Poststratification: A case study

Daniel Quartey

Department of Mathematical Sciences

Montana State University

May 4, 2018

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Daniel Quartey

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

May 4, 2018

John Borkowski
Writing Project Advisor

May 4, 2018

Mark C. Greenwood
Writing Projects Coordinator

Abstract

A simulation study is used to explore the performance of two sampling methods; poststratifcation and simple random sampling. In this paper, estimates from poststratifcation and SRS at three sampling fractions (5%, 10% and 15%) and at different confidence levels (90%, 95% and 99%) are compared. Coverage rates with associated 95% confidence intervals are compared for the two sampling methods. Mean confidence interval widths for both methods are also compared for the three sampling fractions and at different confidence levels (90%, 95% and 99%). Data were sampled from a hypothetical study (Borkowski 2017). Variables collected on each quadrat are severity type of Baddgrass plant densities, the number of Gudgrass plants, and nitrogen level. The response of interest is a count of the number of Gudgrass plants that are present per quadrat and researchers want to estimate the mean number of Gudgrass plants per quadrat in the study area. This study contains an exploratory analysis, with the goal of explaining the benefits of poststratification rather than estimating using a simple random sampling design.

Contents

1 Introduction

Statistical inference is the process of making conclusions about some characteristics of interest for a population based on data collected. Any researcher's sampling goal is to collect a sample that is representative of a study population. The sample mean is typically used as a point estimate for the population mean. A popular sampling design is Simple Random Sampling (SRS). Stratification is also a widely-used technique in sampling which when properly applied, serves dual purposes of providing representative sub-groups of the population and also maximizing gains in precision. Occasionally, researchers are faced with sampling problems where they would like to stratify the population of sampling units on a key variable but cannot assign sampling units to strata until after data has been collected. This paper seeks to help researchers about a useful remedy to this problem and some of its benefits. In this study, poststratification of a sampling design will be presented. I seek to compare coverage rates and mean interval widths of two sampling designs: Simple random sampling and poststratification from a known population.

1.1 Notation

L =number of strata

N_h = number of population units in stratum h. h= 1,2,3,...L

N =$\sum_{h=1}^{L} N_h$= the number of units in the population

n_h = the number of sampled units in stratum h, h= 1,2,....L.

n=$\sum_{h=1}^{L} n_h$= the total number of units sampled

$\hat{\bar{y}}_h$ = the sample mean for stratum h

s_h^2= sample variance for stratum h

2 Simple Random Sampling

Simple random sampling (SRS) without replacement of size n is the probability sampling design for which a fixed number of n sampling units are selected from a population without replacement such that every possible sample of n units has equal probability of being selected. There are $\binom{N}{n}$ possible SRS of size n selected from a population. For any simple random sample S of size n from population size N we have $P(S) = \frac{1}{\binom{N}{n}}$. Simple random sampling without replacement is a sampling procedure in which a sampling unit is randomly selected from the population, its response recorded and is not returned to the population. This process of randomly selecting units without replacement after each stage is repeated n times. Thus a sampling unit cannot be sampled multiple times.

2.1 Estimation of $\bar{y}_U$

The SRS estimator for the population mean $\bar{y}_U$ is the sample mean $\hat{\bar{y}}_U$:

$$\hat{\bar{y}}_U = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1}$$

which has variance

$$V(\hat{\bar{y}}_U) = (\frac{N-n}{N})\frac{S^2}{n} \tag{2}$$

where S^2 is the population variance. The estimated variance is

$$\hat{V}(\hat{\bar{y}}_U) = (\frac{N-n}{N})\frac{s^2}{n} \tag{3}$$

where s^2 is the sample variance.

For larger samples, an approximate $100(1-\alpha)\%$ confidence interval for $\bar{y}_U$ is

$$\bar{y} \pm z^* \sqrt{(\frac{N-n}{N})\frac{s^2}{n}} \tag{4}$$

where z^* is the upper $\alpha/2$ critical value from the standard normal distribution.

For smaller samples, an approximate $100(1-\alpha)\%$ confidence interval for $\bar{y}_U$ is

$$\bar{y} \pm t^* \sqrt{(\frac{N-n}{N})\frac{s^2}{n}} \tag{5}$$

where t^* is the upper $\alpha/2$ critical value from the $t(n-1)$ distribution.

3 Poststratification

In stratified sampling the researcher has knowledge of the strata sizes as well as availability of a frame for drawing a sample in each stratum. What can we do when knowledge of a frame for drawing a sample in each stratum is not available? Practically, it is not possible to know in advance to which stratum a sampling unit belongs until contacted or investigated in the course of the survey itself. Whether we randomly sampled with or without replacement, it is always possible to classify the selected sample to the strata after selection.

Occasionally, it happens that stratum sample sizes are not known until after data collection. Poststratification can be very efficient because after sampling the stratification factors can be chosen in different ways for different sets of variables in order to maximize gains in precision.

In this study, a simulation-based test to show that poststratification can be as precise as proportional allocation when sampling weight $W_h = N_h/N$ is known and $n_h \geq 20$. Another simulation based test compared coverage rates and confidence interval widths between simple random sampling and the poststratification methods.

3.1 Variance and confidence interval estimation with a poststratification sampling design

The estimator for the population mean $\bar{y}_U$ is the weighted mean $\hat{\bar{y}}_{Ustr}$ for poststratifcation is:

$$\hat{\bar{y}}_{Ustr} = \sum_{h=1}^{L} (\frac{N_h}{N}) \hat{\bar{y}}_h \tag{6}$$

In poststratification, the n_h are random variables with $E(n_h)$ =nW_h, where stratum h=1,...,H. If n_h is fixed, then

$$\hat{V}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{s_h^2}{n_h} - \frac{1}{N} \sum_{h=1}^{L} W_h s_h^2 \tag{7}$$

$$W_h = \frac{N_h}{N}$$

where stratum h =1,...,L

Note $W_h = N_h/N$ represents the population proportion in stratum h. Thus, the researcher does not need to know N_h and N but only the relative proportion of the population for each stratum. If needed W_h can be estimated by $\hat{W}_h = \frac{n_h}{n}$

The average value of $\hat{V}(\hat{\bar{y}}_{Ustr})$ in repeated samples of size n must now be calculated. Care must be taken since at least one stratum sample size n_h could be zero. If this happens, two or more strata would have to be combined before calculating the estimate, and a less precise estimate would be produced. With an increasing n, the probability of n_h being zero becomes small so that the contribution to the variance from the source is negligible. If the case in which n_h is zero is ignored, Stephan(1945) provides us with a good approximation of $E(\frac{1}{n_h})$

$$E(\frac{1}{n_h}) \approx \frac{1}{nW_h} + \frac{1 - W_h}{n^2 W_h^2} \tag{8}$$

and the estimated variance is

$$\hat{V}_p(\bar{y}_{st}) = (\frac{1}{n} - \frac{1}{N})\sum_{h=1}^{L} W_h s_h^2 + \frac{1}{n^2}\sum_{h=1}^{L}(1 - W_h)s_h^2 \tag{9}$$

where the subscript p refers to poststratifcation. The first term in $\hat{V}_p(\hat{\bar{y}}_{Ustr})$ in (9) is the variance one would get from a stratified weighted mean under proportional allocation. The second term is always nonnegative and represents the contribution to the variance from post - rather than pre-stratification. Note that the divisor of the second term is n^2 and consequently that term is usually quite small. In summary, the Stephan(1945) approximation only works well when sample size is large and n_h values are relatively large as well. A practical consequence of this is that we should not poststratify too finely.

Notes of caution:

1. If $\frac{N_h}{N}$ is not known or cannot be accurately estimated, then poststratification can cause the estimator to be very imprecise.

If all of the stratum sample sizes n_h are sufficiently large (Thompson(2012) suggests $n_h \geq 30$), an approximate $100(1-\alpha)\%$ confidence interval for $\bar{y}_U$ is

$$\hat{\bar{y}}_{Ustr} \pm z^* \sqrt{\hat{V}_p(\hat{\bar{y}}_{Ustr})} \tag{10}$$

where $\hat{\bar{y}}_{Ustr}$ is the poststratification estimator of the population mean and z^* is the upper $\alpha/2$ critical value from the standard normal distribution.

For smaller samples, an approximate $100(1-\alpha)\%$ confidence interval for $\bar{y}_U$ is

$$\hat{\bar{y}}_{Ustr} \pm t^* \sqrt{\hat{V}_p(\hat{\bar{y}}_{Ust})} \tag{11}$$

where t^* is the upper $\alpha/2$ critical value from the $t(d)$ distribution. In this case, d is Satherthwaite's(1946) approximate degrees of freedom where

$$d = \frac{(\hat{V}_p(\hat{\bar{t}}_{st}))^2}{\sum_{h=1}^{L}(a_h s_h^2)^2/(n_h - 1)} \tag{12}$$

and $a_h = \frac{N_h(N_h - n_h)}{n_h}$

If the stratum sample sizes n_h are all equal and the stratum sizes N_h are all equal, then the degrees of freedom reduces to $d = n - L$ where $n = \sum n_h$ is the total sample size.

4 An Example

4.1 Background of the study

The following is a hypothetical study (Borkowski 2017). A region in central Montana has been infested with Baddgrass, a non-native weed species. A mitigation process to remove the Baddgrass and then revegetate the land was applied to a study area. The research process is to first expose a study region to a herbicide (which we will refer to as Bio-B-Gone) to kill all Baddgrass. Unfortunately, Bio-B-Gone will kill any plant. However, once all plant life is killed in the study area, a native plant, Gudgrass, will be planted with the goal of revegetating the area. A land reclamation scientist wants to summarize the amount of vegetation present two years after application of the Bio-B-Gone and subsequent planting of the Gudgrass. If the mitigation process proves to be successful, then the goal would be to expand its use to larger portions of the central Montana region infested with Baddgrass. The study area is a $400m$ by $400m$ region that is divided into 1600 10m by 10m quadrats. The 1600 quadrats are arranged in a rectangular grid of 40 rows and 4from0 columns. Rows 1 to 40 go north to south, and columns 1 to 40 go from west to east.

Just prior to the application of Bio-B-Gone, an aerial survey was performed of the study area. Based on the aerial photograms, a map was made that classified locations into one of four severity types. Severity types 1, 2, 3, and 4, correspond to low, moderate, high, and very high Baadgrass plant densities. It is suspected that the potential for revegetation may differ across severity types. It is also suspected that the amount of nitrogen in the soil when planting the Gudgrass could affect the revegetation efforts. No prior soil nitrogen levels are available, but will be collected from the quadrats sampled based on the sampling design. For each sampled quadrat from a SRS, the data collected are as follows: severity type, nitrogen level, and the number of Gudgrass plants that were observed after two years since application of Bio-B-Gone and replanting with Gudgrass. The nitrogen levels range from 0 to 28.3 mg/kg.

4.2 The response of interest

One measure of vegetation that is easy to collect is a count of the number of Gudgrass plants that are present per quadrat.

4.3 Parameter to be estimated

The parameter of interest for the researcher is the mean number of Gudgrass plants per quadrat in the study area.

4.4 Exploratory Data Analysis

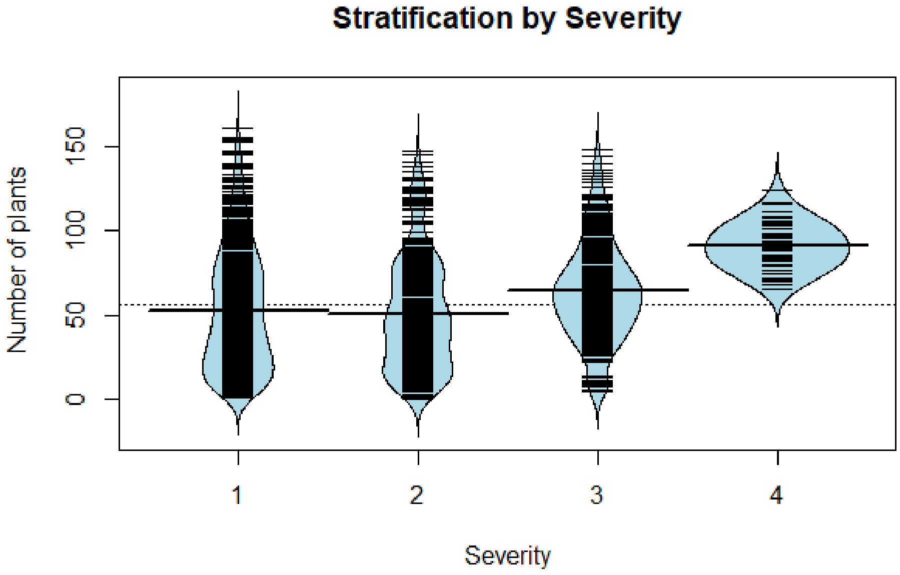


Figure 1: Beanplots of number plants stratified by Severity types

Table 1: Summary statistics after stratifying by Severity types

Severity	Min	Q1	Median	Q3	Max	Mean	sd	N
Severity 1	1	22	48	78	161	52.78	35.52	766
Severity 2	0	23	46	74	147	50.52	33.09	457
Severity 3	4	44	62	81	148	63.866	29.55	321
Severity 4	65	80	90.5	101	124	90.69	13.75	56

Table 1 provides the summary statistics for the number of plants per quadrat when quadrats are stratified by severity levels. Severity types 1, 2 and 3 shows large spread of number of plants for those quadrats. The standard deviation column shows large values meaning the stratified estimator using severity as the stratification variable will have a large variance. Figure 1 shows beanplots for severity types that confirm these figures in Table 1.

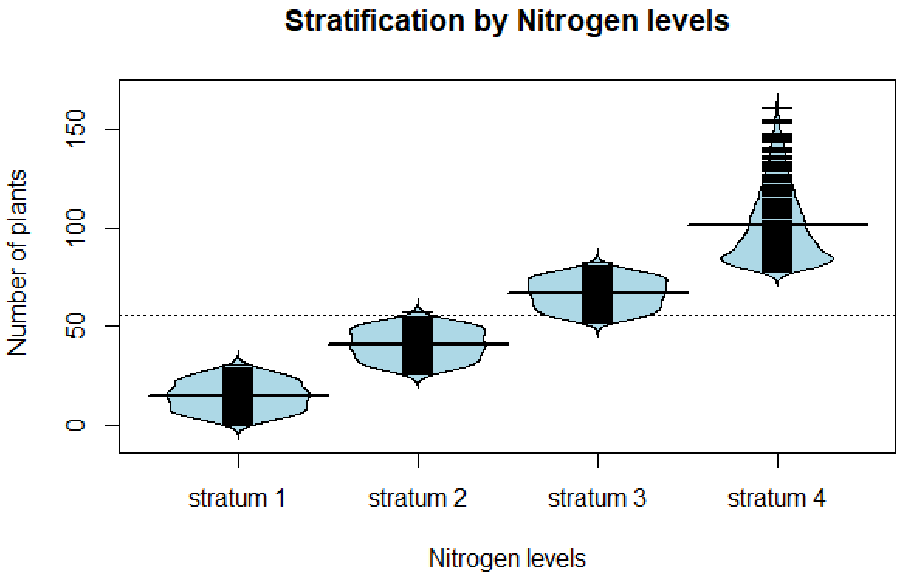


Figure 2: Beanplots of number plants stratified by nitrogen levels

Table 2: Summary statistics after stratifying by nitrogen levels

Strata	Min	Q1	Median	Q3	Max	Mean	sd	N
Stratum 1	0	8	15	21	30	14.578	7.51	401
Stratum 2	26	34	41	47	57	40.594	7.68	404
Stratum 3	52	60	67	73	82	66.43	7.56	396
Stratum 4	78	86	97	112.5	161	101.64	18.64	399

Table 2 provides the summary statistics when quadrats are stratified by nitrogen levels. Strata 1, 2, 3 and 4 have much smaller standard deviations for the number of plants for those quadrats compared to within severity type variation. Figure 2 shows beanplots for nitrogen levels across nitrogen level strata confirms these figures in Table 2.

Table 3: Summary statistics of nitrogen levels

Min	Q1	Median	Q3	Max	Mean	sd	N
0	11.2	16	19.4	28.3	15.31	5.70	1600

Table 4: Summary statistics of number of Gudgrass plants observed

Min	Q1	Median	Q3	Max	Mean	sd	N
0	27	53	79	161	55.68	34.14	1600

Tables 3 and 4 provide summary statistics of nitrogen levels and the number of Gudgrass plants observed, respectively.

4.5 Why poststratify with Nitrogen levels?

The basic motivating principle behind using stratification is to produce an estimator with small variance by partitioning the population so that the units within each stratum are as similar as possible. This is known as the stratification principle. In this study, researchers defined strata for a geographical region into groups of units that will be similar with respect to nitrogen level categories because it is suspected that the number of plants may vary greatly across strata while they will tend to be similar within each stratum.The researchers however, do not know which units belong to each stratum prior to data collection. Nitrogen levels helped stratify the number of plants for each quadrat in four strata because they were stratified based on prior estimates of quartiles. The bean plot in Figure 2 clearly shows there is smaller variability within strata when stratification is done by nitrogen levels and larger variability when stratification is done by severity type. This confirms that we expect to obtain more precise estimates when plant counts are stratified by nitrogen level.

4.6 Method

Using this study, we will compare the estimates, coverage probabilities and precision among the two sampling designs: Simple random sampling and poststratification.

5 Results and Discussion

The coverage rate was computed from 10,000 samples and summarized as the percentage of 10,000 confidence intervals that contained the population mean. For each interval, the difference between the upper bound and the lower bound was recorded. The mean width was calculated as the average of the 10000 confidence interval widths.

5.1 Results: Coverage Rate

Table 5: Coverage rates for 90% confidence level(95% CI for coverage rates)

Method	**80(5% of N)**	**160(10% of N)**	**240(15% of N)**
Poststratification	89.84%(89.24%,90.40%)	89.64%(89.04%,90.38%)	89.79% (89.19%,90.38%)
SRS	89.85%(89.26%,90.44%)	90.04% (89.45%,90.63%)	90.13%(89.50%,90.71%)

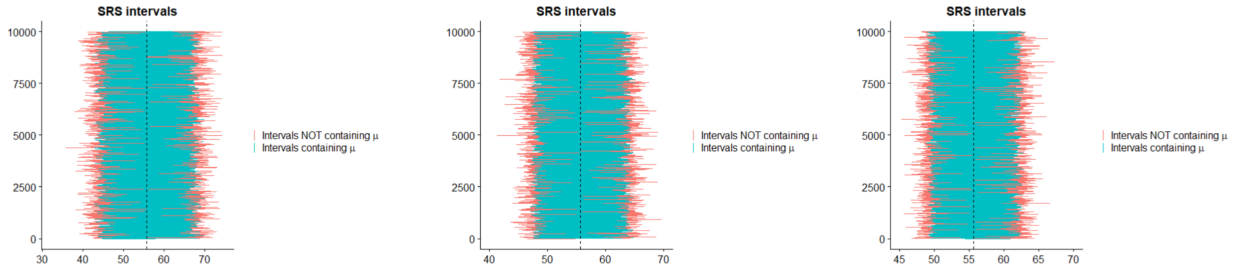
Table 6: Coverage rates for 95% confidence level(95% CI for coverage rates)

Method	**80(5% of N)**	**160(10% of N)**	**240(15% of N)**
Poststratification	93.73%(93.25%,94.21%)	94.66%(94.22%,95.10%)	94.67%(94.23%,95.11%)
SRS	95.23%(94.81%,95.65%)	95.06% (94.64%,95.48%)	95.33% (94.92%,95.74%)

Table 7: Coverage rates for 99% confidence level (95% CI for coverage rates)

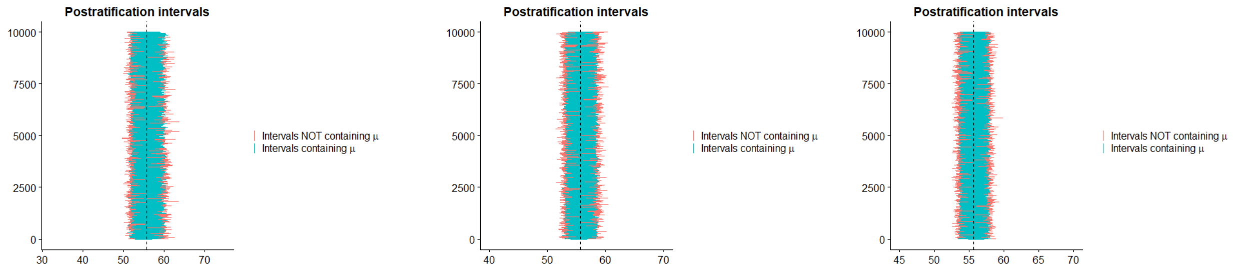
Method	**80(5% of N)**	**160(10% of N)**	**240(15% of N)**
Poststratification	98.51%(98.27%,98.75%)	98.87% (98.66%,99.08%)	98.85%(98.64%,99.06%)
SRS	98.89%(98.68%,99.09%)	99.13%(98.95%,99.31%)	99.08%(98.89%,99.27%)

Tables 5, 6 and 7 above show the coverage rates along with the associated confidence intervals for poststratification and SRS methods. Approximate 90%, 95% and 99% nominal levels of confidence for all sample sizes were attained. It is observed that all coverage rates are within 1% of the nominal confidence level for all three sample sizes (80, 160 and 240). For all three sample sizes, the coverage rates were close to the three nominal levels of confidence for both poststratification and SRS methods.



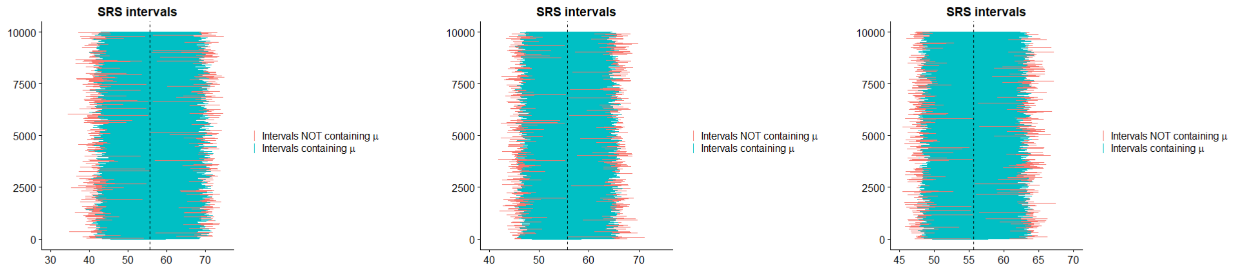
(a) Intervals when n=80 (b) Intervals when n=160 (c) Intervals when n=240

Figure 3: SRS Intervals for a nominal 90% confidence level



(a) Intervals when n=80 (b) Intervals when n=160 (c) Intervals when n=240

Figure 4: Poststratification Intervals for a nominal 90% confidence level

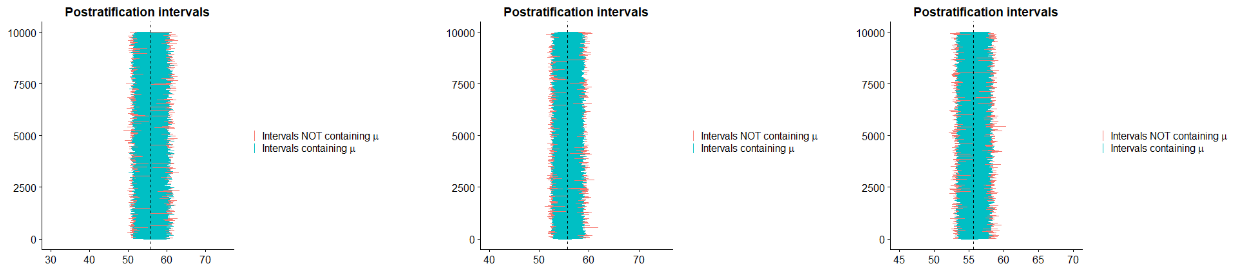


(a) Intervals when n=80 (b) Intervals when n=160 (c) Intervals when n=240

Figure 5: Poststratification intervals for a nominal 95% confidence level

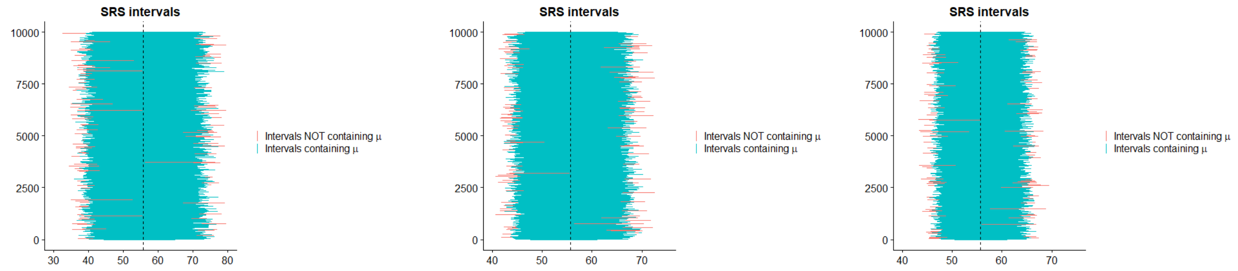
Figures 3 to 8 show a pictorial view of 10,000 simulated intervals and which intervals contain the true mean(μ) or otherwise. Looking at the 10,000 simulated intervals for each sampling design the following observation were made:

- Poststratification intervals are narrower intervals than SRS intervals.
- Predictably, larger confidence levels had wider confidence intervals.
- Finally, as the sample size increased, the interval became narrower for all three confi-



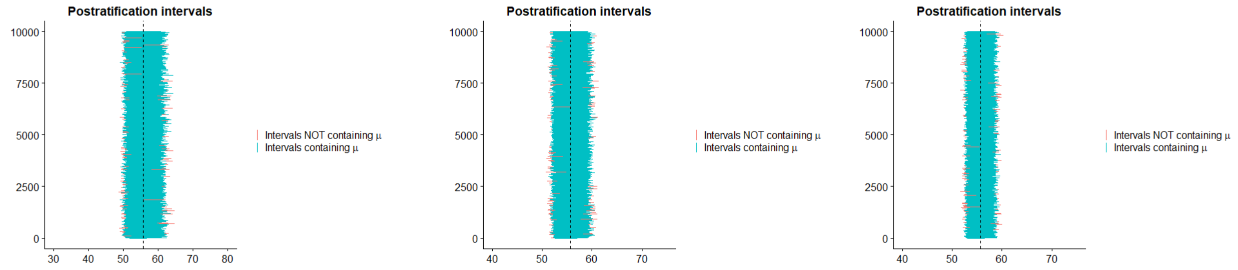
(a) Intervals when n=80 (b) Intervals when n=160 (c) Intervals when n=240

Figure 6: Poststratification intervals for a nominal 95% confidence level



(a) Intervals when n=80 (b) Intervals when n=160 (c) Intervals when n=240

Figure 7: SRS intervals for a nominal 99% confidence level



(a) Intervals when n=80 (b) Intervals when n=160 (c) Intervals when n=240

Figure 8: Poststratification intervals for a nominal 99% confidence level

dence levels.

5.2 Results: Mean Interval Width

Table 8: Mean Interval Width/SE for 90%CI

Method	**80(5% of N)**	**160(10% of N)**	**240(15% of N)**
Poststratification	4.1863/0.00468	2.8489/0.00212	2.2509/0.00131
SRS	12.3556/0.0086	8.4627/0.00397	6.70147/0.00251

Table 9: Mean Interval Width/SE for 95%CI

Method	**80(5% of N)**	**160(10% of N)**	**240(15% of N)**
Poststratification	5.0035/0.00557	3.4010/0.0025	2.6857/0.001565
SRS	14.7853/0.01	10.0996/0.0046	7.9917/0.00293

Table 10: Mean Interval Width/SE for 99%CI

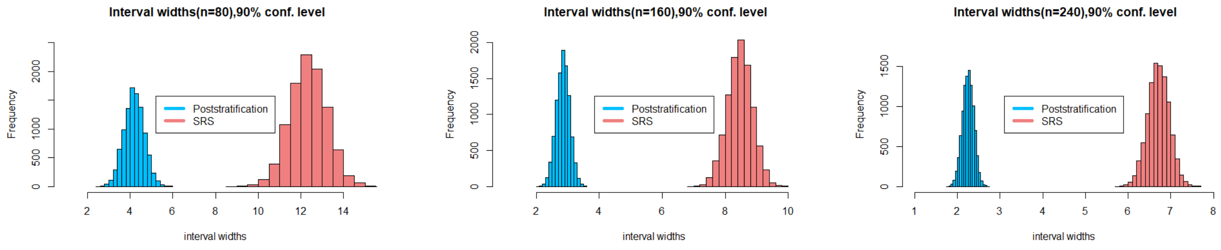
Method	**80(5% of N)**	**160(10% of N)**	**240(15% of N)**
Poststratification	6.6555/0.00737	4.4895/0.00336	3.5388/0.0020699
SRS	19.5949/0.0136	13.3315/0.006217	10.5424/0.0039

Based on Tables 8, 9 and 10, the following observations were made:

- At each confidence level, the mean interval width for the SRS is about three times the mean width of poststratification for the three sample sizes. The SE for poststratification is also about half the the standard error for the SRS.

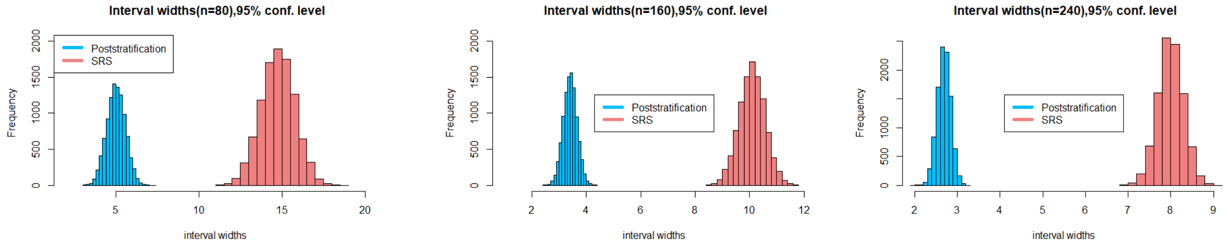
- As sample size increases, the standard error of the mean width decreases at each confidence level for both methods.

- Mean width increases as the nominal confidence level also increases for all three sample sizes.



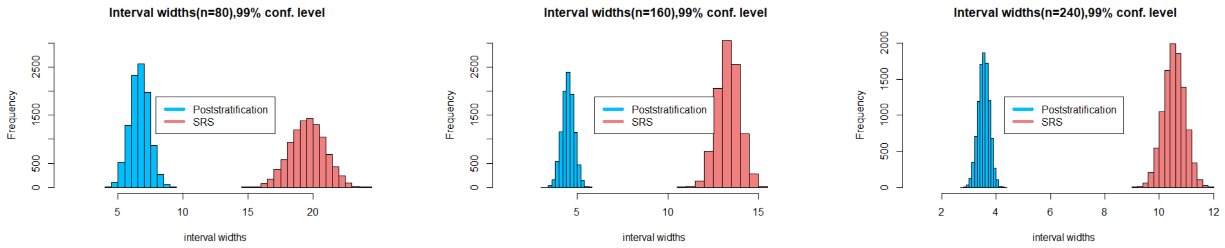
(a) Intervals when n=80 (b) Intervals when n=160 (c) Intervals when n=240

Figure 9: Histograms of Widths for a nominal 90% confidence level for Poststratification and SRS



(a) Intervals when n=80 (b) Intervals when n=160 (c) Intervals when n=240

Figure 10: Histograms of Widths for a nominal 95% confidence level for Poststratification and SRS



(a) Intervals when n=80 (b) Intervals when n=160 (c) Intervals when n=240

Figure 11: Histograms of Widths for a nominal 99% confidence level for Poststratification and SRS

Based on Figures 9, 10 and 11, the following observations were made:

- The histogram of widths for SRS is centered at a larger value than where the center of the poststratification histogram is located.
- The histogram of the SRS has a wider spread than that of the poststratification histogram of widths.
- The space between the two histograms of width of the two methods indicates how different the intervals are for both methods.
- For both methods, the mean width decreases as the sampling fraction (sample size) was increased.

5.3 Discussion

When constructing confidence intervals, the goal is to match the desired level of confidence at least approximately. The results showed the poststratifcation design generates narrower and more precise confidence intervals than the SRS sampling design that approximately match the nominal confidence level. On the other hand, it was also observed that SRS may generate confidence intervals that are wider but still attain the desired level of confidence.

6 Improvement and Future work

Future work could look at including ratio or regression estimation to the poststratification process in the estimation of the population mean if more variables are collected on quadrats.

7 References

1. D.Holt and T.M.F Smith,*Poststratification Journal of the Royal Statistical*, Series A Vol.142 No.1 (1979): 33-46.

2. R.L.Scheaffer, Mendenhall III and Lyman Ott. *Elementary Survey sampling fifth edition*:166-169.

3. *Cochran Sampling Techniques 3rd Edition*: 134-135.

4. P.V and B.V Sukhtame, *Sampling theory of surveys with applications*: 94-95.

5. J.J Borkowki *Sampling notes(Stat 446) 2017.*

6. Steven K. Thompson (2012) *Sampling 3rd Edition*: 148-149.

7. Satterthwaite, F. E. (1946), *An Approximate Distribution of Estimates of Variance Components, Biometrics Bulletin 2*: 110–114.

8. Frederick F. Stephan(1945), *The Expected Value and Variance of the Reciprocal and Other Negative Powers of a Positive Bernoullian Variate:* 50-61.