Survey Sampling: Resampling Methods

Random Group Method:

- The basic survey design is replicated independently R times
- After each sample is drawn, the sampled units are replaced in the population so they are available for later samples
- R replicate samples produce R independent estimates of the quantity of interest
- The variability among those estimates can be used to estimate the variance of $\hat{\theta}$

Let

θ = the parameter of interest

$\hat{\theta}_r$ = estimate of θ calculated from r^{th} replicate

$\tilde{\theta} = \sum_{r=1}^{R} \hat{\theta}_r / R$

$$\hat{V}_1(\tilde{\theta}) = \frac{1}{R} \frac{\sum_{r=1}^{R} (\hat{\theta}_r - \tilde{\theta})^2}{R-1}$$

Dividing the sample into random groups

- Divide the complete sample into R groups, so that each group forms a miniature version of the survey, mirroring the sample design
- The groups are then treated as though they are independent replicates of the basic survey design
- SRS of size n, the groups are formed by randomly apportioning the n observations into R groups, each of size n/R
- If the population size is large relative to the sample size, the groups can be treated as though they are independent repli-

cates

- Cluster sample, the psu's are randomly divided among the R groups. The psu takes all its observation units with it to the random group, so that each random group is still a cluster sample

- Stratified multistage sample, a random group contains a sample of psu's from each stratum. If k psu's are sampled in the smallest stratum, at most k random groups can be formed

Estimator often used,

$$\hat{V}_2(\hat{\theta}) = \frac{1}{R} \frac{\sum_{r=1}^{R} (\hat{\theta}_r - \hat{\theta})^2}{R - 1}$$

Balanced repeated replication (BRR)

Some surveys are stratified to the point that only two psu’s are selected from each stratum. This gives the highest of stratification possible while still allowing calculation of variance estimates in each stratum.

Example:

- Suppose an SRS of two observation units is chosen from each of seven strata
- Arbitrarily label one of the sampled units in stratum h as y_{h1}, and the other as y_{h2}

- If using random group method, randomly select one of the observations in each stratum for group 1, and assign the other to group 2. The groups in this situation are half-samples

- The random group estimate of the variance in this case has only one degree of freedom for a two-psu-per-stratum design and is unstable in practice

- McCarthy (1966, 1969), balanced repeated replication methods

$$\boldsymbol{\alpha}_r = (\alpha_{r1}, \cdots, \alpha_{rH})$$

$$y_{h(\boldsymbol{\alpha}_r)} = \begin{cases} y_{h1} & \text{if} a_{rh} = 1 \\ y_{h2} & \text{if} a_{rh} = -1. \end{cases}$$

or equivalently,

$$y_{h(\boldsymbol{\alpha}_r)} = \frac{\alpha_{rh}+1}{2} y_{h1} - \frac{\alpha_{rh}-1}{2} y_{h2}$$

If $\boldsymbol{\alpha}_1 = (1, -1, -1, -1, 1, -1, 1)$, then group 1 contains observations $\{y_{11}, y_{22}, y_{32}, y_{42}, y_{51}, y_{62}, y_{71}\}$

The set of R replicate half-samples is balanced if

$\sum_{r=1}^{R} \alpha_{rh}\alpha_{rl} = 0$ for all $l \neq h$

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

	s1	s2	s3	s4	s5	s6	s7
α_1	-1	-1	-1	1	1	1	-1
α_2	1	-1	-1	-1	-1	1	1
α_3	-1	1	-1	-1	1	-1	1
α_4	1	1	-1	1	-1	-1	-1
α_5	-1	-1	1	1	-1	-1	1
α_6	1	-1	1	-1	1	-1	-1
α_7	-1	1	1	-1	-1	1	-1
α_8	1	1	1	1	1	1	1

If the set of half-sample is balanced, then

$$\hat{V}_{BRR}(\bar{y}_{str}) = \hat{V}_{str}(\bar{y}_{str})$$

proof:

$$\begin{aligned}
\hat{V}_{BRR}(\bar{y}_{str}) &= \frac{1}{R}\sum_{i=1}^{R}(\bar{y}_{(\alpha_r)} - \bar{y})^2 \\
&= \frac{1}{R}\sum_{r=1}^{R}(\sum_{h=1}^{H}\frac{N_h}{N}[\frac{\alpha_{rh}}{2}(y_{h1} - y_{h2})])^2
\end{aligned}$$

$$
\begin{aligned}
&\bar{y}_{str}(\alpha_i) - \bar{y}_{str} \\
&\qquad = \sum_{h=1}^{H} [\frac{N_h}{N} y_h(\alpha_i) - \frac{N_h}{N} \bar{y}_h] \\
&\qquad = \sum_{h=1}^{H} \frac{N_h}{N} [(\frac{\alpha_{ih}+1}{2} y_{h1} \\
&\qquad - \frac{\alpha_{ih}-1}{2} y_{h2}) - \frac{y_{h1}+y_{h2}}{2}] \\
&\qquad = \sum_{h=1}^{H} \frac{N_h}{N} [\frac{\alpha_{ih}}{2} (y_{h1} - y_{h2})]
\end{aligned}
$$

So

$$\begin{aligned}\hat{V}_{BRR}(\bar{y}_{str}) &= \frac{1}{R}\sum_{r=1}^{R}[\sum_{h=1}^{H}\frac{N_h^2}{N^2}\frac{(y_{h1}-y_{h2})^2}{4}\alpha_{rh}^2 \\ &+ \sum_{i\neq j}2\frac{N_i}{N}(\frac{\alpha_{ri}}{2}(y_{i1}-y_{i2}))\frac{N_j}{N}\frac{\alpha_{rj}}{2}(y_{j1}y_{j2}) \\ &= \sum_{h=1}^{H}\frac{N_h^2}{N^2}\frac{(y_{h1}-y_{h2})^2}{4}\end{aligned}$$

Delete 1 Jackknife

- The Jackknife was introduced by (Quenouille (1964)) as a method of reducing bias; Tukey (1958) used it to estimate variances and calculate confidence intervals

- Extends the random group method by allowing the replicate groups to overlap

- SRS

  let $\hat{\theta}_{(j)}$ be the estimator of the same form as $\hat{\theta}$, but not using observation j

Example: Let $\hat{\theta} = \bar{y}$, then $\hat{\theta}_{(j)} = \bar{y}_{(j)} = \sum_{i \neq j} y_i/(n-1)$.

$$\hat{V}_{JK}(\hat{\theta}) = \frac{n-1}{n} \sum_{j=1}^{n} (\hat{\theta}_{(j)} - \hat{\theta})^2$$

- Stratified multistage cluster sample, the jackknife can be applied separately in each stratum at the first stage of sampling, with one psu deleted at a time

  Suppose there are H strata, from each stratum h, n_h psu's are sampled. Define a new weight variable:

$$w_{i(hj)} = \begin{cases} w_i, & \text{unit } i \text{ is not in stratum } h, \\ 0, & \text{unit } i \text{ is in psu } j \text{ of stratum } h, \\ \dfrac{n_h}{n_h - 1} w_i, & \text{unit } i \text{ is in stratum } h \text{ but not in psu } j. \end{cases}$$

Then use the weights $w_{i(hj)}$ to calculate $\hat{\theta}_{(hj)}$, the jackknife estimator is defined as follows:

$$\hat{V}_{JK}(\hat{\theta}) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} (\hat{\theta}_{(hj)} - \hat{\theta})^2.$$

Advantages:

- The jackknife is an all-purpose method
- Works in stratified multistage samples in which BRR does not apply because more than two psu’s are sampled in each stratum
- Provides a consistent estimator of the variance when θ is a smooth function of population totals

Disadvantages:

- The jackknife may require a large amount of computation
- Performs poorly for estimating the variances of some statistics, such as the variance of quantiles in an SRS
- Little is known about how the jackknife performs in unequal probability, without replacement sampling designs in general

Bootstrap: The bootstrap was one of the first computer-intensive statistical techniques, replacing traditional algebraic derivations with data-based computer simulations. Bootstrap has major impact in the field of statistics and virtually every area of statistical application.

- A computer-based technique for estimating standard errors, biases, confidence intervals and other measures of statistical accuracy
- SRS with replacement, developed by Efron (1979, 1982)
- Suppose S is an SRS of size n

- Treat the sample S as if it were a population, and take resamples from S
- Assume that resamples reproduces properties of the whole population

Example 9.8 (Lohr's book)

- Use the bootstrap to estimate the variance of the median height θ using the sample in file htsrs
- population median (hrpop) is $\theta = 168$
- Sample median is $\hat{\theta} = 169$
- The histogram of the population is similar in shape as the histogram of the sample, so we would expect that taking an SRS of size n with replacement from S would be like taking an SRS with replacement from the population
- Take an SRS of size 200 from S to form the first sample. Re-

peating the process to take a total of $R = 2000$ resamples from S and calculate the sample median from each sample

- The sample variance of these 2000 values is the bootstrap estimator of the variance.
- A 95% CI is calculated by finding the 2.5 percentile and the 97.5 percentile of the bootstrap distribution

Rescaling bootstrap (J.N.K. Rao and Wu (1988))

- For each stratum, draw and SRS of size $n_h - 1$ with replacement from the sample in stratum h. Do this independently for each stratum

- Create a new weight variable for each observation in the sample

$$w_i(r) = w_i \times \frac{n_h}{n_h - 1} m_i(r)$$

where $m_i(r)$ is the number of times that observation i is selected to be in the resample. Calculate $\hat{\theta}_r^*$ (the estimate of θ from the rth bootstrap sample, using the weight $w_i(r)$

- Repeat steps 1 and 2 R times, for R a large number
- Calculate

$$\hat{V}_{BS}(\hat{\theta}) = \frac{1}{R-1}\sum_{i=1}^{R}(\hat{\theta}_r^* - \hat{\theta})^2$$