Lecture 7: Complex Surveys

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

Building blocks

Components of a complex survey: random sampling, ratio estimation, stratification, and clustering.

- Cluster sampling with replacement. Select a sample of n clusters with replacement
- Cluster sampling without replacement. Select a sample of n psus without replacement; π_i is the probability that psu i is included in the sample.
- Stratification.

Stratification. Let $\hat{t}_1, \ldots, \hat{t}_H$ be unbiased estimators of the stratum totals $t_1, \ldots, t_H$, and let $\hat{V}(\hat{t}_1), \ldots, \hat{V}(\hat{t}_H)$ be unbiased estimators of the variances. Then estimate the population total by

$$\hat{t} = \sum_{h=1}^{H} \hat{t}_h$$

and its variance by

$$\hat{V}(\hat{t}) = \sum_{h=1}^{H} \hat{V}(\hat{t}_h).$$

Example: Estimate the prevalence of bed net use

- The sampling frame consisted of all rural villages of fewer than 3000 people in the Gambia.
- The villages were stratified by three geographic regions (eastern, central and western) and by whether the village had a public health clinic (PHC) or not.
- In each region five districts were chosen with probability proportional to the district population as estimated in the 1983 national census.
- In each district four villages were chosen, again with probability proportional to census population: two PHC villages and two non-PHC villages.
- Six compounds were chosen more or less randomly from each village, and a researcher recorded the number of beds and nets, along with other information, for each compound.

Estimation Procedure

- Record the total number of nets for each compound.
- Estimate the total number of nets for each village by (number of compounds in the village) $\times$ (average number of nets per compound). Find the estimated variance of the total number of nets, for each village.
- Estimate the total number of nets for the PHC villages in each district. Repeat for the non-PHC villages in each district.
- Add the estimates from the two strata (PHC and non-PHC) to estimate the number of nets in each district; sum the estimated variances from the two strata to estimate the variance for the district.
- Use two-stage cluster sampling formulas to estimate the total number of nets for each region.
- Add the estimated totals for each region to estimate the total number of bed nets. Add the region variances as called for in stratified sampling.

Ratio Estimation in Complex Surveys

- Combined ratio estimator

$$\hat{t}_{yrc} = \hat{B} t_x,$$

where

$$\hat{B} = \frac{\hat{t}_y}{\hat{t}_x};$$

The mean squared error (MSE) of $\hat{t}_{yrc}$ can be estimated by

$$\hat{V}(\hat{t}_{yrc}) = \left(\frac{t_x}{\hat{t}_x}\right)^2 \left[\hat{V}(\hat{t}_y) + \hat{B}^2 \hat{V}(\hat{t}_x) - 2\hat{B}\widehat{\text{Cov}}(\hat{t}_y, \hat{t}_x)\right].$$

- Separate ratio estimator

$$\hat{t}_{yrs} = \sum_{h=1}^{H} \hat{t}_{yhr} = \sum_{h=1}^{H} t_{xh} \frac{\hat{t}_{yh}}{\hat{t}_{xh}},$$

with

$$\hat{V}(\hat{t}_{yrs}) = \sum_{h=1}^{H} \hat{V}(\hat{t}_{yhr}).$$

Sampling weights

In most large sample surveys, weights are used to calculate point estimates.

- Stratified random sampling

$$\hat{t}_{\text{str}} = \sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} y_{hj},$$

where the sampling weight $w_{hj} = (N_h/n_h)$ can be thought of as the number of observations in the population represented by the sample observation y_{hj}. The probability of selecting the jth unit in the hth stratum to be in the sample is $\pi_{hj} = n_h/N_h$, so the sampling weight is simply the inverse of the probability of selection: $w_{hj} = 1/\pi_{hj}$.

- In cluster sampling with equal probabilities

$$w_{ij} = \frac{NM_i}{nm_i} = \frac{1}{\text{probability that the } j\text{th ssu in the } i\text{th psu is in the sample}}.$$

- For three-stage cluster sampling, the principle extends: Let w_p be the weight for the psu, $w_{s|p}$ be the weight for the ssu, and $w_{t|s,p}$ be the weight associated with the tsu (tertiary sampling unit). Then the overall sampling weight for an observation unit is

$$w = w_p \times w_{s|p} \times w_{t|s,p}.$$

Estimating a Distribution Function

Suppose the values for the entire population of N units are known. Then any quantity of interest may be calculated from the **probability mass function**,

$$f(y) = \frac{\text{number of units whose value is } y}{N}$$

or the **cumulative distribution function** (cdf)

$$F(y) = \frac{\text{number of units with value } \leq y}{N} = \sum_{x \leq y} f(x).$$

In probability theory, these are the probability mass function and cdf for the random variable Y, where Y is the value obtained from a random sample of size one from the population. Then $f(y) = P\{Y = y\}$ and $F(y) = P\{Y \leq y\}$. Of course, $\sum f(y) = F(\infty) = 1$.

Estimating a Distribution Function

Any population quantity can be calculated from the probability mass function or cdf. The population mean is

$$\bar{y}_U = \sum_{\substack{\text{values of y} \\ \text{in population}}} y f(y).$$

The population variance, too, can be written using the probability mass function:

$$\begin{aligned} S^2 &= \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y}_U)^2 \\ &= \frac{N}{N-1} \sum_y f(y) \left[y - \sum_x x f(x) \right]^2 \\ &= \frac{N}{N-1} \left[\sum_y y^2 f(y) - \left(\sum_x x f(x) \right)^2 \right]. \end{aligned}$$

Quantile

θ_q is 100qth quantile if

$$F(\theta_q) = q$$

if such a value exists. otherwise,

$$\theta_q \in [a, b]$$

where a is the largest population value of y with

$$F(y) < q$$

and b is the smallest value of y with

$$F(y) > q.$$

If $q < 1/N$, θ_q is the smallest value of y, and if $q > 1 - 1/N$, θ_q is the largest value of y.

Estimates

- Empirical probability mass function (epmf) to be the sum of the weights for all observations taking on the value y, divided by the sum of all the weights:

$$\hat{f}(y) = \frac{\sum\limits_{i \in \mathcal{S}: y_i = y} w_i}{\sum\limits_{i \in \mathcal{S}} w_i}.$$

- empirical cumulative distribution function (empirical cdf) is the sum of all weights for observations with values $\leq y$, divided by the sum of all weights:

$$\hat{F}(y) = \sum_{x \leq y} \hat{f}(x).$$

Estimated variance and quantiles

$$S^2 = \frac{N}{N-1}\left[\sum_y f(y)\left\{y - \sum_x x f(x)\right\}^2\right] = \frac{N}{N-1}\left[\sum_y y^2 f(y) - \left\{\sum_y y f(y)\right\}^2\right].$$

Then, substitute $\hat{f}(y)$ for every appearance of $f(y)$ to obtain an estimate of the population characteristic. Using this method, then,

$$\hat{\bar{y}} = \sum_y y\hat{f}(y) = \frac{\sum_{i\in\mathcal{S}} w_i y_i}{\sum_{i\in\mathcal{S}} w_i}$$

and

$$\hat{S}^2 = \frac{N}{N-1}\left[\sum_y y^2 \hat{f}(y) - \left\{\sum_y y\hat{f}(y)\right\}^2\right].$$

Since the empirical cdf $\hat{F}$ is a step function, we usually interpolate to find a unique value for the quantile.
Let y_1 be the largest value in the sample for which $\hat{F}(y_1) \le q$ and let y_2 be the smallest value in the sample for which $\hat{F}(y_2) \ge q$. Then

$$\hat{\theta}_q = y_1 + \frac{q - \hat{F}(y_1)}{\hat{F}(y_2) - \hat{F}(y_1)}(y_2 - y_1).$$

Example

Consider an artificial population of 1000 men and 1000 women. Each person's height is measured to the nearest centimeter.

- SRS of size 200: each person in the sample represents $w_i = 10$ persons in the population.
- Stratified sample of 160 women and 40 men. In the stratified sample, each woman has weight $1000/160 = 6.25$ and each man has weight $1000/40 = 25$.

Quantity	Population	SRS	Stratified with Weights
Mean	168.6	168.9	169.0
Median	167.3	168.8	167.6
25th percentile	159.9	159.7	160.7
90th percentile	183.2	183.4	181.5
Variance, S^2	124.5	122.6	116.8

Plotting data

- Self-weighting: To construct a relative frequency histogram for an SRS of size n, divide the range of the data into k bins with each bin having width b. Then the height of the histogram in the jth bin is

$$\text{height}(j) = \frac{\text{relative frequency for bin } j}{b} = \frac{\sum_{i \in \mathcal{S}} u_i(j)}{bn},$$

where $u_i(j) = 1$ if observation i is in bin j and 0 otherwise. If a sample is self-weighting, as with an SRS, a regular histogram of the sample data will estimate the population probability mass function.

- If a sample is not self-weighting a histogram of the raw data may underrepresent some parts of the population in the display. We can use the sampling weights to construct a histogram.

Divide the range of the data into k bins with each bin having width b. Now use the sampling weights w_i to find the height of the histogram in bin j:

$$\text{height}(j) = \frac{\sum_{i \in \mathcal{S}} w_i u_i(j)}{b \sum_{i \in \mathcal{S}} w_i}.$$

Dividing by the quantity $b \sum_{i \in \mathcal{S}} w_i$ ensures that the total area under the histogram equals 1.

Density estimates

Smoothed density estimates are useful for displaying the shape of the estimated population data for a variable that takes on a wide range of values.

The kernel density estimation to survey data by incorporating the weights, with

$$\hat{f}(y;b) = \frac{1}{b\sum_{i\in\mathcal{S}} w_i} \sum_{i\in\mathcal{S}} w_i K\left[\frac{y-y_i}{b}\right].$$

Commonly used kernel functions include the normal kernel function $K_N(t) = \exp(-t^2/2)/\sqrt{2\pi}$ and the quadratic kernel function $K_Q(t) = \frac{3}{4}(1-t^2)$ for $|t| < 1$. The sliding histogram described above corresponds to a box kernel with $K_B(t) = 1$ for $|t| \leq 1/2$ and $K_B(t) = 0$ for $|t| > 1/2$; in that case, $\hat{f}(y;b)$ corresponds to the histogram height for a point y in the middle of a bin of width b.

The choice of b, called the bandwidth, determines the amount of smoothing to be used. Small values of b use little smoothing since the sliding window is small. A large value of b provides much smoothing since each point in the plot represents the weighted average of many points from the data.

Note that R uses local linear smoother with Gaussian kernel weights.

Design Effects

Design effect: the effect of the design on the variance of the estimator.

$$\text{deff(plan,statistic)} = \frac{\text{V(estimator from sampling plan)}}{\text{V(estimator from an SRS with same number of observation units)}}.$$

For estimating a mean from a sample with n observation units,

$$\text{deff}(\text{plan},\hat{\bar{y}}) = \frac{V(\hat{\bar{y}})}{\left(1-\frac{n}{N}\right)\frac{S^2}{n}}.$$

The design effect provides a measure of the precision gained or lost by use of the more complicated design instead of an SRS.

If estimating a proportion, the SRS variance is approximately $p(1-p)/n$; if estimating another type of mean, the SRS variance is approximately S^2/n. So if the design effect is approximately known, the variance of the estimator from the complex sample can be estimated by (deff $\times$ SRS variance).

We can estimate the variance of an estimated proportion $\hat{p}$ by

$$\hat{V}(\hat{p}) = \text{deff} \times \frac{\hat{p}(1-\hat{p})}{n}.$$

Design effect

- Stratified sampling with proportional allocation.

$$\frac{V_{\text{prop}}}{V_{\text{SRS}}} \approx \frac{\sum_{h=1}^{H} \frac{N_h}{N} S_h^2}{S^2}$$

$$\approx \frac{\sum_{h=1}^{H} \frac{N_h}{N} S_h^2}{\sum_{h=1}^{H} \frac{N_h}{N} [S_h^2 + (\bar{y}_{Uh} - \bar{y}_U)^2]}.$$

- Cluster sampling. The design effect for single-stage cluster sampling when all psus have M ssus is approximately

$$1 + (M - 1)\text{ICC}.$$

The intraclass correlation coefficient (ICC) is usually positive in cluster sampling, so the design effect is usually larger than 1; cluster samples usually give less precision per observation unit than an SRS.

Design Effects and Confidence Intervals

If n observation units are sampled from a population of N possible observation units and if $\hat{p}$ is the survey estimate of the proportion of interest, an approximate 95% CI for p is (assuming the finite population correction is close to 1):

$$\hat{p} \pm 1.96\sqrt{\text{deff}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

When estimating a mean rather than a proportion, if the sample is large enough to apply a central limit theorem, an approximate 95% CI is

$$\hat{\bar{y}} \pm 1.96\sqrt{\text{deff}}\sqrt{\frac{\hat{S}^2}{n}},$$