM model with default priors and hyperparameters is working adequately.

2 Error Types

The first experiment we look at aims to answer the question, "Which type of errors are worse–errors in the number of records counted in only one database or errors in the number of records counted in multiple databases?"

In order to answer this, we will add or remove a fixed number, ρ, of records from random buckets in the 2^D contingency table of the captures. We will add or remove these records from the two types of buckets, singleton or multiple inclusions, and compare the results of running the NPLCM for capture-recapture to estimate M. We look at multiple values of ρ as different percentages of singletons from 5% to 50%. The result of this process is shown in Figure 2. Because the process of introducing error is random, we repeat it 100 times and look at the resulting 95% credible intervals for M.

Errors that are introduced to the singleton buckets in the contingency table have a much greater effect on the estimate of M than do errors introduced to the multiple inclusion buckets. This is shown by how many intervals contain the true population size (black) versus those that do not (red). For both adding and removing singletons, the estimates of M are drastically different from reality between 10% and 15% error, whereas for the multiples, an effect is not seen until much later. This result indicates that errors in the singleton buckets have a much higher impact on the capture-recapture method than errors in the multiple inclusions. This could help us tailor a record linkage method to work well for capture-recapture.

3 Record Linkage Setting

The experiments in Section 2, while informative, are not realistic in terms of how errors occur in record linkage. In a record linkage procedure, record will either be classified as singletons (no linkage) or multiple inclusions (linked). The result is that errors in singleton classification will necessarily also result in errors in multiple inclusions. However, it is possible for a record linkage procedure to correctly classify all the singletons, and still have a high rate of overall error due to getting the actual linkage wrong.

In this second set of experiments, we are interested in exploring how much singleton error from the record linkage procedure can the NPLCM handle and still produce meaningful results. To test this, we once again add or remove singletons according to ρ, various proportions of singletons from 5% to 50%, but now we also add or remove those records from randomly selected multiple inclusions buckets (in accordance with a realistic record linkage scenario). The resulting 95% credible intervals are shown in Figures 3 and 4 with intervals that contain the truth $M = 1000$ in black and those that do not in red.

Somewhere between 5% and 10% of singleton error leads to an unacceptable amount of error in the estimation of M (coverage $< 95\%$). We zoom in on these values in Figure 4 to pinpoint how the errors are affecting coverage in the posterior estimate of M. This zoomed in view shows that depending on if singletons are added (over linked) or removed (under linked), the amount of acceptable error in singletons from the record linkage procedure is around 5.5% or 7.5%, respectively. Tables 2 and 3 display these results numerically.

Table 2: Coverage of 95% credible intervals for M from using NPLCM capture-recapture after different error levels are passed from the record linkage procedure. The amount of acceptable error in singletons from the record linkage procedure is somewhere between 5% and 10%.

type	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
add	1	0.00	0	0	0	0	0	0	0	0
remove	1	0.37	0	0	0	0	0	0	0	0

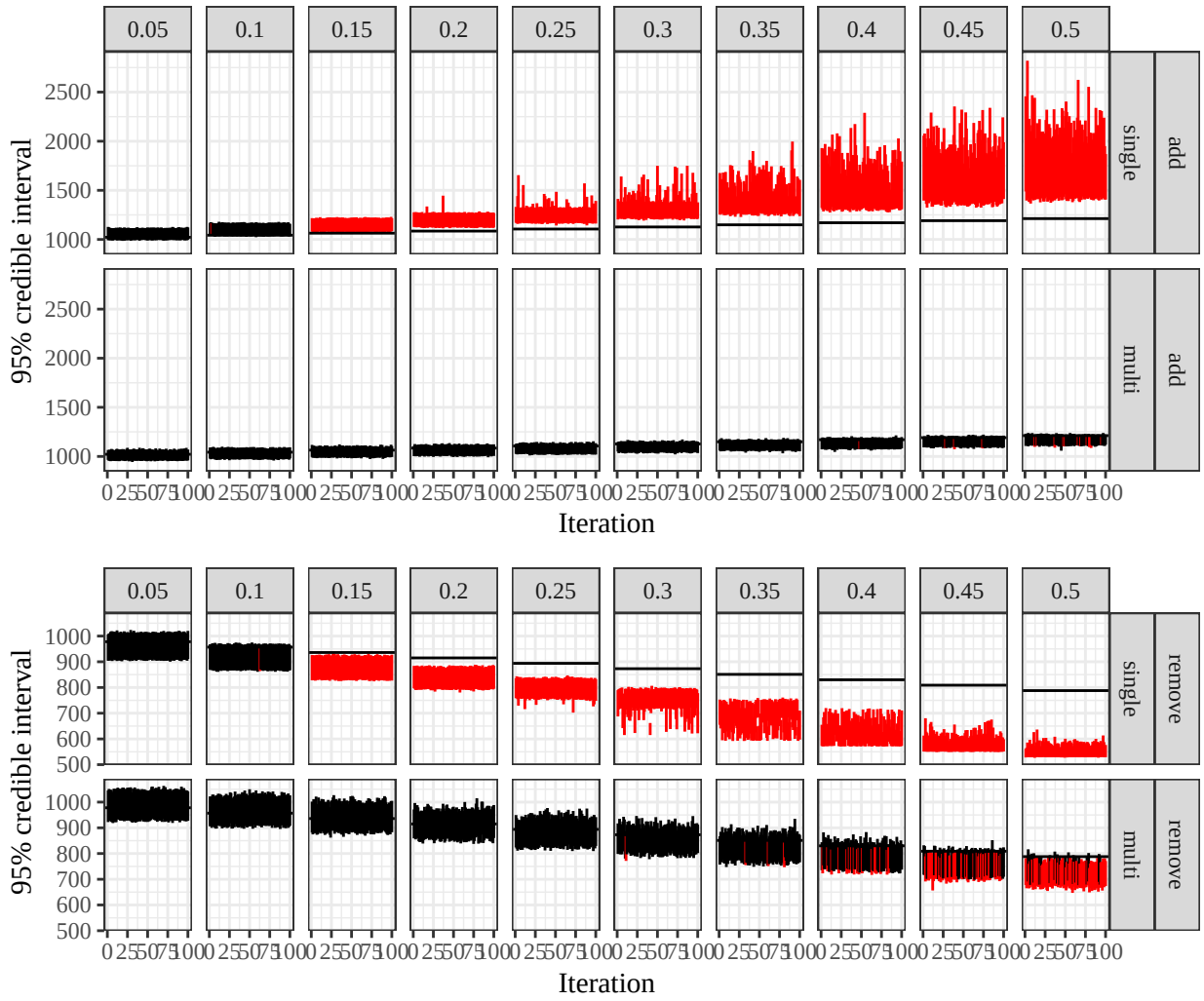

Figure 2: The results of running the NPLCM for capture-recapture to estimate M after adding and removing equal numbers of singletons or multiple inclusions from the contingency table. Because the process of introducing error is random, we repeat it100times and look at the resulting 95% credible intervals for M. Errors that are introduced to the singleton buckets in the contingency table have a much greater effect on the estimate of M than do errors introduced to the multiple inclusion buckets. This is shown by how many intervals contain the true population size (black) versus those that do not (red). For both adding and removing singletons, the estimates of M are drastically different from reality between 10% and 15% error, whereas for the multiples, an effect is not seen until much later.

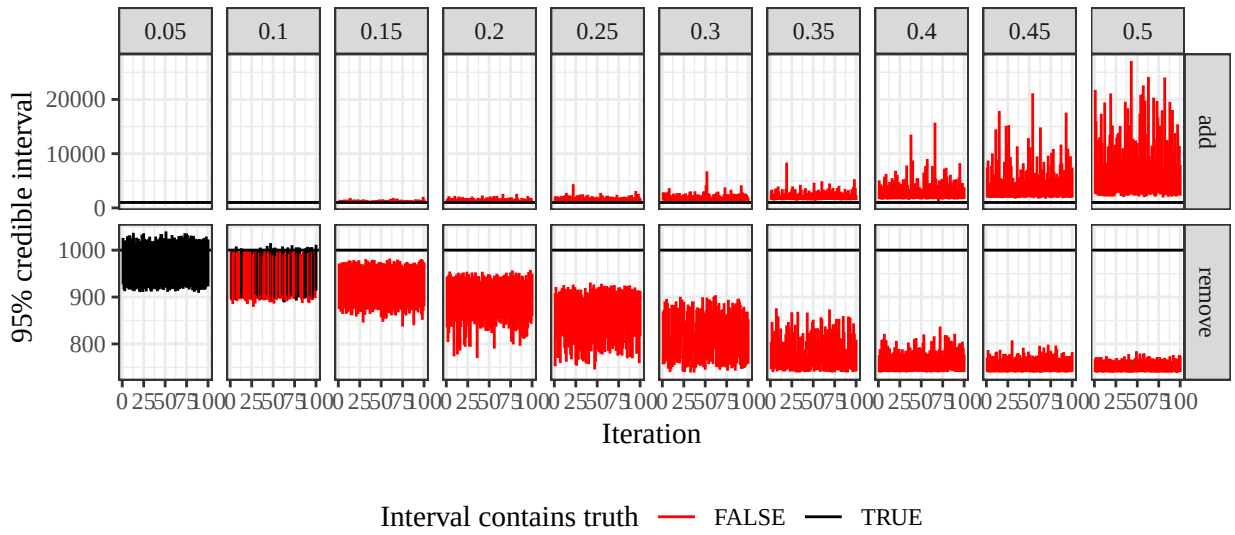

Figure 3: 95% credible intervals resulting from the NPLCM capture-recapture model after removing or adding various proportions of singletons from 5% to 50% and in a record linkage setting. Intervals that contain the truth $M = 1000$ are shown in black and those that do not in red. Somewhere between 5% and 10% of singleton error leads to an unacceptable amount of error in the estimation of M (coverage $< 95\%$).

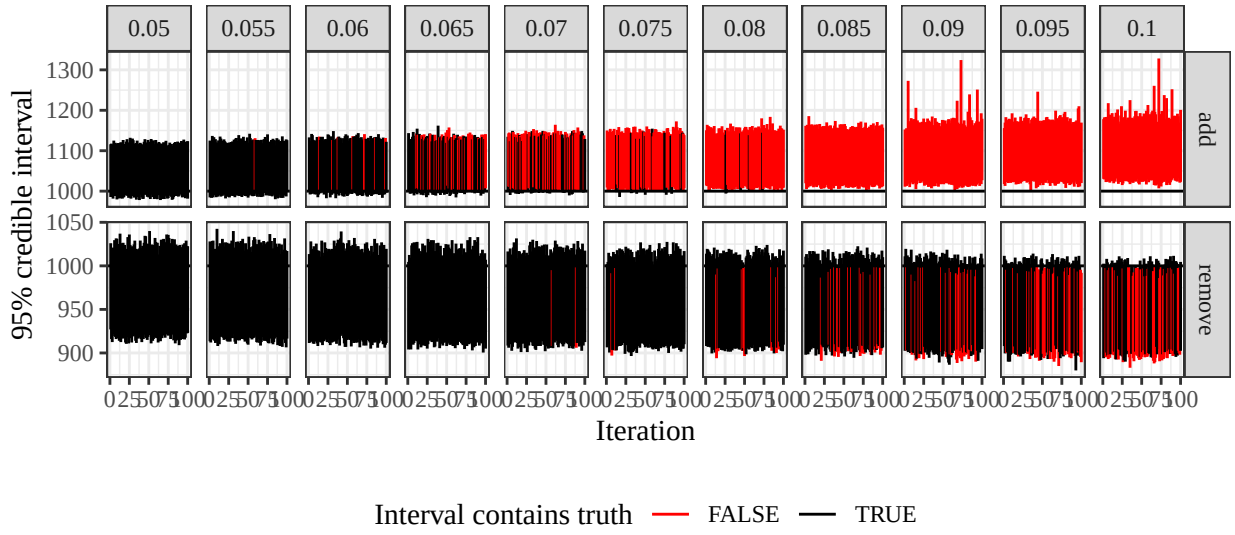

Figure 4: 95% credible intervals resulting from the NPLCM capture-recapture model after removing or adding various proportions of singletons from 5% to 10% and in a record linkage setting. Intervals that contain the truth $M = 1000$ are shown in black and those that do not in red. This zoomed in view shows that depending on if singletons are added (over linked) or removed (under linked), the amount of acceptable error in singletons from the record linkage procedure is around 5.5% or 7.5%, respectively.

Table 3: Coverage of 95% credible intervals for M from using NPLCM capture-recapture after different error levels are passed from the record linkage procedure. Depending on if singletons are added (over linked) or removed (under linked), the amount of acceptable error in singletons from the record linkage procedure is around 5.5% or 7.5%, respectively.

type	0.05	0.055	0.06	0.065	0.07	0.075	0.08	0.085	0.09	0.095	0.1
add	1	0.99	0.92	0.66	0.42	0.16	0.05	0.00	0.00	0.00	0.00
remove	1	1.00	1.00	1.00	0.98	0.98	0.91	0.83	0.77	0.67	0.56

References

Manrique-Vallier, Daniel. 2016. "Bayesian Population Size Estimation Using Dirichlet Process Mixtures." *Biometrics* 72 (4). Wiley Online Library: 1246–54.