



Statistics
Canada

Statistique
Canada

Canada



Statistics Canada
www.statcan.gc.ca

PRASC



Project for the Regional
Advancement of Statistics
in the Caribbean

Projet régional pour
l'avancement de la statistique
dans les Caraïbes

Funded by the
Government
of Canada

Canada



Statistics
Canada

Statistique
Canada

Canada



Statistics Canada
www.statcan.gc.ca

Project for the Regional Advancement of Statistics in the Caribbean - PRASC

Component: Business Survey Infrastructure

Funded by the
Government
of Canada

Canada



Stratification- Part 2

Mike Hidioglou and Wesley Yung
Business Survey Methods Division

January 20, 2016
Port of Spain, Trinidad and Tobago

Outline

- Introduction
- Some notation
- Allocation Criteria
- Allocation Schemes
- Stratification boundaries
- Sampling
- Summary

Introduction

- In part 1, we
 1. Defined stratification
 2. Noted that stratification boundaries (on size) could be
 - **Sample size driven:** Minimise the variance for a fixed sample size
 - **Reliability driven:** Minimise the sample size for a given level of precision
 3. Explained how the stratum boundary between take-all and take-some strata could be computed

Introduction

- In part 2, we
 1. Explain various allocation schemes, as they drive some of the algorithms for determining strata boundaries based on size
 2. Describe how strata boundaries can be obtained for several take-some strata including the take-all stratum

Some notation

- Simple random sampling without replacement is assumed within each stratum
- Finite population of N units divided into L non-overlapping sub-populations or strata of size N_1, \dots, N_L , respectively

$$N = N_1 + N_2 + \dots + N_L$$

Some notation

- A sample of size n_h ($h = 1, \dots, L$) is drawn from each stratum:

$$n = n_1 + n_2 + \dots + n_L$$

- Variable of interest, y , is not usually available
- Use a correlated auxiliary variable, x , as a proxy

Some notation

- Overall and stratum population totals defined as

$$X = \sum_{h=1}^L X_h \text{ and } X_h, \text{ where } X_h = \sum_{i=1}^{N_h} x_{hi}$$

- Population stratum variance S_h^2 defined as

$$S_h^2 = \sum_{i=1}^{N_h} (x_{hi} - \bar{X}_h)^2 / (N_h - 1)$$

with

$$\bar{X}_h = X_h / N_h$$

Some notation

- For stratified sampling, an unbiased estimate for the true population total is

$$\hat{X} = \sum_{h=1}^L N_h \bar{x}_h$$

where

$$\bar{x}_h = \sum_{i=1}^{n_h} x_{hi} / n_h$$

Some notation

- The variance of estimated total \hat{X} is

$$V(\hat{X}) = \sum_{h=1}^L N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_h^2$$

$$= \left(\sum_{h=1}^L A_h / n_h \right) - D$$

$$\text{where } A_h = N_h^2 S_h^2 \text{ and } D = \sum_{h=1}^L N_h S_h^2$$

- Population variance $V(\hat{X})$ estimated from previous surveys

Some notation

- Allocation problem can be formulated in one of two ways:
 1. Minimise the variance of \hat{X} for a fixed sample size
 2. Specify the required precision of \hat{X} as a coefficient of variation, given by c . That is

$$\frac{V(\hat{X})}{X^2} \leq c^2$$

Allocation Criteria

Fixed Sample Size

- The sample size n is allocated to the L strata using a specific allocation (proportional to size, Neyman, etc.)
- The proportion allocated to the h^{th} stratum will be denoted as a_h , where $0 \leq a_h \leq 1$ and $\sum_{h=1}^L a_h = 1$.
- The sample size allocated to each stratum h is

$$n_h = na_h, \text{ for } h = 1, 2, \dots, L$$

Allocation Criteria

Fixed Coefficient of Variation

- Target coefficient of variation, c , specified for total \hat{X}
- Sample size n is **not known** and is computed using the chosen allocation scheme to the strata
- Sample size selected within h -th stratum is $n_h = na_h$ where n is now unknown.

Allocation Criteria

Fixed Coefficient of Variation

- To solve for n and subsequently n_h , we use $V(\hat{X}) = c^2 X^2$
- Substituting $n_h = na_h$ into

$$V(\hat{X}) = c^2 X^2 = \left(\sum_{h=1}^L \frac{A_h}{na_h} \right) - D$$

$$A_h = N_h^2 S_h^2$$
$$D = \sum_{h=1}^L N_h S_h^2$$

- We obtain

$$n_h = a_h \frac{\sum_{h=1}^L A_h / a_h}{c^2 X^2 + D}$$

Allocation Schemes

N - proportional Allocation

- Scheme usually used when
 - Information on stratum variances is not available or
 - One wishes to make the design self-weighting:

$$N_h / n_h = N / n, h = 1, 2, \dots, L$$

- For this type of allocation

$$a_h = N_h / N \text{ for } h = 1, 2, \dots, L.$$

Allocation Schemes

Characteristics of N -proportional Allocation

- Good for safeguarding estimates against the effect of the movement of units between strata, called "**stratum jumping**".
- Superior to simple random sampling of whole population, if the strata averages differ considerably.

Allocation Schemes

X - proportional Allocation

- If the measures of size x_{hi} are available for units in the population, the sample sizes n_h may be found as proportions of X_h (the aggregate measure of size of stratum h).
- For this type of allocation

$$a_h = X_h / X \text{ for } h = 1, 2, \dots, L$$

Allocation Schemes

Characteristics of X-proportional allocation

- X-proportional allocation is used in business surveys because the distributions associated with such surveys are **very skewed**.
- The stratum containing the very large units are often more variable than other strata.
- For X-proportional allocation, the factor X_h / X in the denominator will exert a dampening effect on the variance.

Allocation Schemes

\sqrt{N} and \sqrt{X} proportional allocation

- First proposed by Carroll (1970) and further studied by Bankier (1986).
- The survey data user may be interested in having good reliability attached to the **stratum estimates** \hat{X}_h in addition to the overall estimate \hat{X} .

- For instance, if strata are regions, regional and national estimates are considered important.
- Using \sqrt{N} or \sqrt{X} proportional allocation usually achieves this goal.

\sqrt{N} Proportional	\sqrt{X} Proportional
$a_h = \sqrt{N_h / \sum_{h=1}^L \sqrt{N_h}}$	$a_h = \sqrt{X_h / \sum_{h=1}^L \sqrt{X_h}}$

Allocation Schemes

Neyman Allocation (Neyman, 1934)

- For this type of allocation:

$$a_h = (N_h S_h) / \left(\sum_{h=1}^L N_h S_h \right)$$

- In practice, stratum variances

$$S_h^2 \ (h = 1, 2, \dots, L)$$

may not be known.

Allocation Schemes

Neyman Allocation (Neyman, 1934)

- One way to overcome this limitation is to estimate S_h^2 from a pilot study, historical data or frame information.
- If we cannot estimate S_h^2 , assume that S_h / \bar{X}_h is constant across strata: yields **X- proportional allocation**.

Allocation Schemes

Characteristics

- Neyman allocation provides an allocation of the total sample size to strata that **minimises the overall variance for a given cost or the cost for a given variance.**
- Difficulty is that the stratum population variance or any estimate of it, may not be available.
- Need an auxiliary variable that is highly correlated with the survey variable.
- More units are allocated to the more variable strata.

Allocation Schemes

■ Summary

- Given that we have L take-some strata, allocation schemes can be summarized by a single formula

$$a_h = \frac{N_h^{2q_1} \bar{X}_h^{2q_2} S_h^{2q_3}}{\sum_{h=1}^L N_h^{2q_1} \bar{X}_h^{2q_2} S_h^{2q_3}}$$

- The stratum sample size is then: $n_h = n a_h$
- Hidiroglou, M.A., and Srinath, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic*, 11, 397-405.

Allocation Schemes

N-proportional allocation	$q_1=0.5$ and $q_2=q_3=0$
X-proportional allocation	$q_1=0$ and $q_2=0.5$, and $q_3=0$
Power allocation	$q_1= q_2= p/2$, and $q_3=0$
Neyman allocation	$q_1= q_3= 0.5$, and $q_2=0$

- Baillargeon, S. and Rivest, L.-P.(2011). The construction of stratified designs in R with the package stratification. Survey Methodology, June 2011, Vol. 37, No. 1, 53-65.

Stratification boundaries

- Stratification boundaries options
 - Use boundaries from a previous survey
 - Estimate them using a pilot survey or similar survey

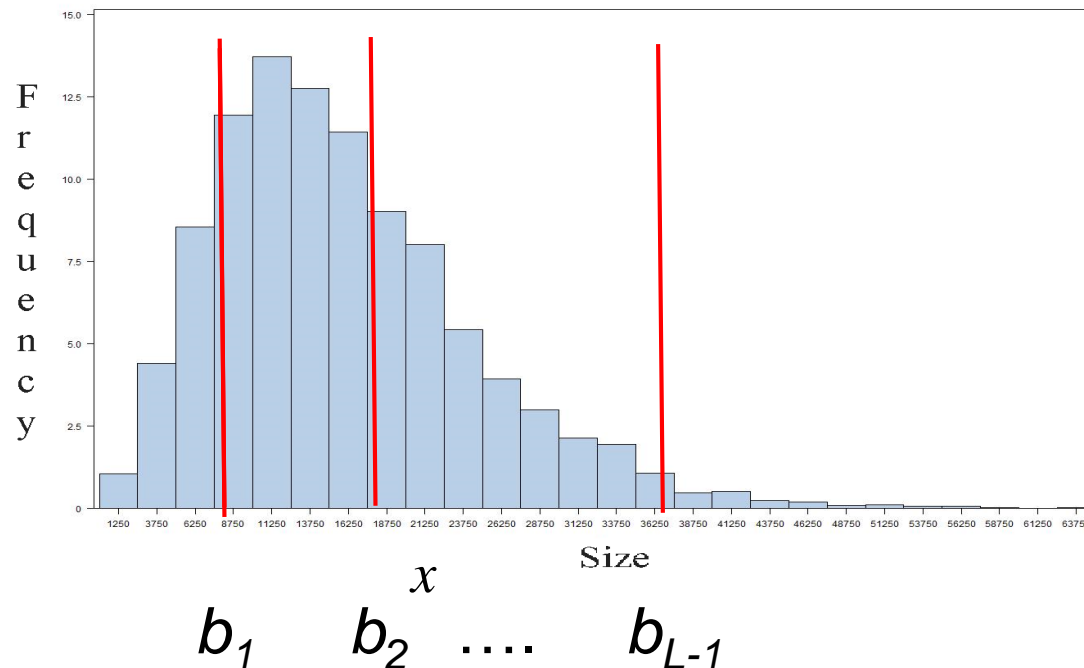
Stratification boundaries

- Determination of stratification boundaries depend on two or more of the following factors
 - Allocation scheme and overall sample size
 - Allocation scheme and coefficient of variation
 - Number of size strata
 - Presence or absence of a take-all stratum

Stratification boundaries

- Define the stratification boundaries as

$$b_1, b_2, \dots, b_{L-1} \text{ where } b_1 < b_2 < \dots < b_{L-1}$$



Stratification boundaries

- How are the stratum boundaries b_1, b_2, \dots, b_L obtained?
 - Dalenius (1950) derived equations to determine stratum boundaries that minimize the variance of the resulting estimator under Neyman allocation

$$\frac{S_h^2 + (b_h - \bar{X}_h)^2}{S_h} = \frac{S_{h+1}^2 + (b_h - \bar{X}_{h+1})^2}{S_{h+1}} \text{ with } h = 1, \dots, L-1$$

- \bar{X}_h : population mean and S_h^2 population variance of stratum h

Stratification boundaries

- Equations are not easy to solve: \bar{X}_h and S_h^2 depend on the stratum boundaries b_h
- Dalenius and Hodge (1957) presented an approximate solution to this problem by assuming that the stratification variable is approximately uniformly distributed between stratification points
- Strata are constructed by taking equal intervals on the cumulative function of the square root of the frequencies
- This method is frequently called the $\text{cum}\sqrt{f}$ (cumrootf)

Stratification boundaries

■ Steps

1. Arrange x 's in ascending order
2. Group the x 's into J intervals of equal size
3. Determine how many x 's are in each interval f_i
4. Compute $Q = \frac{1}{L} \sum_{j=1}^J \sqrt{f_i}$
5. Take the upper boundaries of the x values corresponding to $Q, 2Q, \dots, (L-1)Q$

Stratification boundaries

Frequency distribution of the percentage of bank loans in a population of 13,435 banks of the United States (McEvoy, 1956).

$\frac{\text{Industrial Loan}}{\text{Total Loans}}\%$	$f(x)$	$Cum\sqrt{f}$	$\frac{\text{Industrial Loan}}{\text{Total Loans}}\%$	$f(x)$	$Cum\sqrt{f}$
0-5	3464	58.9	50-55	126	340.3
5-10	2516	109.1	55-60	107	350.6
10-15	2157	155.5	60-65	82	359.7
15-20	1581	195.3	65-70	50	366.8
20-25	1142	229.1	70-75	39	373.0
25-30	746	256.4	75-80	25	378.0
30-35	512	279.0	80-85	16	382.0
35-40	376	298.4	85-90	19	386.4
40-45	265	314.7	90-95	2	387.8
45-50	207	329.1	95-100	3	389.5

Stratification boundaries

- Suppose that we want 5 strata.
 - Total of $\text{cum}\sqrt{x}=389.5$
 - Boundaries should be near 77.9 ($389.5/5$), 155.8, 233.7, 311.6, and 389.5

	1	2	3	4	5
Boundaries	0-5%	5-15%	15-25%	25-45%	45-100%
Near to	77.9	155.8	233.7	311.6	389.5
Q, 2Q, 3Q, 4Q, 5Q	58.9	155.5	229.1	314.7	389.5

Stratification boundaries

- Gunning and Horgan (2004) proposed another approach for identifying stratum boundaries
- Their procedure is to come up with boundaries that has equal CVs in each of the strata

Stratification boundaries

- Maximum value for the x 's is $\max(x)$
- Minimum value is $\min(x)$
- Boundaries b_1, b_2, \dots, b_{L-1} computed as

$$b_h = \min(x) * \left(\frac{\max(x)}{\min(x)} \right)^{\frac{h}{L}} \quad \text{for } h = 1, \dots, L-1$$

- Once the boundaries are determined, the stratum sizes are $n_h = n a_h$ where a_h is the allocation scheme

Stratification boundaries

- Supposing that we want to stratify a population into $L=4$ strata
- $\min(x) = 5$ and $\max(x) = 50,000$

$$b_h = 5 * \left(\frac{50,000}{5} \right)^{\frac{h}{4}} \text{ for } h = 1, \dots, 3$$

- Intervals for strata are:

$(5-50]$, $(50-500]$, $(500-5,000]$, and $(5,000-50,000]$

Stratification boundaries

- Geometric stratification
 - Very simple to implement but has weaknesses
 - If min x is too small, then it can lead to too many 'small' strata
 - It does not account for the allocation /sample size/ and/or overall CV in determining its boundaries
 - **It can lead to inefficient boundaries**

Stratification boundaries

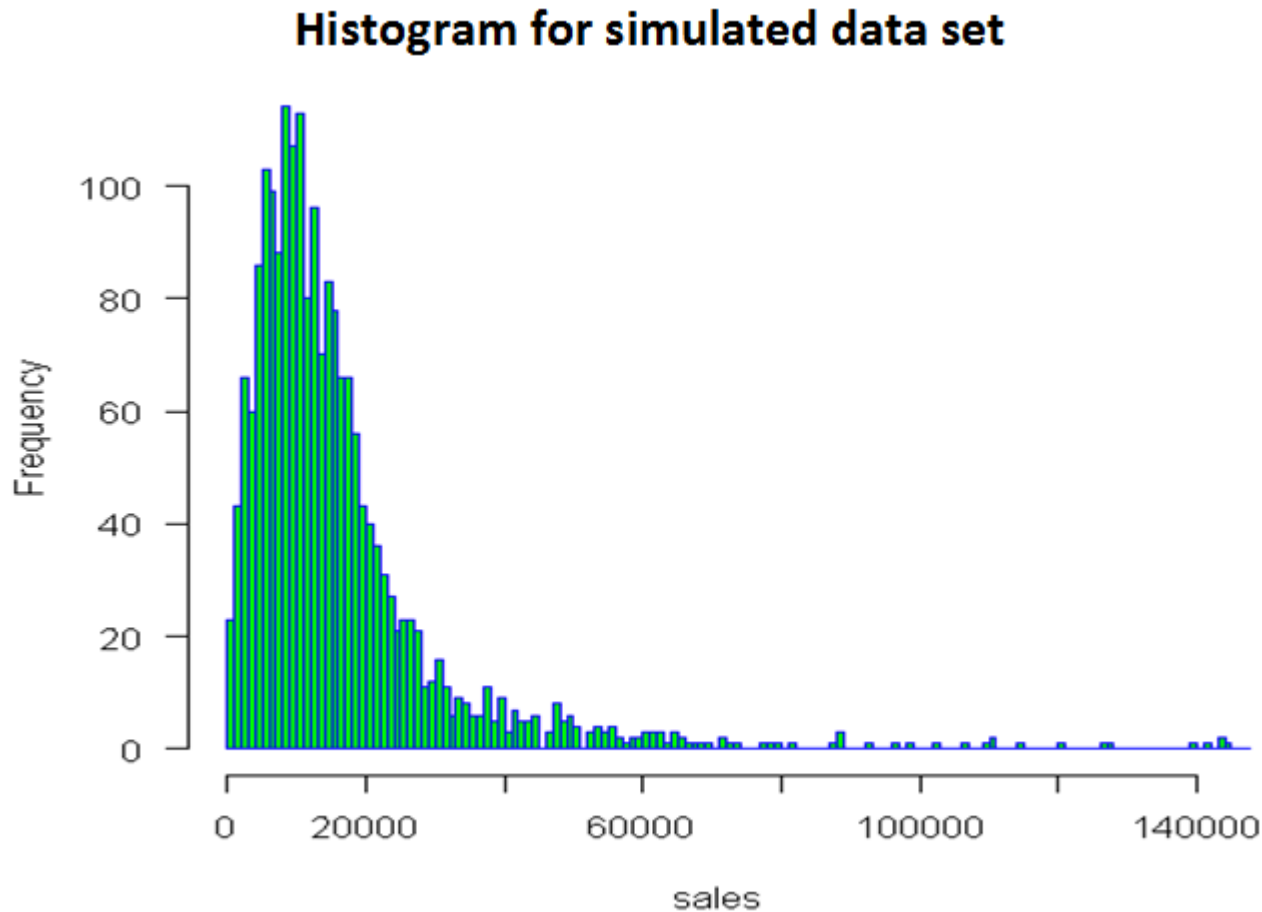
- The previous methods do not force a take-all stratum to ensure that the largest units are selected with certainty
- Lavallée and Hidiroglou (1988) stratified skewed populations by declaring the largest units as take-all (TA)
- They determined b_1, b_2, \dots, b_L that minimizes

$$n = N_L + \frac{\sum_{h=1}^L N_h^2 S_h^2 / N^2 a_h}{c^2 \bar{X}^2 + \sum_{h=1}^L N_h S_h^2 / N^2}$$

Example Stratification boundaries

- Simulated data set
 - Created by Baillargeon and Rivest (CRAN package Univariate Stratification of Survey Populations 2014)
 - Data set simulated realistic stratification variable: the size measure used for Canadian retailers in the Monthly Retail Trade Survey (MRTS) carried out by Statistics Canada.
 - Population created: 2000 businesses

Example Stratification boundaries



Example Stratification boundaries

- How do various boundary procedures compare when CV=1% and allocation is Neyman?

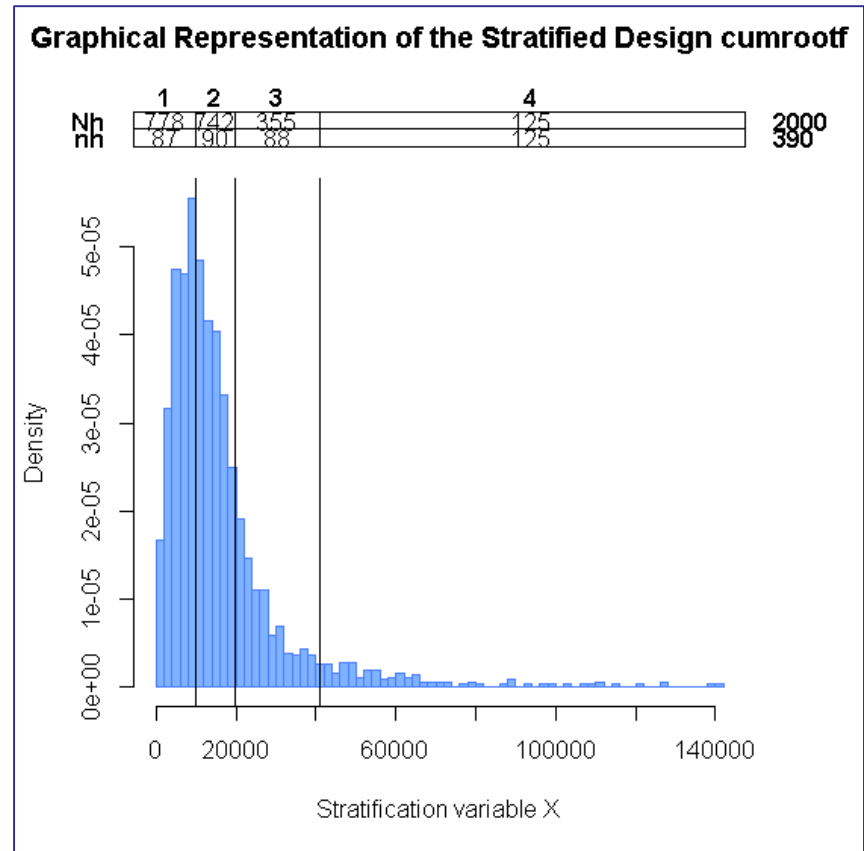
$$a_h = N_h S_h / \sum_{h=1}^L N_h S_h$$

- Size strata created: two, four and ten
- Boundary procedures considered
 - Cumrootf
 - Geometric method
 - Lavallée-Hidiroglou (LH)

Example Stratification boundaries

- Cumrootf for $L=4$
- Required sample size:
 $n= 390$

Boundaries	N_h	n_h
9,865	778	87
19,590	742	90
40,984	355	88
486,367	125	125

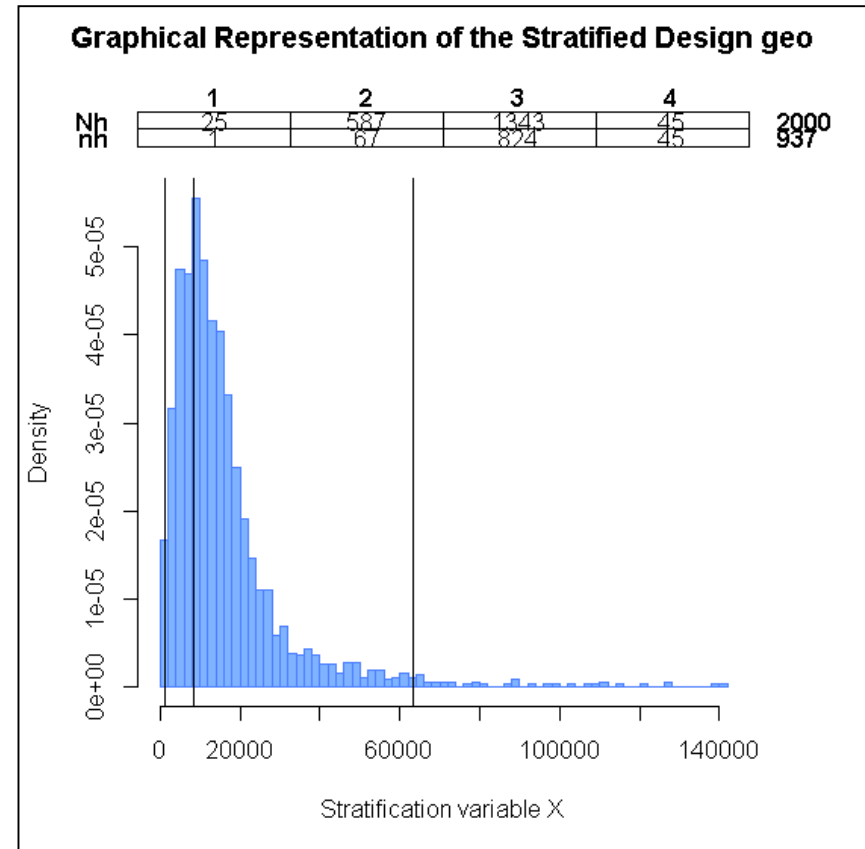


Example Stratification boundaries

- Geometric method for $L=4$
- Required sample size:
 $n=937$

Boundaries	N_h	n_h
1,082	25	1
8,288	587	67
63,490	1343	824
486,367	45	45

Note: Geometric method does not perform well because of some small units

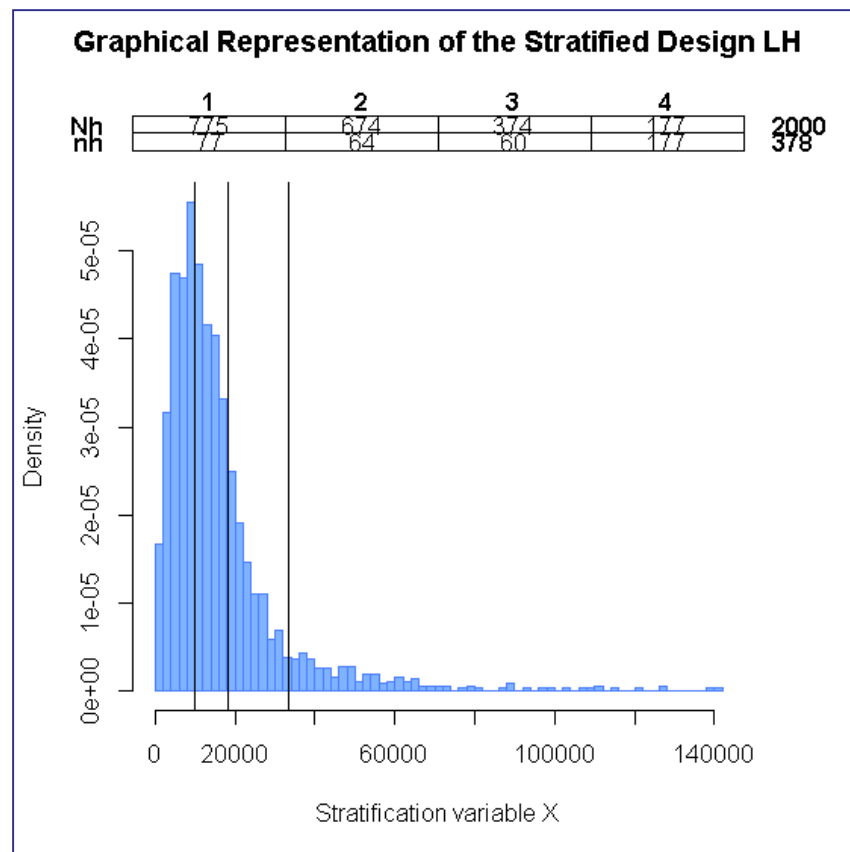


Example Stratification boundaries

- LH method for $L=4$
- Required sample size:
 $n=378$

Boundaries	N_h	n_h
9,801	775	77
18,176	674	64
33,560	374	60
485,637	177	177

Note: *boundaries and sample sizes quite close to Cumrootf*



Example Stratification boundaries

- Summary for MRTS
 - Sample size for different stratification procedures using Neyman allocation for a $CV=0.01$

Number of strata	Stratification Procedure		
	Cumrootf	Geometric	LH
2	852	1310	848
4	390	937	378
10	109	339	106

Example Stratification boundaries

- Summary for MRTS
 - Sample size for different stratification procedures
 N -proportional allocation for a $CV=0.01$

Number of strata	Stratification Procedure		
	Cumrootf	Geometric	LH
2	1696	1754	848
4	1546	1547	384
10	1305	833	125

Sampling

- Once the boundaries and sampling rates (n_h / N_h) have been determined
 - The simplest procedure for sampling the units is to either use stratified simple random sampling or stratified systematic sampling.
 - A major drawback of using these simple methods is that response burden (via rotation of the units) cannot be controlled.
 - There are methods that can do that. They all depend on assigning a uniform random number between $(0,1]$ to each population unit.

Summary

- We have explained the most commonly used allocation schemes
- We have describes how strata boundaries can be obtained for several take-some strata including the take-all stratum
- Numerical examples have been provided
- We have also provided the simplest procedure for sampling the units within each size stratum.



You can contact the PRASC team at:

prasc@statcan.gc.ca

or

statcan.prasc-prasc.statcan@canada.ca

Canada