



Statistics
Canada

Statistique
Canada

Canada



Statistics Canada
www.statcan.gc.ca

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

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 1

Mike Hidioglou and Wesley Yung
Business Survey Methods Division

January 20, 2016
Port of Spain, Trinidad and Tobago



Outline

- Introduction
- Definition
- Stratification variables
- Stratification on size
- Take-all Stratum
- Summary

Introduction

Option1:

Draw a random sample from the frame

Example: No strata

- Population size: $N=20$
- Sample size: $n= 6$

Population	Sample
1	✓
2	✓
3	
4	
5	
6	
7	✓
8	
9	✓
10	
11	
12	✓
13	
14	
15	
16	
17	
18	
19	✓
20	

Introduction

Option 2

Split the frame into homogeneous groups of units called strata, and then draw random samples from each stratum

Example: Two strata

- Stratum 1:  $N_1=15$, $n_1=4$
- Stratum 2:  $N_2=5$, $n_2=2$

Population	Sample
1	
2	✓
3	
4	
5	✓
6	
7	✓
8	
9	
10	
11	
12	✓
13	
14	
15	
16	
17	✓
18	
19	✓
20	

Introduction

- Preferred: Option 2
 - Option 1 will yield estimates that are more variable than those obtained with option 2
 - Option 2 allows us to 'control' outputs at specified levels (i.e.: geography and industry)

Definition

- Stratification is the division of a population into homogeneous, mutually exclusive groups called strata
- Allows us to ‘control’ outputs at specified levels:
 - Stratification should be closely aligned to the level of detail that we will publish
- Samples selected independently in each stratum.

Stratification variables

- What do we stratify on?
 1. Geography (provinces, districts)
 2. Industry (manufacturing, wholesale, transport,...)
 3. Size (employment, revenues, expenditures)

Stratification variables

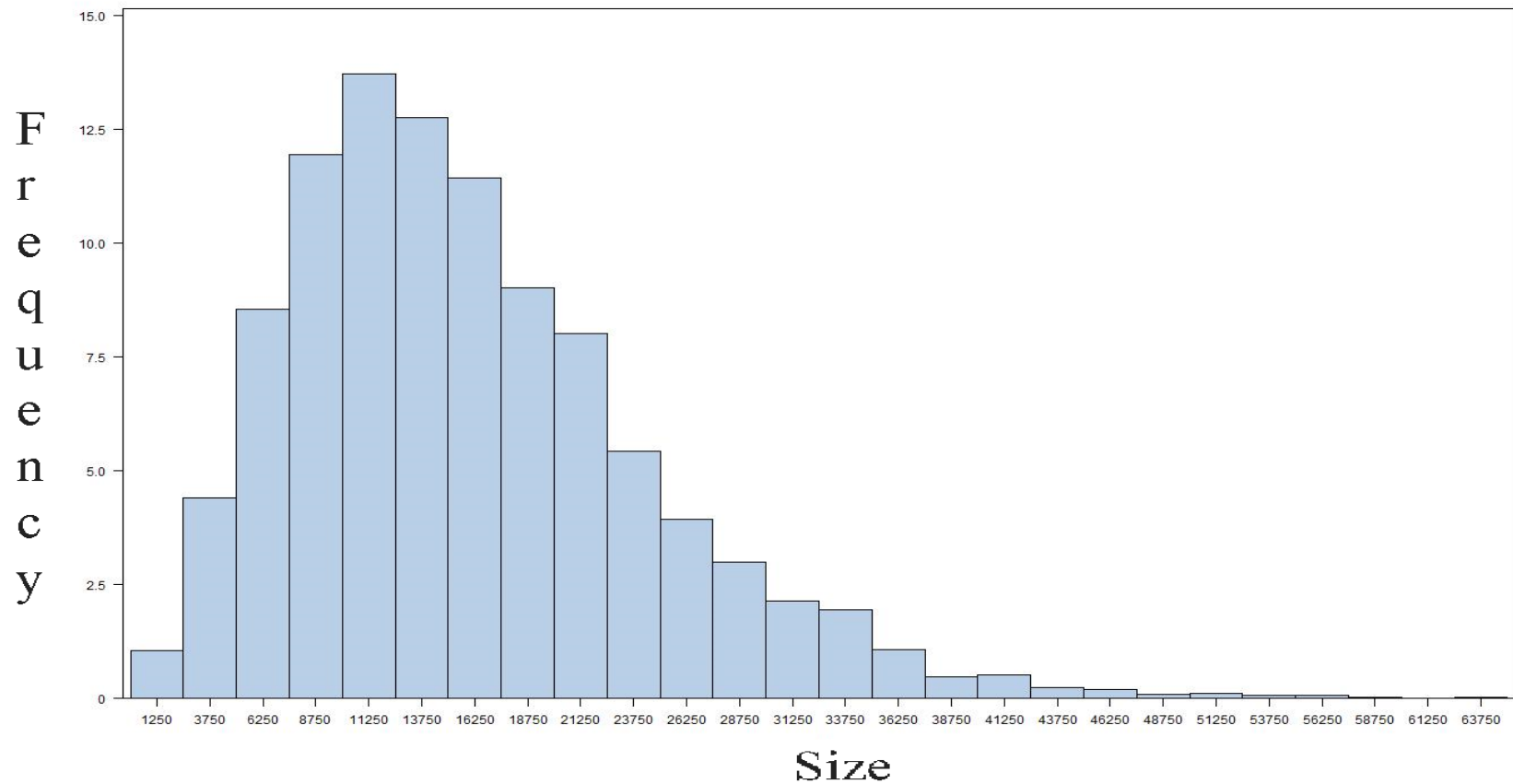
- Stratification on geography and industry
 - Do not necessarily obtain big gains at the overall level
 - Can insure reasonable quality of data for geography by industry cells

Stratification variables

- Third variable is critical
 - Distribution of variables of interest will be quite skewed
 - Few large units account for a large portion of the variable of interest
 - Need to split them into size strata based on skewed variable (employment, sales)

Stratification on size

Before stratification



Stratification on size

- Stratify the population within each geography and industry stratum into
 - Take-all stratum
 - Take-some stratum or several take-some strata
 - Take-none stratum

Stratification on size

- Take-all stratum (TA)
 - Largest units: Sampled with certainty
- Take-some strata (TS)
 - Smaller units : Sampled using simple random sampling
- Take-none stratum (TN)
 - Smallest units : No units are sampled

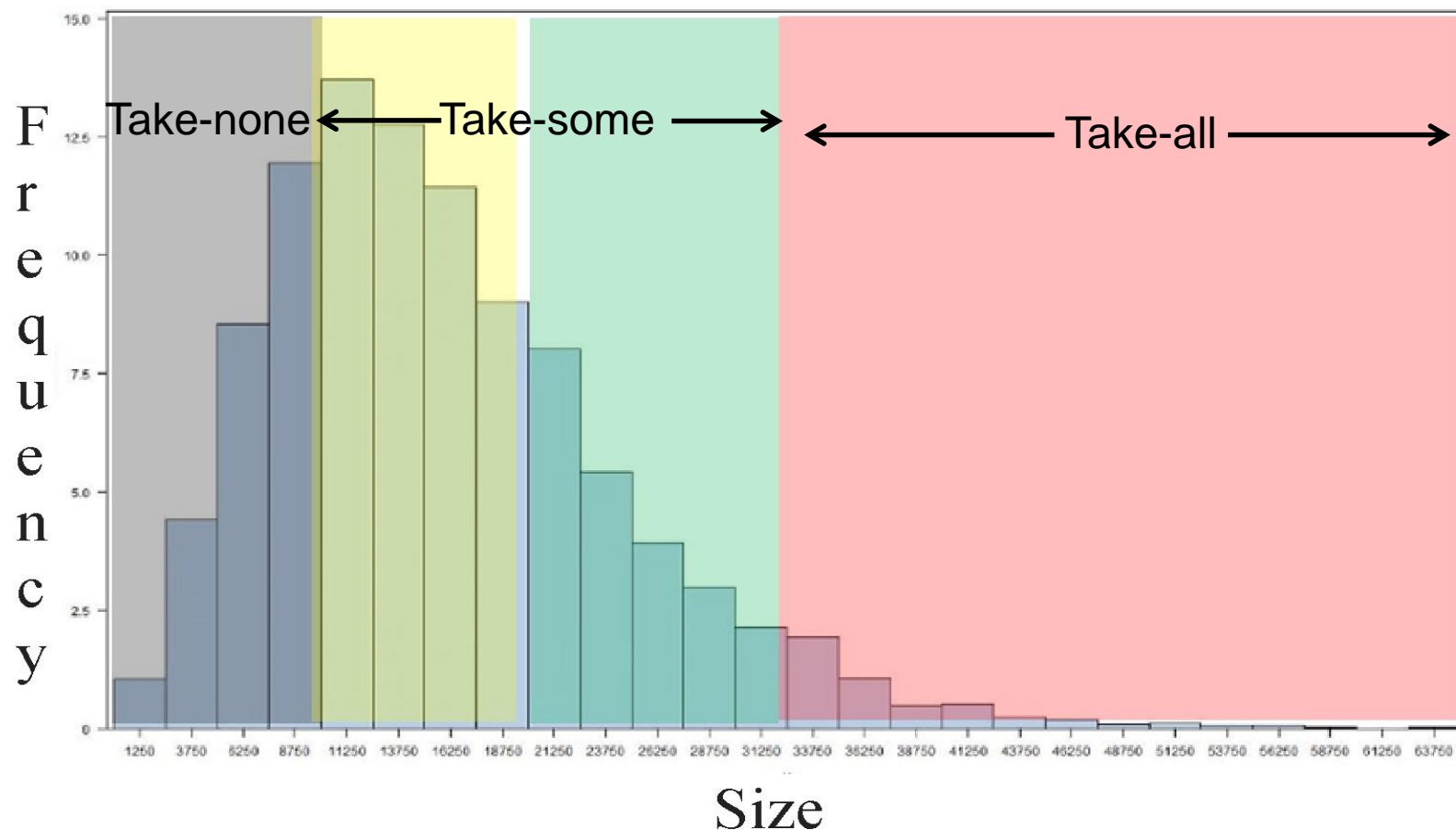
Stratification on size

■ Take-none stratum

- Part of the survey population but not sampled
- Contributes to a very small portion of the estimate
- Benefits: Response burden reduction and cost savings
- Estimated using administrative data or models (ratio estimation)

Stratification on size

After stratification



Stratification on size

- **Reliability:** Expressed as coefficient of variation (CV) for a total

$$CV\% = \frac{\sqrt{\text{Variance of total}}}{\text{Total}} * 100$$

- **Example**
 - Total=20,000
 - Estimated variance =52,900
 - $CV\% = (230/20,000) * 100\%$
=1.15% (excellent)

Stratification on size

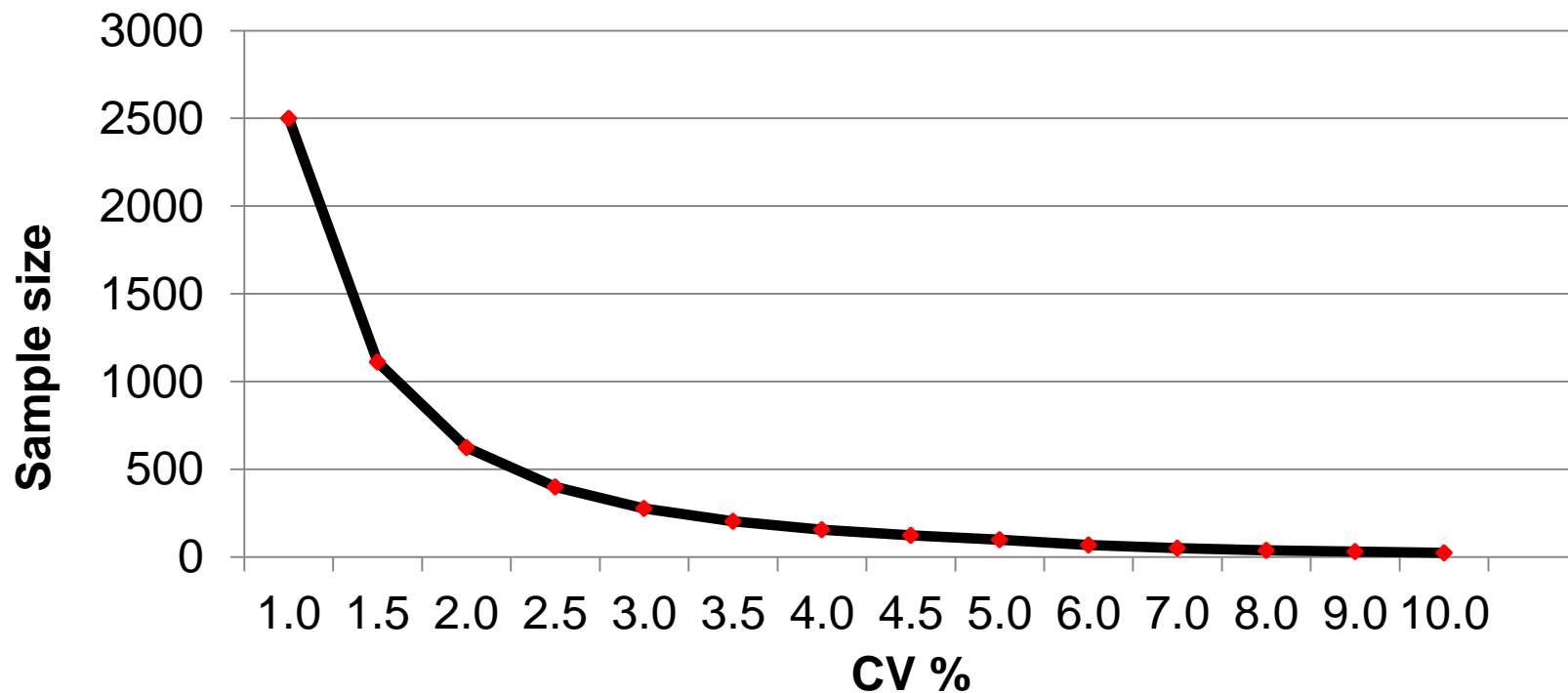
- Stratification boundaries for the total of a given variable of interest obtained with one of the following two criteria:
 - **Sample size driven:** Minimise its variance for a fixed sample size
 - **CV driven:** Minimise the sample size for a given level of precision

Stratification on size

- Sample size criterion is the one given by the clients
 - Bounds the survey costs for data collection
- However, it is sometimes easier to go the CV driven route
 - Given a CV, compute required sample size
 - Plot the sample size versus the CV
 - Choose the CV that yields the required sample size

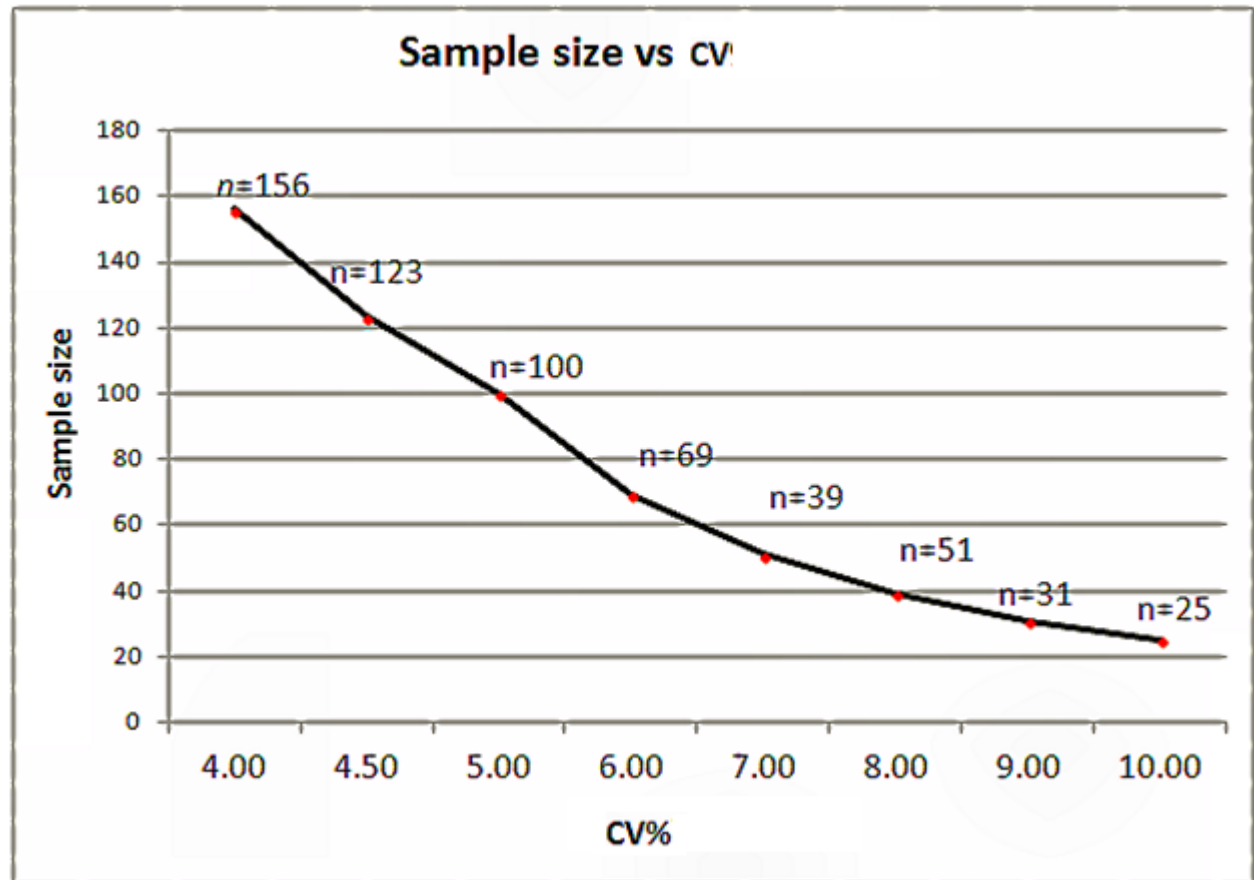
Stratification on size

Sample size vs CV%



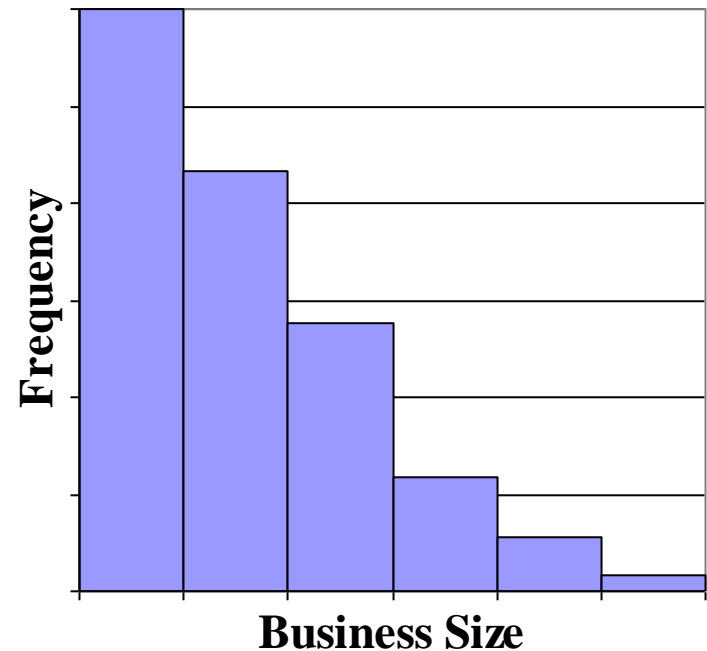
Stratification on size

<i>n</i>	CV%
156	4
123	4
100	5
69	6
51	7
39	8
31	9
25	10



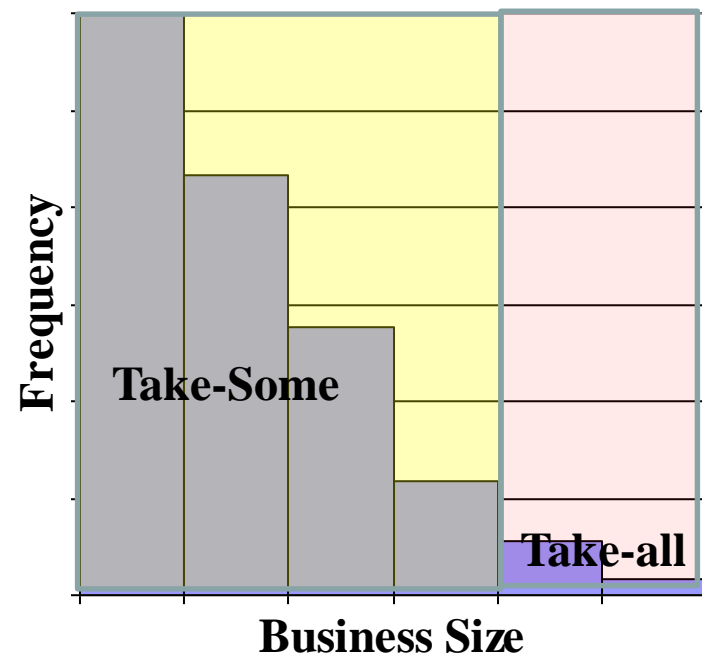
Take-all Stratum

- How do we account for highly skewed nature of data associated with business surveys?



Take-all Stratum

- Stratify the population into a take-all stratum and a take-some stratum (take-some strata)
 - Take-all stratum (TA): the largest units sampled with certainty
 - Take-some strata (TS): the remaining units sampled with a given probability



Take-all Stratum

- Suppose that we want to split a population of size N into one take-all and one take-some stratum
- Assume that x (*known*) is a variable linked to the variable of interest y (*unknown*)

Take-all Stratum

- Order the population units based on x .
- Let the population units be ordered on size from smallest to largest

$$x_{(1)}, x_{(2)}, \dots, x_{(N)}$$

- Denote
 - The population mean as $\bar{X}_N = \sum_{i=1}^N x_i / N$
 - The population variance as $S_N^2 = \sum_{i=1}^N (x_i - \bar{X}_N)^2 / (N - 1)$

Take-all Stratum

Fixed sample size

- Glasser's rule (1962) for determining an optimum cut-off point given a fixed sample size n .
- Compute a cut-off given by

$$\text{cutoff} = \bar{X}_N + \sqrt{\frac{N}{n}} S_N^2$$

- Select all units into the take-all stratum whose x value exceeds this cut-off
- That is , all units i whose $x_{(i)}$ value exceeds the cutoff

Take-all Stratum

- Example: $N = 100$; $\bar{X}_N = 20$; and $S_N^2 = 30$

n	<i>cutoff</i>
10	37
20	32
30	30
40	29

- As n increases
 - The cut-off decreases
 - More take-all units are required

Take-all Stratum

■ Fixed CV

- If the CV is given instead of sample size, compute the cut-off (Hidiroglou approximate rule (1986)) as

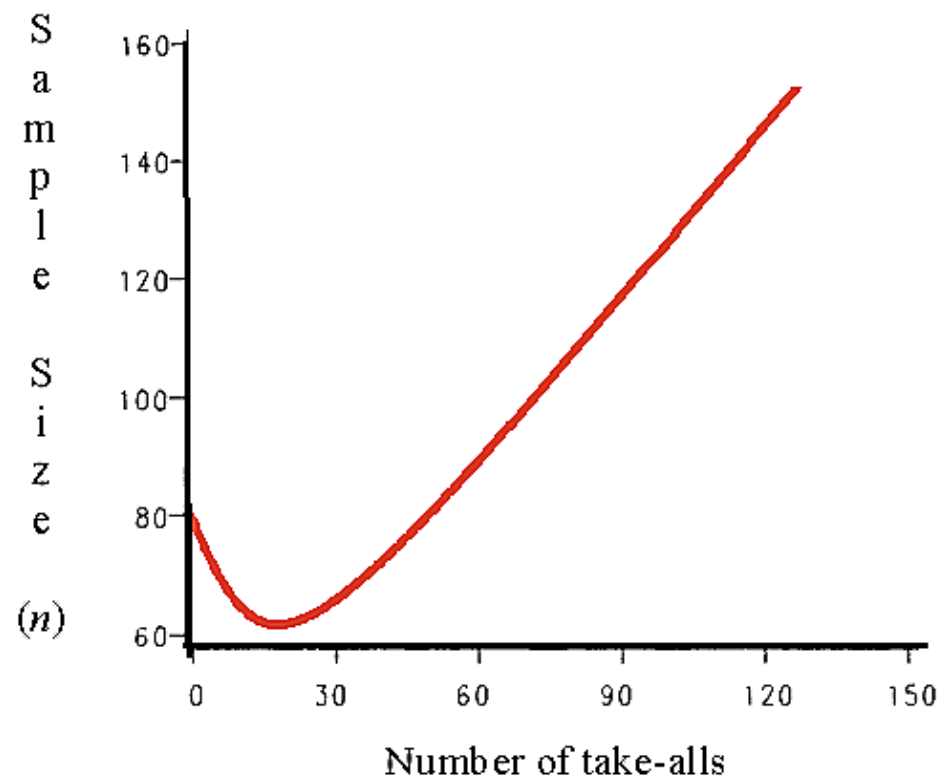
$$\text{cutoff} = \bar{X}_N + \left\{ N c^2 \bar{X}_N^2 + S_N^2 \right\}^{1/2}$$

- Select all units into the take-all stratum whose x value exceeds this cut-off
- As c increases the cutoff increases
 - Fewer take-all units

Take-all Stratum

Fixed CV problem

- For a given level of CV (c) for the estimated total, the cut-off between the take-all and take-some strata is calculated by an iterative process.
- Problem is solved by iteration.



Summary

- So far, we have introduced the term stratification: splitting a population into homogeneous subsets.
- For business surveys, stratification is carried out at the level of geography and industry required for publication.
- Business data are highly skewed and we need to split the population into a take-all stratum and a take some stratum.
- Stratification can be done either across the whole population or for each geography by industry cell.



You can contact the PRASC team at:

prasc@statcan.gc.ca

or

statcan.prasc-prasc.statcan@canada.ca

Canada