SAMPLING
BIOS 662

Michael G. Hudgens, Ph.D.

mhudgens@bios.unc.edu

http://www.bios.unc.edu/∼mhudgens

2008-11-14 15:55

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
- Eg, 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.
- *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
- *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

Preliminaries: Terminology

- *Selection probability*: Likelihood over repeated applications of 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, nonzero selection probability
- *Equal-Probability Sampling*
  - Probability sampling in which all units in the population have the same selection probability
  - AKA "self-weighted" sampling or "epsem" (equal probability of selection method) sampling

Preliminaries: Terminology

- *Non-probability Sampling*
  - Sampling in which subjective judgment (usually by interviewers) is used to ultimately decide who is chosen in the sample
  - Selection probabilities cannot be determined
  - Difficult to determine if sample is representative (i.e., includes members from all relevant segments of the population)

Preliminaries: Terminology

- *Unbiased* estimator: An estimator which, if repeated over all possible samples that might be selected using a 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)
- AKA *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*, 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 (SRSWOR)
- SRS sample can be obtained through a sequence of independent selections from the whole population where each unit has an equal probability of selection at each step, discarding repeat selections and continuing until n distinct units are obtained
- $f \equiv n/N$ sampling rate

Obtaining an SRS 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. Select a second random number between 1 and N. If this number is the same as the first selected number, discard it. Otherwise, record it.
- D. Select another random number between 1 and N. If this number is the same as a previously selected number, discard it. Otherwise, record it.
- E. Continue in this manner until n different numbers between 1 and N have been chosen.
- F. Population units corresponding to selected numbers are an SRS sample of size n.

Key properties of SRS

- All possible SRS 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.

SRS: Estimating population mean

- Denote (finite) population mean by

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

- Denote (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 sample with $Z_i = 1$ if sampled, $Z_i = 0$ otherwise
- Key: y_i's are fixed, Z_i's are random

SRS: Estimating population mean

- Sample mean

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

- Sample variance

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

- Sample mean 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 variance of sample mean,

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

  we need var and cov terms

- Variance easy since Z_i Bernoulli

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

- For SRS, Z_i's not independent

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

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

SRS: Estimating population mean

- Thus

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

- 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

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

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, such that
$$Var(\bar{y}) \approx \frac{\sigma^2}{n}$$
- On the other hand, $Var(\bar{y}) \to 0$ as $n \to N$

SRS: Estimating population variance

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

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

SRS: Estimating population total

- Define 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

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

- Unbiased estimator of variance

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

SRS: Estimating population total

- Often estimator 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

SRS: Finite-population CLT

- Usual CLT 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 $\bar{y}_N$ be the sample mean for a SRS from that population.
- According to the finite-population CLT

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

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

SRS: Finite-population CIs

- This leads to approximate $(1-\alpha)\times 100\%$ 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 population total τ

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

SRS: Example

- Example (Section 2.3 Thompson 1992) Survey of caribou population via aircraft. A 286 mile wide study region divided into exhaustive and mutually exclusive one-mile strips ($N = 286$). A 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 $s^2 = 919.067$
- Thus

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

yielding 95% CI for μ

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

SRS: Example using SAS

```
proc surveymeans total=286;
   var counts;
run;
```

```
                                                    Std Error
Variable                 N            Mean          of Mean
-----------------------------------------------------------
counts                  15       25.933333         7.619553
-----------------------------------------------------------
```

```
           Statistics

Variable       95% CL for Mean
---------------------------------
counts      9.59101702 42.2756497
---------------------------------
```

SRS: Example with SAS

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

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

- SAS

```
proc surveymeans total=286 sum clsum;
  var counts;
  weight wt;  * equals N/n=286/15;
run;
```

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

SRS: Sample Size

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

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

- Assuming

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

  choose n such that

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

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

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

- Note 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

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

- Example: Find n necessary to estimate caribou population total to within 2000 animals of true total with 90% confidence.

  $\sigma^2 = 919$, $d = 2000$, $\alpha = .1 \rightarrow n_0 = 50.9$, $n = 43.2$

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 attribute of interest, $y_i = 0$ otherwise
- Then μ is the proportion of units in the populations with the attribute
- Thus can use methods from before. However, certain special features:
  - Formulae simplify considerably
  - Exact confidence intervals are possible
  - Sample size does not require information about population parameters

SRS: Estimating a proportion

- Let proportion of population with attribute be

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

- Finite population variance

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

- Proportion in sample with attribute

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

- Sample variance

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

SRS: Estimating a proportion

- Since sample proportion is the sample mean of a SRS, all previous results hold. In particular:

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

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

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

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

SRS: CI for a proportion

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

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

  approximation improves as n increases and the closer p is to 0.5

- Exact CI can be computed based on inverting a test and the hypergeometric distribution

SRS: 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

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

- 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 exact $(1-\alpha)\%$ CI given by $(\nu_L/N, \nu_U/N)$

SRS: SS 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 $p = .5$

SRS: Estimating a ratio

- Example: Biologist studying animal population selects a 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: 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
- 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 denominator of estimator is a random variable

SRS: Concluding remarks

- SRS is the simplest probability sampling method
- Only rarely used in practice. Exception: sample size and population small, stratified sampling not possible