Sampling methodology and derivation of survey estimates for tourism-related surveys

Methodology and Evaluation

27 August 2010

Presentation outline

1. Definitions
2. Sampling Methodology
    - 2.1. Example of sampling specifications
    - 2.2. Stratification
    - 2.3. Sample allocation
    - 2.4. Sample selection (using JALES Technique)
3. Estimation - grossing up of sample survey data

1.1. Definition of Tourism

- In Statistics South Africa Tourism is defined as the activities of persons traveling to and staying in places outside their usual environment for not more than one consecutive year for leisure, business and other purposes not related to the exercise of an activity remunerated from within the place visited.

1.2. Supply side of Tourism

- Tourism supply explains the concepts and definitions and tourism units (e.g. establishments industry etc.) relating to the supply of tourism – related and tourism specific products and services.

1.3. Tourist Accommodation Survey

- The survey of Tourist Accommodation is one of the surveys conducted by Statistics South Africa to measure the supply side of tourism, and it covers public and private enterprises registered with the South African Tourism board, involved in the short-stay accommodation industry in South Africa.

2.1 Example of sampling specifications

Table 1: Tourist Accommodation Survey sampling specifications for 2008

Survey title	Proposed sampling SIC level	Collection SIC level	Proposed sample size
Tourist Accommodation	5-digit	5-digit	900

Special requirement:

Group sizes **1, 2 and 3** to be fully enumerated.

Classification of business into size groups

Table 2: DTI cut-off points as stipulated in the National Small Business Amendment Bill of 2003

Column 1	Column 2	Column 3	Column 4	Column 5
Sector or subsector in accordance with the Standard Industrial Classification	Size of class	The total full-time equivalent of paid employees	Total turnover	Total gross asset value (fixed property excluded)
Wholesale Trade, Commercial Agents and Allied Services	Medium	200	R64m	R10m
	Small	50	R32m	R5m
	Very Small	20	R6m	R0.60m
	Micro	5	R0.20m	R0.10m
Catering, Accommodation and other Trade	Medium	200	R13m	R3m
	Small	50	R6m	R1m
	Very Small	20	R5.10m	R1.90m
	Micro	5	R0.20m	R0.10m

Size group 1=Medium, 2=Small, 3=Very small or 4=Micro

Classification of business into size groups (cont'd)

- **Example of Enterprises size grouping:**

Table 3: Distribution of enterprises by size group on the Tourist Accommodation frame 2008

sizeGrp	Frequency	Percent	Cumulative Frequency	Cumulative Percent	Total turnover	Turnover Percentage contribution
1	222	5.22	222	5.22	15,861,560,316.25	71.264
2	216	5.08	438	10.30	1,895,927,409.57	8.518
3	64	1.50	502	11.80	356,170,729.75	1.600
4	3752	88.20	4254	100.00	4,143,680,818.26	18.617

2.2 Stratification

- Stratification is the process of dividing the population (businesses) into homogenous non-overlapping groups called **strata**.
- Strata are formed to reduce variability within the frame
- Strata are formed within economic activity and size of businesses i.e. concatenating the economic activity and size group variables,
  viz. SAS statement **stratum=SIC||sizegrp**

Stratification (cont'd)

- Table 4 below shows example of 1st seven strata formed within the Hotel Accommodation frame.

Table 4: Distribution of enterprises per stratum on Tourist Accommodation frame

stratum	Frequency	Percent	Cumulative Frequency	Cumulative Percent
641011	159	4.71	159	4.71
641012	121	3.58	280	8.29
641013	36	1.07	316	9.36
641014	1143	33.85	1459	43.20
641091	43	1.27	1502	44.48
641092	59	1.75	1561	46.22
641093	19	0.56	1580	46.79

2.3. Sample allocation

- Neyman Allocation is used to allocate sample size within each stratum
- The purpose of the method is to maximize survey precision, given a fixed sample size
- The method assumes that the cost to sample an enterprise is equal across strata.
- Based on Neyman Allocation the best sample size for stratum *h* would be:

$$n_h = n \times \frac{N_h \times \sigma_h}{\sum_{i=1}^{H} N_i \times \sigma_i}$$

where n_h = sample size for stratum h,
n = total sample size,
N_h = population size for stratum h
σ_h = standard deviation of stratum h *and*
H = total number of strata

1.4. Sample selection (JALES Technique)

- Samples are drawn using **Simple Random Sampling (SRS) which applies JALES technique** within each stratum, this takes place by using a unique random number that is available for each enterprise.
- **Unique number** is a random number uniformly distributed over the interval (0,1) to every registered enterprise in the BSF.
- The **JALES technique** uses:
  - **starting point** is any number between (0,1) selected to be different across surveys, to avoid an overlap of sampled enterprises
  - and **sampling direction** chosen across surveys is to the right

Sample selection (Cont'd)

- Every sampled enterprise carries **design weight** = number of enterprises it represents in that stratum
- The **design weight** of an enterprise within a stratum is calculated by dividing the population size by sample size of that particular stratum, e.g.:

Population and sample sizes per STRATUM

stratum	Population size	Sample size	Design weight
641011	159	159	1.000
641012	121	121	1.000
641013	36	36	1.000
641014	1143	145	7.883
641031	11	11	1.000
641032	26	26	1.000
641033	4	4	1.000
641034	684	101	6.772

2. Estimation

- Data reported by sampled businesses are used to compute estimates.
- Data of businesses in the sample are grossed-up by economic activity and size of business.
- For the purpose of this presentation, **collected income** is used as example in computing estimates and assessing the degree of confidence in the estimates.

Estimation (cont...)

- Let $\hat{y}_h$ denote the estimated total income for stratum h, the estimated total income for the overall industry is then computed using the following formula:

$$\hat{y}_w = \sum w_h \hat{y}_h$$

where $\hat{y}_w$ = estimated overall income and w_h is the design weight within a stratum

Estimation (cont...)

- Example of grossing up estimates at the stratum level, to form estimates at the reporting domain, consider the subgroup 64101 which is Hotels (NB: Hotels includes Motels, Botels and Inns).

Table 5: Aggregated total income

Stratum	Total Turnover	Weights	Weighted Total turnover
641011	3,384,865,357.00	1.00	3,384,865,357.00
641012	1,084,905,863.00	1.00	1,084,905,863.00
641013	199,640,268.00	1.00	199,640,268.00
641014	186,711,834.00	7.88	1,471,804,319.05
Gross	4,856,123,322.00		16,141,121,5807.

Estimation (cont...)

- The weighted total income for the Hotel industry is estimated as follows:

$$\sum w_h \hat{y}_h = (1\times 13{,}384{,}865{,}357.00)+(1\times 1{,}084{,}905{,}863.00)+$$
$$(1\times 199{,}640{,}268.00)+(7.88\times 186{,}711{,}834{,}.00)$$
$$=16{,}141{,}215{,}807.05$$

A total income of more than 16 trillion is contributed by the hotel industry towards the total tourist accommodation income.

Estimation (cont...)

- The relative standard error (RSE) is used to the measure the reliability of the estimated income.
- The RSE is computed as $\left[\frac{S.E(\hat{T})}{\hat{T}}\right]\times 100\ \% \leq \alpha$

where

$\hat{T}$ = the estimated total income

α = the required Relative Standard Error.

$S.E(\hat{T})$ = the standard error of the estimated Total income.

- High RSE indicates poor estimate and may result in reviewing the sample size for future survey.

Estimation (cont...)

- As an example, consider again the reporting domain 64101, we estimate the RSE for the estimated total income as follows-:

$$RSE = \left[\frac{101{,}920{,}034.30}{16{,}141{,}215{,}808.00}\right] \times 100\% = 0.63\%$$

Estimation (cont...)

- Two statistics measuring the degree of confidence of the estimated total income relative to the sample size are computed as follows:

  Relative precision = *1.96 x RSE* at 95% confidence level.

  where 1.96 is the critical value at 5 % level of significance and RSE is the relative standard error.

  Confidence Interval = $\hat{T} \pm 1.96 \times S.E(\hat{T})$

- Observing again the reporting domain 64101, we estimate the confidence interval as follows:

  Confidence Interval= 16,141,215,808.00 $\pm$ 1.96*101,920,034.30

  = [15,941,452,540.77 ;16,340,979,075.23]

- We observe a narrower interval resulting in a high degree of confidence in the estimated total income.

Estimation (cont...)

Table 6: Estimates of the sample

Reporting domain	Income estimates	Std Error (SE) of the Income estimate	Relative Std Error (RSE) %	Relative Precision (RP) %	Population size	Lower Limit	Upper limit
6410	402,203,589.00	-	0	0	7	402,203,589.00	402,203,589.00
64101	16,141,215,808.00	101,920,034.30	0.63143	1.2376	1459	15,941,452,540.77	16,340,979,075.23
64102	371,368,423.00	20,894,825.40	5.62644	11.0278	145	330,414,565.22	412,322,280.78
64103	1,262,768,650.00	64,956,850.08	5.144	10.0822	725	1,135,453,223.84	1,390,084,076.16
64109	3,936,897,794.00	115,024,859.23	2.92171	5.7266	1918	3,711,449,069.91	4,162,346,518.09
	22,114,454,262.00				**4254**		

Thank you