Iterative Stratified Sampling with Unknown Variance

Rene Bidart
20637009
STAT 854
Waterloo University

April 24, 2017

Abstract

When doing stratified sampling with unknown stratum variance, there are few options for sample allocation. One method is to obtain auxiliary information, and another is using a pilot study to get a variance estimate, and allocate the remaining samples based on this estimate. In this paper an alternative method is introduced, for use when samples can be iteratively allocated to strata, depending on the previous sample results. This method allocates to the strata based on the expected benefit, and gives allocations that are closer to optimal than the pilot study approach.

1 Introduction

If a population is fully divided into non-overlapping sets, we can increase the effectiveness of our estimators by using stratified sampling, as long as that these sets are relatively homogeneous compared to the overall population. Then the question becomes how do we optimally allocate our samples to each of the strata? If the variances of each stratum are known, we can do this optimally, based on Neyman allocation(Neyman [1934]). The problem is that reality the stratum variances are probably unknown.

Current approaches to deal with with unknown variance involve estimating the variance using outside information such as auxiliary variables or previous surveys. If there is no variance estimate available an alternative strategy is to take a pilot study, and then try to do an approximate Neyman allocation based on the sample variances.

In some cases, such as Internet surveys, it is possible for samples to be taken iteratively, where the stratum is sampled based on the previous samples observed. Each sample will be taken by finding the stratum with the maximum possible benefit in terms of decreasing the overall variation in $\bar{y}$. In this paper we will compare this iterative method to to traditional survey sampling methods.

2 Method

2.1 Traditional Methods

In stratified sampling, our goal is to minimize the variance of out estimate of the true overall mean, μ. Given stratum population sizes of W_h, and stratum means $\bar{y_h}$, the estimator for μ is:

$$\bar{y_{st}} = \sum_{h=1}^{H} \bar{y_h} \tag{1}$$

Our goal is to find a way to minimize the variance of this. An unbiased estimator of this is:

$$v(\bar{y_{st}}) = \sum_{h=1}^{H} W_h^2 (1 - \frac{n_h}{N_h}) \frac{s_{yh}^2}{n_h} \tag{2}$$

A simple form of allocation is **proportional allocation**, where the total number of samples in a given stratum, n are allocated proportionally to the size of the stratum, N_h:

$$n_h = \frac{n}{N} N_h = nW_h \tag{3}$$

This method is simple, but because it does not take into account any information on the stratum variances, it is inefficient. Alternatively, if the stratum variances are known, the optimal allocation is known, and this is **Neyman allocation**. Here we allocate the samples in each stratum, n_h proportional to its standard deviation, σ_{yh}, and its size, W_h:

$$n_h \propto W_h \sigma_{yh} \tag{4}$$

In reality it is unlikely that this will be known, and so alternative methods using auxiliary information or previous surveys have been proposed to estimate the variance. These methods will not be discussed in this paper, as these methods can be treated independently of the optimal allocation method. If necessary, these variance estimation methods can be combined with any other sampling procedure. For the rest of this paper, for simplicity it is assumed that the strata have equal sizes. The following derivations could be adjusted to include differing strata sizes if this was necessary.

If the variances are not know, a pilot study can be used to estimate the stratum variances, and then we can do an approximate Neyman allocation based on these sample standard deviations. The procedure is as follows:

1. Decide on an initial proportion p, and then sample ceiling(n*p) from all strata
2. Compute the hypothetical optimal allocation, assuming these probabilities to be true
3. Check if any strata have already been sampled more than the optimal amount, and if so, exclude them from the next step
4. Considering only the strata that have not already been sampled too much, compute the Neyman allocation for these, and take the sample.

2.2 Iterative Method

We want an iterative sampling procedure to minimize the variance of our mean estimate, $\bar{y}$. Consider the case where we have already taken n samples allocated between the strata, and find the best stratum to sample the next point from.

This means we are trying to find the stratum where sampling from it will cause the largest decrease in variance of $\bar{y}$. We would would like to find i to maximize:

$$v(\bar{y}_{st,i}) - v(\bar{y_{st}}) = \sum_{h=1}^{H} W_{h,i}^2 (1 - \frac{n_{h,i}}{N_{h,i}}) \frac{s_{yh,i}^2}{n_h, i} - \sum_{h=1}^{H} W_h^2 (1 - \frac{n_h}{N_h}) \frac{s_{yh}^2}{n_h} \qquad (5)$$

Here $\bar{y}_{st,i}$ is used to denote this after another sample is taken from stratum i. For simplicity, assume that all strata have the same size, so W_h are equal. Then, we can notice that all the terms in this sum are equal, except for those relating to stratum i:

$$(1 - (\frac{n_i + 1}{N_i}))\frac{s_i^{*2}}{n_i + 1} - (1 - (\frac{n_i}{N_i}))\frac{s_i^2}{n_i} \tag{6}$$

If we assume $s_i^{*2} = s_i^2$, after rearranging we obtain the objective as:

$$\max_i \frac{s_i^2}{n(n+1)} \tag{7}$$

This shows that the expected benefit is increasing in the variance of the sample, as would be expected from Neyman allocation, and is decreasing in the number of samples already taken from that strata, because as more samples are taken from a given strata, the we have a better estimate of the samples mean. We call this the **greedy algorithm**, and it's procedure is as follows:

1. Take a small number of samples from all strata, compute variance, and the expected benefit estimates for all strata.

2. Sample from the strata with the maximum expected benefit

3. Repeat this until a total of n samples are taken

Implementation Detail: To make this computationally efficient, we can use the iterative update formulas for the mean and standard deviation. Using this means previous samples don't have to be stored, and is is faster.

2.3 Non-greedy Strategy

The iterative algorithm above is a greedy strategy, meaning that at every iteration it samples from the mean with highest expected benefit. The problem with this strategy is that there is uncertainty in our estimate of s_i^2. In the above algorithm we have ignored this issue, and assumed $s_i^{*2} = s_i^2$. This situation seems similar to them multi-arm bandit problem (Sutton and Barto [1998]), where there is an exploration-exploitation trade off. Here we want to exploit the stratum with highest expected benefit by sampling from it, but there is a danger that we have estimated the variances incorrectly, and so we may want to sample other strata to improve this estimate.

Instead we can use a strategy from the multi-arm bandits problem, called $\epsilon - greedy$. In this strategy, the procedure is the same as the greedy strategy, but a small portion of the time, ϵ, we will sample randomly from the strata instead of taking from the strata with the highest expected benefit. This will mean that we get better estimates of the variance for each strata, and so there is a smaller chance of making very bad allocations.

3 Simulations and Results

All of the simulations are done using three strata, and to compare the methods we will look at different variance initializations. Each simulation ran with 20 different mean initializations, and for each initialization it is repeated 1000 times. We will run simulations where the variances are the same (1, 1, 1), one were variances are different (1, 2, 3), and one where one variance is much larger than the other two (1, 1, 8). This is because different methods will have different performance depending on the true stratum variance. For example, in the case of equal variance, in our simulation the proportional and the Neyman allocation will give the same allocation.

Table 1: Average MSE after 5000 replications

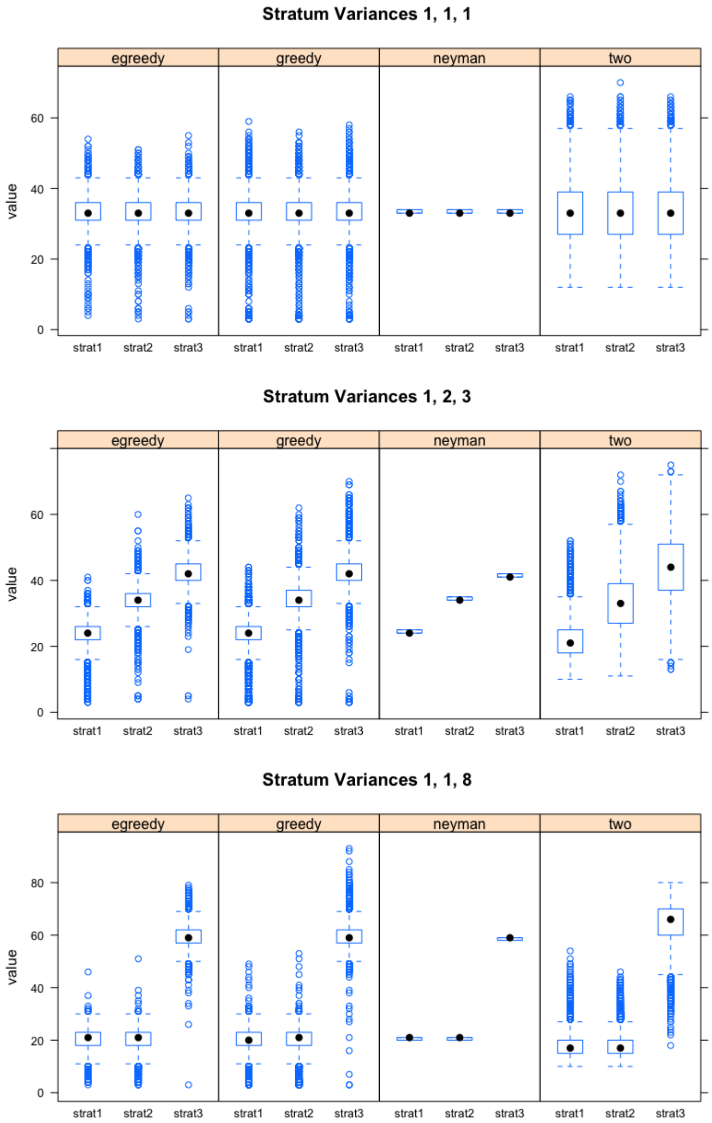
Stratum Variances	Neyman	proportional	two-sample	greedy	$\epsilon - greedy$
1, 1, 1	0.0100	0.0099	0.0107	0.0102	0.0101
1, 2, 3	0.0190	0.0198	0.0206	0.0197	0.0194
1, 1, 8	0.0265	0.0331	0.0274	0.0272	0.0266

The mean squared error of the estimator for $\bar{y}$ was computed for the five different methods: Neyman allocation, proportional allocation, Two-stage allocation, greedy, and epsilon-greedy. For the two-stage allocation an initial sample size of 30% was used, while with the $\epsilon - greedy$ allocation a value of $\epsilon = .1$ of is used.

These simulations are run using a fixed population of size $N = 10,000$, and total sample size of $n = 100$

3.1 Sample Allocation

In figure 1, over all variance initializations we see the that both the greedy and epsilon-greedy algorithms seem to approximate the Neyman allocation reasonably well. It seems that it generates better approximations to the optimal allocation than is given by the two-sample procedure. There are slightly fewer outliers with the $\epsilon - greedy$ strategy, which is to be expected because this strategy uses the non-optimal sampling to help insure against very bad allocations. This seems to be true across all of the variance initializations that were tested.



Stratum Variances 1, 1, 1

Stratum Variances 1, 2, 3

Stratum Variances 1, 1, 8

Figure 1: *Distribution of samples by stratum. Each is run with 20 random mean initializations, and 1000 replications*

Table 2: Random Mean and Variance initializations

Neyman	proportional	two-sample	greedy	e-greedy
0.0288	0.0299	0.0307	0.0298	.0293

3.2 Performance

For each of the variance initializations and each of the sampling methods, we computed the mean squared error, and averaged it over all replications. Results are shown in Table 1.

We see that with equal variances all methods perform similarly, which is to be expected. In the case of different variances we see that the $\epsilon - greedy$ algorithm came closest to reaching the optimal MSE attained by the Neyman allocation. Strangely the proportional allocation performed better than the two-sample. This is unexpected because in this case we know the proportional allocation is quite far from being optimal. With one large variance, the *greedy* and two-sample methods both performed well, but the $\epsilon - greedy$ method was the best, nearly attaining the MSE of the Neyman allocation

In order to get an estimate of the overall performance of the methods, we used both random mean and random variance initializations. The variance was sampled from uniform[1, 5], and the mean from uniform[0,5]. This was done using 100 initializations, with 1000 replications for each initialization, as shown is Table 2. As is expected, the Neyman allocation has performed the best, while the e-greedy is the second best. These results seem reasonable, but the two-sample performance seems unusually bad. It seems that the allocation in the two sample must on average be worse than the proportional allocation. This is unexpected because we know the proportional allocation is quite far from optimal.

4 Conclusion

We have found that using this iterative sampling procedure works reasonably well for a variety of variance initializations. The $\epsilon - greedy$ method performs better than the simple greedy strategy, and it seemed to approximate the Neyman allocation well. It performs better than the pilot study approach, under a variety of different variance initializations. The effective allocations of this strategy also lead it to have a good mean squared error compared to other methods, and at times it approaches the mean squared error of the optimal Neyman allocation.

It would be useful to further look into what the optimal value of epsilon is for the epsilon-greedy algorithm. This may also be dependent on the sample size, which we had fixed during our studies. It could also be useful to check if there is any benefit to using the $\epsilon - greedy$ over increasing the initial sample size taken from each strata.

Bibliography

Jerzy Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.