Introductory Survey Sampling

Training Workshop: Statistics South Africa
November 12 to November 23. 2012

Professor J. N. K. Rao, Consultant, Ottawa, Canada

Mr. G. H. Choudhry, Consultant, Ottawa, Canada

Census: Complete enumeration every 10 or 5 years. Population census collects basic items from all persons. Agricultural census collects information from all farms, etc.

Main use of census: Can produce statistics at any level of disaggregation. Local area statistics are needed.

Limitations: Only few items can be covered, expensive, not current, and not timely.

Sample surveys: Observe only a portion of the population according to a specified sampling design

Advantages: (1) Reduced cost relative to census. It is the sample size that matters and not the population size in terms of precision of estimates.
(2) Greater speed and scope: census may be impractical if highly trained personnel or specialized equipment needed.
(3) Greater accuracy: less measurement errors with well executed sample survey, i.e. less non-sampling errors
(4) Current statistics: South African Quarterly Labour Force Survey (QLFS); Monthly Canadian Labor Force Survey (LFS)
(5) Sampling in census: Long form used to collect more detailed information from a sample (1 in 5 for the Canadian Census).

Main steps in conducting a survey:

(1) Objectives of the survey clearly formulated prior to the survey.
(2) Target population vs. survey population (population to be sampled): under-coverage
(3) Data to be collected: not too many questions
(4) Degree of precision desired: sample size
(5) Method of measurement: self-administered questionnaire, interview. Mode: mail, telephone, personal visit or combination
(6) Questionnaire pretesting, field work organization

(7) Sampling frame: list of units
(8) Selection of the sample
(9) Data collection and data entry
(10) Edit and Imputation
(11) Summary and analysis of collected data

Reference: Cochran, W. G. (1977). Sampling Techniques, 3rd Ed., Wiley.

Questionnaire Design: (1) Pretest before survey. (2) Keep it simple and clear. (3) Use specific questions. (4) Relate questions to the concept of interest. (5) Open-ended questions or specified answer categories. (6) Questions that elicit correct responses. (7) Avoid double negatives. (8) Question wording. (9) Question ordering.

Measurement errors: (1) Not tell the truth: farmers in an area with support program may underreport crop yields hoping for more subsidies. (2) Not understand a question. (3) Telescoping: experience as crime victim in the last 6 months in National Crime Victimization Survey (NCVS). (4) Interviewer effect. (5) Vague questions. (6) Question wording and ordering

Reference: Lohr, S. L. (2010). Sampling: Design and Analysis, 2nd edition, Brooks/Cole.

Errors in surveys: Sampling errors, non-sampling errors: measurement errors, missing data: unit nonresponse and item nonresponse, under coverage.

Sample Survey Setup

Finite population: $U = \{1,2,...,N\}$

Associated values: $\{y_1,...,y_N\}$

Auxiliary values: $\{x_1,...,x_N\}$related to values of interest known from a census or administrative records. Only the total X for the entire population and some sub-groups of the population, e.g. Age/Sex/Race groups, may be known.

Sample: a subset s of U

Sample design: $\{s, p(s)\}$ where s belongs to a set of samples defined by the design.

$p(s)$: known probability of selecting the sample s.

Sampling scheme is used to implement a sample design.

Probability sampling: Design that ensures non-zero first order inclusion probabilities π_i for every unit $i \in U$

Basic sampling methods: Simple random sampling (SRS), stratified SRS, cluster sampling, systematic sampling, probability proportional to size (PPS) sampling. In practice, a combination of the basic methods is used to select a sample s.

Data collected: $\{(i, y_i, x_i), i \in s\}$

Simple parameters of interest:

- Total $Y = y_1 + ... + y_N = \sum_{i \in U} y_i = Y(y)$:operator notation
- Mean $\bar{Y} = Y / N$
- Proportion: binary response
- Median
- Sub-population (domain) total $\sum_{i \in d} y_i \equiv {}_d Y$
- Domain mean: ${}_d \bar{Y} = {}_d Y / {}_d N$ where ${}_d N$ is the unknown domain size. Example of domain: age-sex group

More complex parameters: Regression coefficients, Income inequality: Gini coefficient, low income proportion: proportion of people below poverty line (half median income).

Design-based (or repeated sampling) inference for the total

Design weights: $d_i = 1/\pi_i, i \in s$

Estimator of total: $\hat{Y} = \sum_{i \in s} d_i y_i = \hat{Y}(y)$

$\hat{Y}$ is called the Narain-Horvitz-Thompson (NHT) estimator

Interpretation of $\hat{Y}$: Sum of weight times the value for the units in the sample. Data file will contain columns of weights and corresponding values. Design weight d_i may be interpreted as the number of population units represented by the sample unit i including the sample unit itself.

Alternative expression for $\hat{Y}$:

$\hat{Y} = \sum_{i \in U} d_i a_i y_i$

$a_i = 1$ if $i \in s$ and $a_i = 0$ otherwise

$\pi_i = P(i \in s) = E(a_i)$

NHT estimator is design unbiased if and only if the inclusion probabilities π_i are positive for all the population units: $E_p(\hat{Y}) \equiv Y$ noting that $E(a_i) = \pi_i$

Variance of $\hat{Y}$: Measure of precision and it involves joint inclusion probabilities $\pi_{ij} = E(a_i a_j)$:

$V(\hat{Y}) = \sum_{i<j \in U} (\pi_i \pi_j - \pi_{ij})(z_i - z_j)^2 = V(y)$

if sample size n is fixed, where $z_i = y_i / \pi_i$

Sen-Yates-Grundy (SYG) variance estimator of $\hat{Y}$:

$$v(\hat{Y}) = \sum_{i<j \in s} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} (z_i - z_j)^2 \equiv s^2(\hat{Y}) = v(y)$$

SYG variance estimator is design-unbiased for $V(\hat{Y})$

Estimated coefficient of variation (CV) of $\hat{Y}$: $s(\hat{Y}) / Y$ often expressed in percent.

Large sample $1-\alpha$ level confidence interval (CI) on Y:

$$Y \in \{\hat{Y} - z_{\alpha/2} s(\hat{Y}), \hat{Y} + z_{\alpha/2} s(\hat{Y})\}$$

$z_{\alpha/2}$ = upper $\alpha/2$ point of $N(0,1)$

Interpretation of CI: In repeated sampling according to specified design, proportion $1-\alpha$ of intervals will contain the true value Y.

Efficient sampling strategy: Find a combination of design and estimator that minimizes the variance of the estimator for a given cost or minimizes the cost for a given precision. In practice, this ideal goal is not easy to achieve.

Advantages of design-based approach: No models or distributional assumptions. It provides valid large sample inferences. But inference refers to repeated sampling.

Simple random sampling

Definition: All possible samples of size n have the same probability of selection.

Properties: (1) $\pi_i = n/N, \pi_{ij} = n(n-1)/\{N(N-1)\}$ for all units i and $j(\neq i)$.
(2) $\hat{Y} = N(\text{Sample mean}\,\bar{y}) = (N/n)\sum_{i\in s} y_i$

(3) $V(\hat{Y}) = N^2(1-\frac{n}{N})\frac{S_y^2}{n}$

S_y^2 = Population variance = $\{N/(N-1)\}\sigma_y^2$

$1-n/N$ = Finite population correction (FPC)

(4) Unbiased variance estimator $v(\hat{Y})$ is obtained from (3) by replacing S_y^2 by the sample variance s_y^2

Model-dependent approach: $y_1,...,y_N$ randomly generated from some model. In particular, for SRS we assume them to be IID random variables with mean μ and variance σ_y^2. SRS

design is non-informative and hence the sample also obeys the assumed model.

Write the total as $Y = \sum_{i \in s} y_i + \sum_{i \in r} y_i$ where r denotes the set of non-sampled units. Under the assumed model, the best linear unbiased estimator (BLUE) of μ is the sample mean $\bar{y}$. The best predictor of y_i for $i \in r$ is $\hat{y}_i = \bar{y}$. Therefore the best linear unbiased predictor (BLUP) of Y is

$$\hat{Y}_m = \sum_{i \in s} y_i + \sum_{i \in r} \hat{y}_i = (N/n) \sum_{i \in s} y_i = \hat{Y}$$

The estimator is model-unbiased in the sense $E_m(\hat{Y}_m - Y) = 0$

Note: In this special case both approaches give the same estimator. It is both design unbiased and model unbiased.

Measure of variability = Mean squared prediction error =

$$E_m(\hat{Y}_m - Y)^2 = V_m(\hat{Y} - Y) = N^2(1 - \frac{n}{N})\frac{\sigma^2}{n}$$

We replace σ^2 by s_y^2 which is model-unbiased for σ^2 to get estimator of variance of $\hat{Y}_m$. Hence we also get the same variance estimator (and CI) under the two approaches but interpretations are different.

Interpretation: Under repeated realizations of the population values, proportion $1-\alpha$ of the CI will contain the realized value of Y conditionally given the sample s.

Bayesian approach: An advantage of this approach is that it provides inferences conditional on the data $\{(i, y_i), i \in s\}$, not just on the sample s. But we need to assume a parametric family of distributions for the responses y_i and prior distributions on the model parameters μ and σ^2. We get a predictive distribution of y_i for $i \in r$ given the data. If no prior information is available and we assume normality, then using a non-informative prior, the Bayesian approach will give the same answers as the other two approaches. But the

interpretations are different and the intervals are called credible intervals.

Domain Estimation

Parameters

Domain total ${}_dY = \sum_{i \in d} y_i = \sum_{i \in U} {}_d a_i y_i = Y({}_d ay)$

Domain size $N_d = \sum_{i \in U} {}_d a_i = Y({}_d a)$

Domain mean ${}_d\bar{Y} = {}_dY / {}_dN = Y({}_d ay) / Y({}_d a)$

${}_d a_i = 1$ if unit $i \in d$ and ${}_d a_i = 0$ otherwise

${}_d a_i y_i = y_i$ if unit $i \in d$ and ${}_d a_i y_i = 0$ otherwise

Estimators

Method: Simply replace $Y(.)$ by $\hat{Y}(.)$

Domain total: ${}_d\hat{Y} = \hat{Y}({}_d ay) = \sum_{i \in s(d)} d_i y_i$

$s(d) =$ sample of units belonging to domain d

Data file: Simply multiply weight d_i by the corresponding y_i in the data file and sum over the sample units i belonging to $s(d)$. Equivalently, multiply d_i by ${}_d a_i y_i$ and sum over all sample units $i \in s$

Variance estimator: $v({}_d\hat{Y}) = v({}_d ay)$

Domain size: $_d\hat{N} = \hat{Y}(_d a) = \sum_{i \in s(d)} d_i$

Properties: Estimators $_d\hat{Y}$ and $_d\hat{N}$ are design unbiased. They are called direct estimators or domain-specific estimators because they use data only from the domain of interest.

Domain mean: $_d\hat{\bar{Y}} = {}_d\hat{Y} / {}_d\hat{N}$ = ratio of two estimators
Note: Domain size or domain frame are not known. If domain size is known but not frame we call it post-stratum and if both are known we call it stratum.

Properties: $_d\hat{\bar{Y}}$ in general is design biased but approximately design unbiased for large n.

Variance estimator: $v({}_d\hat{\bar{Y}}) \approx v\{{}_d a(y - {}_d\hat{\bar{Y}})/{}_d\hat{N}\}$

This is obtained by using the approximate formula for the variance of the ratio of two estimators of totals.

Note: $v({}_d\hat{\bar{Y}})$ is obtained by replacing y_i by $z_i = {}_d a_i(y_i - {}_d\hat{\bar{Y}})/{}_d\hat{N}$ in the formula for $v(y)$. Note that z_i is 0 if $i \notin d$ and $z_i = (y_i - {}_d\hat{\bar{Y}})/{}_d\hat{N}$ if $i \in d$.

Note: We can compare the means of two different domains, for example difference of mean incomes in two different age-sex groups.

Simple random sampling: ${}_d\hat{\bar{Y}} = {}_d\bar{y}$ (domain sample mean)

Variance estimator: $v(\bar{y}_d) \approx (1-\frac{n}{N})\frac{{}_d s_y^2}{{}_d n}$

${}_d n$ = domain sample size; ${}_d s_y^2$ sample domain variance

Domain total: ${}_d\hat{Y} = (N/n)(\sum_{i\in s(d)} y_i) = (N\,{}_d n/n)\,{}_d\bar{y} = {}_d\hat{N}\,{}_d\bar{y}$

Note: One can write ${}_d\hat{Y} = N\bar{u}$ where $\bar{u}$ is the mean of $u_i = {}_d a_i y_i$ in the sample. Hence

Variance estimator: $v({}_d\hat{Y}) = N^2(1-n/N)(s_u^2/n) = v(u)$.

Note: s_u^2 will be large because u_i takes the value 0 when unit i is not in the domain d: Price to be paid (in terms of increased variance) for not knowing ${}_d N$.

Calibration Estimators: use of auxiliary information

Suppose that auxiliary information in the form of known population totals $X = (X_1, ..., X_p)^T$ is available and that the auxiliary vector x_i is also observed for $i \in s$: $\{(i, y_i, x_i), i \in s\}$. Aim of calibration is to find new weights $(w_i, i \in s)$ such that a specified distance measure between design weights d_i and revised weights w_i is minimizes subject to $\sum_{i \in s} w_i x_i = X$.

Chi-squared distance: $\Phi = \sum_{i \in s} (w_i - d_i)^2 / (d_i q_i)$

where q_i is a "tuning constant" for unit i.

Solution:

$$w_i = \{1 + (X - \hat{X})^T (\sum_{i \in s} q_i d_i x_i x_i^T)^{-1} (q_i x_i)\} d_i \equiv g_i d_i$$

Note: Weight w_i depends only on x_i. The resulting estimator of Y is called the generalized regression (GREG) estimator:

$$\hat{Y}_{GR}(y) = \sum_{i \in s} w_i y_i = \hat{Y} + (X - \hat{X})^T \hat{B}$$

$$\hat{B} = (\sum_{i \in s} d_i q_i x_i x_i^T)^{-1}) \sum_{i \in s} d_i q_i x_i y_i$$: weighted regression

Note that that $\hat{Y}_{GR}(x) = X$ and GREG ensures consistency of estimators when aggregated over different variables attached to the sample units: $\hat{Y}_{GR}(y_1) + \hat{Y}_{GR}(y_2) = \hat{Y}_{GR}(y_1 + y_2)$

Domain estimation: Domain total Y_d is estimated as ${}_d\hat{Y}_{GR} = \sum_{i \in s(d)} w_i y_i = \hat{Y}_{GR}({}_d ay)$.

Properties of GREG: $\hat{Y}_{GR}$ is approximately design unbiased for large sample sizes. A variance estimator is given by $v(\hat{Y}_{GR}) = v(ge)$ where $e_i = y_i - x_i^T \hat{B}$ are the regression residuals. One can also use $v(e)$ as the variance estimator but $v(ge)$ reduces underestimation caused by $v(e)$. If x explains y through linear regression, then the variability of residuals e_i will be smaller than the variability of y_i and hence GREG estimator will be more efficient than the simple NHT estimator $\hat{Y}$. GREG estimators are extensively used in practice.

Special cases: (1) Suppose x is a scalar and tuning constant $q_i = x_i^{-1}$. Then $\hat{Y}_{GR}$ reduces to the widely used ratio estimator $\hat{Y}_R = (\hat{Y} / \hat{X})X$. Here $g_i = X / \hat{X}$ does not depend on i.

(2) Suppose we partition U into G post strata U_g with known population sizes N_g. Let $x_i = (x_{1i},..., x_{Gi})^T$ where $x_{gi} = 1$ if $i \in U_g$ and $x_{gi} = 0$ otherwise. Then $\hat{Y}_{GR}$ reduces to the post-stratified estimator $\hat{Y}_{PS} = \sum_g (N_g / \hat{N}_g)\hat{Y}_g$ where $\hat{Y}_g$ and $\hat{N}_g$ are the direct estimators of the total Y_g and the size N_g for the post-stratum g: $\hat{Y}_g = \sum_{i \in s(g)} d_i y_i, \hat{N}_g = \sum_{i \in s(g)} d_i$. Note that

$s(g)$ is the sample falling in the post stratum g. The estimator $\hat{Y}_{PS}$ calibrates to the known post-strata sizes: $\hat{Y}_{PS}(x_g) = N_g$ (3) GREG estimator can be used to calibrate to known marginal post-strata sizes of two or more post-stratification variables, for example age and sex marginal counts. (4) Consider SRS, tuning constant $q_i = 1$ and one auxiliary variable $x_i$such that $x_i = (1, x_i)^T$. In this case GREG estimator reduces to the customary regression estimator $\hat{Y}_{REG} = N\{\bar{y} + b(\bar{X} - \bar{x})\}$ where $\bar{y}$ and $\bar{x}$are the sample means and $b = \sum_{i \in s}(y_i - \bar{y})(x_i - \bar{x}) / \sum_{i \in s}(x_i - \bar{x})^2$is the customary least squares estimator of the slope coefficient in the linear regression of y on x.

Model dependent approach: (1) Consider regression through the origin: $E_m(y_i) = \beta x_i$ and $V_m(y_i) = \sigma_i^2 = \sigma^2 x_i$ and assume that the model holds for the sample. In this case the BLUP of the total Y for any sampling design reduces to $\hat{Y}_m = \sum_{i \in s} y_i + \sum_{i \in r} \hat{y}_i$ where $\hat{y}_i = \hat{\beta} x_i$ and $\hat{\beta}$ is the weighted least squares estimator (WLS) of β given by $\hat{\beta} = (\sum_{i \in s} x_i y_i / \sigma_i^2)(\sum_{i \in s} x_i^2 / \sigma_i^2)^{-1} = \bar{y} / \bar{x}$. It is easy to verify that $\hat{Y}_m$ reduces to the ratio estimator $\hat{Y}_m = (\bar{y} / \bar{x}) X$ regardless of the design. Only under SRS the BLUP agrees with the GREG estimator of case 1. (2) Suppose the model is $E_m(y_i) = \beta_0 + \beta_1 x_i, V_m(y_i) = \sigma^2$. In this case the

BLUE of y_i for $i \in r$ is $\hat{y}_i = b_0 + b_1 x_i = \bar{y} + b(\bar{X} - \bar{x})$, where $b_0 = \bar{y} - b\bar{x}$ and $b_1 = b$ are the least squares estimators of β_0 and β_1. BLUP of the total Y is $\hat{Y}_m = N\{\bar{y} + b(\bar{X} - \bar{x})\}$ for any design provided the model holds for the sample. It follows that the BLUP agrees with the GREG estimator only under SRS.

Model-assisted Approach: Here we assume only a "working" model and the resulting design-based inferences are asymptotically valid regardless of the validity of the assumed working model. If the working model is good, then we gain in terms of efficiency of the estimators.

Method: Suppose y_i^0 is the predictor of y_i for $i \in U$ based on x_i and a working model. Then we can write $Y = \sum_{i \in U} y_i^0 + \sum_{i \in U} e_i$, where $e_i = y_i - y_i^0$ are the prediction errors. If the model is good, then the variability of the predictor errors e_i will be small compared to the variability of y_i. Now the population total of the e_i is not known so we replace it by the corresponding NHT estimator $\hat{E} = \sum_{i \in s} d_i e_i$ to get the model assisted estimator $\hat{Y}_{ma} = \sum_{i \in U} y_i^0 + \hat{E}$. This estimator is asymptotically design unbiased regardless of the validity of the working model and yet could lead to increased efficiency if the working model is good.

Working model: $E_m(y_i) = x_i^T \beta,\ \ V_m(y_i) = \sigma_i^2$ for $i \in U$

If the population values y_i and x_i were known then the weighted least squares estimator of β is the WLS estimator $B = T^{-1} t$, where $T = \sum_{i \in U} x_i x_i^T / \sigma_i^2$ and $t = \sum_{i \in U} x_i y_i / \sigma_i^2$.

Now replace T and t by the corresponding NHT estimators $\hat{T}$ and $\hat{t}$ to get $\hat{B} = \hat{T}^{-1} \hat{t}$. This leads to the predictor $y_i^0 = x_i^T \hat{B}$ for $i \in U$ and the resulting model-assisted estimator $\hat{Y}_{ma}$ reduces to $\hat{Y} + (X - \hat{X})^T \hat{B}$. This is a GREG estimator and it is identical to the calibration estimator if we choose the tuning constant as $q_i = \sigma_i^2$, provided the user-specified calibration totals match the $x-$totals corresponding to the auxiliary variables in the working model.

If the working model is different, say includes a quadratic term, then calibration and model assisted estimators do not agree if the user specifies only the total of x. We need to include the total of the x_i^2 values in the calibration in order to match the model-assisted estimator. So the two concepts are fundamentally different. But calibration has the advantage that we use the same adjusted weights w_i for all $y-$variables. As a result, calibration estimator can lead to poor efficiency if the response variable y is not related to the calibration variable.

Projection Estimator: An important special case occurs when $\hat{E} = 0$, leading to $\hat{Y}_{ma} = X'\hat{B} = \sum_{i \in U} y_i^0$. The ratio estimator and the post-stratified estimator are special cases of a projection estimator. A sufficient condition for $\hat{E} = 0$ is that $\sigma_i^2 = \lambda^T x_i$ for some vector λ. For the projection estimator the weights are given by $w_i = d_i g_i$ where $g_i = X^T \hat{T}^{-1} x_i / \sigma_i^2$. Note that we need only write $\sigma_i^2 = \sigma^2 a_i$ where σ^2 is unknown and a_i is known. The estimator $\hat{Y}_{ma}$ does not depend on σ^2. Note that domain GREG does not have the projection form and using the sum of the predicted values y_i^0 in the domain as the domain estimator leads to design-inconsistent estimator.

Stratification

(1) Strata U_h $(h = 1,..., L)$ represent a partition of U . Frames of strata are known. Example: Province X NAICS for business surveys in Canada (NAICS: North American Industry Classification System, 3 or 2 digit code).
(2) Advantages: (a) Administrative convenience: separate field offices. (b) Permits handling sampling problems that differ in different parts of the population. (c) Gain in precision.

Basic theory: Let $Y_h = Y_h(y)$ be the total for stratum h. Then $Y = Y_1 + ... + Y_L$. Sample design consists of independent

samples s_h from the L strata with specified probabilities $p_h(s_h), h = 1,..., L$. Let $\hat{Y}_h = \hat{Y}_h(y)$ be the NHT estimator of Y_h based on the specified design. Then the stratified NHT estimator of Y is $\hat{Y}_{st} = \hat{Y}_{st}(y) = \hat{Y}_1 + ... + \hat{Y}_L$. Because of independent sampling, $V(\hat{Y}_{st}) = \sum_h V(\hat{Y}_h) = \sum_h V_h(y)$. Similarly, $v(\hat{Y}_{st}) = \sum_h v(\hat{Y}_h) = \sum_h v_h(y)$.

Stratified SRS: A sample of size n_h is selected by SRS from the N_h units in stratum h. In this case we have

$$\hat{Y}_h = N_h \bar{y}_h, V_h(y) = N_h^2 (1 - n_h / N_h)(S_{hy}^2 / n_h)$$

where $\bar{y}_h$ is the sample mean and S_{hy}^2 is the population variance in stratum h. The variance estimator $v_h(y)$ is

obtained by replacing S_{hy}^2 with the sample variance s_{hy}^2, assuming $n_h \geq 2$ for all h.

We can write $V(\hat{Y}_{st}) = A_0 + \sum_h A_h / n_h$ where $A_h = N_h^2 S_{hy}^2$ and $A_0 = -\sum_h N_h S_{hy}^2$. A simple cost function is $C = c_0 + \sum_h c_h n_h$ where c_0 is the fixed overhead cost and c_h is the cost per unit in stratum h.

Design issues: (a) Number of strata L. (b) Construction of strata. (c) Sample size allocation to strata. Using a large number of homogeneous strata increases precision of estimation. In large-scale socio-economic surveys,

stratification is done to the extent that two primary sampling units (clusters) are selected from each stratum to permit variance estimation.

Construction of strata: size stratification

Suppose that we know some population size measures, say $x_1 < x_2 < ... < x_N$ related to the variable of interest y (for example size of firms in a business survey). Equal coefficient of variation (CV) within strata: $S_{1x} / \overline{X}_1 = ... = S_{Lx} / \overline{X}_L$ can lead to efficient estimators. Let $k_0 = \min(x_i)$ and $k_L = \max(x_i)$. Then the L strata are defined by the points $k_1,...,k_{L-1}$such that $k_0 < k_1 < ... < k_{L-1} < k_L$, where

$k_h = k_0(k_L / k_0)^{h/L}$ for $h = 1,..., L-1$ (Gunning and Horgan, Survey Methodology, 2004, 159-166).

Example: Suppose $L = 4, k_0 = 5, k_L = 50{,}000$. Then $k_h = 5(10)^h$ which leads to the following stratification boundaries: |5 – 50|50 – 500|500 – 5000|5000 – 50000|

Sample size allocation: stratified SRS

Proportional allocation: Given the strata, let N_h be the size of stratum h such that $\sum_h N_h = N$. Proportional allocation of the total sample size n is given by $n_h = nW_h$ where $W_h = N_h / N$ is the relative size of stratum h. Note that proportional

allocation does not depend on the variable of interest y. Hence, it can be inefficient especially for skewed populations. However, the estimator $\hat{Y}_{st}$ under proportional allocation is more efficient than $\hat{Y}$ under simple random sampling if N_h^{-1} is negligible, although efficiency gains are modest.

Optimal allocation: Given the L strata suppose we want to draw simple random samples of sizes n_h from the strata. We assume a simple cost function $C = c_0 + \sum_h c_h n_h$, where c_0 is the fixed overhead cost and c_h is the cost per unit in stratum h. Objective is to minimize the cost with respect to sample sizes n_h subject to a fixed variance

$$V_0 = V(\hat{Y}_{st}) = A_0 + \sum_h A_h / n_h$$

$$A_h = N_h^2 S_{hy}^2, A_0 = -\sum_h N_h S_{hy}^2.$$

Alternatively, we may want to fix the cost and minimize $V(\hat{Y}_{st})$. In either case, we have

$$n_h(\text{opt}) = \lambda \frac{\sqrt{A_h} / \sqrt{c_h}}{\sum_h \sqrt{A_h}\sqrt{c_h}}.$$

Proportionality factor λ is determined by the constraint used. If cost is fixed, then $\lambda = C - c_0$. In practice, we use the known population values of some variable x correlated with y is used (for example, from the census or administrative records).

Neyman allocation: Suppose $c_h = c$ for all h and cost C is fixed, then the total sample size is $n = (C - c_0)/c$ and $n_h(\text{opt}) = n\sqrt{A_h} / \sum_h \sqrt{A_h}$.

Multiple characteristics: Suppose we have p variables of interest and $\hat{Y}_{l,st}$ is the estimator of the total Y_l for the variable $l = 1,..., p$. Let $V(\hat{Y}_{l,st}) = V_l = A_{0l} + \sum_h A_{hl}/n_h$. Objective is to minimize the cost $C - c_0$ with respect to the n_h subject to fixed tolerances V_{0l} on the variances V_l; that is $V_l \le V_{0l}, l = 1..., p$.

We transform the problem into a convex programming problem by letting $m_h = 1/n_h$. Then we are minimizing a separable convex function $\sum_h c_h / m_h$ with respect to the m_h subject to linear constraints $A_{0l} + \sum_h A_{hl} m_h \le V_{0l}, l = 1,..., p$. Additional constraints on n_h can be imposed. For example, the constraint $2 \le n_h \le N_h$ (equivalently $N_h^{-1} \le m_h \le 1/2$) ensures unbiased variance estimation and stratum sample sizes n_h always below the associated population sizes N_h.

Construction of strata under Neyman allocation: Cumulative $\sqrt{f}$ rule minimizes the variance under Neyman allocation. Rule is to form the square root of frequencies f and choose

the stratification points such that approximately equal sizes on cum $\sqrt{f}$ are obtained.

Take-all stratification: In Business Surveys, it is a common practice to have a take-all stratum of firms in the sense that all the firms are sampled. In our notation, boundary point k_{L-1} creates the take- all stratum L and $n_L = N_L$. Let $a_h = n_h/(n - N_L)$, $h = 1,..., L-1$ so that $\sum_h a_h = 1$. Under SRS within the sampled strata, we have

$$V(\hat{X}_{st}) = \sum_{h=1}^{L-1} N_h^2 (1 - n_h / N_h)(S_{hx}^2 / n_h),$$

where S_{hx}^2 is a function of the stratification points k_{h-1} and k_h. Now substituting $n_h = a_h(n - N_L)$ in $V(\hat{X}_{st})$ and using

Neyman allocation $a_h = N_h S_{hx} / \sum_{h=1}^{L-1} N_h S_{hx}$ and solving for n for a fixed variance $V(\hat{X}_{st})$ we see that n is a function of $k_1,...,k_{L-1}$. Taking the derivative of n with respect to k_h and equating it to 0, we get quadratic equations of the form

$\alpha_h k_h^2 + \beta_h k_h + \gamma_h = 0, h = 1,..., L-1$

where $\alpha_h, \beta_h, \gamma_h$ are functions of k_{h-1}, k_h, k_{h+1}.

Algorithm: (1) Start with Gunning et al. boundary points, say $k_1',...,k_{L-1}'$ and compute the corresponding $\alpha_h', \beta_h', \gamma_h'$. Calculate n and allocate $n - N_L$ to get a_h. (3) Replace $\{k_h'\}$

by $\{k_h''\}$ where k_h'' is the solution of the quadratic equation using the starting values $\{k_h'\}$. (4) Repeat above steps until convergence.

IPPS sampling

To select primary sampling units (or clusters) i varying in size x_i which is related to cluster total Y_i, inclusion probability proportional to size (IPPS) sampling is often used to increase efficiency of estimation. It has also other desirable features in the context of two-stage sampling within strata (discussed later).

We want to select n clusters by IPPS sampling without replacement such that the inclusion probability π_i for cluster i is equal to np_i, where $p_i = x_i / X$ and X is the known population total of the sizes x_i. We assume $np_i < 1$ for all clusters i. In the case of single stage cluster sampling the NHT estimator of the total Y is $\hat{Y} = \sum_{i \in s} d_i Y_i$, where $d_i = (np_i)^{-1}$. It now follows that if the size measures are strongly related to the corresponding cluster totals, the NHT estimator will lead to significant reduction in the variance. There are many methods of IPPS sampling and we consider only one such method that is commonly used to select clusters in stratified two-stage sampling.

Systematic PPS sampling: (1) Arrange the N sizes in some order, say $x_1,...,x_N$. (2) Then the n selected clusters are those whose labels j satisfy

$$\sum_{i=1}^{j-1} x_i < \frac{X}{n}(r+k) \le \sum_{i=1}^{j} x_i$$

for integers k = 0, 1, 2,--- (n-1) , where r is a random number between 0 and 1. This is equivalent to selecting a random number between 0 and X/n and stepping up by X/n.

Example: $n = 3,\ N = 8$ and $X = 300$ so that $X / n = 100.$

i:	1	2	3	4	5	6	7	8
x_i:	15	81	26	42	20	16	45	55
$\sum_{i=1}^{j} x_i$:	15	96	122	164	184	200	245	300

36 136 236

Random number $100r = 36$. This gives 36 for $k = 0$, 136 for $k = 1$ and 236 for $k = 2$. Units 2, 4 and 7are selected for the sample.

Two-stage Sampling

Population consists of N clusters with M_i elements in cluster i ($\sum_i M_i = M_0$). Y_i denotes the total for cluster i. A sample s of n clusters is selected from the N clusters in the population according to IPPS design using size measures p_i so that $\pi_i = np_i$. Suppose m_i elements $s(i)$ are drawn by SRS from cluster i if it is included in the sample s.

If the cluster totals are known then the NHT estimator of the total is $\hat{Y} = \sum_{i \in s} Y_i / (np_i)$. We simply replace Y_i by its estimator $\hat{Y}_i = M_i \bar{y}_i$, where $\bar{y}_i = m_i^{-1} \sum_{j \in s(i)} y_{ij}$ to get

$$\hat{Y}_m = \Sigma_{i \in s} \hat{Y}_i / (np_i) = \Sigma_{i \in s} \Sigma_{j \in s(i)} d_{ij} y_{ij}$$

$d_{ij} = (1/np_i)(M_i / m_i)$: design weight attached to the sample element j in the sample cluster i.

Special case: Suppose all the population cluster sizes are known and we use size measures proportional to cluster sizes: $p_i = M_i / M_0$. In this case $d_{ij} = M_0 / (nm_i)$.

Self-weighting: Suppose we wish to use equal work loads: $m_i = m$. Then all the design weights are equal to $d = M_0 / (nm)$ which is the inverse of the overall sampling

fraction. Estimator $\hat{Y}_m$ is equal to the product of inverse overall sampling fraction and the sample total $\sum_i \sum_j y_{ij}$.

General case of self weighting:

$m_i = (\text{expected overall sampling fraction})\{M_i/(np_i)\}$

Variance estimator: An unbiased variance estimator of $\hat{Y}_m$ is the sum of two components $\hat{T}_1$ and $\hat{T}_2$, where
$\hat{T}_1$ = copy of the variance estimator in single stage sampling of clusters with Y_i replaced by $\hat{Y}_i$: requires joint inclusion probabilities

$\hat{T}_2$ = copy of the estimator in single stage cluster sampling with Y_i replaced by the estimator of conditional variance of $\hat{Y}_i : \hat{V}_i = M_i^2 (1 - m_i / M_i) s_{iy}^2$.

Remark: Unbiased variance estimator is seldom used in practice because of its somewhat complex form. Instead, it is a common practice to use the formula for PPS sampling with replacement as an approximation:

$v_a(\hat{Y}_m) = s_z^2 / n$

where $s_z^2 = \sum_{i=1}^{n} (z_i - \bar{z})^2 / (n-1)$ is the sample variance of the values $z_i = \hat{Y}_i / (np_i) = n \sum_{j \in s(i)} d_{ij} y_{ij}$ for $i = 1, \ldots, n$ are

weighted sample cluster totals. The above approximation forms the basis for re-sampling methods such as the jackknife, balanced repeated replication (BRR) and the Rao-Wu bootstrap, in the sense that the re-sampling variance estimator of $\hat{Y}_m$ reduces to $v_a(\hat{Y}_m)$. Rao-Wu bootstrap is widely used in Statistics Canada.

Remark: In practice, design weights are first adjusted for unit nonresponse; for example within sample clusters as done in the South African QLFS (households within PSUs). In general, weighting classes are formed on the basis of auxiliary information observed for all the sample units and then adjustments are made within each weighting class. After non-response adjustment, the adjusted weights are calibrated

to agree with known auxiliary totals, leading to final weights $\{w_{ij}\}$ which are reported in the data file along with the associated data values $\{y_{ij}\}$. The estimator of Y is then given by $\hat{Y}_{mw} = \sum_i \sum_j w_{ij} y_{ij}$

Stratified two-stage sampling: We introduce the additional subscript h to denote strata and note that $\hat{Y}_m = \sum_h \hat{Y}_{hm} = \sum_h \sum_i \sum_j d_{hij} y_{hij}$

Similarly the final estimator $\hat{Y}_{mw}$ is obtained from the final weights w_{hij}

Approximate variance estimator:

$v(\hat{Y}_m) = \Sigma_h s_{hz}^2 / n_h$

where s_{hz}^2 is the sample variance of the n_h weighted sample cluster totals $z_{hi} = n_h \Sigma_j d_{hij} y_{hij}$ within stratum h.

Efron's Bootstrap: IID case

Suppose $y_1, ..., y_n$ is an IID sample from some distribution and suppose θ is the parameter of interest, for example the mean or the median of the unknown distribution. Suppose we also have an estimator $\hat{\theta}$ of θ based on the random sample. We want to find a computer intensive but routine method to

calculate the variance estimator of $\hat{\theta}$ and confidence intervals for θ.

Bootstrap sampling: (1) Draw a simple random sample with replacement of size n from the original sample of size n and denote this bootstrap sample as $y_1^*,...,y_n^*$ (2) Compute the associated estimator of θ and denote it as θ^*. (3) Repeat (1) and (2) a large number of times, say B to get $\theta_1^*,...,\theta_B^*$. (4) Bootstrap variance estimator of $\hat{\theta}$ is then given by

$$v_{\text{BOOT}}(\hat{\theta}) = B^{-1} \sum_{b=1}^{B} \{\theta_b^* - \hat{\theta})^2$$

Remark: One could replace $\hat{\theta}$ by the average of the bootstrap estimates θ_b^*

Special case: $\hat{\theta} = \bar{y} \Rightarrow v_{\text{BOOT}}(\bar{y}) = \frac{n-1}{n}\{s_y^2 / n\} \approx s_y^2 / n$

(5) Percentile method for CI: Approximate the distribution of $\hat{\theta} - \theta$ by the known bootstrap distribution of $\theta^* - \hat{\theta}$

Suppose $B = 1000$ and let $\theta_{(1)}^* < \ldots < \theta_{(B)}^*$ denote the ordered values. Let $a = \theta_{(50)}^* - \hat{\theta}$ and $b = \theta_{(950)}^* - \hat{\theta}$.Then 90% confidence interval on θ is obtained from $a \le \hat{\theta} - \theta \le b$ which is equivalent to $\hat{\theta} - b \le \theta \le \hat{\theta} - a$.

Rao-Wu bootstrap: stratified two-stage sampling

(1) Select $m_h = n_h - 1$ clusters from the n_h sample clusters in stratum h by SRS with replacement. Let m_{hi}^* be the number of times sample cluster (hi) is selected in the bootstrap sample.

(2) Bootstrap design weights: $d_{hij}^* = d_{hij} m_{hi}^* \dfrac{n_h}{n_h - 1}$.

(3) Substitute d_{hik}^* for d_{hik} in the formula for the final weight w_{hik} to get w_{hik}^*. Use the final bootstrap weights in the estimator $\hat{\theta}$ to get θ^*; for example Y_{mw}^*.

(4) Repeat steps (1) to (3) B times to get $\theta_1^*, ..., \theta_B^*$.

(5) Bootstrap variance estimator of $\hat{\theta}$:

$$v_{\text{BOOT}}(\hat{\theta}) = B^{-1}\Sigma_b(\theta_b^* - \hat{\theta})^2$$

Note: Statistics Canada uses $B = 500$ bootstrap replicates.

Examples of On-going Surveys

Example 1: National Crime Victimization Survey (United States)

Sponsor	U. S. Bureau of Justice Statistics
Collector	U. S. Census Bureau
Purpose	Main Objectives are: • Develop detailed information about the victims and the consequences of crime • Estimate the number and types of crimes not reported to police • Provide uniform measures of selected types of crimes • Permit comparisons over time and by types of areas
Year Started	1973 (previously called the National Crime Survey, up to 1992)
Target Population	Adults and children 12 or older – civilian and non-institutionalized
Sampling Frame	U. S. households, enumerated through counties, blocks, listed addresses, lists of members of the household
Sample Design	Multi-stage, stratified, clustered area probability sample, with sample units rotating in and out of the sample over three years
Sample Size	About 41,00 households (78,600 persons)
Use of Interviewer	Interviewer administered
Mode of Administration	Face-to-face and telephone interviews
Computer Assistance	Paper questionnaire for 70% of the interviews, both face-to-face and telephone interviews (computer assisted) for 30% of the interviews
Reporting Unit	Each person age 12 or older in household reports for self
Time Dimension	Ongoing rotating panel survey of addresses
Frequency	Monthly data collection
Interview per Round of Survey	Sampled housing units are interviewed every six months over the course of three years
Levels of Observation	Victimization incident, person, household
Web Link	http://www.ojp.usdoj.gov/bjs/cvict.htm

Example 2: Quarterly Labour Force Survey (South Africa)

Collector	Statistics South Africa
Purpose	Main Objectives are: • Estimate numbers of persons employed, unemployed and economically inactive, the corresponding rates at the national, provincial and metro levels for the following estimation domains: 1) Age Groups, 2) Sex, 3) Race Categories, 4) Industry Groups, etc.
Year Started	2008
Target Population	Persons 15 – 64 years – civilian and non-institutionalized
Sampling Frame	Households in South Africa, enumerated through PSUs (Census EAs), addresses listed through field listing, lists of members of the household during interviewing
Sample Design	Two-stage, stratified, clustered area probability sample, with sampled dwelling units rotating out of the sample after four quarters
Sample Size	About 30,00 households (98,000 persons)
Use of Interviewer	Interviewer administered
Mode of Administration	Face-to-face interviews with paper questionnaire
Computer Assistance	None during data collection
Reporting Unit	Any adult in the in the household can report for all persons
Time Dimension	Ongoing rotating panel survey of dwelling units
Frequency	Monthly data collection for the survey quarter
Interview per Round of Survey	Sampled dwelling units are interviewed every quarter over the course of 12 months
Levels of Observation	Person, household
Web Link	http://www.statssa.gov.za/qlfs/index.asp