Confidence Intervals for Proportion Estimates in Complex Samples for Performance Audits

James D. Ashley, Danny Lee, Carl Barden
United States Government Accountability Office, Washington, DC

Abstract
Performance audits routinely require estimation of a large number of attributes to support audit findings, make assessments of internal controls, and assess compliance with laws and regulations. Proportion estimates from these audits are often at or near the boundaries (0 or 1.0) and confidence interval estimation requires methods other than normal approximations, especially if they are to be derived from complex sample designs. These methods have been well researched and are now included in several commonly used statistical software packages.

This paper describes and analyzes methods applied by the Government Accountability Office (GAO) to estimate a large number of attribute estimates from a range of complex sample designs. We attempt to assess how the method GAO applies as a confidence interval decision rule (GAO's CI decision rule) affects coverage probabilities across a range of sample designs and population proportions. We further examine potential coverage probability reductions in complex sample designs by relaxing the limitation on the effective sample size increasing beyond the observed sample size for complex sample designs.

Key Words: Performance audit sample estimation, binomial confidence intervals, asymmetric, complex sample design

1. Confidence Interval Methods

For a proportion estimate derived from a complex sample, a standard designed-based 1-α confidence interval (CI) can be constructed using the following formula

$$\hat{p} \pm t_{(1-\alpha/2,d)} SE(\hat{p})$$

where $\hat{p}$ is the weighted sample estimator of the population proportion, $SE(\hat{p})$ is the design-based standard error of $\hat{p}$, $t_{(1-\alpha/2,d)}$ is the 1-α/2 quantile of the t-distribution with d degrees of freedom. When the sample size is large, the sampling distribution of $\hat{p}$ is assumed to be approximately normal as the t-distribution approaches a normal distribution. When the sample size is small, or the proportion estimate is at or near the boundaries (0 or 1.0), this normality assumption does not hold and the performance of the constructed CIs fail to achieve the desired 1-α coverage probability. This is of particular concern for an audit sample because audit standards require sufficient evidence to make assessments of internal controls or compliance with laws or regulations. The decrease in coverage probability corresponds with an increase in the audit risk[1].

[1] According to the 2018 revision of the Government Auditing Standards (Yellow Book), audit risk is defined as the possibility that the auditors' findings, conclusions, recommendations, or assurance may be improper or incomplete as a result of factors such as evidence that is not

In simple random sampling from large populations, this problem can be avoided by using exact binomial methods, as described by Clopper and Pearson (1934). When x number of errors are observed in a simple random sample of size n, the 1-α CI ($p_L(x,n), p_U(x,n)$) can be constructed as follows:

$$p_L(x,n) = \frac{v_1 F_{v_1,v_2(\alpha/2)}}{v_2 + v_1 F_{v_1,v_2(\alpha/2)}}$$

$$p_U(x,n) = \frac{v_3 F_{v_3,v_4(1-\alpha/2)}}{v_4 + v_3 F_{v_3,v_4(1-\alpha/2)}}$$

Where $v_1 = 2x, \quad v_2 = 2(n-x+1), \quad v_3 = 2(x+1), \;\; v_4 = 2(n-x).$

These methods have been shown, for simple random samples, to provide coverage probabilities that are greater than or equal to the desired level. This results in over-coverage which is generally considered acceptable by an auditor.

Korn and Graubard (1998) proposed and evaluated the performance of a modification to this formulation to make it applicable for a proportion estimated from a complex sample. This adjustment involves estimating a degrees-of-freedom adjusted effective sample size, n^*_{df} and using that in place of the sample size, n. Specifically, the degrees-of-freedom adjusted effective sample size is defined by

$$n^*_{df} = \frac{\hat{p}(1-\hat{p})}{\widehat{var}(\hat{p})}\left(\frac{t_{n-1}(1-\alpha/2)}{t_d(1-\alpha/2)}\right)^2$$

where d is the number of Primary Sampling Units (PSUs) minus the number of strata in the complex sample design. Because n^*_{df} is undefined when $\hat{p}$=0 or $\hat{p}$=1.0, n^*_{df} is set to n in these situations reducing the computation to that of the simple random sample.

Several other methods have been proposed and evaluated for a variety of complex sample designs. These methods include, but are not limited to

- Poisson (Breeze) approach,
- Logit transformation,
- Binomial approach,
- Ad-hoc Quadratic/Wilson method,
- Andersson-Nermon method,
- Model-based Wilson method,
- T-adjusted Andersson_Nerman method

The modification to the exact binomial method described above in this paper, often referred to as the binomial approach, has been shown to perform well in most situations and generally provides the expected over-coverage. In addition, the method's similarity to exact binomial methods applicable to simple random samples makes it attractive for performance audits because audit sampling guidance and standards already use such

sufficient or appropriate, an inadequate audit process, or intentional omissions or misleading information because of misrepresentation or fraud.

methods. However, prior studies have suggested there is a risk of over-estimating n^*_{df} under certain outcomes for complex samples that may lead to under-coverage when using this or other similar approaches. As a result, authors have generally suggesting putting a restriction on n^*_{df} such that $n^*_{df} \leq n$. Further, Korn and Graubard highlighted that coverage probabilities can be reduced under certain conditions when $\hat{p}$=0 or $\hat{p}$=1.0 for a particular sample result and there is complete separation within the complex sample design (i.e. perfect homogeneity within clusters).

> "An example with a more serious lack of coverage can also easily be constructed: Suppose that the population consists of clusters of size 100, and that 10% of the clusters have all positive units and the remaining 90% have all zero units. If we sample 10 clusters as a simple random sample, and subsample all units in the sampled clusters, the 35% (=$(1-.1)^{10}$) of the time we will observe no positive units in the sample size of 1000. In this situation, our proposed intervals reduce to the usual binomial ones, so that, e.g., the upper 95% confidence limit for the population proportion is given by .003 (=$1-.05^{1/1000}$). This implies that the upper 95% confidence interval is less that(n) the true value of .10 at least 35% of the time, a serious undercoverage."

2. Confidence Interval Methods Applied by GAO for Performance Audits

In most commonly used statistical software packages, the default methods for estimating CIs rely on the normal approximation[2]. As discussed above, coverage probabilities for this method fall below the desired 1-α level when the sample size is small or the proportions are at or near the boundaries (0 or 1.0). The software packages allow alternative CI estimation methods to be used based on a pre-determined minimum sample size or proportion. However, according to Cochran (1977), the applicability of the normal approximation is a function of both the sample size and the proportion, as shown in Table 1.

Table 1: Smallest Values of np for Use of the Normal Approximation: Cochran (1977)

p	np = Number Observed in the Smaller Class	N = Sample Size
0.5	15	30
0.4	20	50
0.3	24	80
0.2	40	200
0.1	60	600
0.05	70	1400
~0*	80	∞

*This means that p is extremely small, so that np follows the Poisson distribution.

Performance audits often require the estimation of a large number of attributes and CIs to support audit findings. The sample sizes and proportions vary from one attribute to another and it is desirable to have an automated decision rule rather than to choose a CI method one attribute at a time. Rather than basing this rule on either a minimum sample size or a minimum proportion, as the software easily allows, GAO statisticians developed

[2] For this paper we refer to methods using the t-distribution as the normal approximation methods.

an algorithm to extrapolate the smallest values for np given in Table 1 as a decision rule to identify the most appropriate CI estimation method for each attribute being estimated.

In order to allow for estimates derived from complex sample designs, the decision rule algorithm first uses the statistical software to estimate the attribute ($\hat{p}$) and the design-based variance of the attribute ($\widehat{var}(\hat{p}_{Design-Based})$). Next, the algorithm computes the sample size, the number of PSUs and the number of strata individually for each attribute to be estimated[3]. It then estimates the degrees-of-freedom adjusted effective sample size as follows.

$$n_{df}^{*} = \frac{\hat{p}(1-\hat{p})}{\widehat{var}(\hat{p}_{Design-Based})}\left(\frac{t_{n-1}(1-\alpha/2)}{t_{d}(1-\alpha/2)}\right)^{2}$$

The degrees-of-freedom adjusted effective sample size is then used in conjunction with the estimated attribute ($\hat{p}$) to compare to the extrapolated minimum values of np to choose between methods using the normal approximation and the binomial methods described by Korn and Gaubard. Finally, we place two additional restrictions on the decision rule. First, any estimated attribute that is less than 0.05 or greater than 0.95 will always use the binomial methods. Second, and as recommended by the literature, we restrict n_{df}^{*} such that $n_{df}^{*} \leq n$.

To describe the locations of the CI methods this algorithm chooses, we plotted the resulting decision for all possible proportion estimates (0 to 1.0) and degrees-of-freedom adjusted effective sample size ranging from 10 to 1,000. The results of this are shown in Figure 1.

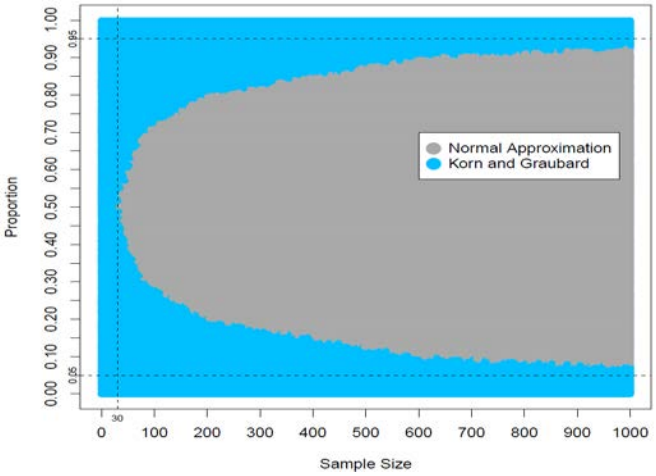


Figure 1: Location of Confidence Interval Method Based on CI Decision Rule

We have developed three functions to apply this method using various software packages. Each of the functions allows for estimation of a large number of attributes at a time and chooses a CI method individually for each possible outcome for each attribute.

[3] This step is necessary to account for subpopulation estimates of various sizes and missing observations for particular attributes.

3. Study Objectives

In this paper, we address the following two study objectives:

1. Assess how the automated CI decision rule affects coverage probabilities across a range of sample designs and population proportions.
2. Examine potential coverage probability reductions in complex sample designs by relaxing the restriction on n_{df}^{*} such that $n_{df}^{*} \leq n$.

4. Simulations

To evaluate the performance of the GAO CI decision rule method, we designed a series of simulations to estimate the coverage probabilities under various conditions. We generated estimates of coverage probabilities from four commonly used sample designs for a variety of sample sizes and attributes (errors) that cover the range of population proportion values (0 to 1.0). For each sample design, sample size and population proportion value, we replicated sample selection and estimation 1,000 times. We estimated CIs at the 95 percent confidence level and compared coverage probabilities obtained from three CI methods (Normal, Korn & Graubard, and the CI decision rule). Additionally, we estimated coverage probabilities using the CI decision rule while relaxing the restriction on n_{df}^{*} such that $n_{df}^{*} \leq n$.

4.1 Simulated Populations & Sample Designs

Design 1: Nominal Simple Random Sample. We generated a population of 20,000 transactions and randomly distributed errors for the 101 attributes (P000 – P100). For this population, we selected 1,000 replicate simple random samples for each of 50 sample sizes (n=20 to 1,000 by 20s).

Design 2: Inefficient Stratified Random Sample. We generated a population of 20,000 transactions with imbalanced stratification by assigning 5 percent of the transactions to stratum 1 (N1=1,000) and the remainder to stratum 2 (N2=19,000). We randomly distributed errors for the 101 attributes (P000 – P100) across the full population such that the expected proportions within each stratum would be equal. For this population, we selected 1,000 replicate stratified random samples for each of 50 sample sizes (n=20 to 1,000 by 20s) equally allocated across the two strata. The expected inefficiency of this design was introduced by the equal allocation to the imbalanced strata.

Design 3: Efficient Stratified Random Sample. We generated a stratified population of 20,000 transactions by dividing the population into two equal-sized strata (N=10,000 each). We distributed errors for the 101 attributes (P000 – P100) by placing 90 percent of errors for each attribute in stratum 2 and 10 percent in stratum 1 until stratum 2 contained transactions that were all assigned as errors. Following that, the remainder of the errors was distributed into stratum 1 until both strata were at 100% errors for attribute P100. For this population, we selected 1,000 replicate stratified random samples for each of 50 sample sizes (n=20 to 1,000 by 20s) proportionally allocated across the two strata. The expected efficiency of this design was introduced by the distribution of the errors in the population creating homogeneity within strata. Table 2 provides a description of the assignment of errors for four attributes included in the simulation.

Table 2: Assignment of Error Rates Within Strata for Design 3

Stratum		Errors Assigned (rate)			
	N	P025	P050	P075	P100
Stratum 1	10,000	500 (.05)	1,000 (.10)	5,000 (.50)	10,000 (1.0)
Stratum 2	10,000	4,500 (.45)	9,000 (.90)	10,000 (1.0)	10,000 (1.0)
Total	20,000	5,000 (.25)	10,000 (.50)	15,000 (.75)	20,000 (1.0)

Design 4: Inefficient 2-Stage Cluster Sample. We generated a population of 20,000 transactions clustered into 200 equal-sized primary sampling units (PSUs) with 100 transactions each. We distributed errors for the 101 attributes (P000 – P100) to maximize homogeneity within PSUs and heterogeneity between PSUs for each attribute. We achieved this first by splitting PSUs into two groups, then randomly placing 90 percent of errors in the second group of PSUs and 10 percent in the first until all PSUs in the second group were at 100% errors. Following that, we placed the remainder of the errors into PSUs in the first group until all clusters were at 100% errors. For this population, we selected 1,000 replicate 2-stage random cluster samples for each of 50 PSU sample sizes (n=2 to 100 by 2s) with an SSU sample size of 10 transactions selected from each PSU. The expected inefficiency of this design was introduced by the cluster design and the distribution of the errors in the PSUs creating homogeneity within clusters.

To summarize the expected performance of the four populations and sample designs, we approximated the design effects for five population proportions (P005, P025, P050, P075 and P095) for samples of 100 transactions selected from each population with the appropriate sample designs. The results of these approximations are given in Table 3.

Table 3: Approximate Design Effects for P005, P025, P050, P075 and P095 for Sample Sizes of 100

Population/Sample Design	P005	P025	P050	P075	P095
1. Nominal SRS	1.00	1.00	1.00	1.00	1.00
2. Inefficient Stratified	1.64	1.41	1.35	1.40	1.35
3. Efficient Stratified	0.98	0.92	0.63	0.88	0.99
4. Inefficient Cluster	1.31	2.93	6.78	4.04	1.49

5. Coverage Probability Results

For each population and sample design, we estimated coverage probabilities obtained from three CI methods (Normal, Korn & Graubard, and the CI decision rule) for 101 population proportions and 50 sample sizes. This created 5,050 estimates of coverage probabilities for each method and design. Given the number of estimates generated, we chose to summarize the results by creating heat maps to show the locations of over or under-coverage. The heat maps for each design show the results for population proportions ranging from 0 to 1.0 (y-axis), sample sizes ranging from 20 to 1,000 (x-axis) and coverage probabilities color-code as defined in Table 4. We generated three heat maps for each design included in our simulation, one heat map for each CI method. Each heat map represents more than 5 million replications. We then used the heat maps to compare areas of over or under-coverage within and between designs.

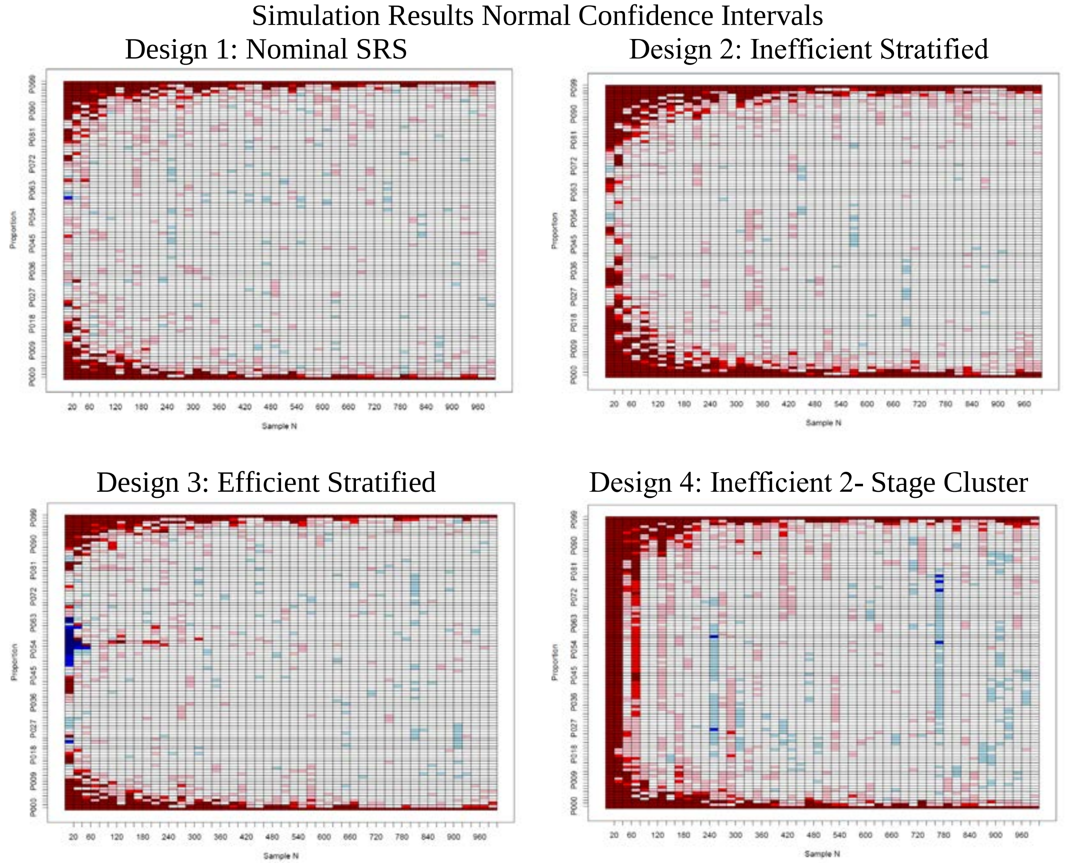
Table 4: Heat Map Display of Estimated Coverage Probability (CP)

CP < 91%	Dark Red	Significant under-coverage
91 <= CP < .92	Red	Under-coverage
.92 <= CP < .9365	Pink	Slight under-coverage
.9365 <= CP < .9635	White	Expected-coverage*
.9635 <= CP < .975	Light Blue	Slight over-coverage
.975 <= CP < .99	Blue	Over-coverage
.99 <= CP	Dark Blue	Significant over-coverage

*Because we limited the number of replications to 1,000 within each cell, the expected coverage range was defined by the 95% CI for a proportion estimate of 95% with a sample size of 1,000.

5.1 Coverage Probability Results – Normal CIs

Across all four designs in our simulation, our results show, as expected that the performance of the normal CIs fail to achieve the desired 95 percent coverage probability when the sample sizes are small and the proportion estimate is at or near the boundaries (0 or 1.0). The results also demonstrate that this failure to achieve desired coverage is a function of both the sample size and the population proportion, as suggested by Cochran. Figure 2 presents heat maps for each design resulting from using the normal CI methods for all values of $\hat{p}$ and n^*_{df}.


Simulation Results Normal Confidence Intervals

Design 1: Nominal SRS

Design 2: Inefficient Stratified

Design 3: Efficient Stratified

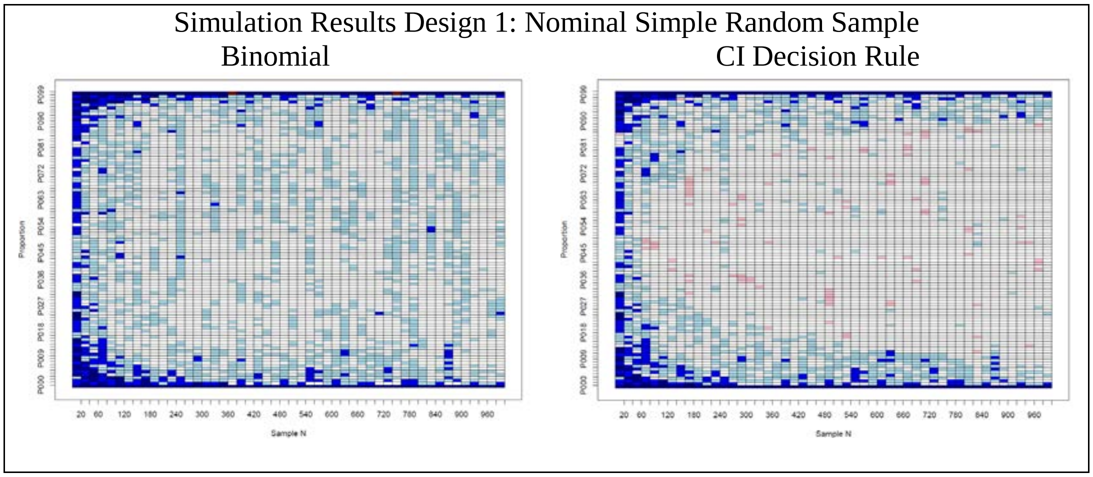
Design 4: Inefficient 2- Stage Cluster

Figure 2: Heat Maps of Simulation Results Using Normal Confidence Intervals

5.2 Coverage Probability Results – Binomial vs. CI Decision Rule

The simulation results when using the binomial approach for all values of $\hat{p}$ and n_{df}^{*} show the expected general over-coverage of the binomial Cis. Results for the CI decision rule show a reduction in the over-coverage in the middle of the distribution and when the sample size is large. Figures 3 through 6, show the comparison of the binomial approach to the CI decision rule for the four designs in our simulation.

The estimated coverage probabilities for Design 1, the nominal simple random sample, show the expected over-coverage of the binomial approach across the range of values of $\hat{p}$ and n_{df}^{*}. Additionally, results for the CI decision rule method show a reduction in the over-coverage in areas where the normal CI methods were used. Areas of slight over coverage (light blue) and slight under-coverage (pink) in the heat map may be a result of having a limited number of replications within each cell.


Simulation Results Design 1: Nominal Simple Random Sample

Binomial

CI Decision Rule

Figure 3: Design 1: Nominal Simple Random Sample - Estimated Coverage Probabilities Comparison between the Binomial and CI Decision Rule Methods

The estimated coverage probabilities for Design 2, the inefficient stratified random sample, show similar results as Design 1 but also show areas of significant under-coverage resulting from the binomial approach. We examined the simulation results and found that this under-coverage was occurring when the estimated $\hat{p}$ for an individual replicate sample was either 0 or 1 and the binomial approach defaulted to CI computations equivalent to a simple random sample. In other words, because the estimate of $\widehat{var}(\hat{p}_{Design-Based})$ is 0 for these outcomes, n_{df}^{*} is set to n for the computation of the CI. In an inefficient sample design, such as this one, that results in CIs that are generally underestimated because the additional variability induced by the sample design has not been incorporated. This is further evidence of the warning provided by Korn and Graubard discussed above and highlights specific areas of concern for performance audit samples because they are often designed to include over-sampling of higher dollar transactions or specific subpopulation of interest to the audit.

Simulation Results Design 2: Inefficient Stratified Random Sample

Binomial CI Decision Rule

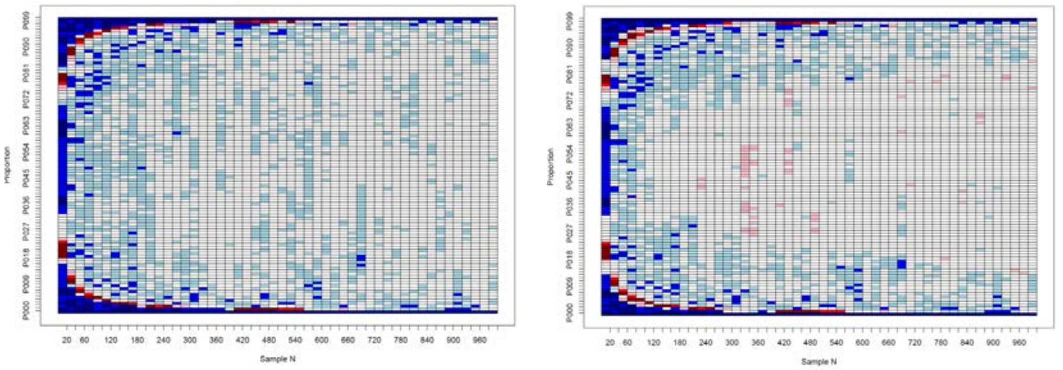


Figure 4: Design 2: Inefficient Stratified Random Sample - Estimated Coverage Probabilities Comparison between the Binomial and CI Decision Rule Methods

The estimated coverage probabilities for Design 3, the efficient stratified random sample, show significant over-coverage for the binomial approach across the range of values of $\hat{p}$ and n^*_{df}, particularly in areas near the middle of the distribution of the population proportion. This is a result of the efficiency of the sample design being maximized in these locations and the restriction of n^*_{df} such that $n^*_{df} \leq n$. In these cases, the binomial approach does not account for the gains in efficiency and generally over-estimate the CI. Results for the CI decision rule show the advantage of choosing the normal CI methods, when appropriate, for efficient sample designs.

Additionally, we observed areas of significant under-coverage or the CI decision rule method when the results for a particular sample replicate yielded perfect homogeneity within strata. Under these conditions, the estimate of $\widehat{var}(\hat{p}_{Design-Based})$ is 0 and n^*_{df} is undefined. The current CI decision rule results in a null CI for these cases and represents the need for an additional rule to apply the binomial methods when this occurs. That does not, however, pose additional risk to the audit in practice because the CIs are always manually reviewed, and these cases would be identified prior to drawing conclusions based on the audit sample results.

Simulation Results Design 3: Efficient Stratified Random Sample

Binomial CI Decision Rule

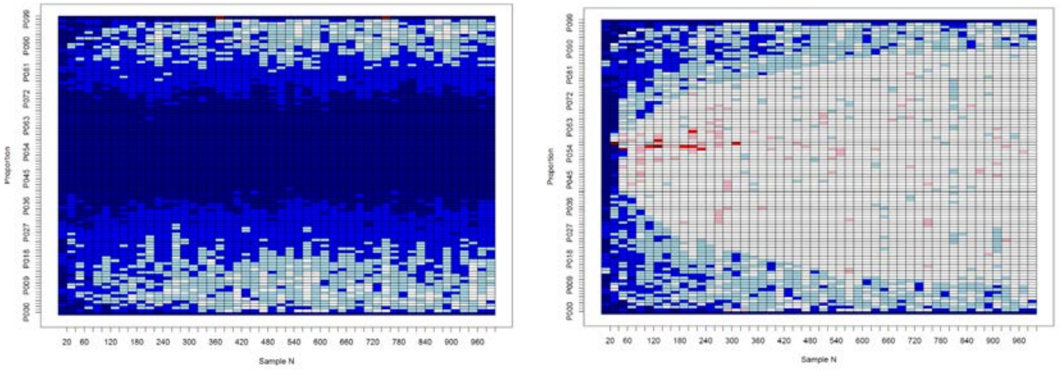
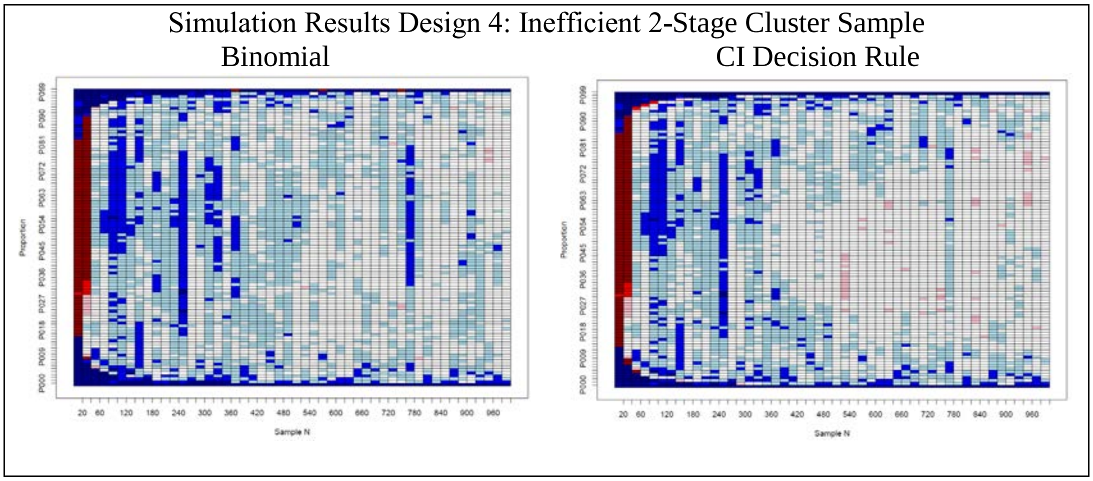


Figure 5: Design 3: Efficient Stratified Random Sample - Estimated Coverage Probabilities Comparison between the Binomial and CI Decision Rule Methods

We observed similar results for Design 4, the inefficient 2-stage cluster sample. In general, using the binomial approach across the range of values of $\hat{p}$ and n_{df}^* shows the expected over-coverage and the use of the CI decision rule reduces the over-coverage when the normal CI methods are appropriate. The areas of significant under-coverage shown in Figure 6 are a result of the small number of PSUs selected at the first stage.


Simulation Results Design 4: Inefficient 2-Stage Cluster Sample

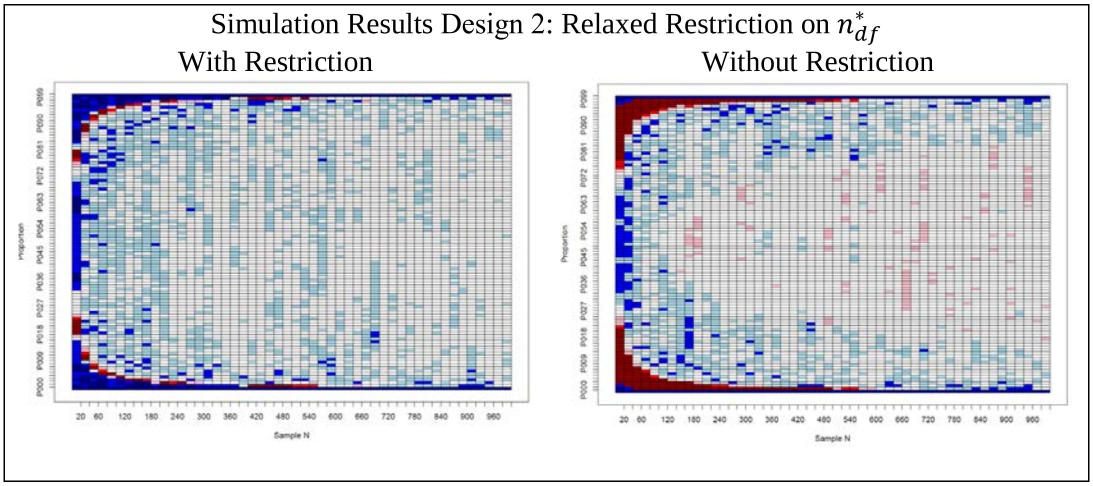
Binomial

CI Decision Rule

Figure 6: Design 4: Inefficient 2-Stage Cluster Sample - Estimated Coverage Probabilities Comparison between the Binomial and CI Decision Rule Methods

5.3 Coverage Probability Results – Relaxed Restriction on n_{df}^*

To test the need to restrict n_{df}^* such that $n_{df}^* \leq n$ due to the possibility of over-estimation of n_{df}^* at or near the boundaries, we relaxed this restriction and applied the CI decision rule methods to Design 2, the inefficient stratified random sample. The results clearly show an increase in the significant under-coverage at or near the boundaries and provide an indication that the restriction is needed in practice.


Simulation Results Design 2: Relaxed Restriction on n_{df}^*

With Restriction

Without Restriction

Figure 7: Estimated Coverage Probabilities for CIs without Restricting n_{df}^* <= n

6. Conclusions and Future Research

Based on this evaluation of estimated coverage probabilities, we concluded that the CI decision rule GAO has implemented works well in most situations. The ability to choose an appropriate CI method automatically provides significant advantages when generating large numbers of proportion estimates in support of performance audits. Furthermore, this evaluation shows that the restriction of n_{df}^{*} such that $n_{df}^{*} \leq n$ is necessary.

We also identified the need for caution in the interpretation of sample results of 0 or 1 when using these methods with inefficient sample designs. As discussed above, these methods default to those applicable to simple random samples under these conditions resulting in CIs that do not account for the additional variation induced by the inefficient design. This results in additional and unmeasured audit risk and further research in this area would be beneficial.

Acknowledgements

We would like to acknowledge the work of Mark Ramage, a long time statistician at GAO who retired in 2015. Mark coded and implemented the CI decision rule methods and introduced it into practice among GAO statisticians.

References

Agresti, A., and Coull, B.A. (1998). “Approximate is better than "exact" for interval estimation of binomial proportions,” The American Statistician, 52, 119-126.

Clopper, C.J., and Pearson, E.S. (1934), “The use of confidence or fiducial limits illustrated in the case of the binomial,” Biometrika, 26, 404-413.

Cochran, W.G. (1977), “Sampling Techniques,” Third Edition, New York : Wiley.

Kott, P. S., Andersson, P.G., and Nerman, O. (2001), “Two-sided Confidence Intervals for Small Proportion Based on Survey Data,” Paper presented at Federal Committee on Statistical Methodology, Washington, D.C.

Korn, E.L., and Graubard, B.I. (1998), “Confidence Intervals for Proportions With Small Expected Number of Positive Counts Estimated From Survey Data,” Survey Methodology, 24, 193–201.

Sukasih, A., and Jang, D. (2016), “An Application of Confidence Interval Methods for Small Proportions in the Health Care Survey of DoD Beneficiaries,” American Statistical Association Section on Survey Research Methods.

Thulin, M. (2014), “The cost of using exact confidence intervals for a binomial proportion,” Electronic Journal of Statistics, 8, 817-840.