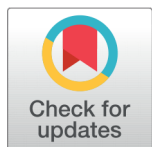


ORIGINAL ARTICLE



OPEN ACCESS

Received: 19/07/2026

Accepted: 10/08/2026

Published: 06/10/2026

Citation: Magistrado JB (2026) Deriving Implied Sampling Costs from Fixed-Budget Stratified Allocation. Indian Journal of Science and Technology 19(34): 2415-2427. <https://doi.org/10.17485/IJST/v19i34.1066>

* **Corresponding author.**jbmagistrado@up.edu.ph**Funding:** None**Competing Interests:** None

Copyright: © 2026 Magistrado. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment ([iSee](#))

ISSN

Print: 0974-6846

Electronic: 0974-5645

Deriving Implied Sampling Costs from Fixed-Budget Stratified Allocation

Jeffrey B Magistrado^{1*}

¹ Institute of Statistics, College of Arts and Sciences, University of the Philippines Los Baños, Laguna, Philippines

Abstract

Objectives: Classical stratified allocation rules optimize statistical precision without regard to per-unit sampling costs. This study inverts that optimization problem to recover a closed-form, per-unit cost structure implied by any fixed-budget allocation, establishing a reverse diagnostic framework for survey auditing. **Method:** The Lagrangian solution to cost-constrained Neyman allocation was inverted to yield a general inverse-cost expression, with equal, proportional, and Neyman weights substituted to produce rule-specific closed-form formulas. The framework was validated through four tiers: algebraic derivation, a five-stratum provincial illustration, a 5,000-replication Monte Carlo simulation, and an empirical post-hoc audit of a published survey (HILDA Wave 18, L = 5 strata). **Findings:** Each allocation rule implied a distinct, non-cost-neutral cost structure. Equal allocation produced a uniform implied cost of ₱11,549.25 across all strata regardless of population size or variability, while proportional and Neyman-weighted allocations produced implied costs ranging from ₱657.52 to ₱75,870.99 depending on stratum size and variability. Monte Carlo results based on 5,000 replications showed mean absolute rounding error ranging from 0.83 percent under equal allocation to 2.90 percent under Neyman allocation by standard deviation, with individual-stratum errors as high as 85.88 percent and error escalating from 1.45 percent to 12.22 percent as the number of strata increased from three to twenty. In the empirical audit, the framework's recovered implied cost diverged from the originally reported cost by more than 500 percent in four of five strata, compared with a stepwise versus full-precision computational gap of no more than approximately 3 percent in the same allocation. Under the cost-only reading of the allocation, this indicates that the divergence reflects the structure of the allocation rule rather than a rounding artifact. **Novelty:** This study introduces the first inverse-cost diagnostic framework for stratified sampling, converting conventional allocation rules, previously treated as statistically neutral defaults, into falsifiable, auditable per-unit cost assumptions, and demonstrates its use as a post-hoc auditing tool for allocations already in the field.

Keywords: Stratified sampling; Neyman allocation; Survey cost; Inverse-cost framework; Sample allocation

1 Introduction

Stratified random sampling requires both a total sample size and its allocation across strata, and cost-constrained allocation continues to receive active attention in the survey methodology literature. Recent work has extended optimal allocation to box-constrained sample sizes through a recursive Neyman algorithm⁽¹⁾, jointly determined strata boundaries and sample sizes under a total cost constraint through dynamic programming⁽²⁾, formulated multivariate compromise allocation as a chance-constrained program that minimizes cost and variance simultaneously⁽³⁾, and proposed compromise allocation techniques that explicitly control measurement cost across multi-character surveys⁽⁴⁾. A parallel and rapidly growing line of work has derived mean estimators that trade estimation accuracy against assumed sampling costs: recent studies have proposed a linear-cost estimator for the population mean under stratified sampling⁽⁵⁾, a related estimator that explicitly balances estimation accuracy against cost⁽⁶⁾, and a family of generalized exponential estimators optimized jointly for cost and precision through integer programming and Lagrangian methods⁽⁷⁾. The same estimator-focused line has continued to expand into closely related survey settings, extending cost- and auxiliary-informed stratified estimation to time-scaled surveys using dual auxiliary variables⁽⁸⁾ and, most recently, to memory-based exponential-cum-product estimation with agricultural survey applications⁽⁹⁾, alongside non-linear and more broadly budget-constrained estimation variants⁽¹⁰⁾,⁽¹¹⁾. Elsewhere, per-unit stratum costs have been used to construct optimum strata boundaries⁽²⁾, optimization algorithms for multivariate allocation under auxiliary variables have been compared⁽¹²⁾, linear inequality constraints have been enforced on balanced integer allocation⁽¹³⁾, and computational stratification methods have been empirically benchmarked against real household and early childhood education survey data⁽¹⁴⁾.

Across this literature, three limitations recur. First, per-unit cost is consistently supplied as a known, exogenous input rather than as a quantity that can be checked against an allocation already chosen, since no study offers a means of recovering cost assumptions after the fact, a gap that persists even in the most recent estimator- and allocation-focused work⁽⁷⁾,⁽¹¹⁾,⁽¹³⁾. Second, where classical allocation rules such as equal, proportional, or Neyman allocation are used at all, they are treated as statistically motivated defaults rather than as choices that encode an implicit, testable claim about cost. Third, empirical validation, where present, evaluates estimator or algorithm performance directly⁽¹⁴⁾, rather than auditing whether an already fielded survey's allocation was consistent with its real per-unit costs.

This study addresses all three gaps by reversing the conventional allocation to cost relationship. Inverting the Lagrangian solution to cost-constrained Neyman allocation yields a general, closed-form expression for the implied per-unit stratum cost consistent with any allocation rule and a fixed budget, requiring no externally supplied cost input. Substituting the equal, proportional, and Neyman weights into this general form demonstrates that each classical rule encodes a distinct, testable cost assumption rather than a cost-neutral default, a distinction further sharpened by showing that conventional stepwise rounding of intermediate sample sizes distorts these implied costs unevenly across strata, thereby motivating full-precision computation as part of the diagnostic itself. The framework is then applied to both a simulated illustration and a published, real-world allocation, demonstrating its use as a post-hoc auditing tool for completed surveys, a capability unavailable to the forward-only approaches described above.

2 Methodology

2.1 Derivation of Cost-optimal Allocation

Cochran's formula provides an initial estimate of the required sample size, $n_{Cochran}$, based on desired precision and estimated variance. Cochran's formula⁽¹⁵⁾,⁽¹⁶⁾ for population variance:

$$n_{Cochran} | \sigma^2 = \frac{Z_{\alpha/2}^2 \sigma^2}{e^2} \quad (1)$$

For Cochran's⁽¹⁵⁾ formula for population proportion (p):

$$n_{Cochran} | p = \frac{Z_{\alpha/2}^2 p(1-p)}{e^2} \quad (2)$$

When the population size N is known and finite, $n_{Cochran}$ is usually rounded up and refined using the finite population correction (FPC)⁽¹⁶⁾:

$$n_{CochranFPC} = \frac{n_{Cochran}}{1 + \frac{n_{Cochran} - 1}{N}} = \frac{N n_{Cochran}}{N + n_{Cochran} - 1} \quad (3)$$

Once $n_{CochranFPC}$ is obtained, it must be allocated across strata with ceiling-rounded sample sizes. Three allocation schemes are commonly used in practice: equal allocation, proportional allocation, and Neyman allocation, expressed as $n_h = w_h \cdot n_{CochranFPC}$, where the allocation weight w_h takes the form:

$$w_h = \begin{cases} \frac{1}{L} & \text{(equal)} \\ \frac{N_h}{N} & \text{(proportional)} \\ \frac{N_h S_h}{\sum_{i=1}^L N_i S_i} & \text{(Neyman)} \end{cases} \quad (4)$$

Let $W_h = \frac{N_h}{N}$, The variance of the stratified mean estimator^{(15), (16), (17)} is given by:

$$Var(\bar{y}_{st}) = \frac{\sum_{h=1}^L W_h^2 S_h^2}{n_h} \quad (5)$$

Assuming a linear total cost function:

$$C = c_0 + \sum_h c_h n_h \quad (6)$$

where c_0 is fixed overhead cost and c_h is the marginal cost per sampled unit in stratum h , minimizes $Var(\bar{Y}_{stratified})$ for fixed C is derived below:

The total cost function (Eq. 6) assumes a linear relationship between sample size and cost within each stratum, with a constant marginal cost c_h per sampled unit. This is the standard assumption in the cost-constrained allocation literature^{(16), (18), (19)} and is retained here for tractability, since it permits the closed-form inversion central to this study's contribution. In practice, per-unit costs may exhibit economies of scale or diseconomies arising from travel and logistics in geographically dispersed strata. The implied-cost expressions should therefore be interpreted as the *linear-equivalent* per-unit cost consistent with a given allocation, rather than as a claim that true marginal costs are constant within each stratum.

To continue, using a Lagrange multiplier λ :

$$\mathcal{L} = \sum_h W_h^2 \frac{S_h^2}{n_h} + \lambda \left(\sum_h c_h n_h - C \right) \quad (7)$$

Setting $\frac{\partial \mathcal{L}}{\partial n_h} = 0$ and solving yields the cost-optimal allocation^{(16), (18), (19)}:

$$n_h \propto \frac{N_h S_h}{\sqrt{c_h}}, \quad n_h = n_{CochranFPC} \cdot \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{i=1}^L \frac{N_i S_i}{\sqrt{c_i}}} \quad (8)$$

$$n_h \propto \frac{N_h \sqrt{P_h(1-P_h)}}{\sqrt{c_h}}, \quad n_h = n_{CochranFPC} \cdot \frac{\frac{N_h \sqrt{P_h(1-P_h)}}{\sqrt{c_h}}}{\sum_{i=1}^L \frac{N_i \sqrt{P_i(1-P_i)}}{\sqrt{c_i}}} \quad (9)$$

$$n_h \propto N_h / \sqrt{c_h}, \quad n_h = n_{CochranFPC} \cdot \frac{\frac{N_h}{\sqrt{c_h}}}{\sum_{i=1}^L \frac{N_i}{\sqrt{c_i}}} \quad (10)$$

2.2 Derivation of Cost-constrained optimal allocation.

While the literature ^{(6), (11), (13)} addresses only the derivation and application of the optimal n_h from known c_h , this study derives the inverse of this approach by determining the per-unit cost c_h consistent with a given allocation n_h under a fixed budget C . Deriving the inverse cost from that condition:

$$c_h = \frac{(C - c_0) N_h^2 S_h^2}{n_h^2 \sum_{i=1}^L \frac{N_i^2 S_i^2}{n_i}} \quad (11)$$

Writing $n_h = w_h n_{CochranFPC}$ for general weight w_h , the implied cost formula was generalized to:

$$c_h = \frac{(C - c_0) N_h^2 S_h^2}{n_{CochranFPC} w_h^2 \sum_{i=1}^L \frac{N_i^2 S_i^2}{n_i}} \quad (12)$$

2.3 Multi-Method Validation Approach

To establish methodological rigor, ensure reproducibility, and address empirical validation requirements, the proposed cost-constrained optimal allocation was evaluated through a four-tiered validation design. All computations were implemented in the Python programming language⁽²⁰⁾ and performed the numerical operations using NumPy⁽²¹⁾ for array-based computation and pseudorandom number generation. This approach integrates algebraic derivations, a synthetic baseline scenario, a Monte Carlo simulation study, and empirical testing on real-world survey data.

2.3.1. Algebraic validation

Each rule-specific inverse-cost formula (Eqs 13–18) was analytically derived by substituting rule-specific allocation weights (w_h) into the general inverse-cost expression (Equation 12). Deductive internal consistency was confirmed by verifying that every rule-specific equation collapses back to the cost-constrained Neyman optimality condition. Furthermore, an algebraic convergence proof was established to demonstrate that Neyman allocation for proportions under maximum variance ($P_h = 0.5$) reduces directly to the proportional allocation cost structure.

2.3.2. Synthetic Baseline Illustration

Since c_h is usually used in optimal allocation in stratified sampling, there is a limited open-access data to be used for comparison. It is understandable since C , c_0 , and c_h are naturally confidential and need internal access to be used in comparative benchmarks. Despite of lack of open-access cost parameters in a survey, this study used a synthetic 5-stratum survey scenario to demonstrate the computational mechanics under controlled population parameters (N_h, S_h, P_h). This scenario established calculation of differences per cost comparing classical stepwise-rounded sample allocations (Eqs. 1-4) against the derived cost-inversion formulas across equal, proportional, and optimal allocation.

2.3.3. Extensive Monte Carlo Sensitivity Analysis

After showing a single synthetic case, the study extended its potential reproducibility when non-cost parameters were simulated, using different values. Specifically, an extensive Monte Carlo simulation study has set using 5,000 independent replications. For each replication, L strata were generated with population sizes N_h drawn uniformly from 500 to 10,000, expenditure standard deviations S_h drawn uniformly from ₱ 20,000 to ₱ 300,000, and access proportions P_h drawn uniformly from 0.05 to 0.95. Cochran's parameters were fixed at $Z = 1.96$, $e = 0.05$, and $p = 0.5$, the conservative default used when no prior variance estimate is available. The budget parameters C and c_0 were held at the values used in Section 3.3 ($C = 5,000,000$; $c_0 = 500,000$) so that results remain comparable to the illustrative example. For each replication, n_h and the corresponding implied cost c_h were computed twice under each allocation rule, equal, proportional, Neyman by S_h , and Neyman by P_h : once via the classical stepwise path with ceiling-rounding applied at $n_{Cochran}$, at the FPC stage, and at each stratum allocation, and once via the direct full-precision cost-inversion formula. The relative error of the classical value against the full-precision value was recorded for every stratum in every replication.

The primary configuration used $L = 5$ strata with 5,000 replications. A secondary set of runs varied L from 3 to 20 strata, with 2,000 replications each, to test whether distortion scales with the number of strata in a design. Framework accuracy and rounding distortions were evaluated using four explicit quantitative performance metrics pooled across all strata and

replications: mean absolute relative error (MAE(%)), median relative error (Median(%)), standard deviation of relative error (SD(%)), and maximum relative error (Max(%)). Comparative benchmarking was performed by computing sample sizes and implied costs twice for every replication, once through the classical stepwise rounded path applying ceiling rounding at the Cochran, finite population correction (Eqs. 1-4), and per stratum allocation stages, and once via the direct full precision cost inversion formula. This directly quantifies the error that rounding induces.

To support reproducibility, all Monte Carlo replications were generated using a fixed random seed for the NumPy pseudorandom number generator⁽²¹⁾. To examine how rounding-induced error changes with the design, the mean relative error of each allocation rule was computed for $L = 3, 5, 8, 12$, and 20 strata, each based on 2,000 replications, and is reported in Table 8. This addresses the concern that a single maximum value drawn from one simulation run may not reliably characterize the tail behavior of the error distribution, particularly for allocation rules whose weights compound stratum-level variability.

2.3.4. Post-hoc Diagnostic Audit

The post-hoc diagnostic audit evaluates how an externally published survey allocation aligns with its implied cost assumptions. The procedure begins by extracting stratum parameters from Wave 18 of the Household, Income and Labour Dynamics in Australia (HILDA) Survey⁽²⁾. Specifically, this study used the five stratum sizes, reported optimal sample sizes, and assumed per-unit costs. Since the stratum-level standard deviations were not reported in the source publication, empirical allocation weights are derived directly from the reported sample distributions. These allocation weights and population sizes are then substituted into the general, rule-agnostic inverse-cost formula under a fixed budget constraint to reverse-derive the exact per-unit stratum costs mathematically required for the reported allocation to be cost-optimal.

To evaluate performance and isolate error sources, the methodology executes a dual comparative benchmark. Implied costs computed via stepwise-rounded sample sizes are first compared against direct full-precision floating-point calculations, establishing that computational rounding accounts for less than three percent error across all strata. Recovered implied costs are then benchmarked directly against the per-unit costs originally reported in the source study. By ruling out numerical rounding as a primary error source, the process demonstrates that the observed divergence exceeding five hundred percent stems from a structural mismatch between rule assumptions and field parameters, validating the framework as a functional diagnostic tool for post-hoc survey auditing.

3 Results and Discussion

3.1 Closed-Form Cost-Inversion Formulas by Allocation Rule

Substituting each classical allocation weight w_h into the general inverse-cost formula (Eq. 12) produced distinct closed-form expressions for the implied per-unit sampling cost c_h under each allocation mechanism:

Let $w_h = \frac{1}{L}$, the Cost-inversion formula using equal allocation is:

$$c_h |_{equal(S_h)} = \frac{L(C - c_0) N_h^2 S_h^2}{n_{CochranFPC} \sum_{i=1}^L N_i^2 S_i^2} \quad (13)$$

Let $w_h = \frac{N_h}{N}$, the Cost-inversion formula using proportional allocation is:

$$c_h |_{prop(S_h)} = \frac{N(C - c_0) S_h^2}{n_{CochranFPC} \sum_{i=1}^L N_i S_i^2} \quad (14)$$

Let $w_h = \frac{N_h S_h}{\sum_i N_i S_i}$, the Cost-inversion formula using Neyman allocation is:

$$c_h |_{Neyman(S_h)} = \frac{(C - c_0)}{n_{CochranFPC}} \quad (15)$$

When the stratum variable of interest is a proportion rather than a continuous measurement, the stratum standard deviation is replaced by the binomial or Bernoulli variance term $S_h = \sqrt{P_h(1 - P_h)}$. Wherein, P_h is the estimated proportion possessing the characteristic of interest in stratum h . Substituting $S_h \rightarrow \sqrt{P_h(1 - P_h)}$, and correspondingly $S_i \rightarrow \sqrt{P_i(1 - P_i)}$ directly into the formula:

Let $w_h = \frac{1}{L}$, the Cost-inversion formula using equal allocation is:

$$c_h |_{equal(P_h)} = \frac{L(C - c_0) N_h^2 P_h (1 - P_h)}{n_{CochranFPC} \sum_{i=1}^L N_i^2 P_i (1 - P_i)} \quad (16)$$

Let $w_h = \frac{N_h}{N}$, the Cost-inversion formula using proportional allocation is:

$$c_h |_{prop(P_h)} = \frac{N(C - c_0) P_h (1 - P_h)}{n_{CochranFPC} \sum_{i=1}^L N_i P_i (1 - P_i)} \quad (17)$$

Let $w_h = \frac{N_h S_h}{\sum_i N_i S_i}$, the Cost-inversion formula using Neyman allocation is:

$$c_h |_{Neyman(P_h)} = \frac{C - c_0}{n_{CochranFPC}} \quad (18)$$

When no prior data is available for both S_h and P_h , maximum variability ($P_h = 0.5$) is applied:

$$c_h |_{equal(P_h=0.5)} = \frac{L(C - c_0) N_h^2}{n_{CochranFPC} \sum_{i=1}^L N_i^2} \quad (19)$$

$$c_h |_{prop(P_h=0.5)} = \frac{N(C - c_0) 0.5(1 - 0.5)}{n_{CochranFPC} N \sum_{i=1}^L 0.5(1 - 0.5)} = \frac{C - c_0}{n_{CochranFPC}} = c_h |_{Neyman(P_h=0.5)} \quad (20)$$

3.2 Algebraic Verification of the Cost-Inversion Framework

The verification of the cost-inversion framework rests on two internal checks. First, each rule-specific formula (Eqs. 13-15) was derived independently by substituting its corresponding allocation weight w_h into the general inverse-cost expression (Eq. 12), then re-verified algebraically against the general cost-constrained Neyman form (Eq. 11). Since the general inverse-cost expression (Eq. 12) nests every classical rule as a special case of w_h , any rule-specific formula that failed to collapse back to general cost-constrained Neyman form (Eq. 11) under its own weight would indicate an error in derivation rather than a property of the allocation rule itself.

Second, the framework's validity is reinforced by a convergence. Neyman allocation by proportion at maximum variance ($p_h = 0.5$ for all strata) has shown to reduce exactly to the proportional-allocation cost structure. Their exact convergence under a $p_h = 0.5$ maximizes the Bernoulli variance term, and is consistent with the framework correctly encoding the underlying statistical theory rather than producing coincidentally similar outputs.

These two checks are justified as the appropriate standard of validity for this study precisely because the object being validated is a mathematical derivation, rather than an estimator or a predictive model fit to data. A cost-inversion formula generates an algebraic identity that either holds or fails to hold under substitution, rather than a testable prediction against an outcome variable. Consequently, the relevant validity criterion is deductive correctness, established by tracing each closed-form expression back to the general optimality condition from which it was derived, rather than a statistical criterion such as predictive accuracy, which would require an external dataset outside the scope of this study.

3.3 Hypothesized Illustration: Household Water Access Survey in Province of Kalayaan

3.3.1. The Scenario

A provincial statistics agency in a fictitious island called the Province of Kalayaan is planning a survey to estimate average monthly household water expenditure across its five administrative regions. The agency has secured a survey budget of $C = 5,000,000.00$ Philippine Peso, of which $c_0 = 500,000$ Philippine Peso is allocated to fixed cost such as training, questionnaire design, sampling frame construction. The budget has 4,500,000.00 Philippine Peso for the allocated households.

Preliminary data from regional records provide the population size (N_h), historical standard deviation of expenditure (S_h , in Philippine Pesos), and estimated proportion of households with piped water access (P_h) for each stratum, shown in Table 1.

Table 1. Stratum characteristics

Region	N_h	S_h	P_h
Matapang	5,000	₱ 120000.00	0.50
Matalino	3,000	₱ 80000.00	0.30
Maasahan	2,000	₱ 150000.00	0.65
Mahinahon	4,000	₱ 500000.00	0.10
Madiskarte	1,000	₱ 200000.00	0.50

3.3.2. Implied cost per Region (c_h) using Equal Allocation

Equal allocation distributes the total sample size uniformly across all regions, regardless of population size, expenditure variability, or water access. Table 2 shows equal allocation produces an identical implied cost across all five regions, regardless of differences in population size, expenditure variability, or water access.

Table 2. Implied cost per region under Equal Allocation

Region	$c_{h Equal}$ Stepwise	$c_{h Equal}$ Cost-inversion	Difference
Matapang	₱ 11,538.46	₱ 11,549.25	– ₱ 10.79
Matalino	₱ 11,538.46	₱ 11,549.25	– ₱ 10.79
Maasahan	₱ 11,538.46	₱ 11,549.25	– ₱ 10.79
Mahinahon	₱ 11,538.46	₱ 11,549.25	– ₱ 10.79
Madiskarte	₱ 11,538.46	₱ 11,549.25	– ₱ 10.79

3.3.3. Implied cost per Region (c_h) using Proportional Allocation and Neyman Allocation where $P = 0.5$

Proportional allocation distributes the sample in direct proportion to each region's population share. This section presents the implied cost per region under this rule. This result also applies to Neyman allocation evaluated at maximum variance ($P_h = 0.5$).

Table 3. Implied cost per region under Proportional Allocation

Region	$c_{h Prop}$ Stepwise	$c_{h Prop}$ Cost-inversion	Difference
Matapang	₱ 3,085.27	₱ 3,034.84	₱ 50.43
Matalino	₱ 8,570.18	₱ 8,430.11	₱ 140.07
Maasahan	₱ 19,282.92	₱ 18,967.75	₱ 315.17
Mahinahon	₱ 4,729.34	₱ 4,741.94	– ₱ 12.60
Madiskarte	₱ 71,524.01	₱ 75,870.99	– ₱ 4,346.98

Consistent with the algebraic equivalence (See Eq. 20), proportional allocation and Neyman allocation at $P_h = 0.5$ yield identical implied-cost values across all regions. In Table 3, this implies that proportional-style allocation may overstate the true cost of reaching small strata, leading to chronic under-sampling of regions relative to what is actually cost-efficient.

3.3.4. Implied cost per Region (c_h) using Neyman Allocation by Proportion

Neyman allocation by proportion distributes the sample using each region's population size weighted by the Bernoulli variance of its access proportion, $\sqrt{P_h(1 - P_h)}$. This section presents the resulting implied cost per region, relevant when the survey's key variable of interest is access-related rather than continuous.

Table 4. Implied cost per region under Neyman Allocation using Proportion

Region	$c_{h Neyman(P_h)}$ Stepwise	$c_{h Neyman(P_h)}$ Cost-inversion	Difference
Matapang	₱ 2,395.76	₱ 2,408.51	– ₱ 12.75
Matalino	₱ 7,824.72	₱ 7,964.66	– ₱ 139.94
Maasahan	₱ 16,591.11	₱ 16,542.00	₱ 49.11
Mahinahon	₱ 10,398.25	₱ 10,453.62	– ₱ 55.37
Madiskarte	₱ 59,893.91	₱ 60,212.86	– ₱ 318.95

In Table 4, the main implication is that when the survey's key variable is access-related rather than purely financial, allocating by P_h better reflects the added fieldwork effort typically required to reach underserved populations, and avoids the extreme cost swings observed under Proportional Allocation.

3.3.5. Implied cost per Region (c_h) using Neyman Allocation by Standard Deviation

Neyman allocation by standard deviation distributes the sample using each region's population size weighted by its expenditure variability, S_h . This section presents the resulting implied cost per region under this rule, computed via both the classical stepwise path and the direct cost-inversion formula.

Table 5. Implied cost per region under Neyman Allocation using Standard Deviation

Region	$c_h \mid \text{Neyman}(S_h)$ Stepwise	$c_h \mid \text{Neyman}(S_h)$ Cost-inversion	Difference
Matapang	₱ 7,299.98	₱ 7,305.78	– ₱ 5.80
Matalino	₱ 43,756.50	₱ 45,661.10	– ₱ 1,904.60
Maasahan	₱ 28,394.46	₱ 29,223.10	– ₱ 828.64
Mahinahon	₱ 672.06	₱ 657.52	₱ 14.54
Madiskarte	₱ 63,887.53	₱ 65,751.98	– ₱ 1,864.45

In Table 5, this implies that a single large, high-variance stratum can distort the implied cost structure for the entire stratum, making $c_h \mid \text{Neyman}(S)$ sensitive to how size and variability interact rather than to variability alone.

3.4 Monte Carlo Sensitivity Analysis Results

To assess whether this distortion is a general property of the cost-inversion framework rather than an artifact of that particular configuration, a Monte Carlo sensitivity analysis was conducted, applying Eq. 1 through 4 (classical, stepwise-rounded) and Eqs. 13-18 cost-inversion to maximum variability to randomly generated stratified populations.

3.4.1. Overall Error Distribution by Allocation Rule

Table 6 summarizes the distribution of relative error, pooled across all strata and replications, for the primary $L = 5$ configuration.

Table 6. Distribution of rounding-induced relative error by allocation rule ($L = 5$, 5,000 replications)

Rule	MAE(%)	Median(%)	SD(%)	Max(%)
Equal	0.83%	0.72%	0.44%	1.53%
Proportional	1.56%	1.03%	1.71%	15.51%
Neyman (P_h)	1.64%	1.04%	1.89%	21.61%
Neyman (S_h)	2.90%	1.43%	4.87%	85.88%

Across all four rules, the median error stays under 1.5 percent, consistent with the Kalayaan illustration's finding that stepwise rounding introduces a modest distortion in typical cases. The maximum errors and standard deviations, however, diverge sharply by rule. Equal allocation, whose weights do not depend on N_h or S_h , produces a tightly bounded error distribution, with a standard deviation of 0.44 percent and a maximum of 1.53 percent. Neyman allocation by S_h produces both the largest average error and a long right tail, with individual strata occasionally distorted by more than 80 percent.

This pattern is consistent, where rules whose weights are sensitive to stratum-level variability compound rounding error through the $(N_h S_h)^2$ term, while equal allocation's constant weight structurally insulates it from this compounding.

3.4.2. Sensitivity to Stratum Size

Pooling across all replications, the correlation between a stratum's population size N_h and its rounding error confirms that smaller strata are more distortion-prone under variance- or size-weighted rules. The correlation between $1/N_h$ and absolute error is positive for proportional allocation ($r = 0.59$, $r^2 = 0.35$) and Neyman by p_h ($r = 0.54$, $r^2 = 0.29$), and weaker but still positive for Neyman by S_h ($r = 0.22$, $r^2 = 0.05$). For equal allocation, the correlation is negligible ($r = -0.09$, $r^2 = 0.01$), since its implied cost does not depend on N_h at all. Reporting r^2 alongside r clarifies that even the strongest observed relationship (proportional allocation) accounts for only about a third of the variance in rounding error, indicating

that stratum size is a meaningful but partial driver of distortion rather than its dominant cause; the remaining variance likely reflects the joint influence of σ_h , p_h , and the total number of strata (See Section 3.4.3.). All reported correlations are based on $n = 25,000$ stratum-level observations (5,000 replications times 5 strata) and are statistically significant at $p < 0.001$, though at this sample size statistical significance is expected even for small effect sizes and should not be read as an indicator of practical importance on its own.

Table 7. Mean relative error for the smallest stratum in a set, by stratum-size quartile (proportional allocation)

Smallest N_h in replication	MAE(%)
500 to 1,051	5.17%
1,051 to 1,744	3.40%
1,744 to 2,806	2.41%
2,806 to 8,241	1.65%

Table 7 reports the mean relative error for the smallest stratum in each replication under proportional allocation, grouped by quartile of that stratum's population size. Error falls steadily as the smallest stratum grows, from 5.17 percent when it has 500 to 1,051 units to 1.65 percent when it has more than 2,806 units, so the smallest quartile shows about three times the error of the largest. This indicates that rounding distortion is concentrated in designs containing very small strata, whose small allocation weights make ceiling rounding a larger share of their sample size and implied cost.

3.4.3. Sensitivity to Number of Strata

Table 8 shows mean error across replications as the number of strata L increases from 3 to 20, holding the per-stratum size range fixed.

Table 8. Mean relative error by number of strata

L	Equal	Proportional	Neyman (P_h)	Neyman (S_h)
3	0.61%	0.95%	1.00%	1.45%
5	0.83%	1.62%	1.71%	2.82%
8	0.90%	2.47%	2.64%	4.89%
12	0.58%	3.68%	3.94%	7.58%
20	4.31%	5.97%	6.26%	12.22%

Mean error increases with L for every rule except an irregular dip at $L = 12$ for equal allocation. This pattern is consistent with a design where each additional stratum draws its own independent N_h from the same range, so the total population grows while any individual stratum's share of the fixed sample shrinks, making per-stratum ceiling-rounding proportionally more consequential.

The effect is most pronounced for Neyman allocation by S_h , whose mean error more than quadruples from $L = 3$ to $L = 20$, rising from 1.45 percent to 12.22 percent, while equal allocation remains comparatively stable until a marked jump at $L = 20$ (Table 8).

3.5 Empirical Post-Hoc Diagnostic Audit

3.5.1. The Dataset

To test the cost-inversion framework against a real, published allocation outside the author's control, the five-stratum ($L = 5$)⁽²⁾ drawn from Wave 18 of the HILDA Survey. Table 9 presents the stratum population sizes (N_h), the study's originally reported optimum sample allocation (n_h), and the corresponding per-unit cost (c_h) assumed in that study.

Since the source study's stratum standard deviations (S_h) were not separately reported, the general, rule-agnostic cost-inversion formula (Eq. 11) was used, applying the empirical allocation weight $w_h = n_h / \sum n_h$ derived directly from the study's own reported allocation

3.5.2. Implied Cost: Stepwise Rounding vs. Proposed Framework

Table 10 presents the implied cost per stratum computed via the classical, stepwise-rounded total sample size and via the proposed framework's full-precision total sample size, both applying the reported allocation weights w_h .

Table 9. Reported stratum characteristics

Region	N_h (HILDA)	n_h (HILDA)	Simulated c_h (HILDA)
Stratum 1	2,775	70	AUD 13.81
Stratum 2	2,825	77	AUD 13.69
Stratum 3	2,277	78	AUD 13.60
Stratum 4	1,236	61	AUD 13.22
Stratum 5	506	214	AUD 13.16

The difference between the two computation paths is small across all five strata, ranging from near-zero to roughly 3 percent of the implied cost value. This is consistent with the sensitivity analysis that confirms under real stratum sizes and population counts that stepwise rounding introduces only a modest distortion relative to full-precision computation.

Table 10. Implied cost per stratum, classical stepwise vs. proposed framework

Stratum	c_h (HILDA) Stepwise	c_h (HILDA) Cost-inversion	Relative Difference
1	AUD 113.04	AUD 112.84	-0.18%
2	AUD 90.85	AUD 93.24	2.56%
3	AUD 90.85	AUD 90.86	0.01%
4	AUD 144.46	AUD 148.59	2.78%
5	AUD 12.09	AUD 12.07	-0.17%

3.5.3. Implied Cost vs. Reported Cost

Table 11 compares the proposed framework's full-precision implied cost against the per-unit cost originally reported in the stratum⁽²⁾. Unlike the near-zero gap between stepwise and full-precision computation in Table 10, the gap between the framework's implied cost and the study's originally reported cost is substantial, exceeding 500 percent for four of the five strata. This divergence should not be interpreted as direct evidence that the reported allocation was cost-inconsistent. The general inversion formula (Eq. 11) assumes the quantity $c_h \sigma_h^2$ is constant across strata, whereas the external study's allocation⁽²⁾ was constructed under a Neyman-type rule in which this quantity instead scales with N_h/S_h , a term that varies substantially across their five strata. Because stratum-level standard deviations (S_h) were not reported in the source publication, this scaling term could not be reconstructed or controlled for here. Consequently, the magnitude reported in Table 11 reflects some unknown combination of (1) genuine misalignment between the reported allocation and its assumed per-unit costs, and (2) formula misspecification arising from applying a constant- $c_h \sigma_h^2$ inversion to an allocation generated under a different, size-and-variability-scaled cost rule. The present data do not allow these two sources to be separated, and the result should therefore be read as an illustration of the framework's mechanics under incomplete external data, rather than a confirmed finding that the reported HILDA allocation departed from its own cost assumptions. A dataset reporting S_h alongside N_h , n_h , and c_h would be required to attribute the divergence with confidence.

Table 11. Proposed framework implied cost vs. reported cost

Stratum	c_h (HILDA) Stepwise	c_h (HILDA) Cost-inversion	Relative Difference
1	AUD 112.84	AUD 13.81	-717.09%
2	AUD 93.24	AUD 13.69	-581.08%
3	AUD 90.86	AUD 13.60	-568.09%
4	AUD 148.59	AUD 13.22	-1023.98%
5	AUD 12.07	AUD 13.16	8.28%

3.6 Interpretation of the Implied-Cost Structures

The three closed-form expressions collectively establish that no classical allocation rule is cost-neutral, each instead encoding a distinct and internally consistent assumption about how per-unit cost varies across strata. Equal allocation (Eq. 13) implies a constant cost per stratum, independent of population size, variability, or proportion, a structure that is administratively

convenient but requires every stratum, however large, remote, or heterogeneous, to be equally cheap to reach. In Table 2, the Kalayaan illustration makes this concrete. All five regions were assigned an identical implied cost of approximately ₱ 11,549.25, despite differing in population size by a factor of five and in expenditure variability by more than a factor of six. By contrast, Proportional allocation (Eq. 14) implies a cost structure inversely proportional to the square of stratum size. In Table 3, the allocation to be cost-optimal, larger strata must be disproportionately cheaper to sample per unit than smaller ones, where the smallest region, Madiskarte, carries an implied cost of roughly ₱ 75,870.99, more than twenty times that of the largest region, Matapang, at roughly ₱ 3,034.84. On the other hand, Neyman allocation (Eq. 15 and Eq. 18) generalizes this further by making implied cost inversely proportional to the squared product of stratum size and variability, so that strata which are simultaneously large and highly variable must also be the cheapest to reach for the allocation to remain optimal, an effect Tables 4 and 5 show directly, since Madiskarte, despite having the smallest population, again carries the highest implied cost under both proportion-weighted and variance-weighted Neyman allocation, its small size only partially offset by its variability.

In Table 3, a further interpretive result is the algebraic convergence between proportional allocation and Neyman allocation evaluated at maximum Bernoulli variance, where the estimated proportion equals 0.5 for every stratum. This is not an artifact of the particular Kalayaan parameters but a structural property of the general inversion formula (Eq. 12), since maximum variance reduces the Neyman weight to a form identical to the proportional weight, meaning that an analyst who defaults to proportional allocation in the absence of prior variance information is implicitly making the same cost assumption as one who explicitly assumes maximum variability under Neyman allocation, even though the two are conventionally presented in the literature as distinct design choices rather than as algebraically equivalent special cases.

Finally, the Monte Carlo results show that the interpretive value of these implied-cost expressions depends on the precision with which they are computed. Stepwise ceiling rounding, applied conventionally at the initial sample size, at the finite population correction stage, and at each stratum allocation⁽¹⁶⁾, introduces relative errors that are structurally uneven across rules. Because equal allocation's implied cost does not depend on stratum size or variability, its error distribution remains tightly bounded, with a mean absolute error of 0.83 percent and a standard deviation of 0.44 percent, whereas Neyman allocation by standard deviation, whose weight compounds both stratum size and variability, produces the largest average error and the longest right tail, including individual-stratum distortions exceeding 80 percent. This confirms that rules whose implied cost is more sensitive to stratum-level heterogeneity are also more vulnerable to rounding-induced misinterpretation, reinforcing the recommendation that full floating-point precision be preserved until final reporting whenever implied cost is used diagnostically.

3.7 Comparison with Existing Methods

Direct numerical comparison with prior studies is possible only where a study reports both an allocation and a per-unit cost figure against which the proposed framework's recovered cost can be checked. Within the cited literature, this condition is met only by⁽²⁾, whose reported five-stratum allocation is examined in this study's empirical validation. For the remaining studies, comparison is necessarily structural rather than numerical, since each addresses the forward problem of deriving sample sizes from an assumed cost, and none reports the reverse quantity that the present framework recovers. The recursive Neyman algorithm for box-constrained allocation⁽¹⁾ treats per-unit cost as a fixed, exogenous parameter supplied prior to optimization, and offers no mechanism for testing whether an allocation already fielded is consistent with the cost assumption that would justify it. The stochastic compromise allocation framework⁽³⁾ extends this forward logic to the multivariate case by treating both cost and variance as random variables to be optimized against, but likewise does not permit cost to be recovered from a completed allocation, and the compromise allocation technique for multi-character surveys⁽⁴⁾ similarly controls measurement cost as an input constraint rather than an output diagnostic. Among the estimator-focused literature, the linear-cost mean estimators^{(5), (6)} and the generalized exponential cost-optimized estimator⁽⁷⁾ derive bias and efficiency properties under an assumed cost function, but the cost function itself is never checked against an allocation independently of the estimator's performance.

. The multivariate allocation comparison across optimization algorithms⁽¹²⁾ and the balanced sampling approach under linear inequality constraints⁽¹³⁾ both operate at the level of the allocation mechanism itself, without reference to the implied cost that mechanism encodes, while the empirical comparison of computational stratification methods applied to South African household and early childhood education data⁽¹⁴⁾ evaluates estimator or algorithm performance directly against known ground truth, a different evidentiary standard from auditing whether a chosen allocation was cost-consistent, and one that the present framework does not require, since it operates only on the allocation and budget already on record.

Applied to the allocation reported⁽²⁾, the proposed framework was used without modification on the reported five-stratum design. The comparison shows that the difference between the classical, stepwise-rounded implied cost and the full-precision cost-inversion value is small, ranging from near zero to approximately 3 percent across strata, consistent with the rounding-

distortion magnitudes observed in the Monte Carlo analysis and supporting the framework's numerical stability under real, externally derived allocation data. Benchmarking against the originally reported cost, however, shows a substantial divergence, exceeding 500 percent in four of five strata, between the framework's recovered implied cost and the per-unit cost originally assumed in that study. This divergence is consistent with a structural mismatch between the general, rule-agnostic inversion formula, which assumes $c_h S_h^2$ is constant across strata, and the Neyman-type allocation actually used in ⁽²⁾, where this quantity instead scales with $N_h S_h$ and therefore varies across strata. However, because the unreported stratum standard deviations prevented this scaling term from being reconstructed, the observed gap cannot be cleanly attributed to rule mismatch alone and may also partly reflect genuine inconsistency in the source allocation, formula misspecification, or both. This audit should therefore be read as a demonstration of how the framework can be applied to an externally published allocation, rather than as a confirmed diagnosis of that allocation's cost consistency. No comparable audit of an already published allocation exists elsewhere in the cited literature, since the forward-only methods reviewed above are not constructed to perform this check in the first place; extending the audit to a dataset with fully reported cost parameters remains necessary before the magnitude of any true divergence can be established.

4 Conclusion

This study set out to invert the classical cost-constrained allocation problem. Its central novelty lies in demonstrating, for the first time, that equal, proportional, and Neyman allocation are not statistically neutral defaults but three distinct, closed-form cost assumptions nested within a single general inversion formula, a reverse derivation not previously offered in the literature. This novelty is supported numerically. The Kalayaan illustration showed equal allocation implying a uniform cost of approximately ₱ 11,549.25 across all five regions, while proportional and Neyman-weighted allocations produced implied costs ranging from approximately ₱ 657.52 to ₱ 75,870.99 depending on stratum size and variability, and the Monte Carlo analysis of 5,000 replications quantified how stepwise ceiling rounding distorts these figures, with mean absolute error ranging from 0.83 percent under equal allocation to 2.90 percent under Neyman allocation by standard deviation, individual stratum errors as high as 85.88 percent, and error rising from 1.45 to 12.22 percent as the number of strata increased from three to twenty. Applied to a real published allocation ⁽²⁾, the framework recovered an implied per-unit cost that diverged from the study's originally reported cost by more than 500 percent in four of five strata, against a stepwise-versus-full-precision computational gap in the same allocation of no more than approximately 3 percent. This large gap between formula-implied and reported cost, set against the much smaller gap attributable to rounding, illustrates that the framework is sensitive to structural assumptions about the cost model rather than to numerical precision; however, because the source study did not report stratum-level variability, the divergence cannot be fully separated into rule-mismatch and genuine-inconsistency components, and this result should be treated as an illustrative application of the framework rather than a validated audit finding.

Three limitations bound these findings and point to future work. First, the framework does not yet extend to non-linear or multivariate cost functions. Second, no reviewed study, including the empirical validation source ⁽²⁾, reports stratum-level standard deviations alongside its allocation and assumed cost, leaving the empirical divergence in Section 3.5 only partially attributable and preventing a clean separation of formula mismatch from genuine cost inconsistency. Third, empirical validation in this study relies on a single external dataset with incomplete cost parameters; a broader audit across multiple surveys with fully reported cost structures would be needed to establish the framework's diagnostic reliability more generally. Future work should extend the inversion framework to non-linear and multivariate cost structures, validate it against surveys with directly measured and fully reported per-unit costs and variability, and formalize the observed rounding-error patterns into a standardized diagnostic threshold for routine survey auditing. In the interim, practitioners using proportional or Neyman allocation, particularly for designs with small strata, are advised to recompute implied cost at full floating-point precision before relying on it diagnostically, since rounding distortion is largest under precisely these conditions.

References

- 1) Wesołowski J, Wieczorkowski R, Wójcik W. Recursive Neyman algorithm for optimum sample allocation under box constraints on sample sizes in strata. *Survey Methodology*. 2024;50(2):487–511. Available from: <http://www.statcan.gc.ca/pub/12-001-x/2024002/article/00003-eng.pdf>.
- 2) Reddy KG, Khan MGM. Constructing efficient strata boundaries in stratified sampling using survey cost. *Heliyon*. 2023;9(10):e21112. Available from: <https://www.sciencedirect.com/science/article/pii/S2405844023086152>.
- 3) Mahfouz MI, Rashwan MM, Khadr ZA. Optimal stochastic allocation in multivariate stratified sampling. *Mathematics and Statistics*. 2023;11(4):676–684. Available from: https://www.hrpub.org/journals/article_info.php?aid=13433.
- 4) Verma MK, Yadav SK, Varshney R, Vishwakarma GK. A new optimum allocation technique in multivariate stratified sampling under measurement cost. *Comput Manag Sci*. 2026;23(24). Available from: <https://link.springer.com/article/10.1007/s10287-026-00571-8>.
- 5) Sharma P, Kumari M, Singh P. Advancing stratified sampling: optimized mean estimation via linear cost structures. *J Stat Theory Pract*. 2025;19:55. <https://doi.org/10.1007/s42519-025-00470-2>.

- 6) Sharma P, Pusa N, Kumari M, Singh P. Balancing accuracy and cost: a new estimator for stratified random sampling. *Commun Stat Simul Comput*. 2025;p. 1–17. <https://doi.org/10.1080/03610918.2025.2552216>.
- 7) Singh P, Sharma P, Singh R, Aloraini B, Akhtar A. Optimizing population mean estimation in stratified sampling using linear cost: a simulation study. *Heliyon*. 2024;10(24):e40878. <https://doi.org/10.1016/j.heliyon.2024.e40878>.
- 8) Kumari M, Sharma P, Singh P. Variance estimation techniques in time scale survey using dual auxiliary variables: mathematical and simulation approach. *REVSTAT Stat J*. 2025. Available from: <https://revstat.ine.pt/index.php/REVSTAT/article/view/922>.
- 9) Kumari M, Sharma P, Singh P. Mean estimation in time-scaled surveys: a dual auxiliary information approach with applications to agricultural data. *REVSTAT Stat J*. 2026. Available from: <https://revstat.ine.pt/index.php/REVSTAT/article/view/1008>.
- 10) Verma MK, Yadav SK, Varshney R, Vishwakarma GK. Computation strategy for population mean in stratified random sampling: a problem of budget constraint. *Stat J IAOS*. 2025;41(2):482–495. <https://doi.org/10.1177/18747655251335756>.
- 11) Verma MK, Sharma DK, Yadav SK, Varshney R, Bundela OS. Optimizing mean estimation in stratified sampling: addressing budget constraints. *J Indian Soc Probab Stat*. 2026;27(1):527–556. <https://doi.org/10.1007/s41096-026-00275-x>.
- 12) Pumputis D, Nekrašaitė-Liegė V. Optimizing multivariate sample allocation using auxiliary variables in stratified sampling. *J Stat Comput Simul*. 2026;96(1):75–104. <https://doi.org/10.1080/00949655.2025.2554294>.
- 13) Tripet A, Tillé Y. Balanced sampling with inequalities: application to category bounding, matrix rounding, and spread sampling. *J Am Stat Assoc*. 2026;121(553):796–806. <https://doi.org/10.1080/01621459.2025.2550667>.
- 14) Borros G, Er Ş, Salau S. Computational methods for optimal stratified sampling of ECE outcomes, fees and household income in South Africa. *S Afr Stat J*. 2026;60(1):1–24. <https://doi.org/10.37920/sasj.2026.60.1.1>.
- 15) Cochran WG. Sampling techniques. 2nd ed. New York. John Wiley and Sons. 1963. 978-0471162384.
- 16) Cochran WG. Sampling techniques. 3rd ed. New York. John Wiley and Sons. 1977. 978-0471162407.
- 17) Kish L. Survey sampling. New York. John Wiley and Sons. 1965. 978-0471489009.
- 18) Neyman J. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J R Stat Soc*. 1934;97(4):558–606. <https://doi.org/10.1111/j.2397-2335.1934.tb04184.x>.
- 19) Levy PS, Lemeshow S. Sampling of populations: Methods and applications. 4th ed. Hoboken (NJ). Wiley. 2008. <https://doi.org/10.1002/9780470374597>.
- 20) Foundation PS. Python Language Reference, version 3.9.6. Wilmington (DE). Python Software Foundation. 2021. Available from: <https://www.python.org/downloads/release/python-396/>.
- 21) Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. *Nature*. 2020;585(7825):357–362. <https://doi.org/10.1038/s41586-020-2649-2>.