Lecture 8: Variance Estimation in Complex Surveys

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

Linearization Methods

For any constants $a_1, \ldots, a_k$, we can define a new variable

$$q_i = \sum_{j=1}^{k} a_j y_{ij}$$

so that

$$\hat{t}_q = \sum_{i \in \mathcal{S}} w_i q_i = \sum_{j=1}^{k} a_j \hat{t}_j$$

and

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

The first-order version of Taylor's theorem states that if the second derivative of h is continuous, then

$$h(x) = h(a) + h'(a)(x-a) + \int_a^x (x-t)h''(t)dt;$$

under conditions commonly satisfied in statistics, the last term is small relative to the first two and we use the approximation

$$\begin{aligned} h(\hat{p}) &\approx h(p) + h'(p)(\hat{p}-p) \\ &= p(1-p) + (1-2p)(\hat{p}-p). \end{aligned}$$

Then,

$$V[h(\hat{p})] \approx (1-2p)^2 V(\hat{p}-p),$$

and $V(\hat{p})$ is known, so the approximate variance of $h(\hat{p})$ can be estimated by

$$\hat{V}[h(\hat{p})] = (1-2\hat{p})^2 \hat{V}(\hat{p}). \quad \blacksquare$$

Replicating the Survey Design

Suppose the basic survey design is replicated independently R times. Independently here 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.

Let

$\theta =$ parameter of interest

$\hat{\theta}_r =$ estimate of θ calculated from rth replicate

$$\tilde{\theta} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_i.$$

If $\hat{\theta}_r$ is an unbiased estimator of θ, so is $\tilde{\theta}$, and

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

is an unbiased estimator of $V(\tilde{\theta})$. Note that $\hat{V}_1(\tilde{\theta})$ is the sample variance of the R independent estimators of θ divided by R—the usual estimator of the variance of a sample mean.

Random grouping approach

- SRS: 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, however, the groups can be treated as though they are independent replicates.
- Cluster sampling: 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.
- Stratified multistage sample: a random group contains a sample of psus from each stratum.

If θ is a nonlinear quantity, $\tilde{\theta}$ will not, in general, be the same as $\hat{\theta}$, the estimator calculated directly from the complete sample. For example, in ratio estimation, $\tilde{\theta} = (1/R)\sum_{r=1}^{R} \hat{\bar{y}}_r/\hat{\bar{x}}_r$, while $\hat{\theta} = \hat{\bar{y}}/\hat{\bar{x}}$. Usually, $\hat{\theta}$ is a more stable estimator than $\tilde{\theta}$. Sometimes $\hat{V}_1(\tilde{\theta})$ is used to estimate $V(\hat{\theta})$, although it is an overestimate. Another estimator of the variance is slightly larger, but is often used:

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

Example: Two variances

Random Group Number	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

The seven estimates of θ are treated as independent observations, so

$$\tilde{\theta} = \frac{1}{7}\sum_{r=1}^{7}\hat{\theta}_r = 16.67$$

and

$$\hat{V}_1(\tilde{\theta}) = \frac{1}{7}\left\{\frac{1}{6}\sum_{r=1}^{7}(\hat{\theta}_r - \tilde{\theta})^2\right\} = \frac{0.1704}{7} = 0.024.$$

Using the entire data set, we calculate $\hat{\theta} = 16.64$ with

$$\hat{V}_2(\tilde{\theta}) = \frac{1}{7}\left\{\frac{1}{6}\sum_{r=1}^{7}(\hat{\theta}_r - \hat{\theta})^2\right\} = \frac{0.1716}{7} = 0.025.$$

We can use either $\tilde{\theta}$ or $\hat{\theta}$ to calculate CIs; using $\hat{\theta}$, a 95% CI for mean age is

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

Balanced Repeated Replication (BRR)

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

A Small Stratified Random Sample, Used to Illustrate BRR

Stratum	$\frac{N_h}{N}$	y_{h1}	y_{h2}	$\bar{y}_h$	$y_{h1} - y_{h2}$
1	0.30	2,000	1,792	1,896	208
2	0.10	4,525	4,735	4,630	−210
3	0.05	9,550	14,060	11,805	−4,510
4	0.10	800	1,250	1,025	−450
5	0.20	9,300	7,264	8,282	2,036
6	0.05	13,286	12,840	13,063	446
7	0.20	2,106	2,070	2,088	36

Using formulas

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

The estimated population mean is

$$\bar{y}_{\text{str}} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = 4451.7.$$

Ignoring the finite population corrections (fpcs) in (3.5) gives the variance estimator

$$\hat{V}_{\text{str}}(\bar{y}_{\text{str}}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h};$$

when $n_h = 2$, as here, $s_h^2 = (y_{h1} - y_{h2})^2/2$, so

$$\hat{V}_{\text{str}}(\bar{y}_{\text{str}}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{(y_{h1} - y_{h2})^2}{4}.$$

Here, $\hat{V}_{\text{str}}(\bar{y}_{\text{str}}) = 55{,}892.75$. This may overestimate the variance if sampling is without replacement.

BRR

To define balance, let's introduce the following notation. Half-sample r can be defined by a vector $\boldsymbol{\alpha}_r = (\alpha_{r1}, \ldots, \alpha_{rH})$: Let

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

Equivalently,

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

If group 1 contains observations $\{y_{11}, y_{22}, y_{32}, y_{42}, y_{51}, y_{62}, y_{71}\}$ as above, then $\boldsymbol{\alpha}_1 = (1, -1, -1, -1, 1, -1, 1)$. Similarly, $\boldsymbol{\alpha}_2 = (-1, 1, 1, 1, -1, 1, -1)$. The set of R replicate half-samples is **balanced** if

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

For replicate r, calculate $\hat{\theta}(\boldsymbol{\alpha}_r)$ the same way as $\hat{\theta}$ but using only the observations in the half-sample selected by $\boldsymbol{\alpha}_r$. For estimating the mean of a stratified random sample, $\hat{\theta}(\boldsymbol{\alpha}_r) = \sum_{h=1}^{H} (N_h/N) y_h(\boldsymbol{\alpha}_r)$. Define the BRR variance estimator to be

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

If the set of half-samples is balanced, then for stratified random sampling

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

Example

		Stratum (h)						
		1	2	3	4	5	6	7
Half-sample (r)	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

Half-sample	$\hat{\theta}(\alpha_r)$	$[\hat{\theta}(\alpha_r) - \hat{\theta}]^2$
1	4732.4	78,792.5
2	4439.8	141.6
3	4741.3	83,868.2
4	4344.3	11,534.8
5	4084.6	134,762.4
6	4592.0	19,684.1
7	4123.7	107,584.0
8	4555.5	10,774.4
average	4451.7	55,892.8

Weights

The sampling weight for observation i in stratum h is $w_{hi} = N_h/n_h$, and

$$\bar{y}_{\text{str}} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{2} w_{hi}y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{2} w_{hi}}.$$

Define

$$w_{hi}(\alpha_r) = \begin{cases} 2w_{hi} & \text{if observation } i \text{ of stratum } h \text{ is in} \\ & \text{the half-sample selected by } \alpha_r \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\bar{y}_{\text{str}}(\alpha_r) = \frac{\sum_{h=1}^{H}\sum_{i=1}^{2} w_{hi}(\alpha_r)y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{2} w_{hi}(\alpha_r)}.$$

Jackknife

The jackknife was introduced as a method of reducing bias.
Delete-1 jackknife:

For an SRS, let $\hat{\theta}_{(j)}$ be the estimator of the same form as $\hat{\theta}$, but not using observation j. Thus, if $\hat{\theta} = \bar{y}$, then $\hat{\theta}_{(j)} = \bar{y}_{(j)} = \sum_{i \neq j} y_i/(n-1)$. For an SRS, define the delete-1 jackknife estimator as

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

Why the multiplier

$$\frac{n-1}{n}?$$

Example: Ratio estimation

j	x	y	$\bar{x}_{(j)}$	$\bar{y}_{(j)}$	$\hat{B}_{(j)}$
1	1365	3747	1580.6	3617.7	2.2889
2	1677	4983	1545.9	3480.3	2.2513
3	1500	1500	1565.6	3867.3	2.4703
4	1080	2160	1612.2	3794.0	2.3533
5	1875	2475	1523.9	3759.0	2.4667
6	3071	5135	1391.0	3463.4	2.4899
7	1542	3950	1560.9	3595.1	2.3032
8	930	4050	1628.9	3584.0	2.2003
9	1340	4140	1583.3	3574.0	2.2573
10	1210	4166	1597.8	3571.1	2.2350

Let's use the jackknife to estimate the ratio of nonresident tuition to resident tuition for the first group of colleges. Here, $\hat{\theta} = \bar{y}/\bar{x}$, $\hat{\theta}_{(j)} = \hat{B}_{(j)} = \bar{y}_{(j)}/\bar{x}_{(j)}$, and

$$\hat{V}_{\mathrm{JK}}(\hat{B}) = \frac{n-1}{n} \sum_{j \in \mathcal{S}} (\hat{B}_{(j)} - \hat{B})^2.$$

For each jackknife group in the Table, omit one observation. Thus, $\bar{x}_{(1)}$ is the average of all x's except for x_1: $\bar{x}_{(1)} = (1/9) \sum_{i=2}^{9} x_i$. Here, $\hat{B} = 2.3288$, $\sum (B_{(j)} - \hat{B})^2 = 0.1043$, and $\hat{V}_{\mathrm{JK}}(\hat{B}) = .09377$.

Jackknife for survey data

- Cluster sample: For a cluster sample, then, we would apply the jackknife variance estimator by letting n be the number of psus, and letting $\hat{\theta}_{(j)}$ be the estimate of θ that we would obtain by deleting all the observations in psu j.
- 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.
  For example, suppose there are H strata, and n_h psus are chosen for the sample from stratum h. Assume these psus are chosen with replacement.

To apply the jackknife, delete one psu at a time. Let $\hat{\theta}_{(hj)}$ be the estimator of the same form as $\hat{\theta}$ when psu j of stratum h is omitted. To calculate $\hat{\theta}_{(hj)}$, define a new weight variable: Let

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

Then use the weights $w_{i(hj)}$ to calculate $\hat{\theta}_{(hj)}$, and

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

Example: variance of the mean egg volume

Jackknife Weights. The values w_i are the relative weights; $w_{i(k)}$ is the set of jackknife weights for the replication omitting psu k.

clutch	*csize*	w_i	$w_{i(1)}$	$w_{i(2)}$	...	$w_{i(184)}$
1	13	6.5	0	6.535519	...	6.535519
1	13	6.5	0	6.535519	...	6.535519
2	13	6.5	6.535519	0	...	6.535519
2	13	6.5	6.535519	0	...	6.535519
3	6	3	3.016393	3.016393	...	3.016393
3	6	3	3.016393	3.016393	...	3.016393
4	11	5.5	5.530055	5.530055	...	5.530055
4	11	5.5	5.530055	5.530055	...	5.530055
⋮	⋮	⋮	⋮	⋮		⋮
183	13	6.5	6.535519	6.535519	...	6.535519
183	13	6.5	6.535519	6.535519	...	6.535519
184	12	6	6.032787	6.032787	...	0
184	12	6	6.032787	6.032787	...	0
Sum	3514	1757	1753.53	1753.53	...	1754.54

We have only one stratum, so h=1 for all observations. 184 PSUS.

Estimate

The mean

$$\hat{\theta} = \bar{y}_r = 4375.947/1757 = 2.49.$$

In fact,

$$\hat{\theta} = \frac{\sum w_i y_i}{\sum w_i}.$$

to find $\hat{\theta}_{(kj)}$, we follow the same procedure but use $w_{i(kj)}$ in place of w_i. Thus, $\hat{\theta}_{(1,1)} = 4349.348/1753.53 = 2.48034$; $\hat{\theta}_{(1\,2)} = 4345.036/1753.53 = 2.47788$; $\hat{\theta}_{(1,184)} = 4357.819/1754.54 = 2.48374$. Then, we calculate $\hat{V}_{\mathrm{JK}}(\hat{\theta}) = 0.00373$. This results in a standard error of 0.061.

Bootstrap

General idea: Suppose S is an SRS with replacement of size n. We hope, in drawing the sample, that it reproduces properties of the whole population. We then treat the sample S as if it were a population, and take resamples from S.

Let R be the number of bootstrap replicates to be created. Typically, $R = 500$ or 1,000, although some statisticians use smaller values of R.

1. For bootstrap replicate r ($r = 1, \ldots, R$), select an SRS of $n_h - 1$ psus with replacement from the n_h sample 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.
2. Create the replicate weight vector for replicate r as

$$w_i(r) = w_i \times \frac{n_h}{n_h - 1} m_{hj}(r), \text{ for observation } i \text{ in psu } j \text{ of stratum } h.$$

   The result is R vectors of replicate weights.
3. 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,

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

Confidence intervals

Roughly speaking, the assumptions for linearization, jackknife, BRR, and bootstrap are as follows:

1. The quantity of interest θ can be expressed as a smooth function of the population totals; more precisely, $\theta = h(t_1, t_2, \ldots, t_k)$, where the second-order partial derivatives of h are continuous.
2. The sample sizes are large: Either the number of psus sampled in each stratum is large, or the survey contains a large number of strata. (See Rao and Wu, 1985, for the precise technical conditions needed.) Also, to construct a CI using the normal distribution, the sample sizes must be large enough so that the sampling distribution of $\hat{\theta}$ is approximately normal.

Consequently, when the assumptions are met, an approximate 95% CI for θ may be constructed as

$$\hat{\theta} \pm 1.96 \sqrt{\hat{V}(\hat{\theta})}.$$

Alternatively, a t_{df} percentile may be substituted for 1.96, with df = (number of groups − 1) for the random group method, and df = (number of psus − number of strata) for the other methods.

Confidence intervals for population quantiles

Let q be between 0 and 1. Then define the quantile θ_q as $\theta_q = F^{-1}(q)$, where $F^{-1}(q)$ is defined to be the smallest value y satisfying $F(y) \geq q$. Similarly, define $\hat{\theta}_q = \hat{F}^{-1}(q)$.

Some of the methods already discussed work quite well for constructing CIs for quantiles. The random group method works well if the number of random groups, R, is moderate. Let $\hat{\theta}_q(r)$ be the estimated quantile from random group r. Then, a CI for θ_q is

$$\hat{\theta}_q \pm t \sqrt{\frac{1}{R(R-1)} \sum_{r=1}^{R} [\hat{\theta}_q(r) - \hat{\theta}_q]^2},$$

where t is the appropriate percentile from a t distribution with $R - 1$ degrees of free-dom. Similarly,

$$\hat{\theta}_q \pm 1.96 \sqrt{\hat{V}(\hat{\theta}_q)},$$

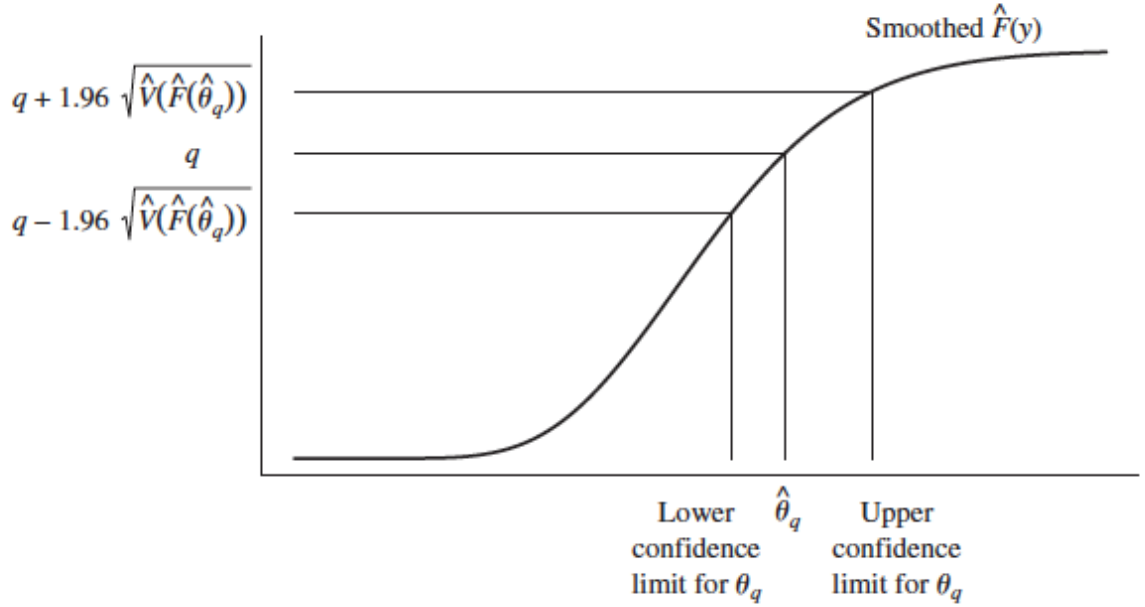
where the variance estimate is calculated using BRR or bootstrap.

Woodruff's CI

An alternative interval can be constructed based on a method introduced by Woodruff (1952). For any y, $\hat{F}(y)$ is a function of population totals: $\hat{F}(y) = \sum_{i \in \mathcal{S}} w_i u_i / \sum_{i \in \mathcal{S}} w_i$, where $u_i = 1$ if $y_i \leq y$ and $u_i = 0$ if $y_i > y$. Thus, a method in this chapter can be used to estimate $V[\hat{F}(y)]$ for any value y, and an approximate 95% CI for $F(y)$ is given by

$$\hat{F}(y) \pm 1.96\sqrt{\hat{V}[\hat{F}(y)]}.$$

Woodruff's confidence interval for the quantile θ_q if the empirical distribution function is continuous. Since $F(y)$ is a proportion, we can easily calculate a confidence interval for any value of y, shown on the vertical axis. We then look at the corresponding points on the horizontal axis to form a confidence interval for θ_q.



Woodruff's CI

Now let's use the CI for $q = F(\theta_q)$ to obtain an approximate CI for θ_q. Since we have a 95% CI,

$$
\begin{aligned}
0.95 &\approx P\left\{\hat{F}(\theta_q) - 1.96\sqrt{\hat{V}[\hat{F}(\hat{\theta}_q)]} \le q \le \hat{F}(\theta_q) + 1.96\sqrt{\hat{V}[\hat{F}(\hat{\theta}_q)]}\right\} \\
&= P\left\{q - 1.96\sqrt{\hat{V}[\hat{F}(\hat{\theta}_q)]} \le \hat{F}(\theta_q) \le q + 1.96\sqrt{\hat{V}[\hat{F}(\hat{\theta}_q)]}\right\} \\
&= P\left(\hat{F}^{-1}\left\{q - 1.96\sqrt{\hat{V}[\hat{F}(\hat{\theta}_q)]}\right\} \le \theta_q \le \hat{F}^{-1}\left\{q + 1.96\sqrt{\hat{V}[\hat{F}(\hat{\theta}_q)]}\right\}\right).
\end{aligned}
$$

So an approximate 95% CI for the quantile θ_q is

$$
\left[\hat{F}^{-1}\left\{q - 1.96\sqrt{\hat{V}[\hat{F}(\hat{\theta}_q)]}\right\}, \hat{F}^{-1}\left\{q + 1.96\sqrt{\hat{V}[\hat{F}(\hat{\theta}_q)]}\right\}\right].
$$

Example: CI for median

The following values were obtained for the empirical distribution function:

y	165	166	167	168	169	170	171
$\hat{F}(y)$	0.3781	0.4438	0.4844	0.5125	0.5375	0.5656	0.6000

We estimated the population median by

$$\hat{\theta}_{0.5} = 167 + \frac{0.5 - 0.4844}{0.5125 - 0.4844}(168 - 167) = 167.6.$$

Note that

$$\hat{F}(\hat{\theta}_q) = \frac{\sum_{h=1}^{2}\sum_{i\in S_h} w_{hi}u_{hi}}{\sum_{h=1}^{2}\sum_{i\in S_h} w_{hi}} = \frac{\sum_{h=1}^{2}\sum_{i\in S_h} w_{hi}u_{hi}}{2000}$$

Example: CI for median hight

The lower confidence bound for the median is then $\hat{F}^{-1}(0.5 - 0.0684)$, and the upper confidence bound for the median is $\hat{F}^{-1}(0.5 + 0.0684)$. We again use linear interpolation to obtain

$$\hat{F}^{-1}(0.4316) = 165 + \frac{0.4316 - 0.3781}{0.4438 - 0.3781}(166 - 165) = 165.8$$

and

$$\hat{F}^{-1}(0.5684) = 170 + \frac{0.5684 - 0.5656}{0.6 - 0.5656}(171 - 170) = 170.1.$$

Thus, an approximate 95% CI for the median is [165.8, 170.1].