BIOS 662 Fall 2018

Survey Sampling, Part I

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

Outline

- Preliminaries
- Simple random sampling
  - Population mean
  - Population total
  - Sample size
  - Proportion

Preliminaries: References

- SK Thompson. *Sampling*. John Wiley and Sons, 1992
- L Kish. *Survey sampling*. Wiley, New York, 1965
- WG Cochran. *Sampling Techniques*. John Wiley and Sons, 1977

Preliminaries: What is "(Survey) Sampling"?

- *Sampling* study: Selecting some part of a population to be observed so that one may estimate something about the whole of the population
- Example: To estimate the amount of lichen in a well-defined area, a biologist collects lichen from selected small plots within the study area
- Typically want to estimate total or mean
- Observational – does not intentionally perturb or disturb population (i.e., not experimental)
- However, one does have control over how the sample is selected

Preliminaries: Terminology

- *Population*: The group of units (e.g., people) we are sampling and studying; assumed to be of known, finite size, rather than infinite
- *Sampling design*: The strategy followed in selecting a sample from a population
- *Sampling unit*: Unit designated for listing and selection in a sample survey (e.g., persons, dwellings, households, area units, pharmacies)
- *Sampling frame*: List of sampling units from which a sample is drawn

Preliminaries: Terminology

- *Variable*: Some measurement taken on members of the sample (e.g., number of children ever born to a woman aged 15-49 years); sometimes call this the y-variable or x-variable
- *Selection probability*: Likelihood, over repeated applications of a sampling design, that a particular unit will be chosen for a sample
- *Probability sampling*
  - Sampling in which the design calls for using random methods to ultimately decide which units are chosen
  - Every unit has a known, non-zero selection probability

Preliminaries: Terminology

- *Equal-probability sampling*
  - Probability sampling in which all units in the population have the same selection probability
  - Also known as "self-weighted" sampling or "epsem" (equal probability of selection method) sampling
- *Non-probability sampling*
  - Sampling in which subjective judgment (usually by interviewers) is used to decide who is chosen in the sample
  - Selection probabilities cannot be determined
  - Difficult to determine if the sample is representative

Preliminaries: Terminology

- *Unbiased* estimator: An estimator which, if repeated over all possible samples that might be selected using the sampling design, would yield estimates which on average equal the parameter being estimated (e.g., sample mean from a simple random sample is an unbiased estimator of the population mean)
- Also known as *design-unbiased*
- Key idea: the randomness in the estimator is induced by the sampling design

Preliminaries: Software

- SAS: proc surveymeans, surveyfreq, ...
- R: "survey" package

Preliminaries: Sampling Designs

- Simple random sampling
- Stratified sampling
- Cluster sampling

Simple Random Sampling (SRS)

- Let N denote the number of units in the population
- *Simple random sampling*, or random sampling *without replacement* (SRSWOR), is the sampling design in which n distinct units are selected from the N units in the population in such a way that every possible combination of the n units is equally likely to be the sample selected
- A simple random sample can be obtained through a sequence of independent selections from the whole population such that each unit has an equal probability of selection at each step, discarding repeat selections and continuing until n distinct units are obtained
- $f \equiv n/N$ is the sampling rate or sampling fraction

Obtaining a Simple Random Sample

A. Number the units in the population (i.e., sampling frame) from 1 to N.

B. Select and record a random number between 1 and N.

C. At each subsequent step, select a random integer between 1 and N. If it is the same as a previously selected number, discard it. Otherwise, record it.

D. Continue in this manner until n different numbers between 1 and N have been chosen.

E. Population units corresponding to the selected numbers form a simple random sample of size n.

Obtaining a Simple Random Sample

Alternative approach:

A. Generate a random number from $U(0, 1)$ for each unit in the population (i.e., sampling frame).

B. Sort in order of the random numbers.

C. Take the first n units in the sorted list.

Key Properties of SRS

- All possible simple random samples have the same chance of being selected
- The probability that any one population unit will be chosen is n/N
- Selection probabilities in an SRS are not statistically independent

$$\Pr[i \text{ in sample}] = \frac{\binom{1}{1}\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{n}{N}$$

$$\Pr[i \text{ and } j \text{ in sample}] = \frac{\binom{2}{2}\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)} \neq \left(\frac{n}{N}\right)^2$$

SRS: Estimating Population Mean

- Denote the (finite) population mean by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$$

- Denote the (finite) population variance by

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2$$

- Let Z_i indicate whether unit i is in the sample, that is $Z_i = 1$ if i is sampled, $Z_i = 0$ otherwise
- Key point: The y_i are fixed, the Z_i are random

SRS: Estimating Population Mean

- Sample mean

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{N} y_i Z_i$$

- Sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{N} (y_i - \bar{y})^2 Z_i$$

- The sample mean is unbiased: Each Z_i is Bernoulli with $E(Z_i) = n/N$, thus

$$E(\bar{y}) = \frac{1}{n} \sum_{i=1}^{N} y_i E(Z_i) = \frac{1}{N} \sum_{i=1}^{N} y_i = \mu$$

SRS: Estimating Population Mean

- To derive the variance of the sample mean,

$$\mathrm{Var}(\bar{y}) = \frac{1}{n^2}\left(\sum_{i=1}^{N} y_i^2 \mathrm{Var}(Z_i) + \sum_{i \neq j} y_i y_j \mathrm{Cov}(Z_i, Z_j)\right)$$

  we need the variance and covariance terms

- The variance is easy because the Z_i are Bernoulli

$$\mathrm{Var}(Z_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right)$$

- For SRS, the Z_i are not independent

$$\begin{aligned}\mathrm{Cov}(Z_i, Z_j) &= E(Z_i Z_j) - E(Z_i)E(Z_j) = \frac{n}{N}\frac{(n-1)}{(N-1)} - \left(\frac{n}{N}\right)^2 \\ &= -\frac{n}{N}\left(1 - \frac{n}{N}\right)\frac{1}{N-1}\end{aligned}$$

SRS: Estimating Population Mean

- Thus

$$\mathrm{Var}(\bar{y}) = \frac{1}{n^2}\left(\frac{n}{N}\right)\left(1-\frac{n}{N}\right)\left(\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i\neq j} y_i y_j\right)$$

- Using the identity

$$\sum_{i=1}^{N}(y_i-\mu)^2 = \sum_{i=1}^{N} y_i^2 - \frac{(\sum y_i)^2}{N} = \frac{1}{N}\left((N-1)\sum_{i=1}^{N} y_i^2 - \sum_{i\neq j} y_i y_j\right)$$

we get

$$\mathrm{Var}(\bar{y}) = \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{\sum(y_i-\mu)^2}{N-1} = \left(1-\frac{n}{N}\right)\frac{\sigma^2}{n}$$

SRS: Estimating Population Mean

- The quantity

$$1 - \frac{n}{N} = \frac{N-n}{N} = 1 - f$$

  is called the *finite population correction factor*

- If the population is large relative to the sample size, n/N will be small, so that

$$\text{Var}(\bar{y}) \approx \frac{\sigma^2}{n}$$

- On the other hand, $\text{Var}(\bar{y}) \to 0$ as $n \to N$

SRS: Estimating Population Variance

- Exercise: Show that $E(s^2) = \sigma^2$, i.e., the sample variance, is an unbiased estimator for the finite population variance
- From this fact, it follows that an unbiased estimator for $\mathrm{Var}(\bar{y})$ is given by

$$\widehat{\mathrm{Var}}(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{s^2}{n}$$

SRS: Estimating Population Total

- Define the population total

$$\tau = \sum_{i=1}^{N} y_i = N\mu$$

- Unbiased estimator

$$\hat{\tau} = N\bar{y} = \frac{N}{n}\sum_{i=1}^{N} y_i Z_i$$

  with variance

$$\mathrm{Var}(\hat{\tau}) = N^2\mathrm{Var}(\bar{y}) = N(N-n)\frac{\sigma^2}{n}$$

- Unbiased estimator of variance

$$\widehat{\mathrm{Var}}(\hat{\tau}) = N^2\widehat{\mathrm{Var}}(\bar{y}) = N(N-n)\frac{s^2}{n}$$

SRS: Estimating Population Total

- The estimator is often written as

$$\hat{\tau} = \sum_{i=1}^{N} w_i y_i Z_i = \sum_{i=1}^{N} \frac{y_i Z_i}{\pi_i}$$

  where $w_i^{-1} = \pi_i = n/N = f$ is the selection probability

- "Inverse probability weighting"
- This is the formulation SAS uses (more below)
- Special case of the *Horvitz-Thompson* estimator

  (Horvitz, D.G. and Thompson, D.J. (1952) A generalization of sampling without replacement from a finite universe. J. Amer. Stastist. Assoc., 47, 663-685.)

SRS: Finite-population CLT

- The usual central limit theorem requires independence
- Imagine a sequence of populations with population size N becoming large along with sample size n. Let μ_N be the population mean and $\bar{y}_N$ the sample mean for an SRS from that population.
- According to the finite-population CLT

$$\frac{\bar{y}_N - \mu_N}{\sqrt{\text{Var}(\bar{y}_N)}} \to Z \sim N(0,1)$$

as both $n \to \infty$ and $N - n \to \infty$

SRS: Finite-population CIs

- This leads to approximate $100(1-\alpha)\%$ CIs for the population mean μ

$$\bar{y} \pm t_{n-1,1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}$$

- Likewise, for the population total τ

$$\hat{\tau} \pm t_{n-1,1-\alpha/2}\sqrt{N(N-n)\frac{s^2}{n}}$$

SRS: Example

- (Section 2.3, Thompson 1992)
- A survey of the caribou population was done by aircraft
- A 286 mile wide study region was divided into exhaustive and mutually exclusive one-mile strips ($N = 286$)
- An SRS of $n = 15$ yielded counts of 1, 2, 4, 4, 5, 7, 10, 15, 21, 21, 29, 36, 50, 86, 98
- Sample mean $\bar{y} = 25.933$
- Sample variance $s^2 = 919.067$

SRS: Example cont.

- Thus

$$\widehat{\mathrm{Var}}(\bar{y}) = \left(1 - \frac{15}{286}\right)\frac{919.067}{15} = 58.058$$

  yielding 95% CI for μ

$$25.933 \pm 2.145\sqrt{58.058} = (9.59, 42.28)$$

- For the population total, $\hat{\tau} = N\bar{y} = 7417$, with 95% CI

$$7417 \pm 2.145\sqrt{286(286-15)\frac{919.067}{15}} = (2743, 12091)$$

SRS: Example Using SAS

```
proc means mean clm sum;  *** Assuming infinite population;
   var counts;
```

Mean	Lower 95% CL for Mean	Upper 95% CL for Mean	Sum
25.9333333	9.1448299	42.7218368	389.0000000

```
proc surveymeans total=286 mean clm sum clsum;
   var counts;
```

Variable	Mean	Std Error of Mean	95% CL for Mean	
counts	25.933333	7.619553	9.59101702	42.2756497

Variable	Sum	Std Dev	95% CL for Sum	
counts	389.000000	114.293298	143.865255	634.134745

SRS: Example Using SAS

```
proc surveymeans total=286 mean clm sum clsum;
   var counts;
   weight wt;  *** wt=286/15 for all;
```

```
                          Std Error
Variable         Mean       of Mean      95% CL for Mean
-----------------------------------------------------------
counts      25.933333     7.619553   9.59101702 42.2756497
-----------------------------------------------------------

Variable           Sum      Std Dev       95% CL for Sum
-----------------------------------------------------------
counts    7416.933333  2179.192221   2743.03087 12090.8358
-----------------------------------------------------------
```

SRS: Example Using R

```
> library("survey")

> caribou <- data.frame(y=c(1,2,4,4,5,7,10,15,21,21,29,36,50,86,98),fpc=15/286)

> design <- svydesign(ids=~1,data=caribou,fpc=~fpc)

> # R uses Z not t for the confidence intervals here
> svymean(caribou$y, design); confint(svymean(caribou$y, design))
         mean     SE
y   25.933333 7.6196

           2.5 %      97.5 %
y   10.99928344 40.86738323


> svytotal(caribou$y, design); confint(svytotal(caribou$y, design))
    total     SE
y    7417 2179.2

       2.5 %   97.5 %
y   3145.795 11688.07
```

SRS: Sample Size

- Suppose we want to choose the smallest sample size n such that

$$\Pr[|\hat{\theta} - \theta| > d] \leq \alpha$$

- Assuming

$$\frac{\hat{\theta} - \theta}{\sqrt{\text{Var}(\hat{\theta})}} \sim N(0, 1)$$

choose n such that

$$z_{1-\alpha/2}\sqrt{\text{Var}(\hat{\theta})} = d$$

SRS: Sample Size

- For example, if $\theta = \mu$, choose n such that

$$z_{1-\alpha/2}\sqrt{\left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}} = d$$

- This implies

$$n = \frac{1}{1/n_0 + 1/N} \quad \text{where} \quad n_0 = \frac{z_{1-\alpha/2}^2\,\sigma^2}{d^2}$$

- Note that if $N >> n$, then $n \approx n_0$

SRS: Sample Size

- If $\theta = \tau$, choose n such that

$$z_{1-\alpha/2}\sqrt{N(N-n)\frac{\sigma^2}{n}} = d$$

implying

$$n = \frac{1}{1/n_0 + 1/N} \quad \text{where} \quad n_0 = \frac{N^2 z_{1-\alpha/2}^2 \sigma^2}{d^2}$$

- Example: Find n necessary to estimate the caribou population total to within 2,000 animals of the true total with 90% confidence

Here $\sigma^2 = 919$, $d = 2000$, $\alpha = 0.1$

So $n_0 = 50.9$, $n = 43.2$

SRS: Estimating a Proportion

- Suppose responses are binary, e.g., want to estimate the proportion of voters favoring a candidate for elected office
- Let $y_i = 1$ if unit i has the attribute of interest, $y_i = 0$ otherwise
- Then μ is the proportion of units in the population with the attribute
- Thus can use methods from before. However, there are some special features now:
  - Formulas simplify considerably
  - Exact confidence intervals are possible
  - Sample size calculation does not require information about population parameters

SRS: Estimating a Proportion

- Let the proportion of the population with the attribute be

$$p = \frac{1}{N}\sum_{i=1}^{N} y_i = \mu$$

- Finite population variance

$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^{N}(y_i - p)^2}{N-1} = \frac{\sum_{i=1}^{N} y_i^2 - Np^2}{N-1} \\ &= \frac{Np - Np^2}{N-1} \\ &= \frac{Np(1-p)}{N-1}\end{aligned}$$

SRS: Estimating a Proportion

- Proportion in the sample with the attribute

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{N} y_i Z_i = \overline{y}$$

- Sample variance

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2 Z_i}{n-1} \\ &= \frac{\sum_{i=1}^{N} y_i^2 Z_i - n\hat{p}^2}{n-1} \\ &= \frac{n\hat{p}(1-\hat{p})}{n-1} \end{aligned}$$

SRS: Estimating a Proportion

- Because the sample proportion is a sample mean of an SRS, all previous results hold; in particular:

$$E(\hat{p}) = p$$

$$\mathrm{Var}(\hat{p}) = \left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n}$$

$$\widehat{\mathrm{Var}}(\hat{p}) = \left(\frac{N-n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}$$

$$E(\widehat{\mathrm{Var}}(\hat{p})) = \mathrm{Var}(\hat{p})$$

SRS: CI for a Proportion

- Approximate $100(1-\alpha)\%$ CI

$$\hat{p} \pm t_{1-\alpha/2,n-1}\sqrt{\widehat{\mathrm{Var}}(\hat{p})}$$

- The approximation improves as n increases and the closer p is to 0.5
- An exact CI can be computed based on inverting a test and the hypergeometric distribution

SRS: Exact CI for a Proportion

- Suppose ν units in the population have the attribute of interest
- Let

$$X = \sum_{i=1}^{N} y_i Z_i$$

- Then

$$\Pr[X = j|\nu] = \frac{\binom{\nu}{j}\binom{N-\nu}{n-j}}{\binom{N}{n}}$$

SRS: Exact CI for a Proportion

- Suppose we observe $X = x$ for one particular SRS (such that $\hat{p} = x/n$)
- Let ν_L be the smallest integer such that

$$\Pr[X \geq x | \nu_L] > \alpha/2$$

and let ν_U be the largest integer such that

$$\Pr[X \leq x | \nu_U] > \alpha/2$$

- Then an exact $100(1-\alpha)\%$ CI is given by

$$(\nu_L/N, \nu_U/N)$$

SRS: Sample Size for a Proportion

- To obtain an estimator $\hat{p}$ having probability at least $1-\alpha$ of being no farther than d from the population proportion

$$n = \frac{Np(1-p)}{(N-1)d^2/z^2_{1-\alpha/2} + p(1-p)}$$

- If $N >> n$

$$n \approx \frac{z^2_{1-\alpha/2}\, p(1-p)}{d^2}$$

- If no a-priori knowledge of p, conservatively assume $p = 0.5$

SRS: Estimating a Ratio

- Example 1: A biologist studying an animal population selects an SRS of plots in the study region. In each selected plot, she counts the number y_i of young animals and the number x_i of adult females, with the object of estimating the ratio of young to adult females in the population
- Example 2: In a household survey to estimate the number of television sets per person in the region, an SRS of households is conducted. For each selected household the number y_i of television sets and the number x_i of people is recorded

SRS: Estimating a Ratio

- Ratio estimator

$$r = \frac{\sum_{i=1}^{N} y_i Z_i}{\sum_{i=1}^{N} x_i Z_i} = \frac{\bar{y}}{\bar{x}}$$

- Note that the denominator of the estimator is a random variable

SRS: Concluding Remarks

- SRS is the simplest probability sampling method
- It is quite rarely used in practice
- Exception: When the sample size and population size are small and stratified sampling is not possible