

quantitative initiative for policy and social research

## **Applied Survey Data Analysis**

## Module 2: Variance Estimation March 30, 2013

## Approaches to Complex Sample Variance Estimation

- In simple random samples many estimators are linear estimators where the sample size n is fixed. A linear estimator is a linear function of the sample observations.
- When survey data are collected using a complex design with unequal size of clusters or when weights are used in estimation, most statistics of interest will not be simple linear function of the observed data.

$$\overline{y}_{w} = \frac{\sum_{h} \sum_{\alpha} \sum_{i} w_{h\alpha i} y_{h\alpha i}}{\sum_{h} \sum_{\alpha} \sum_{i} w_{h\alpha i}} = \frac{\sum_{h} \sum_{\alpha} \sum_{i} u_{h\alpha i}}{\sum_{h} \sum_{\alpha} \sum_{i} v_{h\alpha i}} = \frac{u}{v}$$

Where:

 $y_{h \propto i}$  =Measurement on unit *i* in cluster  $\alpha$  in stratum *h*  $w_{h \propto i}$  =Corresponding weight



qipsr

# Two approaches

- Replication or Resampling Methods technique: -Jackknife Repeated Replication
  Balanced Repeated Replication
- Taylor Series approximation or linearization technique:

-Approximate nonlinear statistics as a linear function of sample totals

-Specific form of the variance estimator for each statistic





# **Replication Methods**

- Replication methods use information on variability between estimates drawn from different subsamples of an overall sample to make inferences about variance in the population
- Steps in Replicated Methods:
  - 1. A defined number (K) of subsets (replicate samples) of the full sample are selected.
  - 2. Create revised weight for replicate sample.
  - 3. Compute weighted estimates of population statistic of interest for each replicate using replicate weight.
  - 4. The variability between these subsample replicate statistics are used to estimate the variance of the full sample statistic.





#### Jackknife Repeated Replication (JRR)

- The Jackknife Repeated Replication (JRR) is applicable to a wide range of complex sample designs including designs in which two or more PSUs are selected from each of h=1,...,H primary stage strata.
- Using Jackknife for unstratified surveys, one PSU at a time is omitted from the sample and the others reweighted to keep the same total weight (known as the JK1 Jackknife).
- For stratified designs, Jackknife removes one PSU at a time, but reweights only the other PSUs in the same stratum.





## JRR: Constructing Replicates, Replicate Weights

- Suppose there are H strata with a<sub>h</sub> clusters.
- Each replicate is constructed by deleting one or more PSUS from a single stratum
- Replicate weight values for cases in the deleted PSUs are assigned a value of "0" or " missing"
- The replicate weight for each replicate multiplies the weights for remaining cases in the deleted stratum by a factor of a<sub>h</sub>/[a<sub>h</sub> -1].
- Replicate 1 weight values remain unchanged for cases in all other strata.





### JRR: Constructing Replicates, Replicate Weights (2)

 Each Stratum will contribute a<sub>h</sub>-1 unique JRR replicates, yielding a total of

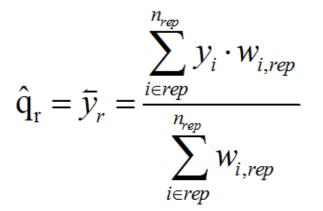
$$R = \sum_{h=1}^{H} (a_h - 1) = a - H = \# clusters - \# strata$$





### **JRR: Constructing Estimates**

• The weighted estimate for each of r=1,...R replicates:



• The full sample estimate of the mean is:

$$\hat{\mathbf{q}} = \frac{\sum_{i=1}^{n} y_i \cdot w_i}{\sum_{i=1}^{n} w_i}$$



### **JRR: Estimating the Sampling Variance**

$$var_{JRR}(\hat{q}) = \sum_{r=1}^{\infty} (\hat{q}_r - \hat{q})^2$$





## **Balanced Repeated Replication(BRR)**

- Balanced Repeated Replication (BRR) is a half-sample method that was developed specifically for estimating sampling variances under two PSU- per-stratum sample designs.
- A half sample is defined by choosing one PSU from each stratum.
- A complement of a half sample is made up of all those PSUs not in the half sample. A complement is also a half sample.
- There are 2<sup>H</sup> possible half samples and their complements. We only need H half samples for variance estimation.





#### Hadamard Matrix for a H=4 strata design

BRR Replicate		Stratum			
	1	2	3	4	
1	+	+	+	-	
2	+	-	-	-	
3	-	-	+	-	
4	-	+	-	-	





### BRR: Constructing Replicates and Replicate Weights

- H replicates are created based on the deletion pattern ( + and -) in the Hadamard matrix
- Replicate weight is then created for each of the h=1,..., H BRR sample replicates.
- Replicate weight values for cases in the complement half-sample PSUs are assigned a value of "0" or " missing"
- Replicate weight values for the cases in the PSUs retained in the half-sample are formed by multiplying the full sample analysis weights by a factor of 2.





### **BRR: Constructing Estimates**

• The weighted estimate for each of r=1,...R replicates:

$$\hat{\mathbf{q}}_{\mathbf{r}} = \bar{y}_{r} = \frac{\sum_{i \in rep}^{n_{rep}} y_{i} \cdot w_{i,rep}}{\sum_{i \in rep}^{n_{rep}} w_{i,rep}}$$

• The full sample estimate is:

$$\hat{\mathbf{q}} = \frac{\sum_{i=1}^{n} y_i \cdot w_i}{\sum_{i=1}^{n} w_i}$$



### **BRR: Estimating the Sampling Variance**

$$var_{BRR}(\bar{y}_w) = var_{BRR}(\hat{q}) = \frac{1}{R} \sum_{r=1}^{R} (\hat{q}_r - \hat{q})^2$$





#### **Balanced Repeated Replication**

- Pro
  - Relatively few computations
  - Asymptotically equivalent to linearization methods for smooth functions of population totals and quantiles
- Con
  - 2 psu per stratum





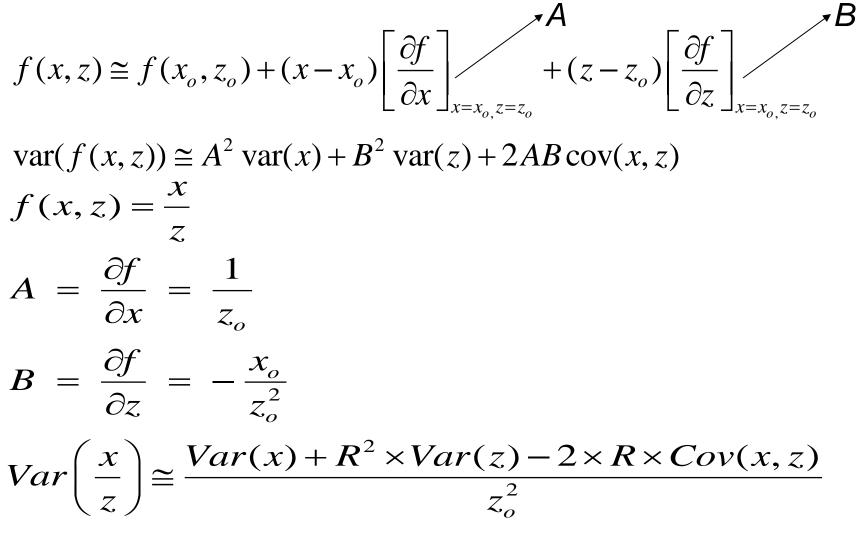
## **Linearization (Taylor Series Method)**

- Linearization techniques make mathematical adjustments so that standard 'linear estimators' can be applied to data.
- Linearization is a widely used technique for estimating variance of any functions of the weighted totals. These include ratios, subgroup differences in the ratios, regression coefficients and correlation coefficients.





#### **Taylor Series Linearization**





## The estimates under TSL

• Consider the weighted estimates of the population mean of variable y

$$\overline{y}_{w} = \frac{\sum_{h} \sum_{\alpha} \sum_{i} w_{h\alpha i} y_{h\alpha i}}{\sum_{h} \sum_{\alpha} \sum_{i} w_{h\alpha i}} = \frac{\sum_{h} \sum_{\alpha} \sum_{i} u_{h\alpha i}}{\sum_{h} \sum_{\alpha} \sum_{i} v_{h\alpha i}} = \frac{u_{i}}{v_{i}}$$

• Rewriting it as a linear combination of weighted sample totals using TSL

$$\bar{y}_{w,TSL} = \frac{u_0}{v_0} + (u - u_0) \left[ \frac{\partial \bar{y}_{w,TSL}}{\partial u} \right]_{u = u_0, v = v_0} + (v - v_0) \left[ \frac{\partial \bar{y}_{w,TSL}}{\partial v} \right]_{v = v_0, u = u_0} + remainder$$

$$\bar{y}_{w,TSL} \cong \frac{u_0}{v_0} + (u - u_0) \left[ \frac{\partial \bar{y}_{w,TSL}}{\partial u} \right]_{u = u_0, v = v_0} + (v - v_0) \left[ \frac{\partial \bar{y}_{w,TSL}}{\partial v} \right]_{v = v_0, u = u_0}$$

$$\overline{y}_{w,TSL} \cong constant + (u - u_0) \cdot A + (v - v_0) \cdot B$$

Where

$$A = \frac{\partial \bar{y}_{w,TSL}}{\partial u} \bigg|_{u=u_0, v=v_0} = \frac{1}{v_0}; B = \frac{\partial \bar{y}_{w,TSL}}{\partial v} \bigg|_{u=u_0, v=v_0} = -\frac{u_0}{v_0^2};$$



### The Variance under TSL

• The approximate variance of the "linearized' form of the estimate  $\bar{y}_{w,TSL}$ 

$$var(\bar{y}_{w,TSL}) \cong var[constant + (u - u_0) \cdot A + (v - v_0) \cdot B]$$
  
$$\cong 0 + A^2 var(u - u_0) + B^2 var(v - v_0) + 2ABcov(u - u_0, v - v_0)$$
  
$$\cong A^2 var(u) + B^2 var(v) + 2ABcov(u, v)$$

Where:

Applied

$$A = \frac{\partial \bar{y}_{w,TSL}}{\partial u} \bigg|_{u=u_0, v=v_0} = \frac{1}{v_0}; B = \frac{\partial \bar{y}_{w,TSL}}{\partial v} \bigg|_{u=u_0, v=v_0} = -\frac{u_0}{v_0^2}; and$$

 $u_0, v_0$  are the weighted sample totals computed from the survey data.

• Therefore, the sampling variance of the nonlinear estimate  $\bar{y}_{w,TSL}$  is approximated by a simple algebraic function of quantities that can be readily computed from the complex sample survey data.

$$var(\bar{y}_{w,TSL}) \cong \frac{var(u) + \bar{y}_{w,TSL}^2 \cdot var(v) - 2 \cdot \bar{y}_{w,TSL} \cdot cov(u,v)}{v_0^2}$$

# Linearization (Taylor Series Methods)

- Pro:
  - Linearization technique is useful if the estimate can be expressed as a function of sample totals
  - Theory is well developed
  - The default is most software package for complex samples
- Con:
  - Finding partial derivatives may be difficult
  - Different method is needed for each statistic
  - The function of interest may not be expressed a smooth function of population totals or means
  - Accuracy of the linearization approximation



