South African Statistical Journal
Vol. 54, No. 2, 153–162
https://doi.org/10.37920/sasj.2020.54.2.2
© 2020 South African Statistical Association

ACCURACY IN INSURANCE BILLING ERROR ESTIMATION USING AUXILIARY INFORMATION

Bhargab Chattopadhyay[1]
Indian Institute of Management Visakhapatnam, A.P., India
e-mail: *bhargab@iimv.ac.in*

Billing fraud by health care providers is a widespread problem to a country's health care system. This article develops a general theory for estimating the billing error in medical claims within pre-specified error bound using auxiliary information on the average payment amount made by all persons in the population. Estimation methods with pre-specified sample size cannot be used to achieve the fixed-width confidence interval for billing error. In this article we propose two two-stage procedures for accuracy in estimating billing error in medical claims using sample standard deviation and sample Gini's mean difference as estimators of population standard deviation. This problem is the same as constructing a fixed-width confidence interval for billing error. In two-stage estimation procedures, the final sample size is not fixed in advance by using supposed unknown population parameter(s). Data in two-stage procedures are collected in two stages in which the final sample size is based on the estimate of the unknown parameter(s) in the first stage. The comparison of the proposed two-stage procedures are examined using a Monte Carlo simulation study.

Key words: Gini's mean difference, Optimal sample size, Stratified sampling, Two-stage procedure.

1. Introduction

Detecting billing fraud by health care providers is a very important and relevant problem. Often there are cases of over-billing by health care service providers (hospitals, doctors and others) for their services. This is a concern for agencies responsible for reimbursing medical bills. The U.S. Department of Health & Human Services website reported that "a nationwide take down by Medicare Fraud Strike Force operations in six cities has resulted in charges against 90 individuals, including 27 doctors, nurses and other medical professionals, for their alleged participation in Medicare fraud schemes involving approximately $260 million in false billings" (e.g. CMS, 2017). Like over-billing, under-billing can also have significant financial effects. Both under-billing and over-billing come under medical care fraud which may result in costly audits and legal consequences (GAO, 2016).

In order to verify possible fraud, the agencies have designed sampling plans to check the submitted bills in detail. Cohen and Naus (2007) proposed a stratified sampling design in order to compute the lower confidence bound for the amount of over-/under-billing in which several strata were formed based on the claim amount. From each stratum samples are independently drawn to determine the necessary size of an audit sample to ensure a 95% confidence level for the absolute error in payment amount. For a discussion on the application of stratified sampling in billing error estimation, one may

[1] Corresponding author.
MSC2010 subject classifications. 62F12, 62F30, 62L12.

refer to WHCA (2014) published in the Washington State Health Care Authority. In fact, stratified sampling design is widely used in the field of healthcare management. For details, we refer to Cohen and Naus (2007), Buddhakulsomsiri and Parthanadee (2008) and Kim et al. (2013).

In general, the length of a $100(1-\alpha)\%$ confidence interval for a parameter decreases if we increase the sample size, but this in turn increases the overall sampling cost. On the other hand, a smaller sample might decrease the sampling cost but it might increase the width of the confidence interval. One way to solve this problem is to fix the length of the confidence interval and try to minimise the sample size or, in other words, the sampling cost. Since accuracy is a matter of concern, a fixed-width $100(1-\alpha)\%$ confidence interval for the absolute error in payment amount is desired under the following scenario:

- observations are drawn from a finite population divided into several strata.
- an auxiliary information on the sum of payment amounts made by all persons in the population.

In general, the problem of finding a fixed-width confidence interval for a parameter cannot be solved with a pre-specified sample size. This problem can only be solved using two-stage or multi-stage sampling methods in which sample sizes are not pre-fixed. The final sample size depends on the statistical analysis carried out on the already collected observations. For an extensive review of the literature on two-stage procedure, one may refer to Mukhopadhyay and De Silva (2009) and Ghosh et al. (1997), among others.

1.1 Contribution of this paper

There exist several articles on detecting insurance billing fraud in medical claims. Examples of work on measurement and billing fraud detection methods include Upadhyaya and Singh (1999), Cohen and Naus (2007), Li et al. (2008), David et al. (2013) and Johnson and Nagarur (2016). However, none of the methods can be used to measure the amount of insurance billing fraud within a pre-specified error bound. In this article we are going beyond the contribution of Cohen and Naus (2007) by proposing two two-stage procedures to obtain the optimal sample size from each stratum and thereby construct a fixed-width confidence interval for the billing fraud amount using the auxiliary information without using any data distribution. Characteristics of both the procedures are discussed using Monte Carlo simulation.

The organisation of the remainder of the paper is as follows. We formulate the fixed-width confidence interval in the payment amount based on an auxiliary information in Section 2. Section 3 presents the two-stage procedures — one based on sample standard deviation and the other is based on sample Gini's mean difference required to compute a fixed-width $100(1-\alpha)\%$ confidence interval for the payment amount based on auxiliary formation. In Section 4, we assess and compare the performance of both two-stage procedures using a simulation study. Section 5 presents an illustration of the two-stage procedure using a dataset followed by concluding remarks in Section 6.

2. Estimation of absolute error in payment amount

We recall that stratified design using simple random sampling without replacement is used in estimating billing error in medical claims. The goal is to construct a $100(1-\alpha)\%$ fixed-width confidence interval for the average amount of billing error in the population, $\bar{Y}$, when the information on the average payment amount ($\bar{X}$) made by all persons in the population are available. The corresponding error rate in the payment amount is $R = \bar{Y}/\bar{X}$.

Suppose there are H strata, with N_h persons in the hth stratum. The variable of interest is the amount of payment error on a sample claim. For the ith person belonging to the hth stratum we define y_{ih} = |Amount billed and actually paid − Amount that should have been billed and paid| and x_{ih} = Actual paid amount.

Also, suppose that X is the sum of payment amounts made by all persons in the population. This quantity is usually known. Also, suppose Y is the unknown total billing error in the population. $\bar{X}$ and $\bar{Y}$ are the corresponding population averages from N individuals. Then the corresponding error rate in payment amount is $R = \bar{Y}/\bar{X}$. Additionally, suppose $\bar{X}_h, S^2_{xh}$ and $\bar{Y}_h, S^2_{yh}$ are population means and variances corresponding to the hth stratum based on $(x_1, \ldots, x_{N_h})$ and $(y_1, \ldots, y_{N_h})$, respectively, and S_{xyh} is the population covariance based on $(x_1, \ldots, x_{N_h})$ and $(y_1, \ldots, y_{N_h})$.

Next, a sample of n_h persons is drawn via simple random sampling without replacement from the hth stratum containing N_h persons in total. Let $\bar{y}_h$ and $\bar{x}_h$ be the sample means from n_h persons drawn from the hth stratum. Also, let $n(= \sum_{h=1}^H n_h)$ be the total sample size. The corresponding ratio estimator for the absolute error in payment amount is given by

$$\bar{Y}_n = \widehat{R}_n \bar{X}, \text{ where } \widehat{R}_n = \frac{\sum_{h=1}^H w_h \bar{y}_h}{\sum_{h=1}^H w_h \bar{x}_h} \text{ and } w_h = \frac{N_h}{N}.$$

For details about ratio estimators, one may refer to Cochran (1977), Jhajj and Walia (2012) and others. In fact, depending on the situation, one may use other suitable estimators. For example, one can use the Sisodia–Dwivedi estimator if the population coefficient of variation of the actual paid amount is known, or the Upadhyaya–Singh estimator when both population kurtosis and the population coefficient of variation of the actual paid amount are known, etc. For details, one may refer to Sisodia and Dwivedi (1981), Upadhyaya and Singh (1999), Singh and Vishwakarma (2008) or Solanki and Singh (2016).

Using Taylor's theorem, similar to equation (A3) of Chattopadhyay and De (2016) or equation (73) of Kelley et al. (2018) or equation (81) of Kelley et al. (2019), we have

$$R_n - R = \frac{\bar{y}_{st}}{\bar{x}_{st}} - \frac{\bar{Y}}{\bar{X}} = \frac{1}{\bar{X}}(\bar{y}_{st} - \bar{Y}) - \frac{\bar{Y}}{\bar{X}^2}(\bar{x}_{st} - \bar{X}) + E_n = \frac{1}{\bar{X}}(\bar{y}_{st} - \bar{Y}) - \frac{R}{\bar{X}}(\bar{x}_{st} - \bar{X}) + E_n,$$

where,

$$E_n = -2(\bar{y}_{st} - \bar{Y})(\bar{x}_{st} - \bar{X})/b^2 + 4a(\bar{x}_{st} - \bar{X})/b^3,$$

with $a = \bar{Y} + p(\bar{y}_{st} - \bar{Y})$, $b = \bar{X} + p(\bar{x}_{st} - \bar{X})$ for $p \in (0, 1)$. Also, $\bar{x}_{st} = \sum_{h=1}^H w_h \bar{x}_h$ and $\bar{y}_{st} = \sum_{h=1}^H w_h \bar{y}_h$. Thus, we have

$$\sqrt{n}(\bar{Y}_n - \bar{Y}) = \sqrt{n}(\bar{y}_{st} - \bar{Y}) - R\sqrt{n}(\bar{x}_n - \bar{X}) + \sqrt{n}E_n = \sqrt{n}\sum_{h=1}^H w_h \left[(\bar{y}_h - \bar{Y}_h) - R(\bar{x}_h - \bar{X}_h)\right] + \sqrt{n}E_n.$$

As $n \to \infty$, $E_n \xrightarrow{p} 1$. By the central limit theorem, if n is large, $\sqrt{n}(\bar{Y}_n - \bar{Y}) \xrightarrow{d} N(0, V)$, where

$$nV = \sum_{h=1}^H \frac{w_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)(S^2_{yh} + R^2 S^2_{xh} - 2RS_{xyh}).$$

We note that $Var(\bar{y}_h) = \sum_{h=1}^{H}(S_{yh}^2/n_h)(1-n_h/N_h)$, $Var(\bar{x}_h) = \sum_{h=1}^{H}(S_{xh}^2/n_h)(1-n_h/N_h)$, and $Cov(\bar{y}_h,\bar{x}_h) = \sum_{h=1}^{H}(S_{xyh}/n_h)(1-n_h/N_h)$.

Thus the approximate $100(1-\alpha)\%$ confidence interval for $\bar{Y}$ is given by

$$\left(\bar{Y}_n - z_{\alpha/2}\sqrt{V(\bar{Y}_n)}, \bar{Y}_n + z_{\alpha/2}\sqrt{V(\bar{Y}_n)}\right),$$

where

$$V(\bar{Y}_n) = \sum_{h=1}^{H} w_h^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right)S_h^2, \text{ where, } S_h^2 = S_{yh}^2 - 2RS_{xyh} + R^2S_{xh}^2. \tag{1}$$

Equation (1) can also be found in Cochran (1977). The assumption that the population size (N_h) of each of the strata is also large is valid due to the nature of application and hence a large sample approximation can be considered. Under optimum sample allocation, the sample size for the hth stratum is $n_h = nw_hS_h/\sum_{h=1}^{H} w_hS_h$. So, under optimal allocation, the total optimal sample size that will be required to construct an approximate $100(1-\alpha)\%$ fixed-width confidence interval of width $2d(>0)$, say, of the average absolute error in billing amount in the population, $\bar{Y}$, is

$$n_d = \frac{N\chi_{1-\alpha;1}^2\left(\sum_{h=1}^{H} w_hS_h\right)^2}{d^2N + \chi_{1-\alpha;1}^2\left(\sum_{h=1}^{H} w_hS_h^2\right)}. \tag{2}$$

Note that the population variance S_h^2 is usually unknown. So, one cannot find the optimal sample size n_d using (2). We use a two-stage procedure to find an estimate of the optimal sample size, n_d, required to get a fixed-width interval for average absolute error in billing amount in the population.

For estimating population variance, we generally use the sample standard deviation (SD). However, the use of alternative estimators of population standard deviation like Gini's mean difference (GMD) is also well studied. In fact, Yitzhaki et al. (2003) Mukhopadhyay and Chattopadhyay (2011), Mukhopadhyay and Chattopadhyay (2012) and Chattopadhyay and Mukhopadhyay (2013) proposed a GMD-based estimator of population variance in case of non-normality and also in the presence of suspect outliers in case of normal distribution. In the absence of any specific data distribution, we may use sample GMD and sample SD as competing estimators of population standard deviation, S_h. In the next section, we propose two two-stage procedures: a SD-based two stage procedure, a procedure in which sample SD is used as an estimator of S_h, and a GMD-based procedure, a procedure in which sample GMD is used as an estimator of S_h.

3. Two-stage methodology

We suggest the following SD-based and GMD-based two-stage procedures:

Stage 1: Draw a pilot sample of pre-specified size, m_h, from the hth ($h = 1,2,...,H$) stratum using simple random sampling without replacement. Thus, we have a sample of $m(=\sum_{h=1}^{H} m_h)$ observations. Obtain estimates of stratum means using pilot sample corresponding to each stratum. Using the m_h observations from the hth, we obtain $\bar{x}_{hm}$ and $\bar{y}_{hm}$ and compute

$$\widehat{R}_m = \frac{\sum_{h=1}^{H} w_h\bar{y}_{hm}}{\sum_{h=1}^{H} w_h\bar{x}_{hm}}.$$

Also, for the hth strata, for $i = 1, \ldots, m_h$, define $v_{ih} = y_{ih} - \widehat{R}_m x_{ih}$ and compute an estimate of S_h, defined as s_{hm}. In this case, use

$$s_{hm} = \begin{cases} \left[\frac{1}{m-1}\sum_{i=1}^{m}(v_{ih} - \bar{v}_h)^2\right]^{1/2} & \text{if sample SD is used as an estimator of } S_h, \\ \binom{m_h}{2}^{-1}\sum_{i=1}^{m_h}\sum_{j=1_{i<j}}^{m_h}|v_{ih} - v_{jh}| & \text{if sample GMD is used as an estimator of } S_h. \end{cases}$$

Compute the combined sample size of the two-stage procedure based on sample SD or sample GMD as

$$n_o = max\left[m, \frac{N\chi^2_{1-\alpha;1}\left(\sum_{h=1}^{H} w_h s_{hm}\right)^2}{d^2N + \chi^2_{1-\alpha;1}\left(\sum_{h=1}^{H} w_h s^2_{hm}\right)}\right]. \tag{3}$$

If the optimal sample size $N\chi^2_{1-\alpha;1}(\sum_{h=1}^{H} w_h s_{hm})^2/(d^2N + \chi^2_{1-\alpha;1}(\sum_{h=1}^{H} w_h s^2_{hm})) < m$, stop sampling and report the final sample size as $n_o = m$. If $n_o > m$, follow Stage 2.

Stage 2: Under optimal allocation, for allocating the remaining $(n_o - m)$ observations among the H strata, we re-define the weights. For the hth stratum, the weight is given by $w_{1h} = (N_h - m_h)/(N - m)$. Thus, the estimated optimal sample size for the hth stratum is

$$n_{oh} = \frac{(n_o - m)w_{1h}s_{hm}}{\sum_{h=1}^{H} w_{1h}s_{hm}}.$$

We collect the remaining $(n_o - m)$ observations in such a way that, in the hth stratum, we collect n_{oh} pairs of observations. Suppose $\bar{y}_{hn_{oh}}$ and $\bar{x}_{hn_{oh}}$ represent the (Stage-2) sample means corresponding to X and Y, then the ratio estimator of the average absolute error in payment amount is

$$\bar{Y}_{n_o} = \widehat{R}_{n_o}\bar{X}, \text{ where } \widehat{R}_{n_o} = \frac{\sum_{h=1}^{H} w_h\left(\frac{m_h\bar{y}_{hm} + n_{oh}\bar{y}_{hn_{oh}}}{m_h + n_{oh}}\right)}{\sum_{h=1}^{H} w_h\left(\frac{m_h\bar{x}_{hm} + n_{oh}\bar{x}_{hn_{oh}}}{m_h + n_{oh}}\right)}.$$

The $100(1 - \alpha)\%$ fixed-width confidence interval for the average of absolute error in billing amount, $\bar{Y}$, is

$$\left(\bar{Y}_{n_o} - d, \bar{Y}_{n_o} + d\right). \tag{4}$$

3.1 Modification of Stage 2

Suppose for the hth $(h = 1, 2, ..., H)$ stratum the Stage 2 optimal sample size is such that $n_{oh} > N_h - m_h$. In such a case, define $n_{oh} = N_h - m_h$. For the stratum $h'(\neq h) = 1, 2, ..., H$, with $n_{oh'} < N_{h'} - m_{h'}$, redefine

$$n_{oh'} = \frac{(n_o - m - n_{oh})w_{1h'}s_{h'm}}{\sum_{h'(\neq h)=1}^{H} w_{1h'}s_{h'm}}, \text{ where } w_{1h'} = \frac{N'_h - m'_h}{N - m - noh}.$$

We continue modifying the optimal sample sizes for each stratum in a similar way until $n_{oh} \leq N_h - m_h$ for all $h = 1, 2, ..., H$ and then draw observations on X and Y to get the fixed-width confidence interval.

Table 1. Parameters of stratified population.

Cases	Stratum (h)	N_h	$(\bar{X}_h, \bar{Y}_h)$	$(S^2_{xh}, S^2_{yh}, S_{xyh})$
1	1	3000	(323.8455, 32.9444)	(10355.4, 549.9568, − 12.0747)
	2	5000	(400.7421, 39.8678)	(16496.62, 802.8175, 166.6778)
	3	2000	(478.7529, 48.0044)	(23602.6, 1096.4360, 137.87)
2	1	3000	(1096.412, 154.313)	(749.8012, 2255.842, − 7.270515)
	2	5000	(2982.361, 153.136)	(10945.95, 1512.257, − 94.35078)
	3	2000	(8102.661, 157.0822)	(56272.53, 3298.059, 600.4041)

3.2 A note

The two-stage procedure discussed here can easily be extended to the sample size planning method given by Cohen and Naus (2007). The authors proposed a method in which the sample size is computed assuming the population standard deviation, which is highly unrealistic. We may easily adopt our two-stage procedure in order to estimate the required stratum sizes.

4. Performance via simulation study

We now evaluate the performance of our proposed two-stage procedures via a simulation study. We consider a population of size $N = 10\,000$ with $H = 3$ strata. We consider two scenarios:

Case 1: For the first stratum, to get observations on y_{i1} $(i = 1, \dots, N_1)$ we randomly selected 3 000 observations from a Gamma distribution with shape parameter 10 and scale parameter 4 and multiplied each observation with 50. Now, to get observations on x_{i1} $(i = 1, \dots, N_1)$, we randomly selected 3 000 observations from a Gamma distribution with shape parameter 10 and scale parameter 2 and multiplied each observation with 40. Thus we have a pair of 3 000 observations for the first stratum. In the same way, we collected 5 000 pairs for the second stratum, and the remaining 2 000 pairs for the third stratum and then we respectively added 10 to each pair of the second stratum and 20 to each pair of the third stratum.

Case 2: In this scenario, we use the following to draw 3 000 pairs of random observations from Stratum 1, 5 000 pairs from Stratum 2 and 2 000 pairs from Stratum 3:

$y_{i1} \sim$ Lognormal $(5, 0.3), y_{i2} \sim$ Lognormal $(5, 0.25), y_{i13} \sim$ Lognormal $(5, 0.35)$, $x_{i1} \sim$ Lognormal $(7, 0.025), x_{i2} \sim$ Lognormal $(8, 0.035), x_{i13} \sim$ Lognormal $(9, 0.03)$.

Considering the data obtained in each of the various scenarios as a population, we detail the stratum-wise population means, variances and the covariances in Table 1.

To implement the two-stage procedure in this simulation study, we fix $d = 2$, $\alpha = 0.1, 0.05$. Suppose from all three strata we select m_h (= 50) pair of observations from the above population as our pilot sample. Thus in the pilot stage we have 300 (= m) observations in total. Based on the pilot sample, we implemented the two stage procedure to compute n_o as given in equation (3) and then the confidence interval given in (4).

This whole process is repeated 10 000 times and the results are given in Table 3 which presents the overall average final sample size, $\overline{\overline{n_o}}$ (which estimates n_d) from 10 000 replications, the ratio of the average of the estimated final sample sizes for each stratum and optimal sample sizes and the overall

Table 2. Characteristics: Two-stage procedure.

Case	α	Estimator	$\overline{\widehat{n}_o}$	n_d	$\overline{\widehat{n}_{o1}}/n_{d1}, \overline{\widehat{n}_{o2}}/n_{d2}, \overline{\widehat{n}_{o3}}/n_{d3}$	$\overline{\widehat{n}_o}/n_d$	CP $(\bar{p})$
		of S_h	**s**$(\overline{\widehat{n}_o})$				**s**$(\bar{p})$
1	0.1	SD	573.7707 **1.0051**	581	1.0664, 0.9046, 1.0755	0.9876	0.8601 **0.0035**
1	0.1	GMD	679.3835 **1.0681**	581	1.2440, 1.0835, 1.2653	1.1693	0.8825 **0.0032**
1	0.05	SD	793.8603 **1.3581**	805	1.0435,0.9292, 1.0483	0.9862	0.9086 **0.0029**
1	0.05	GMD	935.8187 **1.4336**	805	1.2154,1.1031,1.2337	1.1625	0.9231 **0.0027**
2	0.1	SD	1210.378 **1.6377**	1224	0.9949,0.9680, 1.0184	0.9889	0.8920 **0.0031**
2	0.1	GMD	1442.659 **1.7129**	1224	1.1827,1.1663,1.1956	1.1786	0.6796 **0.0047**
2	0.05	SD	1626.137 **2.1054**	1651	0.9893,0.9717 ,1.0033	0.9849	0.9418 **0.0023**
2	0.05	GMD	1920.749 **2.1626**	1651	1.1659 ,1.1587 ,1.1690	1.1634	0.6956 **0.0046**

ratio and the coverage probability, CP $(\bar{p})$ obtained from the total sample of size n_o. Moreover, **s**$(\overline{\widehat{n}_o})$ and **s**$(\bar{p})$ represent the standard errors of $\widehat{n}_o$ and $\bar{p}$, respectively.

The sixth and seventh columns of Table 2 indicate that average sample sizes, both overall and stratum-wise, for the SD-based two stage procedure are almost the same as the optimal sample size n_d, in the sense that the ratio is close to 1. Thus, in table 2, among the two procedures, we find that on average the proposed SD-based two-stage procedure performs well compared to the GMD-based procedure in terms of sampling cost, both overall and stratum-wise.

The last column illustrates that the coverage probability is not significantly different than $1-\alpha$ for the SD-based two-stage procedure. Even though for the first case (when the observations are drawn from the Gamma distribution), the coverage probability of the GMD-based procedure is closer to the target than the SD-based procedure, but performed poorly in the second case (when the observations are drawn from the Lognormal distribution). Therefore for non-normal data as well, the simulation study indicates that the SD-based two-stage procedure is preferred over the GMD-based two-stage procedure for constructing the fixed-width confidence interval for the average billing error.

5. An illustration

In this section, we provide a realistic example based on the dataset (Tables 1 and 2) in Cohen and Naus (2007) in which the population was divided into six strata. As an illustration of our two-stage

Table 3. Estimated average final sample size.

α, d	Strata	m_h	N_h	n_{os}	n_{og}
0.05, 10	1	4	4000	873	1135
	2	6	2200	487	605
	3	6	1000	220	270
	4	6	500	95	113
	5	8	200	112	135
	6	6	100	91	99
0.1, 10	1	4	4000	713	969
	2	6	2200	398	517
	3	6	1000	180	230
	4	6	500	78	97
	5	8	200	91	115
	6	6	100	74	85

procedures, we consider the pairs of observations contained in Table 2 of Cohen and Naus (2007) as the pilot sample observations for each stratum. Using the pilot sample observations and the SD- and GMD-based two-stage procedures, we compute the final sample size for each stratum required to get the fixed-width confidence interval for the average billing error in the whole population. In Table 3, m_h refers to the pilot sample size for each stratum and N_h refers to the total number of observations in the stratum. The last two columns, n_{os} and n_{og} indicate the estimated optimal sample size given by the SD-based two-stage procedure and the GMD-based two-stage procedure for each stratum.

Similar to the simulation study, we find that the GMD-based two-stage procedure requires a larger sample size from each stratum than the SD-based two-stage procedure.

6. Concluding remarks

Billing fraud in medical claims is a common problem in health-care systems in many countries. A stratified sampling design is often used for auditing the medical claims. With no distributional assumption, this article develops two-stage procedures based on sample GMD and sample SD for estimating the billing error in medical claims within a pre-specified error bound, provided the sum of payment amounts made by all persons in the population is known. Several authors proposed that sample GMD is a suitable estimator of the unknown population standard deviation when the data distribution deviates from normality. Apart from sample SD, sample GMD is also considered as an estimator of the finite population standard deviation in order to formulate the two-stage procedure.

Both SD- and GMD-based two-stage procedures are studied using Monte Carlo simulation. The SD-based two stage procedure is found to be more efficient than the GMD-based procedure in terms of final stratum-wise sample sizes. Also the corresponding coverage probability is not significantly different from the target coverage probability unlike in some cases of the GMD-based two-stage procedure. However, both these properties seem to be lacking in the GMD-based procedure.

Unlike the fixed sample size procedures, the basic concept of two-stage procedures revolves around the idea of estimating the required optimal sample size based on the pilot sample. Recent methodological advances in the medical claim fraud domain revolve around the fraud detection methods. However, the estimation of the actual fraud amount with a pre-specified accuracy is also important and we believe that this is the first article, with a practical approach, to make developments in this area.

References

Buddhakulsomsiri, J. and Parthanadee, P. (2008). Stratified random sampling for estimating billing accuracy in health care systems. *Health Care Management Science*, **11**, 41–54.

Chattopadhyay, B. and De, S. K. (2016). Estimation of Gini index within pre-specified error bound. *Econometrics*, **4**, 30.

Chattopadhyay, B. and Mukhopadhyay, N. (2013). Two-stage fixed-width confidence intervals for a normal mean in the presence of suspect outliers. *Sequential Analysis*, **32**, 134–157.

CMS (2017). NHE fact sheet. Website. Accessed: 02.20.2018.
URL: *https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet.html*

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons, New York.

Cohen, A. and Naus, J. (2007). A representative sampling plan for auditing health insurance claims. *In Complex Datasets and Inverse Problems: Tomography, Networks and Beyond*, volume 54. Institute of Mathematical Statistics, 121–131.

David, G., Gunnarsson, C. L., Waters, H. C., Horblyuk, R., and Kaplan, H. S. (2013). Economic measurement of medical errors using a hospital claims database. *Value in Health*, **16**, 305–310.

GAO (2016). Medicare advantage. Website. Accessed: 02.20.2018.
URL: *https://www.gao.gov/assets/680/676441.pdf*

Ghosh, M., Mukhopadhyay, N., and Sen, P. K. (1997). *Sequential Estimation*, volume 904. John Wiley & Sons, New York.

Jhajj, H. and Walia, G. S. (2012). A generalised ratio and product type estimator for the population mean in stratified random sampling: Theory and methods. *South African Statistical Journal*, **46**, 53–64.

Johnson, M. E. and Nagarur, N. (2016). Multi-stage methodology to detect health insurance claim fraud. *Health Care Management Science*, **19**, 249–260.

Kelley, K., Bilson Darku, F., and Chattopadhyay, B. (2019). Sequential accuracy in parameter estimation for population correlation coefficients. *Psychological Methods*, **24**, 492–515.

Kelley, K., Darku, F. B. D., and Chattopadhyay, B. (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods*, **23**, 226–243.

Kim, Y. J., Oh, Y., Park, S., Cho, S., and Park, H. (2013). Stratified sampling design based on data mining. *Healthcare Informatics Research*, **19**, 186–195.

Li, J., Huang, K.-Y., Jin, J., and Shi, J. (2008). A survey on statistical methods for health care fraud detection. *Health Care Management Science*, **11**, 275–287.

Mukhopadhyay, N. and Chattopadhyay, B. (2011). Estimating a standard deviation with u-statistics of degree more than two: the normal case. *Journal of Statistics: Advances in Theory and*

Applications, **5**, 93–130.

Mukhopadhyay, N. and Chattopadhyay, B. (2012). Asymptotic expansion of the percentiles for a sample mean standardized by GMD in a normal case with applications. *Journal of the Japan Statistical Society*, **42**, 165–184.

Mukhopadhyay, N. and De Silva, B. M. (2009). *Sequential Methods and their Applications*. Chapman and Hall/CRC, Boca Raton, Florida.

Singh, H. P. and Vishwakarma, G. K. (2008). A family of estimators of population mean using auxiliary information in stratified sampling. *Communications in Statistics – Theory and Methods*, **37**, 1038–1050.

Sisodia, B. and Dwivedi, V. (1981). Modified ratio estimator using coefficient of variation of auxiliary variable. *Indian Society of Agricultural Statistics*, **33**, 13–18.

Solanki, R. S. and Singh, H. P. (2016). An improved estimation in stratified random sampling. *Communications in Statistics – Theory and Methods*, **45**, 2056–2070.

Upadhyaya, L. N. and Singh, H. P. (1999). Use of transformed auxiliary variable in estimating the finite population mean. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, **41**, 627–636.

WHCA (2014). Report on assessment of audit sampling and extrapolation process. Website. Accessed: 02.20.2018.
URL: *https://www.bidnet.com/bneattachments?/341233066.pdf*

Yitzhaki, S. et al. (2003). Gini's mean difference: A superior measure of variability for non-normal distributions. *Metron*, **61**, 285–316.

Manuscript received 2019-02-20, revised 2019-12-26, accepted 2020-02-13.