Stat472/572 Sampling: Theory and Practice

Instructor: Yan Lu

University of New Mexico

Chapter 9: Variance Estimation in Complex Surveys

Topics

- Linearization (Taylor Series) Methods
- Random Group Methods
- Resampling and Replication Methods
- Generalized variance functions (GVF)

Linearization for estimating variances

Most of the variance formulas in Chapters 2 through 6 were for estimators of means and totals. Those formulas can be used to find variances for any linear combination of estimated means and totals.

$$V\left(\sum_{i=1}^{k} a_j \hat{t}_j\right) = \sum_{j=1}^{k} a_j^2 V(\hat{t}_j) + 2\sum_{j=1}^{k-1}\sum_{l=j+1}^{k} a_j a_l \text{Cov}(\hat{t}_j, \hat{t}_l)$$

Suppose we are interested in a population quantity θ that is a function of population means or totals

- In general, let θ be a parameter of interest,

$$\theta = h(t_1, t_2, \cdots, t_k), \quad \hat{\theta} = h(\hat{t}_1, \hat{t}_2, \cdots, \hat{t}_k)$$

- Taylor's theorem allows us to linearize a smooth nonlinear function $h(t_1, t_2, \cdots, t_k)$ of the population totals
  By first-order Taylor expansion
  $\hat{\theta} - \theta \approx \sum_{j=1}^{k} \frac{\partial h}{\partial t_j}(\hat{t}_j - t_j)$

Example: Ratio estimator $\theta = B = \frac{t_y}{t_x}, \quad \hat{\theta} = \hat{B} = \frac{\hat{t}_y}{\hat{t}_x}$

$$h(x,y) = \frac{y}{x}, \quad \frac{\partial h}{\partial x} = \frac{-y}{x^2}, \quad \frac{\partial h}{\partial y} = \frac{1}{x}$$

$$h(t_x,t_y) = \frac{t_y}{t_x}, \quad \frac{\partial h}{\partial t_x} = \frac{-t_y}{t_x^2}, \quad \frac{\partial h}{\partial t_y} = \frac{1}{t_x}$$

$$\begin{aligned}
\hat{B} - B &= h(\hat{t}_x,\hat{t}_y) - h(t_x,t_y) \\
&\approx \frac{\partial h}{\partial x}|_{(t_x,t_y)} \cdot (\hat{t}_x - t_x) + \frac{\partial h}{\partial y}|_{(t_x,t_y)} \cdot (\hat{t}_y - t_y) \\
&= -\frac{t_y}{t_x^2}(\hat{t}_x - t_x) + \frac{1}{t_x}(\hat{t}_y - t_y) \\
&= -\frac{1}{t_x}[B(\hat{t}_x - t_x) - (\hat{t}_y - t_y)]
\end{aligned}$$

$$\begin{aligned}\hat{B} - B &\approx -\frac{1}{t_x}[B(\hat{t}_x - t_x) - (\hat{t}_y - t_y)] \\ &= \frac{1}{t_x}[\hat{t}_y - t_y - B(\hat{t}_x - t_x)] \\ &= \frac{1}{t_x}[\hat{t}_y - B\hat{t}_x - (t_y - Bt_x)]\end{aligned}$$

Let $z_i = y_i - Bx_i$,

$$\begin{aligned}\hat{t}_z &= \sum_{i \in S} w_i z_i = \sum_{i \in S} w_i (y_i - Bx_i) \\ &= \hat{t}_y - B\hat{t}_x\end{aligned}$$

$$\hat{B} - B \approx \frac{1}{t_x}[\hat{t}_z - t_z]$$

$$V_L(\hat{B}) \approx V\left(\frac{\hat{t}_z}{t_x}\right) = \frac{1}{t_x^2} V(\hat{t}_z)$$

$z_i = y_i - Bx_i$, $e_i = y_i - \hat{B}x_i$

Linearization variance

$$\begin{aligned} \hat{V}_L(\hat{B}) &\approx \hat{V}(\hat{t}_e)/\hat{t}_x^2 \\ &= \frac{N^2\left(1-\frac{n}{N}\right)\frac{s_e^2}{n}}{N^2\bar{x}^2} \\ &= \left(1-\frac{n}{N}\right)\frac{s_e^2}{n\bar{x}^2} \end{aligned}$$

Recall Chapter 4, SRS

$$\hat{B} = \bar{y}/\bar{x} = \hat{t}_y/\hat{t}_x$$

$$\hat{V}(\hat{B}) = \left(1-\frac{n}{N}\right)\frac{s_e^2}{n\bar{x}^2}$$

Advantages:

- If the partial derivatives are known, linearization almost always gives a variance estimate for a statistic and can be applied in general sampling designs
- Theory and softwares are well developed for linearization

Disadvantages:

- Calculations can be messy, and the method is difficult to apply for complex functions involving weights
- Not all statistics can be expressed as a smooth function of the population totals, such as median and other quantiles
- Need large sample size for the accuracy of the linearization approximation

Random Group Methods

1. **Replicating the Survey Design**

- the basic survey design is replicated independently for R times. —- Independently means that each of the R sets of random variables used to select the sample is independent of the other sets after each sample is drawn, the sampled units are replaced in the population so they are available for later samples.
- the R replicate samples produce R independent estimators of the quantity of interest; the variability among those estimates can be used to estimate the variance of $\hat{\theta}$.

Let θ represent the parameter of interest.
Define:

- $\hat{\theta}_r$: the estimator of θ from the rth replicate.
- $\tilde{\theta} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r$: the average of the replicate estimators.
- $\hat{V}_1(\tilde{\theta}) = \frac{1}{R}\cdot\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\theta}_r - \tilde{\theta})^2$: an estimator of the variance of $\tilde{\theta}$.

If $\hat{\theta}_r$ is an unbiased estimator of θ, then $\tilde{\theta}$ is also an unbiased estimator of θ, and $\hat{V}_1(\tilde{\theta})$ is an unbiased estimator of $V(\tilde{\theta})$.

Example 9.3.

- Objective: Estimate the ratio of out-of-state tuition to in-state tuition for public colleges and universities in the United States, using data from `college.csv` (see Example 3.12 for details).
- Method: Implement the random group method by selecting five simple random samples (SRSs) of size 10 each.
  —- The SRSs are drawn without replacement, but the same college may appear in more than one sample.
  —- Data for this example are in `collegerg.csv`.

Table 1: Summary Statistics for Five SRSs of Colleges, Used in Example 9.3. $\hat{\theta}_i =$ average of nonresident tuitions for sample i/average of resident tuitions for sample i

Replicate sample r	Average In-state tuition $\bar{x}_r$	Average Out-of-state tuition $\bar{y}_r$	$\hat{\theta}_r$
1	8913.3	21614.7	2.4250
2	9542.0	21497.5	2.2529
3	10210.6	21323.4	2.0884
4	9004.7	18469.0	2.0510
5	9467.1	22844.0	2.4130

- The sample average of the five independent estimates is

$$\tilde{\theta} = \sum_{i=1}^{5} \hat{\theta}_r/5 = 2.246.$$

- The variability across the estimates i,e., the sample standard deviation of the five estimates is 0.175, so the standard error of $\tilde{\theta}$ is

$$\sqrt{\hat{V}_1(\tilde{\theta})} = 0.175/\sqrt{5} = 0.0784.$$

- A 95% CI for the ratio is

$$2.246 \pm 2.78 * (0.0784) = [2.03, 2.46],$$

where where 2.78 is the t critical value with 4 degrees of freedom (df).

2. Dividing the Sample into Random Groups

- In practice, the complete sample is selected according to the survey design.
- The complete sample is then divided into R groups, so that each group forms a miniature version of the survey, mirroring the sample design.
  —-If the sample is an SRS of size n, the groups are formed by randomly apportioning the n observations into R groups, each of size n/R.
  —-In a cluster sample, the psus are randomly divided among the R groups. The psu takes all its observation units with it to the random group, so each random group is still a cluster sample.
  —-In a stratified multistage sample, a random group contains a sample of psus from each stratum. Note that if k psus are sampled in the smallest stratum, at most k random groups can be formed.
  —-The groups are then treated as though they are independent replicates of the basic survey design.

- These pseudo-random groups are not quite independent replicates because an observation unit can only appear in one of the groups; if the population size is large relative to the sample size, however, the groups can be treated as though they are independent replicates.

Definitions and Variance Estimation

- Let θ be the parameter of interest.
- $\hat{\theta}_r$: Estimator of θ calculated from the rth random group.
- $\tilde{\theta} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r$: Average estimate across the R random groups.
- $\hat{\theta}$: Estimate of θ calculated from the complete sample.
  — Typically, $\hat{\theta}$ is a more stable estimator than $\tilde{\theta}$.
- Variance estimators:

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

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

  — $\hat{V}_2(\tilde{\theta})$ is slightly larger but is often preferred.

Example 9.4: The 1987 Survey of Youths in Custody was divided into seven random groups.

- The survey design had 16 strata. Strata 6-16 each consisted of one facility (=psu), and these facilities were sampled with probability one. In strata 1-5, facilities were selected with probability proportional to number of residents in the 1985 Children in Custody census.

Table 2: Survey of Youth in Custody Stratum Information

Stratum	CIC size number of residents	Number of psu's in frame	Number of residents in CIC	Number of eligible psu's in sample
1	1-59	99	2881	11
2	60-119	39	3525	7
3	120-179	30	4355	7
4	180-239	13	2594	7
5	240-359	14	4129	7

- Seven random groups were formed because strata 2 through 5 each have seven psus.
- For each self-representing facility in strata 6-16, random group numbers were assigned as follows: The first resident selected from the facility was assigned a number between 1 and 7. Let's say the first resident was assigned number 6. Then the second resident in that facility would be assigned number 7, the third resident 1, the fourth resident 2, and so on.

- In strata 1-5, all residents in a facility (psu) were assigned to the same random group. Thus for the seven facilities sampled in stratum 2, all residents in facility 33 were assigned random group number 1, all residents in facility 9 were assigned random group number 2, and so on.
- After all random group assignments were made, each random group had the same basic design as the original sample.
  —- Random group 1, for example, forms a stratified sample in which a (roughly) random sample of residents is taken from the self representing facilities in strata 6-16, and an unequal-probability sample of facilities is taken from each of strata 1-5.

Table 3: Estimates of mean age of residents for each random group, $\hat{\theta}_r = \sum w_i y_i / \sum w_i$, where w_i is the final weight for resident i, and the summations are over observations in random group r.

Random group	Estimate of mean age $\hat{\theta}_r$
1	16.55
2	16.66
3	16.83
4	16.06
5	16.32
6	17.03
7	17.27

- $\tilde{\theta} = \frac{1}{7}\sum_{r=1}^{7}\hat{\theta}_r = 16.67$, $\hat{\theta} = 16.64$ (using entire data)
- $\hat{V}_1(\tilde{\theta}) = \frac{1}{7} \cdot \frac{1}{7-1}\sum_{r=1}^{7}(\hat{\theta}_r - \tilde{\theta})^2 = \frac{0.1704}{7} = 0.024$
- $\hat{V}_2(\tilde{\theta}) = \frac{1}{7} \cdot \frac{1}{7-1}\sum_{r=1}^{7}(\hat{\theta}_r - \hat{\theta})^2 = \frac{0.1716}{7} = 0.025$
- using $\hat{\theta}$, a 95% CI for mean age is

$$16.64 \pm 2.45\sqrt{0.025} = [16.3, 17.0],$$

where 2.45 is the t critical value with 6 df.

Advantages of random group methods:

- Easy to calculate the variance estimate.
- The method is well-suited to multiparameter or nonparametric problems. It can be used to estimate variances for percentiles and nonsmooth functions as well as variances of smooth functions of the population totals.
- Random group methods are easily used after weighting adjustments for nonresponse and undercoverage.

Disadvantages of random group methods:

- The number of random groups is often small. This gives imprecise estimates of the variances.
  —-The survey design may limit the number of random groups that can be constructed. If two psus are selected in each stratum, then only two random groups can be formed.
  —-Generally one would like at least ten random groups to obtain a more stable estimate of the variance and to avoid inflating the CI by using a critical value from a t distribution with few df.
- If $\hat{\theta}$ is a nonlinear statistic, $\tilde{\theta}$ can have large bias if the number of observations in each group is small.
- Setting up the random groups can be difficult in complicated designs, as each random group must have the same design structure as the complete survey.

Resampling and Replication Methods

- Random group methods are easy to compute and explain but are unstable if a complex sample can only be split into a small number of groups.
- Resampling methods treat the sample as if it were itself a population; we take different samples from this new "population" and use the subsamples to estimate the variance.
- All of the methods in this section calculate variance estimates for a sample in which psus are sampled with replacement. If psus are sampled without replacement, these methods may still be used, but are expected to overestimate the variance and result in conservative CIs.

1. **Balanced Repeated Replication (BRR)**
   Restrict to surveys that are stratified to the point that only two psus are selected from each stratum. This design gives the highest degree of stratification possible while still allowing calculation of variance estimates in each stratum.

- Half-samples:
  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.

Altogether 2^H (H stratum) possible half-samples could be formed. McCarthy (1966, 1969) suggest using a balanced sample of the 2^H possible half-samples to estimate the variance.

—- Balanced repeated replication uses the variability among R replicate half-samples that are selected in a balanced way to estimate the variance of $\hat{\theta}$.

- Defining balance,
  Half-sample r can be defined by a vector
  $\boldsymbol{\alpha}_r = (\alpha_{r1}, \cdots, \alpha_{rH})$, let

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

The set of R replicate half-samples is balanced if

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

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

Table 4: Illustration of balanced repeated replication

	Stratum(h)						
	1	2	3	4	5	6	7
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

2. **The Jackknife (delete-1 jackknife)**
   $\hat{\theta}_{(j)}$: the estimator without observation j
   $\hat{\theta}$: estimator with full data set

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

Example: $\hat{\theta} = \bar{y}$

$$\begin{aligned}
\hat{\theta}_{(j)} = \bar{y}_{(j)} &= \frac{1}{n-1}\sum_{i \neq j} y_i = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i - y_j\right) \\
&= \bar{y} \cdot \frac{n}{n-1} - \frac{y_j}{n-1} \\
&= \bar{y} - \frac{1}{n-1}(y_j - \bar{y})
\end{aligned}$$

where $\bar{y}_{(j)} = \dfrac{\text{sum of everyone except } y_j}{n-1}$

$$\begin{aligned}\sum_{j=1}^{n}(\bar{y}_{(j)}-\bar{y})^2 &= \sum_{j=1}^{n}\left(\bar{y}-\frac{1}{n-1}(y_j-\bar{y})-\bar{y}\right)^2 \\ &= \frac{1}{(n-1)^2}\sum_{j=1}^{n}(y_j-\bar{y})^2 \\ &= \frac{1}{n-1}s_y^2\end{aligned}$$

thus

$$\hat{V}_{JK}(\bar{y})=\frac{n-1}{n}\sum_{j=1}^{n}(\bar{y}_{(j)}-\bar{y})^2=\frac{s_y^2}{n},$$

which is equal to the with-replacement estimator of the variance of $\bar{y}$.

Example 9.7. Let's use the jackknife method to estimate the ratio of out-of-state tuition (y) to in-state tuition (x) described in Example 9.3.

$$\hat{\theta} = \frac{\bar{y}}{\bar{x}}, \quad \hat{\theta}_{(j)} = \hat{B}_{(j)} = \frac{\bar{y}_{(j)}}{\bar{x}_{(j)}}$$

The jackknife variance estimator is given by:

$$\hat{V}_{\hat{B}}(\hat{\theta}) = \frac{n-1}{n} \sum_{j \in S} \left(\hat{B}_{(j)} - \hat{B}\right)^2$$

For each jackknife group (Table 9.6 page 374), omit one observation.

For instance, for the first replicate group, $\bar{x}_{(1)}$ is the average of all x values in the sample except x_1:

$$\begin{aligned}\bar{x}_{(1)} &= \frac{1}{9}\sum_{i=2}^{10} x_i \\ &= \frac{1}{9}(7140 + 9808 + \cdots + 11976 + 8935 + 8316) \\ &= 8802\end{aligned}$$

Please refer to Table 9.6 on page 374 for Jackknife calculation. We can calculate,

$$\hat{B} = 2.425, \sum(\hat{B}_{(j)} - \hat{B})^2 = 0.0595,$$

and

$$\hat{V}_{JK}(\hat{B}) = (0.9)(0.0595) = 0.05358$$

Example 9.7 using R

```
data(collegerg)
# replicate group 1
collegerg1<-collegerg[collegerg$repgroup==1,]
collegerg1[,23:24]
> collegerg1[,24:25]
   tuitionfee_in tuitionfee_out
 1          9912           23640
 2          7140           14810
 3          9808           26648
 4          8987           35170
 5          7930            8674
 6          7200           17550
 7          8929           21692
 8         11976           22488
 9          8935           27199
10          8316           18276
```

Linearization estimates (default)

```
collegerg1$sampwt<-rep(500/10,10)
dcollegerg1<-svydesign(id=~1, weights=~sampwt,
data=collegerg1)
> svymean(~tuitionfee_in+tuitionfee_out, dcollegerg1)
                    mean      SE
tuitionfee_in    8913.3  454.46
tuitionfee_out 21614.7 2325.15
```

Calculte estimate and SEs of ratio using linearization

```
> ratio.lin<-svyratio(~tuitionfee_out,~tuitionfee_in,
dcollegerg1)
> ratio.lin
Ratio estimator: svyratio.survey.design2(~tuitionfee_out,
~tuitionfee_in, dcollegerg1)
Ratios=
                tuitionfee_in
tuitionfee_out      2.424994
SEs=
                tuitionfee_in
tuitionfee_out     0.2311776
> confint(ratio.lin,df=degf(dcollegerg1))
                                    2.5 %   97.5 %
tuitionfee_out/tuitionfee_in 1.902034 2.947954
```

Jackknife estimates

```
> ## define jackknife replicate weights design object
> dcollegerg1jk <- as.svrepdesign(dcollegerg1, type="JK1")
> dcollegerg1jk
Call: as.svrepdesign(dcollegerg1, type = "JK1")
Unstratified cluster jacknife (JK1) with 10 replicates.
# JK estimates are same as linearization estimates
# since we are estimating mean from an SRS and
# we didn't include fpc in design object
> svymean(~tuitionfee_in + tuitionfee_out, dcollegerg1jk)
                    mean      SE
tuitionfee_in    8913.3  454.46
tuitionfee_out 21614.7 2325.15
```

```
# jackknife SE for ratio
svyratio(~tuitionfee_out, ~tuitionfee_in,
design = dcollegerg1jk)
> svyratio(~tuitionfee_out, ~tuitionfee_in,
design = dcollegerg1jk)
Ratio estimator: svyratio.svyrep.design(~tuitionfee_out,
~tuitionfee_in, design = dcollegerg1jk)
Ratios=
                tuitionfee_in
tuitionfee_out      2.424994
SEs=
           [,1]
[1,] 0.2314828
```

$\hat{B} = 2.425, SE = 0.2314828$ and $\hat{V}_{JK}(\hat{B}) = .05358.$

Using weights for Jackknife

$$\bar{y} = \frac{\sum_{i \in S} w_i y_i}{\sum_{i \in S} w_i}, \quad \bar{y}_{(i)} = \frac{\sum_{k \in S} w_{k(i)} y_k}{\sum_{k \in S} w_{k(i)}}$$

SRS: $w_i = N/n$.

$$w_{k(i)} = \begin{cases} \dfrac{n}{n-1} w_k & k \neq i \\ 0 & k = i \end{cases}$$

Extension to a complex survey data:

- Cluster samples: delete one psu instead of deleting one unit
- Stratified multistage cluster sample: the jackknife is applied separately in each stratum at the first stage of sampling, with one psu deleted at a time

a. H strata, n_h psus are chosen for the sample from stratum h. Assume these psus are chosen with replacement

b. n: number of psus

c. $\hat{\theta}_{(j)}$: estimate of θ that would be obtained by deleting psu j

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

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

Example 9.8 coots data using R

Delete one psu at a time

```
> data(coots)
> coots$relwt<-coots$csize/2
> dcoots<-svydesign(id=~clutch,weights=~relwt,data=coots)
> dcootsjk <- as.svrepdesign(dcoots, type="JK1")
> dcootsjk
Call: as.svrepdesign(dcoots, type = "JK1")
Unstratified cluster jacknife (JK1) with 184 replicates.
> svymean(~volume,dcootsjk)
         mean    SE
volume 2.4908 0.061
> confint(svymean(~volume,dcootsjk),df=degf(dcootsjk))
          2.5 %   97.5 %
volume 2.370354 2.611203
```

Advantages:

- The jackknife is an all-purpose method. The same procedure is used to estimate the variance for every statistic for which jackknife can be used.
- The jackknife provides a consistent estimator of the variance when θ is a smooth function of population totals (Krewski and Rao, 1981).
- Replication methods such as the jackknife can be used to account for some of the effects of imputation on the variance estimates (Rao and Shao, 1992).

Disadvantages:

- For some sampling designs, the jackknife may require a large amount of computation.
- The jackknife performs poorly for estimating the variances of some statistics that are not smooth functions of population totals. For example, the jackknife does not give a consistent estimator of the variance of quantiles and median.

3. **Bootstrap**
   Treat sample as population, then draw "resamples" with replacement from the original sample
   Take R bootstrap resamples, obtaining $\hat{\theta}_1^*, \cdots, \hat{\theta}_R^*$.

Example 9.9, estimate the variance of the median height θ in the population, using data from *htsrs* (refer to example 7.4).

- The median height in population *htpop* is $\theta = 168$
- Sample median from *htsrs*, $\hat{\theta} = 169$
- Resample 2000 times, $R = 2000$.

Median of resample	165	166	166.5	167	167.5	$\cdots$	172
frequency	1	5	2	40	15	$\cdots$	4

Sample mean: $\dfrac{\hat{\theta}_1^* + \cdots + \hat{\theta}_{2000}^*}{2000} = 169.3$

Sample variance:

$$\hat{V}_B(\hat{\theta}) = \frac{1}{R-1}\sum_{r=1}^{R}\left[\hat{\theta}_r^* - \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r^*\right]^2 = 0.9148$$

An approximate 95% CI may be constructed using the bootstrap variance as

$$169 \pm 1.96 * \sqrt{0.9148} = [167.4, 171.2]$$

or using 2.5% and 97.5% quantiles as

$$[q_{2.5\%}, q_{97.5\%}] = [167.5, 171]$$

Direct code to find bootstap variance and CI for median of height

```
# bootstrap by direct coding
# number of iteration
R <- 10000
# init location for bootstrap theta
thetahat <- rep(NA, R)
# draw R bootstrap resamples
for (i in 1:R) {
  #
  resam <- sample(htsrs$height, 199, replace = TRUE)
thetahat[i] <- median(resam)
}
```

Variance and CI estimate by normal approximation using bootstrap variance

```
> # variance and CI estimate by normal approximation
> sebs<-sqrt(var(thetahat))
> sebs
[1] 0.9685265
> m<-median(htsrs$height)
> m
[1] 169
> CI.bs1 <- c(m-1.96*sebs,m+1.96*sebs)
> CI.bs1
[1] 167.1017 170.8983
```

An approximate 95% CI of height median may be constructed using the bootstrap variance as

$$169 \pm 1.96 * 0.9685 = [167.1017, 170.8983]$$

Sort the bootstrap estimates to obtain bootstrap CI

```
> # 0.025th and 0.975th quantile gives equal-tail
> # bootstrap CI
> thetahat.sorted <- sort(thetahat)
> CI.bs2 <- c(thetahat.sorted[round(0.025*R)],
thetahat.sorted[round(0.975*R+1)])
> CI.bs2
[1] 167 171
```

The 0.025th and 0.975th quantile of the R sorted bootstrap estimates gives equal-tail bootstrap CI for median as $[167, 171]$.

The following is bootstrapping mean of height

```
> data(htsrs)
> nrow(htsrs)
[1] 200
> head(htsrs)
     rn height gender
1   257    159 F
2  1016    174 M
3  1264    186 M
4   817    158 F
5   374    178 F
6  1063    177 M
> wt<-rep(10,nrow(htsrs))
> dhtsrs<-svydesign(id=~1, weights=~wt,data=htsrs)
> dhtsrs
Independent Sampling design (with replacement)
svydesign(id = ~1, weights = ~wt, data = htsrs)
```

```
> set.seed(9231)
> dhtsrsboot <- as.svrepdesign(dhtsrs, type="subbootstrap",
replicates=1000)
# linearization
> svymean(~height,dhtsrs)
         mean     SE
height 168.94 0.7831
> svymean(~height,dhtsrsboot)
         mean     SE
height 168.94 0.7978
> degf(dhtsrsboot) # 199 = n - 1
[1] 199
> confint(svymean(~height,dhtsrsboot),df=degf(dhtsrsboot))
          2.5 %   97.5 %
height 167.3667 170.5133
```

An approximate 95% CI may be constructed using the bootstrap variance as $168.94 \pm 1.971957 * 0.7978 = [167.4, 171.5]$

Distribution of the $\hat{\theta}_r^*$s should be like distribution of $\hat{\theta}$ from the original sample.

$$\hat{F}^*(x) - \hat{F}(x) \to 0$$

Estimate quantiles of distribution of $\hat{\theta} - \theta$ using histrogram of bootstrap values.

$$\hat{V}_B(\hat{\theta}) = \frac{1}{R-1} \sum_{r=1}^{R} \left[\hat{\theta}_r^* - \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}_r^* \right]^2$$

or

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

The rescaling bootstrap of Rao and Wu (1988) for a stratified multistage sample

— n_h: number of psus sampled from stratum h

— R: number of bootstrap replicates to be created. Typically, $R = 500$ or $1,000$

— bootstrapping is applied within each stratum

a. For bootstrap replicate $r(r = 1, \cdots, R)$, select an SRS of size $n_h - 1$ psus with replacement from the n_h psus in stratum h. Do this independently for each stratum.
Let $m_{hj}(r)$ be the number of times psu j of stratum h is selected in replicate r.

b. Create the replicate weight vector for replicate r as

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

for observation i in psu j of stratum h.
The result is R vectors of replicate weights.

c. Use the vectors of replicate weights to estimate $V(\hat{\theta})$. Let $\hat{\theta}_r^*$ be the estimator of θ, calculated the same way as $\hat{\theta}$ but using weights $w_i(r)$ instead of the original weights w_i. Then,

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

Example 9.10. We use the bootstrap to estimate variances from the data in file *htstrat.csv*, discussed in Example 7.6. The bootstrap weights are constructed by taking 1000 stratified random samples with replacement from the data set;

- select 159 women ($n_{women} = 160$)
- select 39 men ($n_{men} = 40$)

with replacement for each resample.

```
> data(htstrat)
> nrow(htstrat)
[1] 200
> head(htstrat)
     rn height gender
1   201    166 F
2   965    163 F
3   490    166 F
4   249    155 F
5   260    154 F
6   324    160 F
> dhtstrat <- svydesign(id = ~1, strata = ~gender,
fpc = c(rep(1000,160),rep(1000,40)), data = htstrat)
> dhtstrat
Stratified Independent Sampling design
svydesign(id = ~1, strata = ~gender,
fpc = c(rep(1000, 160), rep(1000, 40)), data = htstrat)
```

```
> set.seed(982537455)
> dhtstratboot <- as.svrepdesign(dhtstrat,
 type="subbootstrap",replicates=1000)
> svymean(~height,dhtstratboot)
         mean     SE
height 169.02 0.7296
> degf(dhtstratboot)
[1] 198
> confint(svymean(~height,dhtstratboot),
df=degf(dhtstratboot))
           2.5 %   97.5 %
height 167.5769 170.4543
```

The average height is estimated as $\bar{y}_{str} = 169.02$ with bootstrap standard error of 0.7296; the standard error calculated using the stratified sampling formula (equation (3.6), page 85), ignoring the fpc, is 0.739.

Advantage:

- The bootstrap works for smooth functions of population means and for some nonsmooth functions such as quantiles in general sampling designs.
- The bootstrap is well suited for finding CIs directly.

Disadvantages:

- In some settings, the bootstrap may require more computations than BRR or jackknife, since R is typically a very large number.
- Less theoretical work has been done on properties of the bootstrap in complex sampling designs.

Generalized variance functions (GVF)

- In many large government surveys such as the U.S. Current Population Survey (CPS) or the Canadian Labour Force Survey, hundreds or thousands of estimates are calculated and published.
- The agencies could calculate standard errors for each published estimators, but that would add greatly to the labor involved in publishing timely estimates from the surveys.

- In addition, other analysts of the public-use data files may wish to calculate additional estimates, and the public-use files may not provide enough information to allow calculation of standard errors.
  —— Want a flexible formula to allow calculating standard errors for most estimators wanted
- Generalized variance functions (GVFs) are provided in a number of surveys to calculate standard errors. They have been used for the CPS since 1947. Villiant (1987), Wolter (2007, Chapter 7) describes the theory underlying GVFs.

Given a survey statistics $\hat{t}_i$, such as the estimated number of persons employed.

- Let $\hat{p}_i$ be an estimated proportion, $\hat{p}_i = \hat{t}_i/N$, where N is a control total (i.e. constant), provided by the Census Bureau,
- Let d_i be the design effect related to variable i. We have

$$V(\hat{p}_i) = d_i \times \frac{p_i(1-p_i)}{n}.$$

$$V(\hat{t}_i) = N^2 d_i \times \frac{p_i(1-p_i)}{n} = a_i t_i^2 + b_i t_i$$

where $a_i = -\frac{d_i}{n}, b_i = d_i \frac{N}{n}$. If the deffs are similar for different estimates, we have $a_i \approx a$ and $b_i \approx b$.

The general procedure for constructing a generalized variance function is as follows:

a. Using replication or some other method, estimate variances for k population totals of special interest, $t_1, t_2, \cdots, t_k$. Let v_i be the relative variance for t_i, for $i = 1, 2, \cdots, k$.

$$\begin{aligned} v_i &= V(\hat{t}_i)/\hat{t}_i^2 \\ &= \frac{a_i \hat{t}_i^2 + b_i \hat{t}_i}{\hat{t}_i^2} \\ &= a_i + \frac{b_i}{\hat{t}_i} \end{aligned}$$

b. Postulate a regression model relating a set of v_i to $\hat{t}_i$ by

$$v_i = \alpha + \beta/\hat{t}_i$$

to find the parameter estimates a and b for α and β.
This is a linear regression model with response variable v_i and explanatory variable $1/\hat{t}_i$. Valliant (1987) found that this model produces consistent estimators of the variances for the class of superpopulation models he studied.

c. Use regression techniques to estimate α and β by a and b. Valliant (1987) suggests using weighted least squares to estimate the parameters, giving higher weight to items with small v_i.

d. The GVF estimate of relative variance is predicted from the regression function $a + b/\hat{t}_i$ or

$$\hat{V}(\hat{t}_{\text{new}}) = a\hat{t}^2_{\text{new}} + b\hat{t}_{\text{new}}$$

Advantages of Generalized Variance Functions (GVFs):

- The GVF may be used when insufficient information is provided in the public-use data files to allow direct calculation of standard errors. The data collector can calculate the GVF, and often has more information for estimating variances than is released to the public.
- Valliant (1987) found that if design effects for the k estimated totals are similar, the GVF variances are often more stable than the direct estimates of variance, as they smooth out some of the fluctuations from item to item.
- A GVF saves a great deal of time and speeds production of annual reports. It is also useful for designing similar surveys in the future.

Disadvantages of Generalized Variance Functions (GVFs)

- If the design effects (deffs) vary significantly across different responses, the simple GVF model may produce inaccurate variance estimates for some responses.
- Caution is needed when applying GVFs to subpopulations, especially if the subpopulation exhibits an unusually high level of clustering (and thus a high design effect). In such cases, the GVF estimate of variance may be substantially underestimated because it does not account for the higher clustering.
- GVFs may not perform well for variables not included in the original calculation of regression parameters. Such estimates may not align with the assumptions of the GVF model, leading to unreliable results.

Disclaimer

Slides are intended for a course based on the book: *Sampling: Design and Analysis, Third Edition* by Sharon L. Lohr
All examples, datasets, and pages directly scanned and included in this chapter are from the textbook.
References to papers or books cited in these slides can be found in the Bibliography section of the textbook.