Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Model-based stratification of payment populations in Medicare integrity investigations

SURVEY
METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

by Don Edwards, Piaomu Liu and Alexandria Delage

Release date: January 3, 2024

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “Contact us” > “Standards of service to the public.”

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Model-based stratification of payment populations in Medicare integrity investigations

Don Edwards, Piaomu Liu and Alexandria Delage[1]

Abstract

When a Medicare healthcare provider is suspected of billing abuse, a population of payments X made to that provider over a fixed timeframe is isolated. A certified medical reviewer, in a time-consuming process, can determine the overpayment $Y = X -$ (amount justified by the evidence) associated with each payment. Typically, there are too many payments in the population to examine each with care, so a probability sample is selected. The sample overpayments are then used to calculate a 90% lower confidence bound for the total population overpayment. This bound is the amount demanded for recovery from the provider. Unfortunately, classical methods for calculating this bound sometimes fail to provide the 90% confidence level, especially when using a stratified sample.

In this paper, 166 redacted samples from Medicare integrity investigations are displayed and described, along with 156 associated payment populations. The 7,588 examined (Y, X) sample pairs show (1) Medicare audits have high error rates: more than 76% of these payments were considered to have been paid in error; and (2) the patterns in these samples support an "All-or-Nothing" mixture model for (Y, X) previously defined in the literature. Model-based Monte Carlo testing procedures for Medicare sampling plans are discussed, as well as stratification methods based on anticipated model moments. In terms of viability (achieving the 90% confidence level) a new stratification method defined here is competitive with the best of the many existing methods tested and seems less sensitive to choice of operating parameters. In terms of overpayment recovery (equivalent to precision) the new method is also comparable to the best of the many existing methods tested. Unfortunately, no stratification algorithm tested was ever viable for more than about half of the 104 test populations.

Key Words: Medicare fraud; All-or-nothing mixture model; Dalenius-Hodges stratification; Anticipated moments; Neyman allocation.

1. Introduction

According to the U.S. Centers for Medicare and Medicaid Services (CMS) 2022 Trustees Report (https://www.cms.gov/files/document/2022-medicare-trustees-report.pdf), the Medicare Trust Fund Hospital Insurance trust is estimated to be depleted by 2028. In an effort to extend the depletion date, over the last decade the CMS and the legislatures have focused efforts on eliminating fraud, improving quality and reducing overall cost of care (Huffman, 2021; Salmond and Echevarria, 2017). The CMS uses the Comprehensive Error Rate Testing (CERT, https://www.cms.gov/Research-Statistics-Data-and-Systems/Monitoring-Programs/Improper-Payment-Measurement-Programs/CERT) program to estimate the Medicare Fee-for-Service (FFS) program's improper payment rate each year by sampling claims to determine whether they were paid properly under Medicare coverage, coding and payment rules. In 2022, the improper payment rate was estimated at 7.46 percent, representing $31.46 billion in improper payments. It is clear that even with the focused effort on eliminating fraud, waste, and abuse from the Medicare FFS program that there are still program integrity concerns related to improperly paid claims, and that these will continue (Clemente, McGrady, Repass, Paul III and Coustasse, 2018).

[1] Don Edwards, 6 Carpenter Road, Tybee Island GA 31328. E-mail: edwards@stat.sc.edu; Piaomu Liu, Dept. of Math. Sciences, Bentley University, 175 Forest Street, Waltham, MA 02452; Alexandria Delage, Palmetto GBA, 17 Technology Circle, Columbia, SC 29203.

In this paper, a “provider” is any entity that bills Medicare: physicians, home health providers, hospitals, hospice providers, durable medical goods providers, ambulance services, etc. The CMS uses contractors at multiple levels to investigate providers suspected of abusing the system. The current guidelines governing the investigation process are given in the Medicare Program Integrity Manual (MPIM), Chapter 8: https://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/Downloads/pim83c08.pdf

When there is reason to suspect that a provider is billing improperly to Medicare FFS, a Unified Program Integrity Contractor (UPIC), often a subsidiary of an insurance company, conducts a sampling investigation. First, the UPIC obtains detailed information on all Medicare claims paid to that provider for a specified period, usually 1-2 years: a population of payments. A sampling unit is chosen and a probability sample is designed and implemented. A Certified Medical Reviewer then examines the evidence in support of each sampled payment, obtaining the overpayment amount Y as

$$Y = \text{ overpayment} = (\text{amount paid X}) - (\text{amount justified by the evidence}).$$

Note that since the amount justified by the evidence is non-negative, the overpayment amount is bounded above by the payment amount. The medical review usually requires several months for a sample of moderate size such as those displayed in Section 2. The UPIC then uses the overpayment amounts to “extrapolate”: the UPIC calculates a 90% lower confidence bound for the total overpayment made to the provider over the specified time period. This lower bound is the amount demanded for recovery from the provider. (If the point estimate achieves “high precision”, it can be used as the demand amount. This is very rare).

The sampling unit in these investigations is usually either the paid claim or all claims paid for services to a particular beneficiary (more precisely, to a Medicare identification number, formerly called a Health Insurance Claim Number HICN). The sampling plans are nearly always simple random samples or stratified random samples with strata determined by payment amounts. The extrapolation method is usually the “mean per unit” approximate method based on the finite population Central Limit Theorem (Hájek, 1964; Li and Ding, 2017).

Specifically, consider a stratified random sample with L strata, denote stratum sizes by N_l and sample sizes by n_l, for $l = 1, 2, \ldots, L$. Let $N = \sum_{l=1}^{L} N_l$ and $n = \sum_{l=1}^{L} n_l$. For each l, let $W_l = N_l / N$ and let $\overline{Y}_l$ and S_l be the sample mean and standard deviation of stratum l overpayments, respectively. The CLT-based extrapolation has the form

$$N \sum_{l=1}^{L} W_l \overline{Y}_l \; - \; t_v N \sqrt{\sum_{l=1}^{L} W_l^2 \left(\frac{S_l^2}{n_l} \right) \left(1 - \frac{n_l}{N_l} \right)} \tag{1.1}$$

where t_v is the 90th percentile of Student’s t-distribution with v degrees of freedom. Some contractors use the approximate degrees of freedom due to Satterthwaite (1946), though some use the less conservative choices $(n - L)$ or infinite degrees of freedom (i.e., the standard Normal percentile point). The Satterthwaite degrees of freedom are at most $(n - L)$, so it is the most conservative of the three choices. For simple random sampling, equation (1.1), suppressing subscript l, reduces to

$$N\overline{Y} \ - \ t_{n-1} N\left(\frac{S}{\sqrt{n}}\right)\sqrt{1-\frac{n}{N}}. \tag{1.2}$$

The provider has the right to appeal the reviewer's overpayment determinations and the sampling methods employed by the UPIC. In this event, an independent Medicare Administrative Contractor (MAC) conducts a review of the sampling plan and overpayment determinations as the first level of appeal. If the provider further appeals the MAC's decision, a Qualified Independent Contractor conducts another review as the second level of appeal. If the provider is still unsatisfied, the matter can be appealed to the third level and heard in a formal Administrative Law Judge hearing. Further appeals are also possible. At any level of appeal, the provider may submit additional documentation to support payment on their claims.

Section 2 of this paper examines the raw data for 166 redacted samples examined by a single reviewer at a MAC during the period 2013-2020. This data motivates a statistical model for Medicare sample overpayments, discussed in Section 3. Section 4 discusses appropriate Monte Carlo testing methods for sampling plans given the model. Section 5 reviews existing methods for stratifying populations and proposes a new class of stratification methods. Section 6 gives the results of a Monte Carlo efficiency study comparing these stratification methods, conducted using 104 of the payment populations.

2. Honor the data: An exploration of 166 Medicare samples

The redacted samples, in rough chronological order 2013-2020, are numerically summarized and graphically displayed using scatterplots of overpayment versus payment amount at: https://drive.google.com/drive/folders/1CIzQKzN4-RIIY38WonSAU7ydICwb76wh. The reader is encouraged to take a few minutes to page through the plots to form his/her own opinions on patterns in the data. The investigations summarized here were conducted by four different program integrity contractors. Sixty-eight percent (113) of the samples were simple random samples and the remaining 32% (53) were stratified by payment amount. Total sample sizes varied between 25 and 159 with 90% of the samples having between 30 and 70 total payments. Seventy-five percent (124) of these samples were from home health providers, 16% (27) were from hospice providers, and the remaining 9% (15) included physicians, skilled nursing facilities, ambulance services, etc. Nine of the samples used the beneficiary's HICN as the sampling unit, 149 used the paid claim, and 8 were drawn and analyzed as pennysamples (Edwards, Gilliland, Ward-Besser and Lasecki, 2015).

Define the "taint" for a sample payment as the ratio of overpayment to payment. Figure 2.1 provides a schematic classifying the 7,588 sample points from the 166 samples into four zones based on taint value: Zone 1 (taint $\geq$ 0.95), Zone 0 ($0 \leq$ taint $\leq$ 0.05), the Negative Zone (taint $<$ 0), and the Partial Zone. Some observations:

1. More than 76% of these payments were adjudged to have been paid either partially or totally in error. The conventional "wisdom" that auditing error rates are typically low does not apply to Medicare investigations.

2. Only one negative overpayment occurred in 7,588 reviewed payments.
3. Except for "complete error" samples (overpayment equals payment, 24 samples), overpayment amount is not *linearly related* to payment amount in any of these samples, though the two variables are sometimes highly correlated (the important distinction between "correlated" and "linearly related", so eloquently sung by Anscombe's (1973) quartet, is unfortunately blurred at times in the literature, and by some data analysts). A handful of samples have numerous partial overpayments, in which case the overpayment-payment relationship is essentially formless within the triangle formed by Zones 0 and 1. For the majority of these samples, though, the relationship is best described as approximately *bilinear*, with a horizontal line in Zone 0, a 45º line in Zone 1, and possibly a handful of partial overpayments in between.
4. It is less obvious to the naked eye, but there is no evidence in this data that larger payments are more likely to have larger taints. Figure 2.2, admittedly crowded, shows a plot of taint versus the logarithm base 10 of payment amount, showing no obvious monotone trend. Figure 2.3 displays the results of a generalized additive model fit of taint on $\log_{10}$ (payment) using a separate intercept for each sample and a common smooth function for the taint vs. $\log_{10}$ (payment) relationship. The figure includes Scheffé-style simultaneous 95% confidence bands for the true regression function; these bands include a horizontal line. Analogous results were obtained using payment amounts instead of $\log_{10}$ (payment amounts) as regressor.

Figure 2.1 A classification of the 7,588 overpayment-payment pairs in the 166 samples.

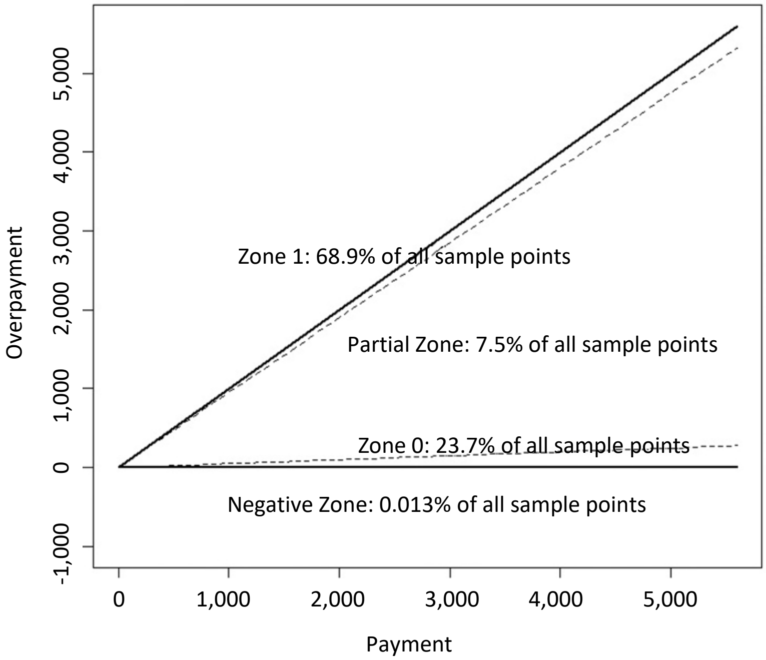


Figure 2.2 Taint versus log₁₀ (Payment amount) for 7,588 sample points.

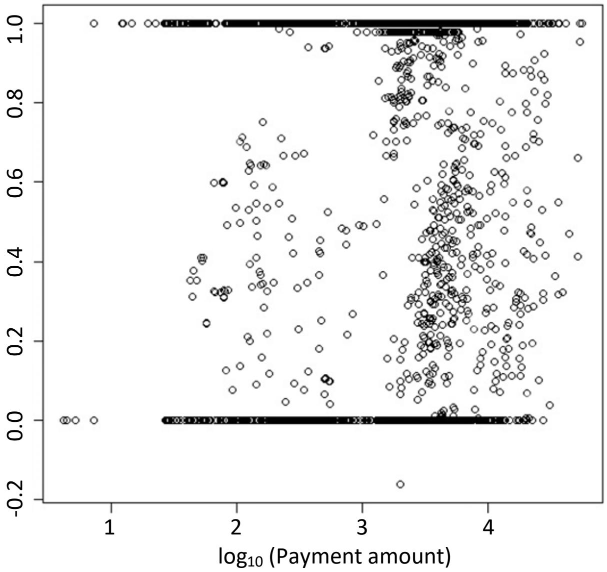
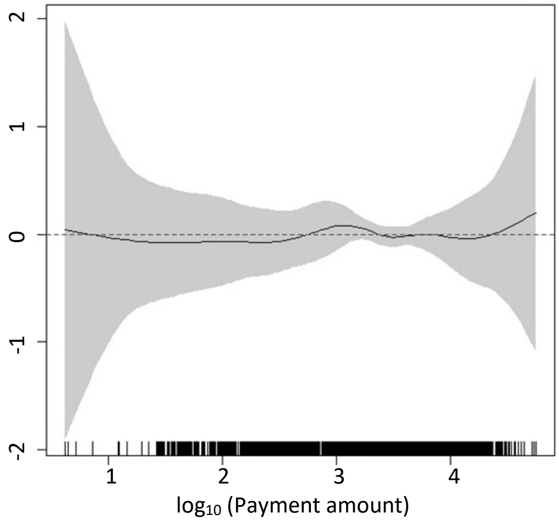


Figure 2.3 Graphical summary of a generalized additive model analysis of taint vs. log₁₀ (payment). Gray shading is a 95% Scheffé-style confidence region for the true nonparametric regression function.



3. A model for Medicare sample data

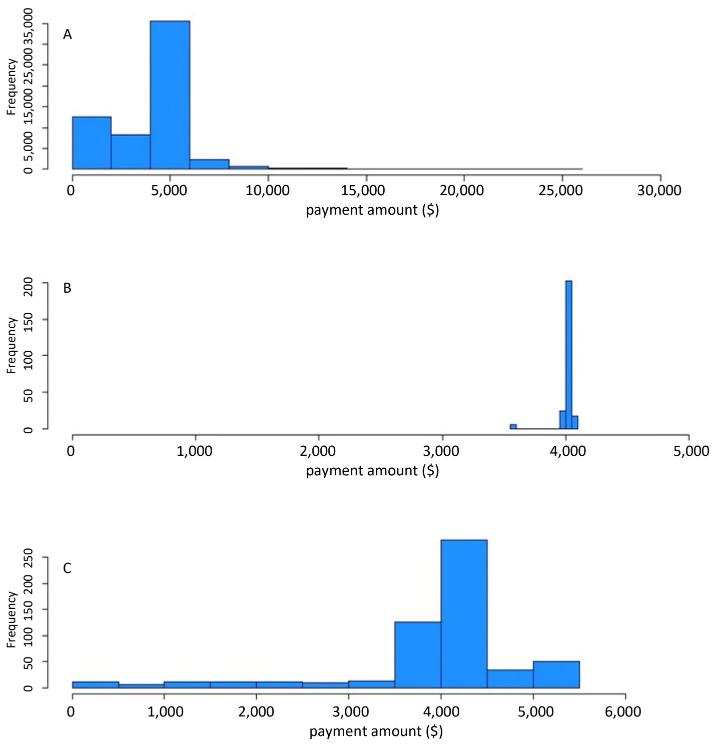
The patterns in the 166 samples suggest a simple mixture model for Medicare sample data. This model has previously been discussed by several authors, including King (1996), Edwards, Ward-Besser, Lasecki, Parker, Wieduwilt, Wu and Moorhead (2003), and King and Madansky (2013). The population payments $X_i, i = 1, 2, \ldots, N$ are known constants. Let $Z_i = 1$ with probability P_E, $0 \leq P_E \leq 1$, and $Z_i = 0$ otherwise, regardless of the value of X_i. The "All or Nothing" mixture model with error rate P_E, here abbreviated AN (P_E), has overpayment $Y_i = Z_i X_i$. That is, with probability P_E the overpayment equals the payment, and with probability $1 - P_E$ the overpayment is zero. The majority of the 166 samples discussed in Section 2 conform to this model with negligible numbers of partial or negative overpayments.

Let μ_X and σ_X^2 denote the known population mean and variance of payment amounts. Under the AN (P_E) model it is straightforward to derive the mean and variance of overpayment Y for a randomly selected X (King, 1996):

$$\begin{aligned} E(Y) &= \mu_Y = P_E \mu_X \\ \operatorname{Var}(Y) &= \sigma_Y^2 = \mu_X^2 P_E (1 - P_E) + P_E \sigma_X^2. \end{aligned} \tag{3.1}$$

Note that the variance of the overpayment amounts can be dramatically different from the variance of the payment amounts. To illustrate, Figure 3.1 shows histograms of three payment populations. We include 0 on the horizontal axis of these histograms in order to help envision the shape of the corresponding overpayment population under a given error rate P_E, as explained below. For population B, for values of P_E near $1/2$, the variance of overpayment amounts σ_Y^2 is more than 800 times the variance of payment amounts σ_X^2. For populations C and A the ratio σ_Y^2 / σ_X^2 reaches a maximum of 4.7 and 1.5 respectively. Hence, the common practice of using the variance of payment amounts as an estimate of the variance of overpayment amounts in sample-size determination formulas has no relevance except when P_E is very near to 1.

Figure 3.1 Example payment populations.



Notes: (A) N = 59,804 hospice HICN payments; (B) N = 249 power wheelchair claim payments; (C) N = 570 hospice claim payments.

4. Monte Carlo testing of sample designs under the AN model

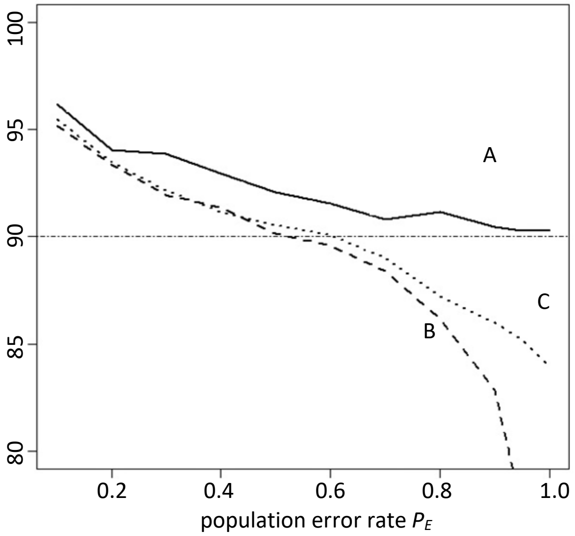
Consider a sampling-and-extrapolation plan proposed for a particular payment population. The $\mathrm{AN}(P_E)$ model motivates a simple Monte Carlo testing procedure for the plan, implemented by the freely-distributed **R** (**R** Core Team, 2021) function *samptest* available from the authors. The testing proceeds as follows: for each choice of P_E on a grid spanning the interval (0,1],

1. A plausible overpayment population is created by randomly choosing a proportion P_E to have overpayment = payment, leaving the remaining proportion $1 - P_E$ to have zero overpayment.
2. The sampling plan is applied to the created overpayment population. The extrapolation is computed and compared to the total overpayment for the population.

Steps 1 and 2 are repeated independently a large number of times, enough to estimate achieved confidence level to high accuracy. The plan investigated can be a simple random sample or a stratified random sample. The function summarizes the testing with plots of the estimated achieved confidence level, and the average percent of overpayment recovery for the plan, versus error rate P_E (the latter quantity is equivalent to average precision, as explained in Section 6). For most payment populations, the *samptest* results are returned in a few seconds, even using a computer of modest computing power.

For example, Figure 4.1 shows *samptest* results on achieved confidence levels for a simple random sample of 30 payments using the standard extrapolation (1.2) applied to the Figure 3.1 payment populations. The plot tells us that this modest sampling plan provides confidence level above or near the MPIM-required 90% level for population A for all error rates, but for populations B and C it fails to provide the 90% confidence level if the error rate P_E exceeds 0.6.

Figure 4.1 Achieved confidence levels (%) versus error rate for the Figure 3.1 populations, using a simple random sample of 30 payments with the standard extrapolation (1.2).



Note: Estimates are accurate to ±1% with 95% confidence.

These results illustrate the strong effect of skewness in the overpayment population on the achieved confidence level of the bound (1.2). Cochran (1977, pages 39-44) noted that a right-skewed population yielded a conservative lower bound and a liberal upper bound for a two-sided confidence interval for the population mean. Using this "principle of skew", which in our experience is pervasive, the conservatism of the lower bound (1.2) for Population A at all error rates can be anticipated: at any error rate P_E the overpayment population has a "spike" of zeros of (approximate) size $\mathrm{N}(1-P_E)$, with the remaining $\mathrm{N}P_E$ overpayments having the same shape as the payment population. Hence at any error rate the overpayment population for population A is also right skewed; hence the bound (1.2) is conservative. At any error rate P_E, overpayment population B has two spikes, with $\mathrm{N}(1-P_E)$ zeros and $\mathrm{N}P_E$ payments near $4,000. For values of P_E above $1/2$ the spike at 0 is the shorter of the two, hence the overpayment population is left-skewed, hence the bound (1.2) is liberal. Finally, payment population C is left-skewed. For small values of P_E the large spike at 0 in the overpayment population will counterbalance this left skew and yield a conservative lower bound, but the spike at 0 gradually disappears as P_E grows, yielding a liberal lower bound at high error rates.

If problems with the achieved confidence level occur using extrapolation (1.2) they invariably occur at high error rates due to left skew in the overpayment population. For this reason, the presence of partial overpayments improves the achieved confidence level: at any particular $P_E > 1/2$, partial overpayments reduce the severity of any left skew. This has been repeatedly confirmed in our experience using *samptest*, which has the capability to flexibly model the occurrence of partial overpayments. We conjecture that any sampling plan which achieves the 90% confidence level under the $\mathrm{AN}(P_E)$ model for a given error rate P_E will also achieve it in the presence of partial overpayments at that value. That is, the presence of partial overpayments tends to improve the confidence level of the bound (1.2). This effect is usually modest since the frequency of partial overpayments is usually modest.

Right-skewed payment populations such as Population A occur frequently in Medicare investigations. For such a population, a simple random sample will usually achieve confidence level above 90% for all error rates. This conservatism signals the potential for improving the sampling plan via stratification. Care must be taken, however, for in a sample stratified by payment amount the effects of skew are mixed. For example, particularly for $L \geq 3$, stratification by payment amount creates payment subpopulations similar in shape to that shown in Figure 3.1B. These strata give rise to left-skewed overpayment strata at high error rates, which will not preserve the confidence level. King and Madansky (2013) say it well: "it is quite possible to destroy the achieved confidence level by stratifying carelessly. We therefore underscore the cautionary note that a misapplication of stratification (e.g., not using the appropriate stratification boundaries) will do more than invalidate the precision of the bounds; it can invalidate the normal confidence coefficient as well". Any stratification scheme must be tested using Monte Carlo.

Note that in some situations a transformation (e.g., the log transformation) may normalize the distribution well enough to provide a viable 90% confidence bound for the population mean of log(Y), but this cannot be back-transformed to obtain a confidence bound for the population mean of Y: the mean of the logs is not the log of the mean. If the transformation symmetrizes the distribution, the back-transformed lower bound is a confidence bound for the population median of Y, not the population mean, and when multiplied by N does not provide a bound for the population total, but something much smaller. The same argument applies to other non-linear transformations, e.g., the square root transformation, inverse transformation, etc.

5. Methods for constructing strata using an auxiliary variable X

Assume that X is known for all sampling units. In Medicare investigations, X is usually the total paid amount for the sampling unit. This section reviews some existing methods for stratifying the population using X and defines a new method.

In some designs, a few of the largest payment amounts are examined in their entirety in a "take all" or "certainty" stratum. This reduces the severity of right skew in the remaining population. If such a certainty stratum is used, the question remains as to how one should stratify the remaining population, and that is our focus here.

It should be stated at the outset that there is no need to use an algorithm from the peer-reviewed literature to determine values of X ("cutpoints"), $XC_1 < XC_2 < \ldots < XC_{L-1}$ defining the $L \geq 2$ strata. In practice, any *a priori* choice achieving confidence level very near to or above 90% for all error rates is viable, regardless of how it was obtained. The time required to find such cutpoints by trial and error, testing each choice using Monte Carlo, may be prohibitive, however. Also, it is sometimes useful to cite the use of an existing algorithm from the peer-reviewed literature when creating a stratified design.

5.1 Existing methods

Several popular methods to determine cut points or stratum boundaries using X exist. With Y the measured variable of interest, Tschuprow (1923) and Neyman (1934) proved that the variance of the linear population mean estimator $\overline{y}_s = \sum_{l=1}^{L} W_l \overline{y}_l$ is minimized for a fixed total sample size n when the sample size n_l in stratum l is proportional to the product of stratum size N_l and stratum standard deviation $\sigma_{Yl}, l = 1, 2, \ldots, L.$ This "Neyman allocation" is the goal for several stratification algorithms under various assumptions. Methodological research in this area dates back to Dalenius and Gurney (1951) and Cochran (1977, Chapter 5A). Relevant literature is extensive, including Ekman (1959), Serfling (1968), Singh (1971), Wang and Aggarwal (1984), Hidiroglou and Srinath (1993), Hidiroglou (1994), Hedlin (2000), Kozak and Verma (2006), Jurina and Gligorova (2017), Hidiroglou and Kozak (2018) and Reddy and Khan (2019).

When the auxiliary variable X is linearly related to $Y,$ and highly correlated with $Y,$ the well-known Dalenius and Hodges (1959) method can achieve approximate Neyman allocation. The method begins with a frequency distribution f for $X.$ Under the additional assumption that the distribution of X within the frequency distribution cells is approximately uniform, choosing stratum boundaries that equate the cumulative $\sqrt{f}$ between strata achieves approximate Neyman allocation for equal stratum sample sizes.

Lavallée and Hidiroglou (1988) take an iterative approach to choosing stratum cutpoints to minimize the total sample size n given a specified relative coefficient of variation for the point estimator of the population total; this problem is equivalent to minimizing the relative coefficient of variation for fixed $n,$ which is our goal. Their approach assumes that the auxiliary variable X is "closely related to" $Y,$ and they warn that their method will not achieve desired efficiency if X and Y are not "highly correlated". The iterative algorithm they use, due to Sethi (1963), was improved by Kozak (2004).

For sampling plans where the stratification variable X is not linearly related to the survey variable $Y,$ the "anticipated moments" under a model (for example, Dayal (1985), Sigman and Monsour (1995) and Sweet and Sigman (1995)) is a critical concept. Generalizing the algorithm in Lavallée and Hidiroglou (1988) and using "anticipated moments" of Y given $X,$ Rivest (1999, 2002) proposed stratification algorithms and models that account for discrepancies between Y and $X,$ in particular (1) a Log-Linear (LL) model, i.e., $\log(Y)$ is linearly related to $\log(X)$ and (2) a "random replacement" model. The latter models $Y = X$ with high probability, but otherwise Y is equal to a randomly selected value of $X.$ The algorithms in Rivest (1999, 2002) choose sample size to achieve a pre-specified level of precision, or to maximize precision for a fixed sample size, allowing for different sample allocation rules. The statistical models and anticipated moment approach in Rivest (1999, 2002) provide a very general and flexible method covering a wide range of survey scenarios.

Baillargeon and Rivest (2009) extend the Log-Linear (LL) model in Rivest (2002) to include a survival probability that models the probability of survey variable Y taking on value 0 (equation (5.1)). The survival probability can be specified to vary by stratum. This model is particularly helpful for business surveys where a business is no longer in operation when the survey takes place but the X variable has been collected resulting in a zero value for $Y.$ Let p_l denote survival probability for the l^{th} stratum; assuming $\epsilon \sim N(0, \sigma^2),$ the LL model is

$$Y = \begin{cases} \exp(\alpha + \beta \log X + \epsilon), & \text{with probability } p_l \\ 0, & \text{with probability } 1 - p_l. \end{cases} \tag{5.1}$$

For business surveys, the p_l values typically increases with $X,$ which means a greater chance of business survival or Y being non-zero when the value of X is large. When a constant survival probability is specified across all strata, the probability of Y being zero is the same for all sampling units. In this case, if $\beta = 0,\ \sigma^2 = 0,$ the model in equation (5.1) is equivalent to the All-or-Nothing model in Section 2 of this paper. This is a testimony to the generality of the models proposed by Rivest (1999, 2002) and Baillargeon and Rivest (2009).

A simpler method for determining cut points of stratification for right-skewed populations is due to Gunning and Horgan (2004) and Gunning, Horgan and Yancey (2004). The method is based on an observation of Cochran (1961) that the coefficients of variation across different strata are comparable in near-optimum stratification. Assuming the coefficient of variation is constant across all strata, boundary points can be written as terms in a geometric series. Once the minimum and maximum values of the set of cut points are specified, the geometric relationship produces cutpoints for all strata. Gunning and Horgan state that their method assumes that Y is "highly correlated with" X. It also assumes that distributions within strata are uniform for efficiency.

5.2 Sample allocations

For a given sample size n or a precision level, values of the boundary points are affected by the sample allocation scheme (Lavallée and Hidiroglou, 1988; Hidiroglou and Srinath, 1993; Horgan, 2006). Common allocation schemes include but are not limited to the power allocations, such as the Y-proportional and N-proportional power allocations (Lavallée and Hidiroglou, 1988; Hidiroglou and Kozak, 2018). A general expression of allocation schemes is described in equation (5.2) below, which is included in the **R** package *stratification*. For stratum l, the sample size is $n_l = na_l$. Combinations of parameters q_1, q_2 and q_3 produce different sample allocations, where $0 \leq 2q_1 \leq 1$, $0 \leq 2q_2 \leq 1$ and $0 \leq 2q_3 \leq 1$. $N_l, \bar{Y}_l$ and S_l are the size, mean and standard deviation of stratum l, respectively.

$$a_l = \frac{N_l^{2q_1} \bar{Y}_l^{2q_2} S_l^{2q_3}}{\sum_{l=1}^{L} N_l^{2q_1} \bar{Y}_l^{2q_2} S_l^{2q_3}}. \tag{5.2}$$

For instance, using the Dalenius-Hodges method where equal sample size achieves the goal of approximating the Neyman allocation, setting $q_1 = q_2 = q_3 = 0$ imples allocating an equal number of observations across the strata. For a power allocation where the power is 0.7, one specifies $q_1 = q_2 =$ 0.35 and $q_3 = 0$. For a Y-proportional power allocation, the combination of parameters $q_1 = q_3 = 0$, and $q_2 =$ 0.35 yields a power of 0.7.

To implement the aforementioned stratification methods under various sample allocation schemes, the **R** package *stratification* (see Baillargeon and Rivest, 2011) provides functions *strata.cumrootf()*, *strata.LH()* and *strata.geo()* to implement the Dalenius-Hodges, Lavallée-Hidiroglou, and Gunning-Horgan methods to produce cut points for a specified number of strata, respectively. The package also implements the alternatives studied by Rivest (2002) and Baillargeon and Rivest (2009).

5.3 A simple new method using anticipated AN-model moments and equal sample sizes

The well-known stratification method due to Dalenius and Hodges, under the model $\mathrm{Y} \approx \mathrm{X}$, chooses stratum cutpoints to achieve near-equality of the quantities $N_l \sigma_{Yl}$. In that case, using equal sample sizes corresponds to Neyman allocation. The relatively simple new method described in this section mimics this approach using the AN model at a particular value P_E. For a given value P_E, the $\mathrm{ESS}(P_E)$ algorithm determines cutpoints by nearly equalizing the quantities $N_l \sigma_{Yl}$ determined by the AN-model moments (3.1).

Thus, for equal sample sizes and this value of P_E, the $\text{ESS}(P_E)$ method provides Neyman allocation, maximizing precision for a given n.

For example, when $L = 2$, a specified cutpoint XC_1 determines N_1 and N_2 as well as μ_{1X}, σ_{1X}^2, μ_{2X} and σ_{2X}^2. Specifying P_E, we then use equation (3.1) to compute σ_{1Y}^2 and σ_{2Y}^2. The final cutpoint is found by iterating on XC_1 until $(N_1\sigma_{1Y} - N_2\sigma_{2Y})^2$ is minimized. For general L, our algorithm (available from the second author) uses a nonlinear search through the R function *optimize* for univariate optimization when $L = 2$, and *constrOptim* for higher-dimensional optimization when $L \geq 3$, to find cut points to minimize the corrected sum of squares for the quantities $N_l\sigma_{lY}$. Exhaustive searches are available but computationally feasible only for $L \leq 3$.

The question remains: how does one choose P_E? Fortunately, as will be seen in the next section, the operating characteristics of $\text{ESS}(P_E)$ seem rather robust to the choice. An idea worth exploring, suggested by the Associate Editor, is to initially choose P_E to maximize (3.1). We recommend exploring the properties of a grid of choices using Monte Carlo prior to making the final choice.

6. Efficiency comparisons for the stratification methods

Equations (1.1) and (1.2) can be written informally as

$$\text{(overpayment recovery)} = \text{(point estimate)} - \text{(margin of error)}. \tag{6.1}$$

The precision of a sample is defined to be its margin of error expressed as a percentage of its point estimate, with precision being considered "high" if this percentage is small. Some would not define precision to include the t critical point; the gist of what is said below (that the average percent of overpayment recovered is inversely related to average precision) still holds under different definitions of precision. Traditionally, analysts seek to create a sample design having high average precision. Averaging (6.1) over all possible samples, and using the fact that the point estimator is unbiased, we obtain

$$\text{(Average overpayment recovery)} = \text{(true totol overpayment)} - \text{(average margin of error)}.$$

Dividing both sides of the above by true total overpayment and multiplying by 100%, we obtain

$$\text{(Average overpayment recovery percent)} = 100\% - \text{(average precision)},$$

Therefore, seeking a design with high average precision is equivalent to seeking a design with high average overpayment recovery. For example, a sample design with 10% average precision has 90% average overpayment recovery. For any sample design, both quantities will vary dramatically depending on the error rate P_E.

We prefer quantifying the efficiency of a design in terms of the average overpayment recovery as opposed to average precision. This facilitates determination of the cost-effectiveness of certain design decisions, such as increasing sample size or paying an analyst to spend an extra hour searching for efficient stratification schemes. In this section, we compare the efficiency of the major stratification algorithms discussed in Section 5.

Of the 166 samples discussed in Section 2, a total of 156 payment populations, matched by number, were available for study. These populations are numerically summarized and graphically depicted at https://drive.google.com/drive/folders/1-7M-4R3KPcCgfPmPWOo24Qc5AOAdOfV7?usp=sharing. Of these, 104 payment populations were selected as candidates for stratification. For these 104 test populations, simple random samples (SRS) of size 30 achieved estimated confidence level above 89% across all 7 tested error rates $P_E = 0.10$, 0.30, 0.50, 0.70, 0.9, 0.95, and 1.0 using 10,000 generated overpayment populations, each with its sample and extrapolation. It is assumed without loss of generality that, if a certainty (or "take-all") stratum is to be used that these few largest payments have already been removed, to be examined in their entirety. Stratification strategies apply only to the remaining payments

In our testing, designs with total sample sizes 30, 60, or 90 were considered, with L = 2, 3, or 4 strata, for a total of 104*9 = 936 test cases. The stratification methods due to Delanius-Hodges (DH), Lavallée-Hidiroglou (LH), Gunning-Horgan (Geo) and the Log-Linear (LL) models were tested using the **R** package *stratification* to generate cutpoints. Three sample allocations were tested for the LH, Geo and LL methods: Y-proportional and N-proportional power allocations with power $p = 0.7$ as well as a power allocation where $q_1 = q_2 = 0.35,\ q_3 = 0$ are specified in equation (5.2). The stratum sample sizes were determined by the **R** functions that generated the cutpoints. For the other methods, stratum sample sizes were taken to be equal, except when $L = 4$ and $n = 30$ or 90, where slight inequalities occurred.

To fully define an ESS stratification method, we must choose a value for P_E. Hopefully the choice will be viable and have good recovery properties regardless of the true value of P_E, and that is what the simulation testing is meant to answer. In this study, the ESS cutpoint algorithm was tested at $P_E = 0.2$ (ESS20), $P_E = 0.5$ (ESS50), $P_E = 0.8$ (ESS80). The LL models are also tested with survival probabilities 0.2, 0.5 and 0.8. The Monte Carlo testing for all cases and methods with 10,000 iterations at each error rate required approximately 27 hours on Bentley University's High Performance Computing cluster. The conservative Satterthwaite (1946) degrees of freedom were used for all methods.

For each of the 936 test cases, a stratification method was considered viable if it achieved estimated confidence above 89% for all 7 tested error rates. Table 6.1 shows, for each of the nine sample size – number of strata combinations, the number of test populations for which each method produced a viable stratification scheme. Only simulation results from the Y-proportional power allocation are reported as these allocation methods yielded better viability than the other two sample allocations. The table shows that stratification is more likely to be viable for small L and large n. The Geo method is viable more often than any other stratification method; unfortunately, it will also be seen that it is usually less efficient than a simple random sample of the same size. Of the other stratification methods, the LL20 method, the ESS20 and the ESS50 methods were comparable in terms of viability and noticeably better in this respect than the other methods tested (see the last column of Table 6.1). Notably, these three methods were viable almost twice as often as the Dalenius-Hodges method, and almost three times as often as the Lavallée-Hidiroglou method. Unfortunately, in no case did any algorithm (except Geo) achieve viability for more than about half of the test populations.

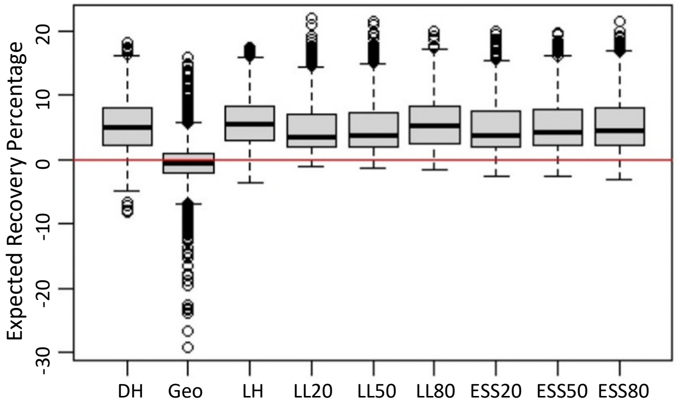
In order to measure the efficiency of (for example) the Dalenius-Hodges method relative to simple random sampling, for each sample size, L-value, and population for which the DH method was viable, the difference in average overpayment recovery in the order DH-SRS was computed for each error rate. Analogous differences were computed for the other stratification methods. Figure 6.1 shows boxplots of these differences. It is evident from this graphic that the Geo method improved on SRS for less than half of the cases for which it was viable. Additionally, when the DH, LH, LL and ESS methods were viable, they improved the overpayment recovery compared to SRS by a median amount of 4-6%. In more than a few cases they improved the recovery dramatically, by 10-22%. However, the recovery using DH can sometimes be worse than SRS by more than 9%. In contrast, SRS was never better than any LL method by more than 1.5%, and never better than any ESS method by more than 3%. Both the LL and ESS methods can improve on SRS by up to 22%.

Table 6.1
Method viability: number of test populations for which a stratification method (see text) achieved estimated confidence at least 89% over all tested error rates, versus number of strata L and total sample size n under Y-proportional power allocation with power value 0.7.

Methods	L = 2			L = 3			L = 4			Total viability frequency
	n = 30	n = 60	n = 90	n = 30	n = 60	n = 90	n = 30	n = 60	n = 90	
DH	22	30	37	5	8	10	1	3	4	120
LH	19	22	25	4	5	5	0	0	0	80
Geo	90	91	83	52	62	61	5	15	23	482
LL20	31	43	57	19	19	26	9	12	15	231
LL50	28	42	47	14	18	20	5	11	10	195
LL80	21	28	36	7	9	14	2	4	5	126
ESS20	33	41	51	17	22	27	10	13	13	227
ESS50	33	44	45	18	23	24	8	8	12	215
ESS80	30	33	44	11	13	17	5	5	6	164
SRS	104	104	104	104	104	104	104	104	104	936

Notes: DH = Dalenius-Hodges; ESS = Equal sample sizes; Geo = Gunning-Horgan; LH = Lavallée-Hidiroglou; LL = Log-linear; SRS = Simple random samples.

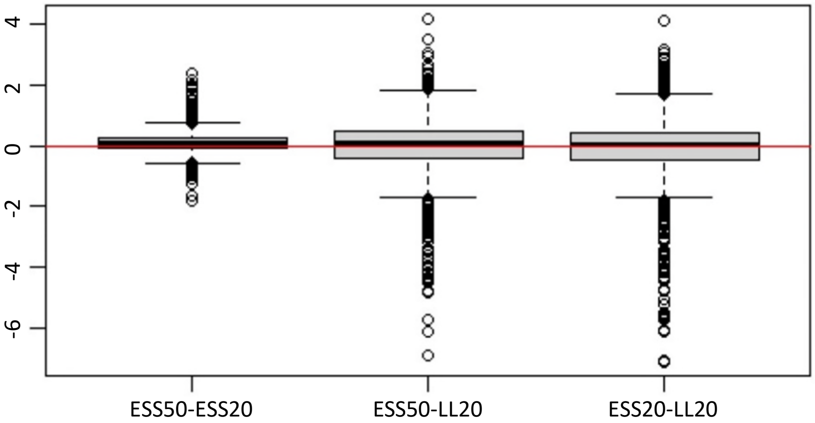
Figure 6.1 Improvement in overpayment recovery versus a simple random sample of the same size for each viable method (see text for details).



Notes: DH = Dalenius-Hodges; ESS = Equal sample sizes; Geo = Gunning-Horgan; LH = Lavallée-Hidiroglou; LL = Log-linear.

The top three algorithms were further compared via calculating differences in the order ESS50-ESS20, ESS50-LL20, and ESS20-LL20 whenever the differenced algorithms were both viable, for all error rates and test cases. These differences are displayed in Figure 6.2. The figure shows that the ESS50 and ESS20 methods had similar efficiency, suggesting that choice of P_E in this range is not of critical importance for efficiency. Overall, the ESS methods and LL20 yielded similar expected overpayment recovery. On average, the ESS methods produced slightly higher medians than those of the LL20. However, the LL20 method had slightly higher mean expected overpayment recovery. In some cases, the ESS methods outperformed the LL20 method by over 4% while in other cases they were worse-off than the LL20 method by a little bit over 6%.

Figure 6.2 Relative efficiency of the ESS50, ESS20 and LL20 methods (see text for details).



Notes: ESS = Equal sample sizes; LL = Log-linear.

7. Discussion and conclusion

This paper’s major accomplishment is its unprecedented sharing of raw data from Medicare investigations. Despite this, these 166 shared samples cannot be considered representative of all such investigations. In particular these samples are taken from the first level of appeal; error rates tend to decrease at higher levels of appeal as providers challenge the overpayment determinations. The samples displayed here are also primarily from home health or hospice providers, a major shortcoming.

We see no reason why similar sharing of thoroughly redacted sample and payment population data could not be done by all MACs and at higher levels of appeal. If this was done, say, on a triannual basis, the anonymity of the individual providers and UPICs would be protected. More data sharing would allow for further refinement of models for the overpayment – payment relationship. It would also be very useful to know if negative overpayments are as rare and negligible at other levels of appeal, and for other provider lines of business, as they were for our samples.

Our data showed 24 “complete error” samples, i.e., samples where every overpayment equaled its payment. Use of the bounds (1.1) or (1.2) based on the finite population Central Limit Theorem can be problematic in these cases: at very high error rates, it is not unusual for these lower confidence bounds for total overpayment to exceed the total payment amount, which renders the bound indefensible. The bounds (1.1) and (1.2) can also fail to provide confidence level at least 90% at high error rates for payment populations similar to populations B and C in Figure 3.1. Alternative extrapolation methods based on the hypergeometric probability distribution (Edwards et al., 2003; Gilliland and Feng, 2010; Edwards et al., 2015) have been developed for situations where the bounds (1.1) and (1.2) tend to fail. These alternative methods always provide lower 90% confidence bounds less than the total payment amount. They are mathematically conservative – guaranteed to provide confidence level at least 90% – as long as negative overpayments do not occur with any great frequency or severity. More sharing of data is needed to shed light on the safety of this assumption.

With Medicare data, the frequency for which extrapolations using (1.1) and (1.2) are not viable is disappointing, to say the least. Alternative approaches using empirical likelihood confidence bounds (Chen, Chen and Rao, 2003; Rao and Wu, 2009) hold great promise. Unfortunately, simulation studies on empirical likelihood methods to date have not included error rates greater than 40%, which are the norm for Medicare samples. A comprehensive study of empirical likelihood approaches using Medicare data is beyond the scope of this paper but is the focus for an ongoing project by the authors.

This paper proposes a new stratification method as a special case of the anticipated moment method based on the All-or-Nothing model. The efficiency study provided here found this new ESS method to be competitive with the best of the many existing methods tested from the **R** package *stratification*. Specifically, the Log-Linear model with survival probability of 0.2 (LL20) under the Y-proportional power allocation ($p = 0.7$) showed a slight advantage in viability over the best ESS methods. Finding this particular choice of operating parameters in the wide range available in *stratification* required the advice of an expert and hours of testing, however. In contrast, the near-equivalent ESS choices ($P_E = 0.2$ or 0.5) were among the first methods we tested. However, no one method dominated all other methods in all situations. And, for every choice of n and L, there were many test populations for which no stratification algorithm was viable.

As is often the case, these results lead to new questions. Are there aspects of the payment population that can provide clues as to when stratification will lead to improved efficiency? Are there aspects of the payment population that can provide clues as to which method is the best choice? The authors will be pursuing these and other questions and invite collaboration from others.

Acknowledgements

The authors thank the CMS for permission to use and display the data from these samples and populations, and thank the associate editor and reviewers for their hard work and helpful suggestions.

Piaomu Liu thanks Bentley University for providing High Performance Computing capabilities to perform efficiency studies in the research.

References

Anscombe, F.J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1), 17-21.

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

Baillargeon, S., and Rivest, L.-P. (2011). The construction of stratified designs in R with the package *stratification*. *Survey Methodology*, 37, 1, 53-65. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011001/article/11447-eng.pdf.

Chen, J., Chen, S.-Y. and Rao, J. (2003). Empirical likelihood confidence intervals for the mean of a population containing many zero values. *Canadian Journal of Statistics*, 31(1), 53-68.

Clemente, S., McGrady, R., Repass, R., Paul III, D.P. and Coustasse, A. (2018). Medicare and the affordable care act: Fraud control efforts and results. *International Journal of Healthcare Management*, 11(4), 356-362.

Cochran, W.G. (1961). Comparison of methods for determining stratum boundaries. *Bulletin of the International Statistical Institute*, 35, 345-358.

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

Dalenius, T., and Gurney, M. (1951). The problem of optimum stratification. ii. *Scandinavian Actuarial Journal*, 1951(1-2), 133-148.

Dalenius, T., and Hodges Jr., J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54(285), 88-101.

Dayal, S. (1985). Allocation of sample using values of auxiliary characteristic. *Journal of Statistical Planning and Inference*, 11(3), 321-328.

Edwards, D., Gilliland, D., Ward-Besser, G. and Lasecki, J. (2015). Conservative penny sampling. *Journal of Survey Statistics and Methodology*, 3(4), 504-523.

Edwards, D., Ward-Besser, G., Lasecki, J., Parker, B., Wieduwilt, K., Wu, F. and Moorhead, P. (2003). The minimum sum method: A distribution-free sampling procedure for medicare fraud investigations. *Health Services and Outcomes Research Methodology*, 4(4), 241-263.

Ekman, G. (1959). An approximation useful in univariate stratification. *The Annals of Mathematical Statistics*, 30(1), 219-229.

Gilliland, D., and Feng, W. (2010). An adaptation of the minimum sum method. *Health Services and Outcomes Research Methodology*, 10(3), 154-164.

Gunning, P., and Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, 30, 2, 159-166. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2004002/article/7749-eng.pdf.

Gunning, P., Horgan, J.M. and Yancey, W. (2004). Geometric stratification of accounting data. *Contaduría y Administración*, (214), 0.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 1491-1523.

Hedlin, D. (2000). A procedure for stratification by an extended ekman rule. *Journal of Official Statistics*, 16(1), 15.

Hidiroglou, M.A. (1994). Sampling and estimation for establishment surveys: Stumbling blocks and progress. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 693-698.

Hidiroglou, M.A., and Kozak, M. (2018). Stratification of skewed populations: A comparison of optimisation-based versus approximate methods. *International Statistical Review*, 86(1), 87-105.

Hidiroglou, M.A., and Srinath, K. (1993). Problems associated with designing subannual business surveys. *Journal of Business & Economic Statistics*, 11(4), 397-405.

Horgan, J.M. (2006). Stratification of skewed populations: A review. *International Statistical Review*, 74(1), 67-76.

Huffman, M. (2021). Value-based care: An executive briefing. *Nurse Leader*, 19(1), 82-86.

Jurina, I., and Gligorova, L. (2017). Determination of the optimal stratum boundaries in the monthly retail trade survey in the croatian bureau of statistics. *Romanian Statistical Review*, (4).

King, B. (1996). Sampling design issues when dealing with zeros. *Proceedings of the Survey Research Methods Section*, 400-405.

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

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

Kozak, M., and Verma, M.R. (2006). Geometric versus optimization approach to stratification: A comparison of efficiency. *Survey Methodology*, 32, 2, 157-163. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9550-eng.pdf.

Lavallée, P., and Hidiroglou, M.A. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 1, 33-43. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1988001/article/14602-eng.pdf.

Li, X., and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520), 1759-1769.

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

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http: //www.R-project.org/.

Rao, J., and Wu, C. (2009). Empirical likelihood methods. Elsevier, *Handbook of Statistics*, 29, 189-207.

Reddy, K.G., and Khan, M.G. (2019). Optimal stratification in stratified designs using weibull-distributed auxiliary information. *Communications in Statistics-Theory and Methods*, 48(12), 3136-3152.

Rivest, L. (1999). Stratum jumpers: Can we avoid them. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 64-72.

Rivest, L.-P. (2002). A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, 28, 2, 191-198. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2002002/article/6432-eng.pdf.

Salmond, S.W., and Echevarria, M. (2017). Healthcare transformation and changing roles for nursing. *Orthopedic Nursing*, 36(1), 12.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6), 110-114.

Serfling, R. (1968). Approximately optimal stratification. *Journal of the American Statistical Association*, 63(324), 1298-1309.

Sethi, V. (1963). A note on optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5(1), 20-33.

Sigman, R.S., and Monsour, N.J. (1995). Selecting samples from list frames of businesses. *Business Survey Methods*, 133-152.

Singh, R. (1971). Approximately optimum stratification on the auxiliary variable. *Journal of the American Statistical Association*, 66(336), 829-833.

Sweet, E.M., and Sigman, R.S. (1995). Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data. *Proceedings of the Section on Survey Research Methods*, 1, 491-496.

Tschuprow, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron*, 2(461-493), 646-683.

Wang, M., and Aggarwal, V. (1984). Stratification under a particular pareto distribution. *Communications in Statistics-Theory and Methods*, 13(6), 711-735.