Sample Design for Audit Populations

Michelle Norris *
Department of Mathematics and Statistics
California State University, Sacramento
February 13, 2018

Abstract

We develop several tools for the determination of sample size and design for Medicaid audits. This audit setting involves a population of claims for reimbursement by a healthcare provider which need to be reviewed by an auditor to determine the correct amount for each claim. The existing literature regarding sample planning for audits is incomplete and often includes restrictive assumptions. To fill these gaps, we exploit the special relationship between the known claim amounts and the unknown post-audit amounts. We propose a hypergeometric generative process for audit populations which we use to derive estimators of variances needed for sample size determination. We further develop a criterion for choosing between simple expansion and ratio estimation and an efficient method for determining exact optimal strata breakpoints in populations with repeated values. We also derive a variance estimator under a more general “partial error” model than previous researches have used. These tools apply more generally to audits where an overstated book/claim amount is the primary concern and estimation of the total dollar value of the claim errors is the goal. The sample design methods we develop are illustrated on two simulated audit populations.

Keywords: Medicaid, sampling, ratio estimator, hypergeometric model, stratification

*Michelle Norris is Associate Professor, Department of Mathematics and Statistics, California State University, Sacramento, 6000 J Street, Sacramento, CA 95819 (e-mail: norris@csus.edu). The author thanks Martin Hauser for his careful proofreading of the manuscript and several helpful discussions.
1 Background and Motivation

According to the Medicaid program website (Centers for Medicare and Medicaid Services (2018a)),

“Medicaid provides health coverage to millions of Americans, including eligible low-income adults, children, pregnant women, elderly adults and people with disabilities...The program is funded jointly by states and the federal government.”

In 2016, $566 billion in Medicaid payments were disbursed to healthcare providers such as pharmacies, medical offices, and school districts in the US (Centers for Medicare and Medicaid Services (2018b)). In California, MediCal is the name for the Medicaid program, and the California State Controller’s Office is charged with conducting audits to ensure that MediCal funds paid to organizations conform to the requirements of the MediCal program and are of the appropriate amount.

In planning a MediCal audit, auditors typically have access to a population of MediCal claims which they are charged with auditing for correctness. For example, if the organization is a medical clinic, a single claim may represent a single visit by a single patient, and the population may contain a million claims from a three-year period. The population may account for tens of millions of dollars in disbursed MediCal payments. Because a complete examination of all claims is not feasible, auditors typically select a sample of claims, then, based on documentation, determine the appropriate amount of MediCal reimbursement that should have been paid for each claim in the sample. There are three possible outcomes for each sampled/audited claim:

1. None of the amount claimed is disallowed, and the entire claimed amount is deemed allowable for reimbursement (as shown in lines 1 and 5 in Table 2.
Table 1: Portion of Hypothetical Data for a MediCal Audit

| Line | Patient ID | Date of Service | Claimed Amount | Disallowed/Error Amount |
|------|------------|-----------------|----------------|-------------------------|
| 1    | 33457      | Jan 15, 2017    | $52.50         | $0                      |
| 2    | 31415      | March 10, 2017  | $78.90         | $30.00                  |
| 3    | 44478      | Oct 27, 2016    | $25.90         | $25.90                  |
| 4    | 67841      | May 5, 2016     | $105.00        | $50.00                  |
| 5    | 55112      | Nov 20, 2016    | $125.00        | $0                      |
| 6    | 98765      | May 1, 2016     | $66.00         | $66.00                  |

1. The entire amount claimed is deemed disallowed, and none is deemed allowable for reimbursement (lines 3 and 6 in Table 1).
2. A portion of the total amount claimed is deemed disallowed and only the remaining portion is allowable for reimbursement (lines 2 and 4 in Table 1). This case is also called a partial payment or partial error.

While the claim amounts are known for the entire population prior to an audit, the disallowed amounts are only known after the audit and only for the sampled claims. We will use both the terms ‘disallowed amount’ and ‘error amount’ to refer to the portion of a claim total that is not allowable for reimbursement.

The total disallowed amount found in the sample is extrapolated from the sample to the population, and the organization is required to pay that amount or some related amount back to the MediCal fund. Clearly, maintaining a small margin of error in estimating the total disallowed amount is of interest to all parties. Thus, it is important to design audit samples which estimate the total disallowed amount with a reasonable margin of error while minimizing the sample size. In addition, since a pilot sample is typically an inconvenience to the organization being audited, audit samples must frequently be designed with
little to no information about the population of disallowed amounts – making it difficult to determine sample size and efficient stratification.

To date, there are few results in the literature pertaining to sample planning for audit populations. In this paper, we will aim to fill some of the gaps and reconcile some differing results in the literature on sample planning for audit populations. Perhaps the most comprehensive treatment on this topic is contained in the book *Statistical Auditing* by Roberts (1978). In particular, Roberts derives estimates of the population variances needed for sample size determination under both simple expansion and ratio estimation. His estimates do not require data from pilot samples. However, they do require estimating the ‘error rate,’ defined as the proportion of claims in the population containing some error amount or disallowed amount. He uses a Bernoulli generative model to derive his estimates. King & Madansky (2013) also propose a Bernoulli model to estimate the variance of the disallowed amounts under simple expansion but arrive at a slightly different estimate. In this paper, we propose a third method of estimating the variance and reconcile the three different estimators. In addition, since all currently available methods of determining sample size depend on estimating the error rate or the variance of the population of disallowed values, we also propose a method for determining a conservative sample size which is based solely on the claimed values and does not require any information about the population of disallowed values.

We also consider the question of choosing between the simple expansion and ratio estimators in simple random sampling. The general advice on p.157 of Cochran (1977) is to use the ratio estimator instead of simple expansion when
$X$, the claim amount, and $Y$, the disallowed amount, satisfy:

$$\frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y} > 1 \frac{\sigma_x}{\mu_x} \frac{\sigma_y}{\mu_y} \tag{1}$$

We specialize this inequality to the audit population setting, and derive a formula for the probability that ratio estimation will outperform simple expansion. Since our formula only relies on the error rate and parameters for the claim population, it can be used for sample planning prior to collecting any information about the population of disallowed amounts.

A common assumption in existing literature on audit sample design is the ‘all-or-nothing error assumption’ which states that the error/disallowed amount in a claim equals the entire claim amount or zero. The all-or-nothing error assumption precludes the possibility of partial errors but greatly simplifies theoretic calculations. Realistically, partial errors do occur in some audit populations so generalizing existing results and deriving new results that apply to more general error models is desirable. We note that Neter & Loebbecke (1977) do consider more general error models in their empirical study, but our research has not revealed any theoretic work on sample design under more general error models. The methods proposed by Roberts (1978) for sample size determination under simple expansion and ratio estimation assume all-or-nothing errors. We extend our method for sample size determination to a population generative model which allows for partial errors where all partial errors are the same proportion of the claim amount. We prove that, under this simple partial error model, the sample sizes determined will be less than or equal to the conservative sample size for an all-or-nothing error population.

For some audit populations, stratification can significantly reduce the sample
| Population Size | Number of Unique Values |
|-----------------|-------------------------|
| 21,000          | 100                     |
| 77,000          | 500                     |
| 1,700,000       | 4000                    |

Table 2: Population Size and Number of Unique Values in MediCal Populations

size needed to attain a desired margin of error (compared to simple random sampling). Often, stratification of audit populations is based on the claim amount; the Cum \( \sqrt{J} \) Rule of Dalenius & Hodges Jr (1959) is used to determine strata boundaries; and optimal allocation is used to allocate the total sample to strata (see Neter & Loebbecke (1977) and Buddhakulsomsiri & Parthanadee (2008)). However, there is no guarantee that this method will produce the optimal stratification since the Cum \( \sqrt{J} \) Rule relies on the assumption that the population is uniformly distributed within each stratum. In the all-or-nothing errors case and assuming that errors are independent of claim amount, even if the strata are formed so that the claim amounts are roughly uniformly distributed in each strata, the disallowed amounts will not be. In the best case, the disallowed amounts in a strata will consists of a point mass at zero and a roughly uniform distribution of the amounts for claims that are fully in error.

In the audit populations we have had the opportunity to work with, we have observed that even though the population may be very large, some values are repeated many times, making the number of distinct values orders of magnitude smaller than the population size as shown in the Table 2. We prove a theorem with states that when stratifying by claim amount, there can be at most four candidates for optimal strata breakpoints within a run of repeated values in an audit population with a fixed error rate. This makes it more feasible to find true optimal breakpoints through a complete search when the number of strata and number of unique values in a population is not too large and an estimate
of the error rate is available. King & Madansky (2013) have studied the gain in stratification assuming the audit population follows a gamma distribution. Assuming all-or-nothing errors, they study the gain under a two-strata sampling plan with proportional allocation under three different models for the probability a claim is in error. Their findings suggest that the gain from stratification with proportional allocation and optimal breakpoint choice is most marked in high error rate scenarios.
2 Notation and Estimators

We now summarize two common estimators used to extrapolate the total disallowed amount in an audit. We start with some notation:

\[ N = \text{the population size} \]
\[ \{x_1, x_2, \ldots, x_N\} = \text{the population of known claimed amounts} \]
\[ \{y_1, y_2, \ldots, y_N\} = \text{the population of unknown disallowed/error amounts} \]
\[ \tau_x = \sum_{i=1}^{N} x_i \]
\[ \tau_y = \sum_{i=1}^{N} y_i \]
\[ R = \frac{\tau_y}{\tau_x} \]
\[ \mu_x = \frac{1}{N} \tau_x \]
\[ \mu_y = \frac{1}{N} \tau_y \]
\[ \sigma^2_x = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_x)^2 \]
\[ \sigma^2_y = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu_y)^2 \]

(2)
\( n \) = the sample size  
\( \{x(1), x(2), \ldots, x(n)\} \) = a sample random sample of claims without replacement  
\( \{y(1), y(2), \ldots, y(n)\} \) = disallowed values corresponding to sampled claims  
\( \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y(i) \)  
\( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x(i) \)  
\( \hat{r} = \frac{\bar{y}}{\bar{x}} \)  
\( \hat{\tau}_{se} = N \bar{y} = \) the simple expansion estimator of total error amount  
\( \hat{\tau}_{r} = \hat{r} \tau_x = \) the ratio estimator of total error amount  
\( \sigma^2_{\tau_{se}} = N^2 \cdot \frac{\sigma^2_y}{n} \cdot \frac{N - n}{N - 1} \)  
\( \sigma^2_{\tau_{r}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - Rx_i)^2 \)  
\( \sigma^2_{\tau_{r}} = N^2 \cdot \frac{\sigma^2_R}{n} \cdot \frac{N - n}{N - 1} \)  

The rest of this paper is organized as follows. In Section 3, we give the sample size formula of interest. In Section 4, we discuss issues pertaining to sample size determination under the simple expansion estimator. In particular, we propose a generative model for audit populations; propose an estimator of the variance needed for sample size calculations under an all-or-nothing error model and compare it to existing estimators; extend the procedure for estimating variance to a population with partial payments; and prove a computational shortcut for determining optimal strata boundaries. In Section 5, we consider the ratio estimator. We start with a criteria for deciding between simple expansion and ratio estimation; continue by developing a estimator for the variance needed in sample determination under ratio estimation; and finish with com-
ments about stratification and partial errors under ratio estimation. In Section 6, we apply the sample design tools developed in this paper to two audit populations simulated to reflect actual audit populations in the literature. We offer some concluding remarks and avenues for further research in Section 7.

3 Sample Size Formula

Under simple expansion, the \((1 - \alpha) \times 100\%\) large-sample confidence level margin of error of \(\hat{\tau}_{se}\) is

\[
E = z_{(1 - \frac{\alpha}{2})} \cdot \sigma_{\hat{\tau}_{se}}
\]

where \(z_p\) denotes the \(p\)th percentile of the standard normal distribution. Substituting equation (3) into equation (4) and solving for \(n\), we obtain the following sample size formula under simple expansion

\[
n = \frac{z^2_{(1 - \frac{\alpha}{2})} \cdot N^3 \sigma_y^2}{E^2(N - 1) + z^2_{(1 - \frac{\alpha}{2})} N^2 \sigma_y^2}
\]

The sample size formula will give the sample size required to attain a chosen margin of error and confidence level provided that the variance of disallowed amounts, \(\sigma_y^2\), is known. However, \(\sigma_y^2\) is typically not known in the planning stages of an audit. One could obtain an estimate of \(\sigma_y^2\) using a pilot sample, but this is an inconvenience to an audited organization since they would have to pull records twice – once for the pilot sample and again for the actual full audit. In the next section, we propose a generative model for audit populations which permits estimation of \(\sigma_y^2\) in cases where the error rate can be approximated. Under ratio estimation, the sample size formula is equation (5) with \(\sigma_R^2\) substituted for \(\sigma_y^2\). We propose methods for estimating \(\sigma_R^2\) during the planning.
stages of an audit in Section 5.2.

4 The Model

To determine variation in the disallowed amounts based on the known claimed amounts, we now introduce a simple model for audit results which assumes that the audit of any single MediCal claim gives an all-or-nothing error – that is, that the claim payment is either entirely allowable or entirely disallowable. Thus, payments that are partially allowed would not occur in this simplified case. Later, we will generalize the results to a partial payment scenario. This simplified model also assumes that the probability a claimed payment is disallowed is independent of the amount of the claim. We use the following notation:

\[ \pi = \text{the population proportion of error claims}, \quad \pi \in \{0, \frac{1}{N}, \frac{2}{N}, ..., \frac{N-1}{N}, 1\} \]

\[ N_e = \pi \cdot N = \text{the number of error claims in the population} \]

\( \pi \) represents the “error rate” of the audit population as defined in Section 1. We can conceptualize an actual audit population with error rate \( \pi \) as having been generated by choosing a simple random sample without replacement (SR-SWOR) of size \( N_e \) from the \( N \) claims in the population and designating these as the errors. Let the random vector \( \vec{X}_s = \{X_{(1)}, X_{(2)}, ..., X_{(N_e)}\} \) denote the SR-SWOR of size \( N_e \) from the population of claimed amounts \( U = \{x_1, x_2, ..., x_N\} \). An observed value of \( \vec{X}_s \) will be denoted with lowercase letters, i.e. \( \vec{x}_s = \{x_{(1)}, x_{(2)}, ..., x_{(N_e)}\} \).

Since all errors are all-or-nothing errors, a random variable \( Y \) which is useful
for estimation of $\sigma_y^2$ is defined as follows:

$$Y = \begin{cases} 
X(i) & \text{with probability } \frac{1}{N} \text{ for } i = 1, 2, \ldots, N_e \\
0 & \text{with probability } 1 - \pi 
\end{cases}$$

(6)

$Y$ could be interpreted as a draw from a hypothetical audit population where the claimed amounts are fixed, but the group of $N_e$ claims which are in error has yet to be determined. Hence, the erroneous claims are denoted as random variables $X(i)$. Under the proposed model, probabilities concerning the $N_e$ claims which are chosen to be erroneous follow a hypergeometric distribution as opposed to the Bernoulli Distribution used in Roberts (1978) and King & Madansky (2013) and defined in equation (8). We argue that, at the time of an audit, a fixed number of errors already exist in the population, even though it is not known specifically which claims are in error. Hence, if the error rate is known, an audit population is like a coin that has already been tossed a finite number of times and is best modeled by the hypergeometric distributions as opposed to a coin that will be tossed a finite number of times and should be modeled by a Binomial distribution.

### 4.1 Three Potential Estimators of $\sigma_y^2$

We now identify and compare three potential estimators of $\sigma_y^2$. To formulate our proposed estimator of $\sigma_y^2$, we observe that the unknown audit population would be completely known if the realized vector of erroneous claims $\vec{X}_s$ were known. Suppose that $\vec{X}_s = \vec{x}_s^*$ resulted in the audit population, then

$$\mu_y = E(Y|\vec{X}_s = \vec{x}_s^*)$$

$$\sigma_y^2 = \text{Var}(Y|\vec{X}_s = \vec{x}_s^*)$$

(7)
But \( \bar{x}_s \) is unknown, so we cannot use equation (7) to determine \( \sigma^2_y \). We propose using the expected value of \( \text{Var}(Y|\bar{X}_s) \) given in Theorem 4.1c to estimate \( \sigma^2_y \) since it minimizes the mean square prediction error. Theorem 4.1 is proven in the Supplement to this paper.

**Theorem 4.1 (Conditional Expected Value and Variance)** Let \( \bar{X}_e = \frac{1}{N_e} \sum_{i=1}^{N_e} X_{(i)} \), \( \bar{X}_e^{(2)} = \frac{1}{N_e} \sum_{i=1}^{N_e} X_{(i)}^2 \) and \( \mu_{x}^{(2)} = \frac{1}{N} \sum_{i=1}^{N} x_i^2 \). Under the model for \( Y \) given in equation (6):

a. \( \text{E}(Y|\bar{X}_s) = \pi \bar{X}_e \)

b. \( \text{Var}(Y|\bar{X}_s) = \pi \bar{X}_e^{(2)} - (\pi \bar{X}_e)^2 \)

c. \( \text{E}(\text{Var}(Y|\bar{X}_s)) = \pi \mu_{x}^{(2)} - (\pi \mu_{x})^2 - \pi (1 - \pi) \frac{\sigma^2_x}{N-1} \)

Roberts (1978) derives a slightly different estimator of the variance \( \sigma^2_y \) by assuming the following Bernoulli generating process for the audit population. Letting \( X_i = \) the \( i \)th claim in the population, \( Y_i = \) the disallowed value of the \( i \)th claim for \( i = 1, 2, ..., N \) and \( 0 \leq \pi \leq 1 \), we assume the audit population is generated as follows:

\[
Y_i = \begin{cases} 
X_i & \text{with probability } \pi \\
0 & \text{with probability } 1 - \pi
\end{cases}
\]  
(8)

Using the notation of this paper, Roberts derives the following estimator of \( \sigma^2_y \):

\[
\hat{\sigma}^2_{(R,y)} = \text{E}_e \left[ \frac{1}{N} \sum_{i=1}^{N} Y_i^2 - \frac{1}{N^2} \left( \sum_{i=1}^{N} Y_i \right)^2 \right] = \pi \mu_{x}^{(2)} - (\pi \mu_{x})^2 - \pi (1 - \pi) \frac{\sigma^2_x + \mu_{x}^2}{N}
\]

Another option is to use the total variance of \( Y \), which can be found using iterated expectations, to estimate \( \sigma^2_y \). This result is derived in King & Madansky
(2013) using the same Bernoulli framework as Roberts (1978) and is also shown in the Supplement.

**Theorem 4.2 (Total Expected Value and Variance)** Under the model for $Y$ given in equation (8):

a. $E(Y) = \pi \mu_x$

b. $\text{Var}(Y) = \pi \cdot \mu_x^2 - (\pi \mu_x)^2$

The federal Office of Inspector General’s RAT-STATS software also uses the total variance of $Y$ to estimate $\sigma_y^2$ (RAT-STATs Companion Manual, Rev 5/2010, p. 4-9). The total variance, however, represents the variation in $Y$ as the audit population and the sample from it vary. We would argue, however, that the audit population is fixed but unknown so that including variation due to a varying audit population in our estimation of $\sigma_y^2$ is not conceptually satisfying.

The three potential estimators of $\sigma_y^2$ are related by the following inequality:

$$\widehat{\sigma_y^2} \leq E(\text{Var}(Y|\bar{X}_s)) \leq \text{Var}(Y)$$

However, if the population size, $N$ is large relative to $\sigma_x^2$ and $\mu_x^2$, the term $\pi(1-\pi)\frac{\sigma_x^2}{N}$ in Theorem 4.1; and the term $\pi(1-\pi)\frac{\sigma_x^2 + \mu_x^2}{N}$ in $\widehat{\sigma_y^2}$ will both be close to 0 so that all three estimators will be roughly equal. This is the case in several audit populations we have reviewed.

In cases where there is a difference in the estimate of $\sigma_y^2$ under the binomial and hypergeometric models, it could be argued that the binomial model would be easier to implement since it only requires estimation of the rate at which the process of preparing claims generates errors, as opposed to the hypergeometric model where an estimate of the exact proportion of errors in the realized audit
population is needed. In a more familiar context, this difference is similar to estimating the rate at which the process of tossing a fair coin will generate heads as 0.50 instead of trying to estimate the proportion of heads that will occur in 20 tosses. Admittedly, the binomial model allows for a variety of realized error rates which follow a binomial probability distribution. The constraint that the realized error rates follow binomial probabilities, however, seems unnecessarily restrictive. When there is uncertainty in the pre-audit estimate of the error rate, we would propose instead that the hypergeometric model be used and sample sizes be calculated under a range of feasible error rates and the largest resulting sample size be used. The advantage of the hypergeometric model is that sample sizes can be determined under explicit and flexible assumptions about the distribution of potential realized error rates.

4.1.1 Estimating \( \pi \)

The formula for \( E(\text{Var}(Y|\bar{X}_s)) \) in Theorem 4.1c only depends on the known population of claimed amounts and the error rate, \( \pi \). If an estimate of \( \pi \) is available from a past survey or a pilot survey, we can estimate \( \sigma^2_y \) using Theorem 4.1c., then substitute the result into equation (5) to determine the sample size needed to achieve a given margin of error and confidence level.

If an estimate of \( \pi \) is not available, we can obtain a conservative sample size by maximizing \( h(\pi) = E(\text{Var}(Y|\bar{X}_s)) = \pi \mu_x^{(2)} - (\pi \mu_x)^2 - \pi(1 - \pi)\frac{\sigma^2_x}{N - 1} \) as a function of \( \pi \). Taking the derivative of \( h(\pi) \) and setting it equal to 0 gives:

\[
  h'(\pi) = \mu_x^{(2)} - 2\pi \mu_x^2 - (1 - 2\pi)\frac{\sigma^2_x}{N - 1} = 0
\]
Solving equation (9), we obtain

$$\pi_{\text{crit}} = \frac{1}{2} \cdot \frac{\mu_x^{(2)} - \frac{\sigma_x^2}{N - 1}}{\mu_x^2 - \frac{\sigma_x^2}{N - 1}} \approx \frac{\mu_x^{(2)}}{2\mu_x^2}$$

(10)

In order to maximize \(h(\pi)\) over \(\pi \in [0, 1]\), we must check \(h(0), h(1)\) and \(h(\pi_{\text{crit}})\). Since \(h(0) = 0\) and \(h(1) = \sigma_x^2\), the maximum value of \(h(\pi)\) is \(h_{\text{max}} = \max\{\sigma_x^2, h(\pi_{\text{crit}})\}\). The sample size obtained by substituting \(h_{\text{max}}\) for \(\sigma_y^2\) in equation (5) will be the maximum sample size needed for a specified margin of error and confidence level over all possible error rates, \(\pi\).

### 4.2 Partial Payments

Thus far, we have considered a model with all-or-nothing errors. We now wish to consider sample size determination under simple expansion when there are partial payments in the population, i.e. only a portion of the amount claimed is deemed allowable and the remaining portion is disallowable. We will consider a population of claims where any partial payment is a fixed proportion, \(q\) with \(0 < q < 1\), of the claimed amount. Let \(\vec{X}_p = \{X_{(1)}, X_{(2)}, \ldots, X_{(p)}\}\) represent the vector of claims partially in error and \(\vec{X}_T = \{X_{(1)}, X_{(2)}, \ldots, X_{(T)}\}\) where \(T > p\) represent claims with any error (partial or full). Thus, \(\vec{X}_T \cap \vec{X}_p^{\mathcal{C}} = \{X_{(p+1)}, X_{(p+2)}, \ldots, X_{(T)}\}\) consists of the claims totally in error. Now, the random variable \(Y\) which represents a random draw from a random claim population
with partial payments is defined as:

\[
Y = \begin{cases} 
qX(i) & \text{with probability } \frac{1}{N} \text{ for } i = 1, 2, \ldots p \\
X(i) & \text{with probability } \frac{1}{N} \text{ for } i = p + 1, p + 2, \ldots T \\
0 & \text{with probability } 1 - \frac{T}{N}
\end{cases}
\]  

Letting \(\vec{x}_T^*\) and \(\vec{x}_p^*\) be the vectors of realized errors and partial errors, we need to estimate \(\sigma^2_y = \text{Var}(Y|\vec{x}_T^*, \vec{x}_p^*)\) in order to determine a sample size using (5). We recommend using \(E(\text{Var}(Y|\vec{x}_T^*, \vec{x}_p^*))\) to estimate \(\text{Var}(Y|\vec{x}_T^*, \vec{x}_p^*)\). In the Supplement, we show that

\[
E_{\vec{x}_T^*, \vec{x}_p^*}(\text{Var}(Y|\vec{x}_T^*, \vec{x}_p^*)) = \left[\pi_T - \pi_p(1 - q^2)\right]\mu_x^{(2)} - \left[\pi_T - \pi_p(1 - q)\right]^2\mu_x^2
- \frac{\sigma_x^2}{N-1}\left[\pi_p(1 - q)\right]^2(1 - \frac{p}{T})
+ (1 - \pi_T)(1 - \frac{p}{T}(1 - q))(\pi_T - \pi_p(1 - q))\tag{12}
\]

\[
< \left[\pi_T - \pi_p(1 - q^2)\right]\mu_x^{(2)} - \left[\pi_T - \pi_p(1 - q)\right]^2\mu_x^2\tag{13}
\]

where the upper bound in line (35) is obtained by dropping the second term in the right-hand side of equation (35) (which is negative). This bound will be sharp if \(\frac{\sigma_x^2}{N-1} \approx 0\) since the coefficient of \(\frac{\sigma_x^2}{N-1}\) is less than 2. This condition is frequently true in MediCal audits so the bound in line (35) will be useful for sample size determination provided estimates of \(\pi_T, \pi_p,\) and \(q\) are available. If
estimates of these quantities are not available, we note that

\[ E_{\bar{X}_T, \bar{X}_p} (\text{Var}(Y|\bar{X}_T, \bar{X}_p)) < [\pi_T - \pi_p(1 - q^2)] \mu_x^{(2)} - [\pi_T - \pi_p(1 - q)]^2 \mu_x^2 \]
\[ < [\pi_T - \pi_p(1 - q^2)] \mu_x^{(2)} - [\pi_T - \pi_p(1 - q^2)]^2 \mu_x^2 \]
\[ = \pi^* \mu_x^{(2)} - (\pi^*)^2 \mu_x^2 \]
\[ \text{where } \pi^* = \pi_T - \pi_p(1 - q^2) \] (14)

Taking the derivative and setting it equal to 0, we find that (14) is maximized at \( \pi^* = \frac{\mu_x^{(2)}}{2\mu_x^2} \) so that if no information is available about the audit population, a conservative sample size which will certainly be large enough, but may be larger than necessary, can be calculated by estimating \( \sigma_y^2 \) with formula (14) and \( \pi^* \). This is close to the estimate of the maximum of \( \sigma_y^2 \) in the all-or-nothing errors case. (It may be possible to maximize the expression in (35) over \( \pi_T, \pi_p, \) and \( q \), but thus far we have not been able to do so.) It remains an open question to obtain a useful estimator of the variance of \( Y \) under more general partial error models.

### 4.3 Stratified Sampling under Simple Expansion

It is often the case that stratified sampling reduces the total sample size needed to attain an estimate of \( \tau_y \) having a fixed margin of error (compared to simple random sampling). In stratified sampling, an auxiliary variable which is known to be highly correlated with the study variable is used to partition the population into \( L \) disjoint groups called strata. Then, a simple random sample is selected within each stratum, and, using the following notation, \( \tau_y \) is estimated
using formula (15).

\[ h = 1, 2, \ldots, L \] denotes the stratum

\[ N_h = \text{the number of population units in stratum } h \]

\[ x_{1h}, x_{2h}, \ldots, x_{Nh,h} = \text{denotes the claim amounts in stratum } h \]

\[ y_{1h}, y_{2h}, \ldots, y_{Nh,h} = \text{denotes the error/disallowed amounts in stratum } h \]

\[ n_h = \text{the sample size for stratum } h \]

\[ s_h \subseteq \{1, 2, \ldots, N_h\} \] denotes the subscripts of claims in the sample from stratum \( h \)

\[ \mu_{yh} = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{ih} \]

\[ \sigma_{yh}^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} (y_{ih} - \mu_{yh})^2 \]

\[ \bar{Y}_h = \frac{1}{n_h} \sum_{i \in s_h} y_{ih} \]

The stratified simple expansion estimator of \( \tau_y \) and its variance are:

\[
\hat{\tau}_{se,st} = \sum_{h=1}^{L} N_h \bar{Y}_h \quad (15)
\]

\[
\text{Var}(\hat{\tau}_{se,st}) = \sum_{h=1}^{L} N_h^2 \cdot \frac{\sigma_{yh}^2}{n_h} \cdot \frac{N_h - n_h}{N_h - 1} \quad (16)
\]

The factor \( \frac{N_h - n_h}{N_h - 1} \) in equation (16) is the finite population correction factor which is needed when sampling without replacement from a finite population and will be close to 1 if \( N_h \gg n_h \). Thus, ignoring the finite population correction factor,

\[
\text{Var}(\hat{\tau}_{se,st}) \approx \sum_{h=1}^{L} N_h^2 \cdot \frac{\sigma_{yh}^2}{n_h}
\]

19
The more homogeneous each stratum is relative to the study variable, the smaller the values of $\sigma^2_{y_h}$ and the greater the overall sample size reduction due to stratification (for a fixed precision level).

A number of methods for allocating the overall sample to the strata exist. These include proportional, equal, or optimal allocation. Optimal allocation is the allocation method that minimizes $\text{Var}(\hat{\tau}_{se,st})$ for fixed strata breakpoints and is given in equation (17). Under optimal allocation and letting $n$ be the overall sample size, the sample size in stratum $h$ and $\text{Var}(\hat{\tau}_{se,st})$ are:

$$ n_h = \frac{N_h \sigma_h}{\sum_{h=1}^{L} N_h \sigma_h} n $$  \hspace{1cm} (17) $$

$$ \text{Var}(\hat{\tau}_{se,st}) \approx \frac{1}{n} \left( \sum_{h=1}^{L} N_h \sigma_{y_h} \right)^2 $$  \hspace{1cm} (18) $$

Typically, the auxiliary variable used to construct the strata is numeric. Thus, it is common to define strata by reordering the population based on ascending values of the auxiliary variable then choosing $L - 1$ breakpoints for the auxiliary variable to partition the population into $L$ disjoint groups. In most populations, the choice of breakpoints can have a substantial effect on $\text{Var}(\hat{\tau}_{se,st})$. Since the functional relationship between the strata breakpoints and $\text{Var}(\hat{\tau}_{se,st})$ is complicated, no simple, general method exists for finding the optimal breakpoints, given $L$, the number of strata.

Since the 1950’s several researchers have proposed methods for finding optimal strata breakpoints. Early work includes the “cum $\sqrt{f}$ rule” of Dalenius & Hodges Jr (1959) discussed in Section 1. Sethi (1963) used an iterative, compute-intensive numerical optimization procedure to implement a condition necessary for optimization put forth by Dalenius (1957). However, this method can get stuck in local minima and fail to find the breakpoints yielding the global
minimum of \( \text{Var}(\hat{\tau}_{se,st}) \).

With advances in computing speed, recent proposals have primarily focused on random search algorithms (Baillargeon & Rivest (2009) and Kozak (2004)). These methods are usually guaranteed to converge to a local minimum, but not necessarily to the global minimum. In addition, there is no way to know for certain whether the resulting stratification gives a local or global minimum of the objective function, or, if it gives a local minimum, how far “off” it is from the global minimum. Moreover, it is often not easy to set up the parameters of the iterative algorithm to enable it to run well or at all.

To ensure the global minimum is found, one could, in theory, perform a complete enumeration by calculating the objective function for every possible choice of the \( L - 1 \) breakpoints and choosing the set of breakpoints which gives the minimum of the objective function. If the population size is \( N \), then there are \( C_{N-1}^{L-1} \) possible sets of breakpoints, which is of the order \( N^{L-1} \) and the number of calculations grows exponentially as the number of strata increases. The calculation quickly becomes computationally infeasible for anything but the smallest populations. However, MediCal audit populations often contain a substantial number of repeated values. As an example, Table 2 shows the approximate population size and number of unique values for three MediCal populations.

While strata breakpoints can occur within a run of repeated values in audit populations, we will next prove a theorem which states that if \( \sigma_y^2 \) is adequately approximated by Theorem 4.2b, then there will be at most four potential breakpoints within a run where the objective function in (18) could assume a minimum value. More specifically, we assume there is a block of \( n' \) repeats of a value, say \( y \), within a listing of the stratification variable. We use \( x_i \)’s to denote the values in this list before the \( y \)’s and \( z_i \)’s to denote the values after the run.
of repeated $y$’s, i.e. this list is denoted $x_1, x_2, \ldots, x_N, y, y, \ldots, y, z_1, z_2, \ldots, z_M$. We show that, of all possible breakpoints within the run of $y$’s, there are at most four which are candidates for minimizing $\sum_{h=1}^{2} N_h \sigma_{y_h}$. Since finding the optimal strata breakpoints now reduces down to checking only four candidates for the entire run of repeats, in populations like those in Table 2, the compute time required for a complete enumeration of all possible choices of strata breakpoints will be reduced by roughly $1/50$ to $1/100$ when our theorem is applied.

**Theorem 4.3 (Minimum Values of $\sum_{h=1}^{2} N_h \sigma_{y_h}$ within a Run of Repeated Values)**

Let $x_1, x_2, \ldots, x_N, y, y, \ldots, y, z_1, z_2, \ldots, z_M$ be a set of positive real numbers. Let $0 \leq k \leq n'$. Let stratum 1 consist of $x_1, x_2, \ldots, x_N, y, y, \ldots, y$ and stratum 2 consist of $z_1, z_2, \ldots, z_M, y, y, \ldots, y$. If $\sigma_{y_h}^2$ is well-approximated by $\pi \mu_{z_h}^{(2)} - (\pi \mu_{x_h})^2$ then the two real values of $k$ which minimize the objective function $\sum_{h=1}^{2} N_h \sigma_{y_h}$ are the roots of the quadratic:

$$\frac{1}{2} c_3 [c_1^2 - c_2^2 + 2c_3(c_4 - c_5)]k^2 + [-c_1c_2(c_1 + c_2) + 2c_3(c_1c_4 + c_2c_5)]k + (c_1^2c_4 - c_2^2c_5) = 0 \tag{19}$$

where

- $c_1 = \tau_{x}^{(2)} + Ny^2 - 2\pi \tau_x y$
- $c_2 = \tau_{z}^{(2)} + My^2 + 2n'y^2 - 2\pi y(\tau_z + n'y)$
- $c_3 = 2y^2(1 - \pi)$
- $c_4 = (M + n')(\tau_{z}^{(2)} + n'y^2) - \pi(\tau_z + n'y)^2$
- $c_5 = N\tau_{x}^{(2)} - \pi \tau_x^2$

and $\tau_x = \sum_{i=1}^{N} x_i$, $\tau_{x}^{(2)} = \sum_{i=1}^{N} x_i^2$, $\tau_z = \sum_{i=1}^{M} z_i$ and $\tau_{z}^{(2)} = \sum_{i=1}^{M} z_i^2$. 

22
To find the integer value of $k$ which minimizes the objective function, we only need to check the two integer values surrounding the each root of equation (38). The proof of this theorem is given in the Supplement.

5 Ratio Estimation

In this section, we give a criterion for deciding whether ratio estimation or simple expansion should be used in an audit. We derive a reasonable estimator of the variance $\sigma^2_R$ needed under ratio estimation. Although the proposed estimator of $\sigma^2_R$ depends on the error rate $\pi$, we are able to show that $\pi = 0.50$ maximizes the estimated value of $\sigma^2_R$ as in the case of estimation of a binomial proportion. Thus, in cases where $\pi$ is unknown, a conservative sample size can be computed. We also comment on stratification and partial errors under ratio estimation.

5.1 Choosing Between Ratio Estimation and Simple Expansion

We now derive a method for determining whether ratio estimation or simple expansion will be more efficient for extrapolating data from an audit sample. Rearranging the criteria for choosing between these two estimators given in inequality (1) gives

$$\text{Cov}(X, Y) - \frac{\sigma^2_x}{2\mu_x} \mu_y > 0$$

(20)

In the audit setting, it is not difficult to show that $\text{Cov}(X, Y | \tilde{X}_s) = \pi \tilde{X}_e(2) - \pi \mu_x \tilde{X}_e$ and $\mu_y = \pi \tilde{X}_e$. Making these substitutions into inequality (20) and
simplifying, we obtain:

\[
g(\bar{X}_e, \bar{X}_e^{(2)}) = \bar{X}_e^{(2)} - (\mu_x + \frac{\sigma^2_x}{2\mu_x})\bar{X}_e
\]

\[
= \frac{1}{N_e} \sum_{i=1}^{N_e} [X_i^2 - (\mu_x + \frac{\sigma^2_x}{2\mu_x})X_i]
\]

\[
= \frac{1}{N_e} \sum_{i=1}^{N_e} U_i \quad \text{where } U_i = X_i^2 - (\mu_x + \frac{\sigma^2_x}{2\mu_x})X_i
\]

\[
= \bar{U}_e > 0
\]

(21)

The probability that \(g(\bar{X}_e, \bar{X}_e^{(2)}) > 0\) will represent our confidence that the ratio estimator will have smaller variance than the simple expansion estimator. In order to compute this probability, we determine the distribution of \(g(\bar{X}_e, \bar{X}_e^{(2)})\). If \(N_e\) is large, the sample mean \(\bar{U}_e = g(\bar{X}_e, \bar{X}_e^{(2)})\) will be approximately normal by the Central Limit Theorem. In the Supplement, we show that the mean and variance of \(g(\bar{X}_e, \bar{X}_e^{(2)})\) are:

\[
E(g(\bar{X}_e, \bar{X}_e^{(2)})) = \frac{1}{2}\sigma^2_x
\]

\[
\text{Var}(g(\bar{X}_e, \bar{X}_e^{(2)})) = \left(\frac{1}{\pi} - 1\right)\frac{1}{N-1} (\sigma_x^{(2)2} + k^2\sigma_x^2 - 2k\mu'_{12})
\]

where \(k = \mu_x + \frac{\sigma^2_x}{2\mu_x}\)

and \(\mu'_{12} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_x)(x_i^2 - \mu_x^{(2)})\)

and \(\sigma_x^{(2)2} = \frac{1}{N} \sum_{i=1}^{N} (x_i^2 - \mu_x^{(2)})^2\)
Let $Z$ represent the standard normal variate. If $N_e$ is large,

$$P(g(\bar{X}_e, \bar{X}_e^{(2)}) > 0) \approx P(Z > \frac{0 - \frac{1}{2} \sigma_x^2}{\sqrt{\left(\frac{1}{\pi} - 1\right) \frac{1}{N - 1} (\sigma_x^{(2)2} + k\sigma_x^2 - 2k\mu_1')}}) \quad (24)$$

where the last line is true since the numerator of the right side of line (24) is negative. The last line implies that ratio estimation is always favored to outperform simple expansion in any claim population provided we can assume $\bar{U}_e$ is normally distributed. If normality of $\bar{U}_e$ is not reasonable, a Monte Carlo estimate of the probability that inequality (21) is true would give a more accurate estimate of our confidence that ratio estimation will outperform simple expansion.

As noted in Section 1, ratio-estimator-based confidence intervals can fail to attain the nominal confidence level when applied to audit populations even if the standard large-sample criteria for using ratio estimation are met. The excess zeros and skewness often found in audit populations require one to check normality assumptions under either estimator to ensure nominal confidence levels are likely to be met with the proposed sample size. This can be done through Monte Carlo simulation under a range of potential error rates prior to starting an audit.

5.2 Estimating $\sigma_R^2$

We propose $E(\sigma_R^2 | \bar{X}_s)$ as an estimator of $\sigma_R^2$ which can be substituted into the sample size formula in equation (5) for $\sigma_y^2$. In the Supplement to this paper, we
show that:

\[ E(\sigma^2_R|\bar{X}_s) = \pi(1 - \pi)(\mu_x^{(2)} + \frac{\tau_x^{(2)}}{\tau_x^2} \sigma_x^2 - \frac{2\mu_{12}}{\tau_x}) \] (25)

Note that equation 25 depends on the error rate \( \pi \). As in the case of estimating a binomial proportion, (25) is maximized if \( \pi = \frac{1}{2} \). If the error rate cannot be estimated beforehand, using \( \pi = \frac{1}{2} \) will yield a sample size which is sufficiently large for any value of \( \pi \) but may be larger than necessary.

Roberts (1978) derives the following analog of this estimator under the binomial generative model:

\[ E(\sigma^2_R|\pi) = \pi(1 - \pi)\mu_x^{(2)}[1 + \frac{1}{N}(\frac{\sigma_x^2}{\mu_x^2} + \frac{4}{1 + (\frac{\mu_x}{\sigma_x})^2} - \frac{G_1}{\mu_x}[1 + (\frac{\mu_x}{\sigma_x})^2] - 5)] \] (26)

where \( G_1 = \sum_{j=1}^{N}(X_j - \mu_x)^3 \)

For large audit populations, \( N \) and \( \tau_x \) will both be large relative to \( \mu_x \) and \( \sigma_x^2 \) so that equations 26 and 25 will give, \( E(\sigma^2_R|\pi) \approx E(\sigma^2_R|\bar{X}_s) \approx \pi(1 - \pi)\mu_x^{(2)} \).

### 5.3 Stratification Under Ratio Estimation

It is useful to note that optimal stratification under simple expansion by minimizing Formula (18) depends on the error rate, \( \pi \), since \( \pi \) cannot be factored out of the estimated value of \( \sigma_y^2 \). The analogous problem of finding optimal strata breakpoints under ratio estimation is independent of \( \pi \) since formula (25) only includes \( \pi \) in the multiplicative constant \( \pi(1 - \pi) \). Thus, the optimal strata breakpoints under ratio estimation will not be incorrect if the estimated value of \( \pi \) is incorrect. Under simple expansion, however, incorrect estimation of \( \pi \) can lead to a suboptimal choice of strata breakpoints. In addition, the computational shortcut for finding strata breakpoints for populations with a substantial
number of repeated values under simple expansion has an analogous result for the ratio estimator.

5.4 Partial Errors under Ratio Estimation

Under the partial error model defined in (11), we have derived the expected value of $\sigma_R^2$. It is a complicated expression which is not likely to be of practical use since it requires estimates of $\pi_T = \text{proportion of all errors (partial or full)}$, $\pi_p = \text{proportion of partial errors}$ and $q = \text{the portion of each partial payment in error (assumed to be the same for all partial errors)}$ and is not easily bounded or approximated by a more feasible expression.

It is possible on a case-by-case basis to use a simulation study to ascertain whether partial errors are likely to cause $\sigma_R^2$ to increase over the all-or-nothing error model estimator given in equation (25). For the two example populations in Section 6, we found that the estimated value of $\sigma_R^2$ will be lower under a range of partial error models unless the error rate is in the 0.80 or 0.90 range.

6 Audit Example

Since actual MediCal audit data are confidential, we demonstrate these sample design tools using two simulated audit populations. The Edwards Population was simulated to resemble the home health services population in Edwards (2011). This population has a low variance and is right skewed with a spike of values in the $100-150$ range. The population size is 9000, and it represents a paid amount of about $1.1$ million. The Neter Population was simulated to resemble Population 4 on p. 502 of Neter & Loebbecke (1977). This population is also right skewed but with higher variance than the Edwards population. It contains 4033 items and represents $7.5$ million. Histograms of these popula-
6.1 Ratio Estimation versus Simple Expansion

First, we consider the choice between ratio estimation or simple expansion for the example populations. Using the criteria in inequality (24), we can calculate the confidence that ratio estimation will outperform simple expansion over a range of potential error rates. Figure 2 shows the results of this calculation with a separate graph for each population. Unless error rates are quite low, ratio estimation should be used for either population, assuming the assumptions for ratio estimation hold.
6.2 Sample Size Under Ratio Estimation

Suppose that for the Edwards Population representing $1.1 million in paid claims, we wish to estimate the total error with maximum margin of error $110,000 (10% of the total amount paid) at 90% confidence level. For the Neter Population, representing $7.5 million, we wish to estimate the total error with maximum margin of error $750,000 at 90% confidence. The sample sizes required over a range of potential error rates is shown for each population in Figure 3. For comparison, the sample size is shown for both estimators even though the ratio estimator is the preferred estimator. Under ratio estimation, maximal sample sizes occur at an error rate of $\frac{1}{2}$ for any population. This behavior is apparent in Figure 3. However, under simple expansion, the error rate at which the maximal sample size occurs depends on the claim population data. For the Edwards Population, the error rate at which the maximal sample size occurs is 0.67 using formula (10). For the Neter Population, the error rate yielding maximal sample size is 2.72 which is outside the range of error rates.
so $\pi = 1$ will give the conservative sample size.

### 6.3 Effect of Partial Errors on Sample Size Determination

To determine if partial errors might call for increased sample size under ratio estimation, we conducted a simulation study under a variety of partial error models. For each partial error model, we estimated $\sigma^2_R$ over a range of overall error rates (defined as the proportion of claims fully or partially in error). We simulated four audit populations under the following partial error scenarios:

- **Scenario 1**: All errors are full errors (all-or-nothing errors).
- **Scenario 2**: 20% of errors are full errors, and 80% are partial errors.
- **Scenario 3**: 50% of errors are full errors, and 50% are partial errors.
Figure 4: Realized Variances, $\sigma^2_{Rt}$, of Simulated Audit Populations from the Edwards Population Under Several Partial Error Models for Various Error Rates (Error bars represent 90% of the realized variances)

- Scenario 4: 80% of errors are full errors, and 20% are partial errors.

To generate the partial errors, the proportion of the claim amount in error was randomly selected from a uniform distribution between 0.2 and 0.8. Figure 4 shows that the realized value of $\sigma^2_{Rt}$ for the Edwards Population will be lower (than the all-or-nothing value) under all partial error models with high probability until the error rate exceeds about 0.80. The graph for the Neter Population is similar but with larger error bars due to more variability in the original population so we omit it.

6.4 Stratification of Example Populations

We now consider the problem of finding optimal strata breakpoints for the example populations to determine if the precision gain under stratification with two strata is significant. Figure 5 shows the standard error of the optimal stratification found through an exhaustive search for both the simple expansion and ratio estimators for the Edwards Population. We compare the standard
errors across a range of error rates and include the standard error under the Cum $\sqrt{f}$ Rule for comparison.

7 Conclusions and Further Research

Using a hypergeometric generative model and assuming all-or-nothing errors, we developed a method for choosing between ratio and simple expansion estimators. We further showed that, for any audit population, ratio estimation is likely to outperform simple expansion as long as the assumptions for ratio-estimator-based confidence intervals are valid. We further derived estimators of the variances needed for sample size calculation under both simple expansion and ratio estimation and found these estimators to give results similar to the binomial-based estimators in the existing literature when the population size is large. Additionally, we developed a computational shortcut for determining
exact optimal strata breakpoints in populations with substantial repeats. Notably, these methods can be implemented without pilot study data, requiring only the known claim data and an estimated error rate. Moreover, in the absence of an estimated error rate, conservative sample sizes can be calculated by maximizing the variance over the error rate, $\pi$. A simple model for partial errors was considered, and we showed that the conservative sample size under the all-or-nothing-error assumption is sufficient under this partial error model.

A useful avenue for further research would be to extend the results in this paper to more general partial error models which allow the proportion of the claim that is in error to vary and/or which allow for both understatements and overstatements of the claim amount. In addition, since the standard ratio-estimator-based confidence interval may fall short of the nominal confidence level under ratio estimation in audit populations, it would be interesting to investigate whether the estimator of $\sigma_R^2$ in equation (25) would improve the attained confidence level. Finally, since the generative models in this paper all assume that the probability of a claim being in error is independent of the claim amount, it would be useful to rederive these results under a generative model which allows for dependency between the probability of an error and the claim amount.

**SUPPLEMENTARY MATERIAL**

**Proofs:** Detailed proofs of theorems in the paper. (ProofsSupplement.pdf)

**R Code:** R code for the simulations and calculations in Section 6.

(CodeForSampleDesignForAuditPopulations.R, functionsForAuditPaperFigs.R)
Proofs Supplement to Sample Design for Audit Populations

Theorem 4.1 (Conditional Expected Value and Variance) Let \( \bar{X}_e = \frac{1}{N_e} \sum_{i=1}^{N_e} X_{(i)} \), \( \bar{X}_e(2) = \frac{1}{N_e} \sum_{i=1}^{N_e} X_{(i)}^2 \) and \( \mu_{x(2)} = \frac{1}{N} \sum_{i=1}^{N} x_i^2 \). Under the model for \( Y \) given in equation (6):

a. \( E(Y|\bar{X}_s) = \pi \bar{X}_e \)

b. \( \text{Var}(Y|\bar{X}_s) = \pi \bar{X}_e(2) - (\pi \bar{X}_e)^2 \)

c. \( E(\text{Var}(Y|\bar{X}_s)) = \pi \mu_{x(2)} - (\pi \mu_x)^2 - \pi(1 - \pi) \frac{\sigma_x^2}{N - 1} \)

Proof:

a.

\[
E(Y|\bar{X}_s) = \frac{1}{N} \left( \sum_{i=1}^{N_e} X_{(i)} + (N - N_e) \cdot 0 \right) \\
= \frac{N_e}{N} \left( \frac{1}{N_e} \sum_{i=1}^{N_e} X_{(i)} \right) \\
= \pi \bar{X}_e
\]

b.

\[
\text{Var}(Y|\bar{X}_s) = E(Y^2|\bar{X}_s) - [E(Y|\bar{X}_s)]^2 \\
= \frac{1}{N} \left[ \sum_{i=1}^{N_e} X_{(i)}^2 + (N - N_e) \cdot 0^2 \right] - [\pi \bar{X}_e]^2 \\
= \frac{N_e}{N} \cdot \frac{1}{N_e} \sum_{i=1}^{N_e} X_{(i)}^2 - [\pi \bar{X}_e]^2 \\
= \pi \bar{X}_e(2) - (\pi \bar{X}_e)^2
\]
c.

\[
E(\text{Var}(Y|\bar{X}_s)) = \pi E(X_e^{(2)}) - \pi^2 E(X_e^2)
\]

\[
= \pi \mu_x^{(2)} - \pi^2 [\text{Var}(X_e) + [E(X_e)]^2]
\]

\[
= \pi \mu_x^{(2)} - \pi^2 \left[\frac{\sigma_x^2}{N_e} \left(\frac{N - N_e}{N - 1}\right) + \mu_x^2 \right]
\]

\[
= \pi \mu_x^{(2)} - \pi^2 \mu_x^2 - \pi^2 \frac{\sigma_x^2}{N_e} \left(\frac{N - N_e}{N - 1}\right)
\]

\[
= \pi \mu_x^{(2)} - \pi^2 \mu_x^2 - \pi(1 - \pi) \frac{\sigma_x^2}{N - 1}
\]

**Theorem 4.2 (Total Expected Value and Variance)** Under the model for $Y$ given in equation (8):

a. $E(Y) = \pi \mu_x$

b. $\text{Var}(Y) = \pi \cdot \mu_x^{(2)} - (\pi \mu_x)^2$

**Proof:**

a.

\[
E(Y) = E E(Y|\bar{X}_s)
\]

\[
= E(\pi \bar{X}_s)
\]

\[
= \pi E(\bar{X}_s)
\]

\[
= \pi \mu_x
\]
b. Using Theorem 4.1a and c in line two below,

\[
\text{Var}(Y) = E[\text{Var}(Y|\vec{X}_s)] + \text{Var}[E(Y|\vec{X}_s)] \\
= \mu_x^{(2)} - \pi^2[\text{Var}(\bar{X}_e) + [E(\bar{X}_e)]^2] + \text{Var}(\bar{X}_e) \\
= \mu_x^{(2)} - \pi^2[\text{Var}(\bar{X}_e) + [E(\bar{X}_e)]^2] + \pi^2\text{Var}(\bar{X}_e) \\
= \mu_x^{(2)} - \pi^2[\text{Var}(\bar{X}_e)]^2 \\
= \mu_x^{(2)} - \pi^2\mu_x^2
\]

Proof of Estimated Variance under Partial Error Model, Equation (Sample Design for Audit Populations, Formula (12)):

\[
\text{Var}(Y|\vec{X}_T, \vec{X}_p) = E(Y^2|\vec{X}_T, \vec{X}_p) - [E(Y|\vec{X}_T, \vec{X}_p)]^2 \\
E_{\vec{X}_T, \vec{X}_p}(\text{Var}(Y|\vec{X}_T, \vec{X}_p)) = E_{\vec{X}_T, \vec{X}_p}(E(Y^2|\vec{X}_T, \vec{X}_p)) - E_{\vec{X}_T, \vec{X}_p}([E(Y|\vec{X}_T, \vec{X}_p)]^2)
\]

We first compute the terms in equation (27). We use the following notation:
\[
\bar{X}_T = \frac{1}{T} \sum_{i=1}^{T} X_{(i)} \quad \text{and} \quad \bar{X}_p = \frac{1}{p} \sum_{i=1}^{p} X_{(i)} \]

The total and partial error rates are
\[
\pi_T = \frac{T}{N} \quad \text{and} \quad \pi_p = \frac{p}{N}
\]

To compute the second term, we calculate
\[
E(Y|\bar{X}_T, \bar{X}_p) = \frac{1}{N}(\sum_{i=p+1}^{T} X(i) + \sum_{i=1}^{p} qX(i))
\]
\[
= \frac{1}{N}(\sum_{i=1}^{T} X(i) - \sum_{i=1}^{p} X(i) + \sum_{i=1}^{p} qX(i))
\]
\[
= \frac{1}{N}(\sum_{i=1}^{T} X(i) - (1-q)\sum_{i=1}^{p} X(i))
\]
\[
= \frac{T}{N} \sum_{i=1}^{T} X(i) - \frac{(1-q) p}{N} \sum_{i=1}^{p} X(i)
\]
\[
= \pi_T \bar{X}_T - \pi_p (1-q) \bar{X}_p
\]  
(29)

Taking the iterated expectation of (29) and letting \( \sigma^2_{x,T} \) = \( \sum_{i=1}^{T} (x(i) - \bar{X}_T)^2 \), we obtain

\[
E_{\bar{X}_T, \bar{X}_p}(E(Y|\bar{X}_T, \bar{X}_p)^2)
= E_{\bar{X}_T}E_{\bar{X}_p}[(\pi_T \bar{X}_T - \pi_p (1-q) \bar{X}_p)^2 | \bar{X}_T]
= E_{\bar{X}_T}Var_{\bar{X}_p}(\pi_T \bar{X}_T - \pi_p (1-q) \bar{X}_p | \bar{X}_T) + E_{\bar{X}_T}[E_{\bar{X}_p}(\pi_T \bar{X}_T - \pi_p (1-q) \bar{X}_p | \bar{X}_T)]^2
= E_{\bar{X}_T}[\pi_p^2 (1-q)^2 (T-p)(\frac{T-p}{T-1}) \frac{\sigma^2_{x,T}}{p} + \bar{X}_T^2 (\pi_T - \pi_p (1-q))^2]
= \frac{\pi_p^2}{N} (1-q)^2 (1 - \frac{p}{T})(\frac{1}{p})(\frac{N}{N-1}) \sigma^2_x + (\frac{\sigma^2_x}{T})(\frac{N-T}{N-1})
+ \mu_x^2 (\pi_T - \pi_p (1-q))^2
\]  
(30)

Letting \( \bar{X}_T^{(2)} = \frac{1}{T} \sum_{i=1}^{T} X(i)^2 \) and \( \bar{X}_p^{(2)} = \frac{1}{p} \sum_{i=1}^{p} X(i)^2 \) A similar calculation, gives the first term of equation (27):

\[
E(Y^2|\bar{X}_T, \bar{X}_p) = \pi_T \bar{X}_T^{(2)} - \pi_p (1-q^2) \bar{X}_p^{(2)}
\]  
(32)
Taking the iterated expectation of (32), we obtain

\[ E_{\vec{X}_T} E_{\vec{X}_p}(Y^2|\bar{X}_T) = E_{\vec{X}_T} E_{\vec{X}_p}(\pi_T \bar{X}_T^{(2)} - \pi_p (1 - q^2) \bar{X}_p^{(2)} | \bar{X}_T) \]

\[ = E_{\vec{X}_T}(\pi_T \bar{X}_T^{(2)} - \pi_p (1 - q^2) \bar{X}_T^{(2)}) \]

\[ = \pi_T \mu_x^{(2)} - \pi_p (1 - q^2) \mu_x^{(2)} \] (33)

Substituting (33) and (31) into (28) and simplifying, we obtain our estimator of \( \sigma_y^2 \) when there are partial payments.

\[ E_{\vec{X}_T,\vec{X}_p}(\text{Var}(Y|\bar{X}_T, \bar{X}_p)) \]

\[ = [\pi_T - \pi_p (1 - q^2)] \mu_x^{(2)} - \pi_p (1 - q) \left( 1 - \frac{p}{T} \right) \left( \frac{N}{N-1} \right) \sigma_x^2 - \left( \frac{\sigma_x^2}{T} \right) \frac{N - T}{N - 1} \]

\[ + \mu_x^2 (\pi_T - \pi_p (1 - q))^2 \] (34)

Rearranging (34), we obtain

\[ E_{\vec{X}_T,\vec{X}_p}(\text{Var}(Y|\bar{X}_T, \bar{X}_p)) \]

\[ = [\pi_T - \pi_p (1 - q^2)] \mu_x^{(2)} - \pi_p (1 - q) \left( 1 - \frac{p}{T} \right) \]

\[ - \frac{\sigma_x^2}{N-1} [\pi_p (1 - q)]^2 (1 - \frac{p}{T}) \]

\[ + (1 - \pi_T) (1 - \frac{p}{T} (1 - q)) (\pi_T - \pi_p (1 - q)) \]

\[ < [\pi_T - \pi_p (1 - q^2)] \mu_x^{(2)} - [\pi_T - \pi_p (1 - q)]^2 \mu_x^2 \]
Theorem 4.3 (Minimum Values of $\sum_{h=1}^{2} N_h \sigma_{yh}$ within a Run of Repeated Values)

Let $x_1, x_2, ..., x_N, y, y, y, y, z_1, z_2, ..., z_M$ be a set of positive real numbers. Let $0 \leq k \leq n'$. Let stratum 1 consist of $x_1, x_2, ..., x_N, y, y, y, y, y, n'_k$ and stratum 2 consist of $z_1, z_2, ..., z_M, y, y, y, y, y, n'_{-k}$. If $\sigma_{yh}^2$ is well-approximated by $\pi \mu_x(2) - (\pi \mu_x h)^2$ then there are at most two real values of $k$ which minimize the objective function $\sum_{h=1}^{2} N_h \sigma_{yh}$. These critical values can be calculated using claim data and shall be derived in this proof. (To find the $k$ which minimizes the objective function, we only need to check the two integer values surrounding the two critical values $k$.)

Proof: Let $\tau_x = \sum_{i=1}^{N} x_i, \tau_x(2) = \sum_{i=1}^{N} x_i^2, \tau_z = \sum_{i=1}^{M} z_i$ and $\tau_z(2) = \sum_{i=1}^{M} z_i^2$. We wish to minimize the objective function with respect to $k$.

$$\sum_{h=1}^{2} N_h \sigma_{yh} = \sqrt{N_1^2 \sigma_{y1}^2 + N_2^2 \sigma_{y2}^2}$$

$$= \sqrt{(N + k)^2 [\pi (\tau_x(2) + ky^2) - \pi^2 (\tau_x + ky)^2]}$$

$$+ \sqrt{(M + n' - k)^2 [\pi (\tau_z(2) + (n' - k)y^2) - \pi^2 (\tau_z + (n' - k)y)^2]}$$

Let $h(k) = \sum_{h=1}^{2} N_h \sigma_{yh}$. Multiplying out the expressions under the radicals and collecting like terms in $k$, we obtain
\[ h(k) = \sqrt{\pi} \left( \sqrt{\frac{1}{2} c_3 k^2 + c_1 k + c_3} + \sqrt{\frac{1}{2} c_3 k^2 - c_2 k + c_4} \right) \]  

where

\[ c_1 = \tau_x^{(2)} + N y^2 - 2\pi \tau_x y \]
\[ c_2 = \tau_z^{(2)} + M y^2 + 2n'y^2 - 2\pi y(\tau_z + n'y) \]
\[ c_3 = 2y^2(1 - \pi) \]
\[ c_4 = (M + n')(\tau_z^{(2)} + n'y^2) - \pi(\tau_z + n'y)^2 \]
\[ c_5 = N\tau_x^{(2)} - \pi \tau_x^2 \]

Differentiating the right side of (35) with respect to \( k \) and setting it equal to zero gives:

\[ h'(k) = \frac{c_3 k + c_1}{2\sqrt{\frac{1}{2} c_3 k^2 + c_1 k + c_3}} + \frac{c_3 k - c_2}{2\sqrt{\frac{1}{2} c_3 k^2 - c_2 k + c_4}} = 0 \]  

\[ \frac{c_3 k + c_1}{\sqrt{\frac{1}{2} c_3 k^2 + c_1 k + c_5}} = \frac{-c_3 k + c_2}{\sqrt{\frac{1}{2} c_3 k^2 - c_2 k + c_4}} \]
\[ \frac{(c_3 k + c_1)^2}{\frac{1}{2} c_3 k^2 + c_1 k + c_5} = \frac{(-c_3 k + c_2)^2}{\frac{1}{2} c_3 k^2 - c_2 k + c_4} \]  

Cross-multiplying (37) and simplifying, the quartic and cubic terms in \( k \) conveniently drop out, leaving a quadratic equation in \( k \).

\[ \frac{1}{2} c_3 [c_1^2 - c_3^2 + 2c_3(c_4 - c_5)]k^2 + [-c_1c_2(c_1 + c_2) + 2c_3(c_1c_4 + c_2c_5)]k + (c_1^2c_4 - c_2^2c_5) = 0 \]  

(The roots looks complicated but are quickly calculated on a computer.) The
two roots of Equation (38) along with \( k = 0 \) and \( k = n' \) are the only possible locations of the minimum of \( h(k) \) in the interval \([0, n']\) since \( h \) and \( h' \) are both defined and continuous on this interval (assuming \( \sigma_{yh} \neq 0 \)). We need only evaluate \( h(k) \) at the integer values adjacent to each root of (38) to find the integer value of \( k \) which minimizes \( h(k) \). This is because we can show that the expressions under the radicals in equation (36) are positive over the domain of \( h' \), and, thus, \( h' \) can be shown to be continuous on \([0, n']\). Because \( h' \) is continuous on \([0, n']\), it must be monotone on the intervals determined by the endpoints and critical points. Hence, the minimum of \( h \) over integer values of \( k \) can only occur at integers adjacent to the critical points or at the endpoints of the interval \([0, n']\).

Mean and Variance of Criteria for Choosing Between Simple Expansion and Ratio Estimation, *(Sample Design for Audit Populations, Formula (22))*

\[
E(g(\bar{X}_e, \bar{X}_e^{(2)})) = \frac{1}{2} \sigma_x^2 \\
\text{Var}(g(\bar{X}_e, \bar{X}_e^{(2)})) = \left( \frac{1}{\pi} - 1 \right) \frac{1}{N-1} \left( \sigma_x^{(2)} + k^2 \sigma_x^2 - 2k\mu_{12} \right) 
\]

where \( k = \mu_x + \frac{\sigma_x^2}{2\mu_x} \)

and \( \mu_{12} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_x)(x_i^2 - \mu_x^{(2)}) \)

and \( \sigma_x^{(2)} = \frac{1}{N} \sum_{i=1}^{N} (x_i^2 - \mu_x^{(2)})^2 \)
Proof of Variance:

\[
\text{Var}(g(\bar{X}_e, \bar{X}_e^{(2)})) = \text{Var}(\bar{U}_e) = \frac{\text{Var}(U_{(i)})}{N_e} \cdot \frac{N - N_e}{N - 1} = \frac{\text{Var}(X^2_{(i)}) + k^2 \text{Var}(X_{(i)}) - 2k \text{Cov}(X^2_{(i)}, X_{(i)})}{N_e} \cdot \frac{N - N_e}{N - 1} = \frac{\sigma^2_x^{(2)} + k^2 \sigma^2_x - 2k \mu_{12}}{N_e} \cdot \frac{N - N_e}{N - 1}
\]

Proof of Estimator of Variance under Ratio Estimation, \textbf{(Sample Design for Audit Populations, Formula (25))}:

Let \( s = \{(1), (2), \ldots (N_e)\} \subseteq \{1, 2, \ldots, N\} \), i.e. \( s \) contains the subscripts of a random subset of size \( N_e \) representing the erroneous claims in the population \( \{x_1, x_2, \ldots, x_N\} \). In this notation, \( R = \frac{\tau_y}{\tau_x} = \frac{\sum_{i \in s} x_i + \sum_{i \notin s} 0}{\tau_x} \) so that

\[
N \sigma^2_R | \bar{X}_s = \sum_{k=1}^{N} (y_k - Rx_k)^2
\]

\[
= \sum_{k=1}^{N} (y_k - \frac{\sum_{i \in s} x_i}{\tau_x} x_k)^2
\]

\[
= \sum_{k \notin s} (0 - \frac{\sum_{i \in s} x_i}{\tau_x} x_k)^2 + \sum_{k \in s} (x_k - \frac{\sum_{i \in s} x_i}{\tau_x} x_k)^2 \quad (41)
\]

To make \( E(N \sigma^2_R | \bar{X}_s) \) easier to compute, we multiply out (41) and write the resulting expression in terms of the random variables \( \bar{X}_e \) and \( \bar{X}_e^{(2)} \).
\[
N\sigma^2_{\bar{X}^2} = \frac{\tau^2}{\tau^2_x} N^2 e x^2 + N e \hat{X}^2_e - \frac{2N^2 e}{\tau_x} X_e \hat{X}^{(2)}_e
\]  

Using a result of Espejo (1997), we find that

\[
E(\bar{X}_e \hat{X}^{(2)}_e) = \left(1 - \pi\right) \frac{1}{N_e} \mu'_{12} + \mu_e \mu^{(2)}_x
\]

where \( \mu'_{12} \) was defined in (40). Substituting this last equation into equation (42) along with the expectations of the remaining random variables then simplifying yields Formula (25) in the manuscript.

References

Baillargeon, S. & Rivest, L.-P. (2009), ‘A general algorithm for univariate stratification’, International Statistical Review 77(3), 331–344.

Buddhakulsomsiri, J. & Parthanadee, P. (2008), ‘Stratified random sampling for estimating billing accuracy in health care systems’, Health care management science 11(1), 41–54.

Centers for Medicare and Medicaid Services (2018a), ‘Medicaid’, https://www.medicaid.gov/medicaid/index.html.

Centers for Medicare and Medicaid Services (2018b), ‘NHE fact sheet’, https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet.html. Accessed: 2018-02-03.

Cochran, W. G. (1977), Sampling Techniques, John Wiley & Sons, New York.
Dalenius, T. (1957), *Sampling in Sweden: contributions to the methods and theories of sample survey practice*, Almqvist and Wiksell.

Dalenius, T. & Hodges Jr, J. L. (1959), ‘Minimum variance stratification’, *Journal of the American Statistical Association* **54**(285), 88–101.

Edwards, D. (2011), ‘On stratified sampling and ratio estimation in medicare and medicaid benefit integrity investigations’, *Health Services and Outcomes Research Methodology* **11**, 79–94.

Espejo, M. R. (1997), Covariance of sample moments for finite populations, in ‘Proceedings of the Royal Irish Academy. Section A: Mathematical and Physical Sciences’, JSTOR, pp. 163–167.

King, B. & Madansky, A. (2013), ‘On sampling design issues when dealing with zeros’, *Journal of Survey Statistics and Methodology* **1**(2), 144–170.

URL: + http://dx.doi.org/10.1093/jssam/smt006

Kozak, M. (2004), ‘Optimal stratification using random search method in agricultural surveys’, *Statistics in Transition* **6**(5), 797–806.

Neter, J. & Loebbecke, J. K. (1977), ‘On the behavior of statistical estimators when sampling accounting populations’, *Journal of the American Statistical Association* **72**(359), 501–507.

URL: http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1977.10480603

Roberts, D. M. (1978), *Statistical auditing*, American Institute of Certified Public Accountants.

Sethi, V. (1963), ‘A note on optimum stratification of populations for estimating the population means’, *Australian & New Zealand Journal of Statistics* **5**(1), 20–33.