

Estimation of Parameters and Variance

Dr. A.C. Kulshreshtha

U.N. Statistical Institute for Asia and the Pacific (SIAP)

Second RAP Regional Workshop on
Building Training Resources for Improving Agricultural & Rural Statistics
Sampling Methods for Agricultural Statistics-Review of Current Practices
SCI, Tehran, Islamic Republic of Iran
10-17 September 2013

Estimation of Parameters

Survey Objectives:

- Are usually met by producing estimates of parameters of survey variable(s)
 - (Population) Mean
 - (Population) Total
 - (Population) Proportion
 - (Population) Ratio, Regression, correlation
- Which estimates are produced depends on the objectives of the survey
 - Your examples

Two aspects of sampling theory

- Sample selection through Sampling Design
- Estimation of Parameters and their Properties
 - *Efficiency*: provide estimates at lowest cost and reasonable enough precision
 - *Sampling distribution*: precision of estimators are judged by the frequency distribution generated for the estimate if the sampling procedure is applied repeatedly to the same population

Estimation of Parameters

Estimator is...

- a function (formula) of observations by which an estimate of some population characteristic (parameter), say, population mean is calculated from the sample
- a random variable and is defined on a random sample. Each random sample will yield one of its possible values

SRS- Providing Estimator

Estimator of Population Mean, \bar{Y} is $\hat{Y} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$

Where: y_i = sample response for variable y , unit i ; n is sample size

Estimator of Population Total, Y is $\hat{Y} = N \times \bar{y} = \sum_{i=1}^n \frac{N}{n} y_i = \sum_{i=1}^n w y_i$

Where, N = population size

w = base (sampling) weight for each sample unit or *inflation factor*

Sampling Distributions of Estimators

- To obtain the sampling distributions of estimators, the following probability sampling mechanism is considered:
 - It is possible to define the set of distinct samples which the sampling procedure is capable of selecting from the population
 - This further implies that it is possible to identify units belonging to different samples

Sampling Distribution of Estimators (Contd.)

- Each of the possible samples is assigned a known probability of selection
 - Out of all possible samples, a sample, selected by a random process determined by the probability of selection
- Method of computing estimate from the sample (estimator) is pre-specified

Example

- Population of $N=6$ households
- Survey variable (y) = household size
- Variance of $y = 1.667$ & $S^2 = 2$
- Population mean = 5 persons
- Population total = 30 persons
- Proportion of HHs with 6 or more members = 0.33

Table 1: Population of 6 units

HH ID	HH Size (y)
a	5
b	7
c	6
d	4
e	5
f	3

Table 2. Possible SRSWOR Samples of $n=2$

Sample No.	Sample Units	y_1	y_2
1	a,b	5	7
2	a,c	5	6
3	a,d	5	4
4	a,e	5	5
5	a,f	5	3
6	b,c	7	6
7	b,d	7	4
8	b,e	7	5
9	b,f	7	3
10	c,d	6	4
11	c,e	6	5
12	c,f	6	3
13	d,e	4	5
14	d,f	4	3
15	e,f	5	3

Show the sampling distribution of ...

- sample mean
- sample proportion
- estimator of population total

Table 3. Estimates from Samples

Sample No.	y_1	y_2	Sample Mean	Sample Proportion	Sample Total	Sampling weight	Estimate of Population Total
1	5	7	6.0	0.5	12	3	36
2	5	6	5.5	0.5	11	3	33
3	5	4	4.5	0.0	9	3	27
4	5	5	5.0	0.0	10	3	30
5	5	3	4.0	0.0	8	3	24
6	7	6	6.5	1.0	13	3	39
7	7	4	5.5	0.5	11	3	33
8	7	5	6.0	0.5	12	3	36
9	7	3	5.0	0.5	10	3	30
10	6	4	5.0	0.5	10	3	30
11	6	5	5.5	0.5	11	3	33
12	6	3	4.5	0.5	9	3	27
13	4	5	4.5	0.0	9	3	27
14	4	3	3.5	0.0	7	3	21
15	5	3	4.0	0.0	8	3	24

Sampling Distribution of \bar{y}

- Example, Table 3,

Table 4. Sampling distribution of \bar{y}

\bar{y}	frequency	probability
3.5	1	0.07
4.0	2	0.13
4.5	3	0.20
5.0	3	0.20
5.5	3	0.20
6.0	2	0.13
6.5	1	0.07
	15	1.00

Probability Sampling

- To calculate the frequency distribution of the estimator, following probability sampling mechanism is considered:
- It is possible to define the set of distinct samples , S_1, S_2, \dots, S_v (all possible samples) which the sampling procedure is capable of selecting from the population. This further implies that it is possible to identify units belonging to different samples

Probability Sampling (Contd.)

- Each of the possible sample S_i is assigned a known probability of selection, say, P_i
- Out of the all possible samples, a sample, S_i , is selected by a random process whereby each sample S_i has a probability of being selected
- The method of computing estimate from the sample is pre specified

Probability Sampling (Contd.)

- Process of generating all possible samples is laborious, particularly for large populations, thus
- Procedure usually followed is
 - Specify inclusion probability of all the units of the population
 - Select units one by one by predetermined probabilities until sample of desired size n is selected (Random Sample)
- The availability of sampling frame is a pre-requisite

Properties of Estimators

- Unbiasedness
- Precision
- Accuracy
- Consistency
- Sufficiency
- Efficiency

Basic Ideas: Unbiased Estimator

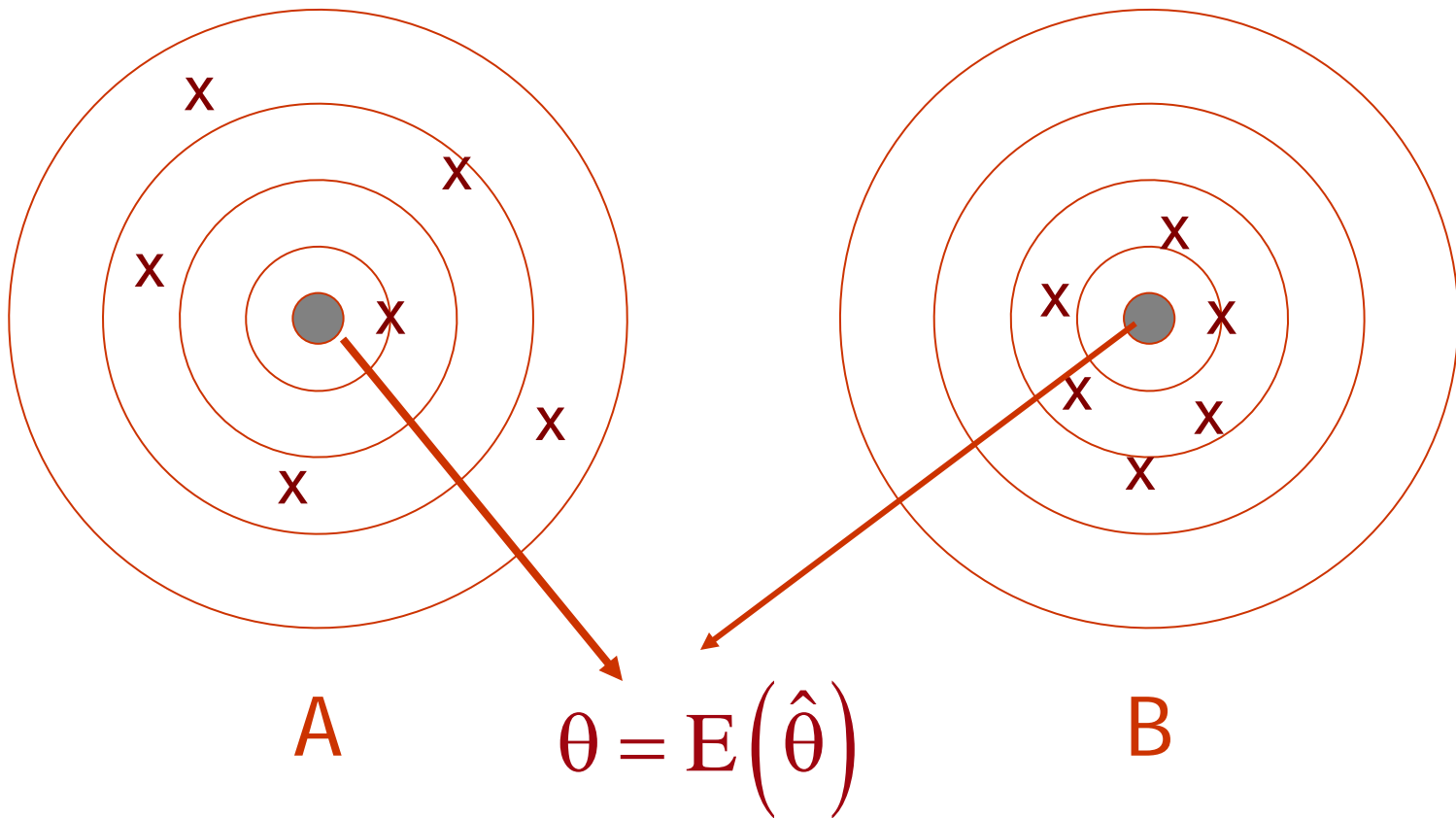
- An estimator $\hat{\theta}$ is an unbiased estimator for the parameter θ if the mean of its sampling distribution is equal to θ .

$$E(\hat{\theta}) = \theta \Rightarrow E(\hat{\theta}) - \theta = 0$$

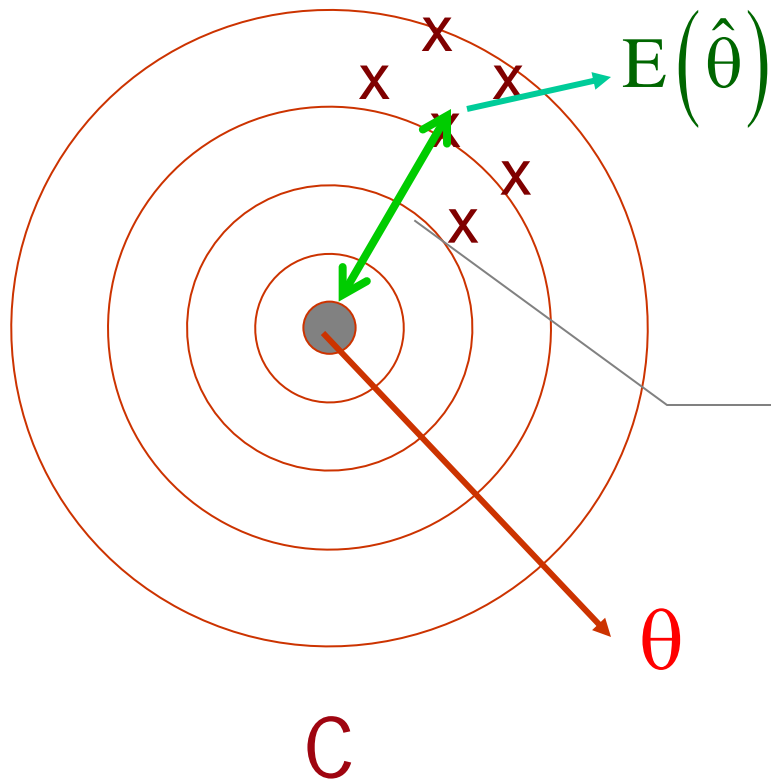
- Bias of an estimator

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Unbiased Estimators



Biased Estimators



$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Properties of Sample Mean:

Illustration

Sample No.	y_1	y_2	Sample Mean	$e = \text{Sampling Error} = \text{Sample mean} - 5$	e^2
1	5	7	6.0	1.0	1.00
2	5	6	5.5	0.5	0.25
3	5	4	4.5	-0.5	0.25
4	5	5	5.0	0.0	0.00
5	5	3	4.0	-1.0	1.00
6	7	6	6.5	1.5	2.25
7	7	4	5.5	0.5	0.25
8	7	5	6.0	1.0	1.00
9	7	3	5.0	0.0	0.00
10	6	4	5.0	0.0	0.00
11	6	5	5.5	0.5	0.25
12	6	3	4.5	-0.5	0.25
13	4	5	4.5	-0.5	0.25
14	4	3	3.5	-1.5	2.25
15	5	3	4.0	-1.0	1.00
Mean of sample mean			5.0	Mean of e^2	0.667
Standard error			0.667		

Example 1- Design-based Estimator

Sampling design SRSWR or SRSWOR

- Population parameter $\bar{Y} = \frac{1}{N} \sum_i^N Y_i$
- Sample mean $\bar{y} = \frac{1}{n} \sum_i^n y_i$ is an unbiased estimator of population mean

Example 2- Design-based Estimator

- Sampling design Stratified SRSWOR

- Population mean $\bar{Y} = \frac{1}{N} \sum_{h=1}^H \sum_{i=1}^{n_h} y_{hi}$

- Sample mean

$\bar{y}_{st} = \frac{1}{N} \sum_{h=1}^H \sum_{i=1}^{n_h} \frac{N_h}{n_h} y_{hi}$ is an unbiased estimator of population mean

(Assuming there are 'H' strata, h-th strata of size N_h and a sample of size n_h drawn from the h-th strata by SRSWOR)

Example 3- Ratio Estimator

Sampling design SRSWOR

- Population parameter $\bar{Y} = \frac{1}{N} \sum_i^N Y_i$
- Estimator $\hat{Y}_R = \bar{y}_r = \frac{\bar{y}}{\bar{x}} \bar{X}$ (ratio estimator) is a biased estimator
- Bias of the ratio estimator is $\left(\frac{N-n}{Nn} \right) \bar{Y} (C_x^2 - \rho C_x C_y)$

Sampling Error


- Sampling error of $\hat{\theta}$ is the difference between the estimate and the parameter it is estimating

$$e = \hat{\theta} - \theta$$

Variance of Unbiased Estimator

- Variance of unbiased estimator $\hat{\theta}$ $e = \hat{\theta} - \theta$

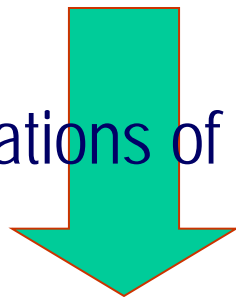
'Average' of squared deviations of all possible estimates


$$\text{var}(\hat{\theta}) = E \left[(\hat{\theta} - \theta)^2 \right]$$

Variance of Estimator, General

- Variance of estimator $\hat{\theta}$

$$\hat{\theta} - E(\hat{\theta})$$

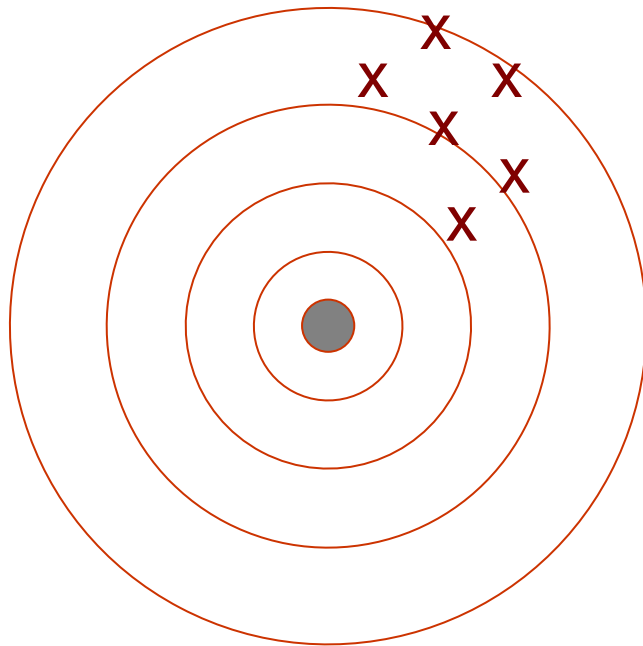


'Average' of squared deviations of estimates from their mean

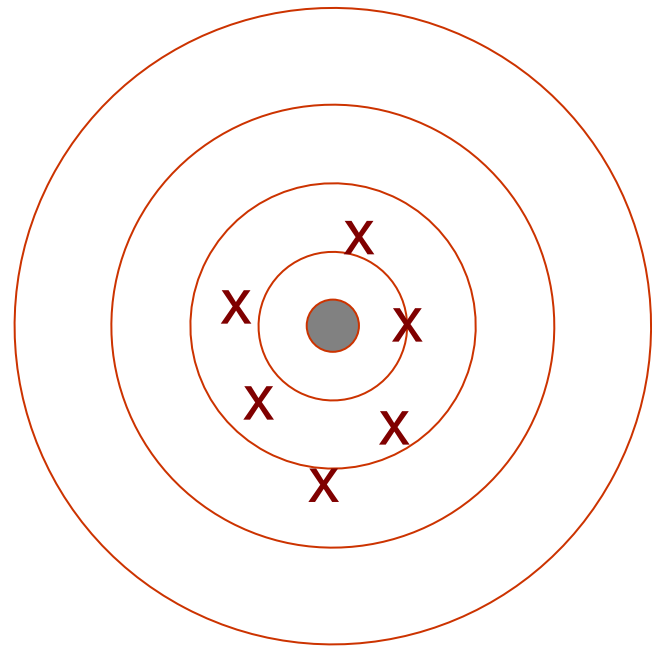
$$\text{var}(\hat{\theta}) = E \left[\left(\hat{\theta} - E(\hat{\theta}) \right)^2 \right]$$

Precise Estimators

- Variance is small of precise estimator
- Smaller the variance, more precise the estimator



C

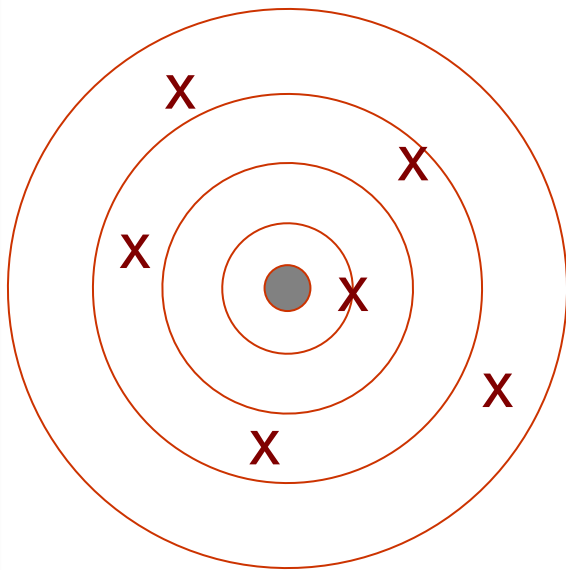


B

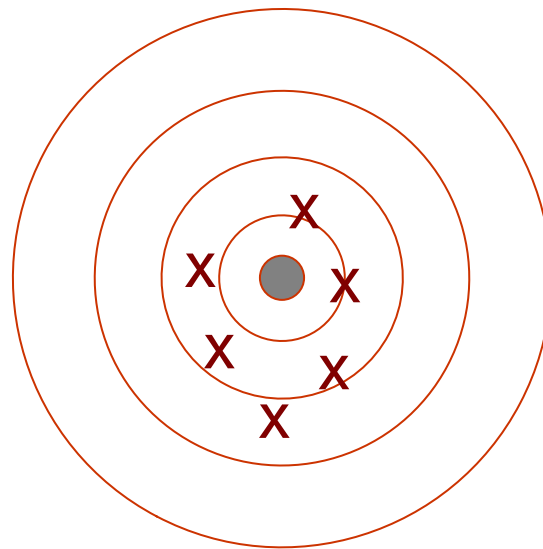
Accurate Estimator

An estimator is said to be accurate if both bias and variance are small

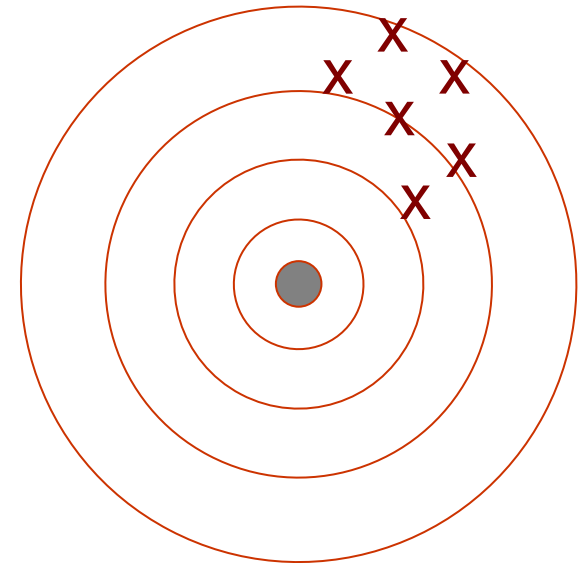
- Which estimator is the most accurate?



A



B



C

Mean Squared Error (MSE)

- Total error (simple model) =

$$\hat{\theta} - \theta = [\hat{\theta} - E(\hat{\theta})] + [E(\hat{\theta}) - \theta]$$

- Measure of accuracy is Mean Squared Error

$$= E\{[\hat{\theta} - E(\hat{\theta})]\}^2 + \{E(\hat{\theta}) - \theta\}^2$$

Variable error

Bias²

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2$$

Example: Ratio Estimator in SRSWOR

- Population parameter $\bar{Y} = \frac{1}{N} \sum_i^N Y_i$
- Estimator $\hat{Y}_R = \bar{y}_r = \frac{\bar{y}}{\bar{x}} \bar{X}$
- Bias of the ratio estimator is $\left(\frac{N-n}{Nn}\right) \bar{Y} (C_x^2 - \rho C_x C_y)$
- MSE of the ratio estimator is

$$\left(\frac{N-n}{Nn}\right) \bar{Y}^2 (C_y^2 + C_x^2 - 2\rho C_x C_y) + \left[\left(\frac{N-n}{Nn}\right) \bar{Y} (C_x^2 - \rho C_x C_y)\right]^2$$

Efficiency

- Given two estimators of the population parameter, one estimator is said to be more efficient than the other if its mean square error is less than that of the other

- Measure of efficiency =
$$\frac{MSE(\hat{\theta}_1)}{MSE(\hat{\theta}_2)}$$

Consistency

- An estimator is said to be a consistent estimator if its value approaches parameter, statistically
- the probability of the difference $\hat{\mu} - \mu$ being less than any specified small quantity tends to unity as n is increased
- Also, when n is increased to 'N' the estimator attains the value of the parameter

Example 1

- Sampling design SRSWR
- Population parameter $\bar{Y} = \frac{1}{N} \sum_i^N Y_i$
- Sample mean $\bar{y} = \frac{1}{n} \sum_i^n y_i$ is a consistent estimator of the population mean

Example 2

- Sampling design SRSWOR
- Population parameter $\bar{Y} = \frac{1}{N} \sum_i^N Y_i$
- Sample mean $\bar{y} = \frac{1}{n} \sum_i^n y_i$ is a consistent estimator of the population mean

Example 3

- Sampling design SRSWOR
- Population parameter $\bar{Y} = \frac{1}{N} \sum_i^N Y_i$
- $\frac{\bar{y}}{\bar{x}} \bar{X}$ Ratio estimator is a consistent estimator of population mean

Confidence interval

- Large sample sizes
- Sampling distribution of estimates is normally distributed
- It is possible to construct a confidence interval for the parameter of interest

Example

- Sampling design SRSWR

- Population parameter $\bar{Y} = \frac{1}{N} \sum_i^N Y_i$

- Sample mean $\bar{y} = \frac{1}{n} \sum_i^n y_i$

- 5% CI $\bar{Y} \pm 1.96 \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S^2}$

- 1% CI $\bar{Y} \pm 2.58 \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S^2}$

Sufficiency

- Non Completeness of sample mean
- Non existence of UMVUE (Uni Min Var Unbiased Estimator)
- Involvement of main stream statisticians to the problem of finite population sampling
- Frame work for finite population inference
- Admissibility and hyper admissibility

Other approaches

- Likelihood function approach
- Model based approach
- Robustness aspect
- Model assisted approach
- Use of models but inferences are design based
- Conditional design based approach

Estimation of Variance

Variance Estimation in Complex Surveys

- Linearization (Taylor's series)
- Random Group Methods
- Balanced Repeated Replication (BRR)
- Re-sampling techniques
 - Jackknife, Bootstrap

Taylor's Series Linearization Method

- Non-linear statistics are approximated to linear form using Taylor's series expansion
- Variance of the first-order or linear part of the Taylor series expansion retained
- This method requires the assumption that all higher-order terms are of negligible size
- We are already familiar with this approach in a simple form in case of ratio estimator

Random Group Methods

- Concept of replicating the survey design
- Interpenetrating samples
- Survey can also be divided into R groups so that each group forms a miniature version of the survey
- Based on each of the R groups estimates can be developed for the parameter θ of interest
- Let $\hat{\theta}_r$ be the estimate based on rth sample

$$\hat{V}(\hat{\theta}) = \frac{1}{R(R-1)} \sum_{r=1}^R (\hat{\theta}_r - \hat{\theta})^2$$

BRR method

- Consider that there are H strata with two units selected per stratum
- There are 2^H ways to pick 1 from each stratum
- Each combination could be treated as a sample
- Pick R samples
- Which samples should we include?

BRR method (Contd.)

- Assign each value either 1 or -1 within the stratum
- Select samples that are orthogonal to one another to create balance
- One can use the design matrix for a fraction factorial
- Specify a vector of 1, -1 values for each stratum

BRR method (Contd.)

- An estimator of variance based on BRR method is given by

$$\hat{V}_{BRR}(\hat{\theta}) = \frac{1}{R(R-1)} \sum_{r=1}^R \left(\hat{\theta}(\alpha_r) - \hat{\theta} \right)^2$$

where
$$\hat{\theta} = \frac{1}{R} \sum_{r=1}^R \hat{\theta}(\alpha_r)$$

Jack-knife Method

- Let $\hat{\theta}^i$ be the estimator of θ after omitting the i^{th} observation. Define

$$\tilde{\theta}^i = n\hat{\theta} - (n-1)\hat{\theta}^i$$

$$\hat{\theta}_J = \frac{1}{n} \sum \tilde{\theta}^i$$

$$\hat{V}_J(\hat{\theta}_J) = \frac{1}{n(n-1)} \sum_{i=1}^n (\tilde{\theta}^i - \hat{\theta}_J)^2$$

Soft-wares for Variance estimation

- OSIRIS – BRR, Jackknife
- SAS – Linearization
- STATA – Linearization
- SUDAAN – Linearization, Bootstrap, Jackknife
- WesVar – BRR, Jackknife, Bootstrap

THANKS