Statistics 522: Sampling and Survey Techniques

Topic 7

Topic Overview

This topic will cover

- Complex surveys
- Examples
- Plots
- Sampling and experimental design

Building blocks for surveys

- Cluster sampling with replacement
- Cluster sampling without replacement
- Stratification
- Ratio estimation
- Weights used to find estimates; computer intensive methods used to SE's.

Cluster sampling with replacement

- Select a sample of n clusters with replacement.
- Cluster i is selected with probability ψ_i.
- Estimate the total for cluster i by an unbiased estimate $\hat{t}_i$.
- Treat the n values of $u_i = \hat{t}_i/\psi_i$ as observations.
- Estimate the population total by $\bar{u}$.
- Estimate the variance of the estimated total by s_u^2/n.

Cluster sampling without replacement

- Select a sample of n clusters without replacement.
- Cluster i is selected in the sample with probability π_i.
- Estimate the total for cluster i by an unbiased estimate $\hat{t}_i$.
- Use the Horvitz-Thompson estimate of the population total

$$\hat{t}_{HT} = \sum \hat{t}_i/\pi_i$$

- Use an exact formula from Chapter 5 or 6, or a method from Chapter 9 to estimate the variance.

Stratification

- Estimate the strata totals by $\hat{t}_1$, $\hat{t}_2$, . . . , $\hat{t}_H$.
- The estimated variances for the strata totals are $\hat{\mathrm{V}}(\hat{t}_1)$, $\hat{\mathrm{V}}(\hat{t}_2)$, . . . , $\hat{\mathrm{V}}(\hat{t}_H)$.
- The estimate of the population total is $\hat{t} = \sum \hat{t}_i$.
- The estimate of the variance is $\hat{\mathrm{V}}(\hat{t}) = \sum \hat{\mathrm{V}}(\hat{t}_i)$.

Ratio estimation

- Let $\hat{t}_x$ and $\hat{t}_y$ be estimators of t_x and t_y, respectively .
- The ratio is estimated by $\hat{B} = \hat{t}_y/\hat{t}_x$.
- The estimated variance is

$$\hat{\mathrm{V}}(\hat{B}) = \frac{\hat{B}}{t_x^2}\hat{\mathrm{V}}(\hat{t}_x) + \frac{1}{t_x^2}\hat{\mathrm{V}}(\hat{t}_y) - 2\frac{\hat{B}}{t_x^2}\hat{\mathrm{Cov}}(\hat{t}_x, \hat{t}_y).$$

- Details are given in Section 9.1
- The ratio estimator of the population total is $\hat{B}t_x$.
- The estimated variance is $t_x^2\hat{\mathrm{V}}(\hat{B})$.

Comments

- We often use ratio estimators for means, letting the auxiliary variable x be an indicator (1 or 0) variable for whether or not unit i is in the sample.
- Here, $\hat{t}_x$ is an estimator of the population size and the ratio is the estimate of the population total divided by the estimated population size.

Malaria in The Gambia

- Malaria is a major health problem.
- Bed nets impregnated with insecticide can be effective in prevention of this disease.
- A sample survey was used to estimate the prevalence of bed net use in rural areas.

The frame and stratification

- All rural villages of fewer than 3000 people in The Gambia.
- Districts were stratified by three geographic regions: eastern, central, and western.
- Villages were stratified based on presence of a public health clinic (PHC).

Regions

In each region (eastern, central, western), five districts were chosen with probability proportional to the district population (used the 1983 census).

Districts

In each district, four villages were chosen, with probability proportional to the 1983 census population.

- two PHC (public health clinic) villages
- two non-PHC villages

Compounds

- Six compounds were chosen *more or less randomly* from each village.
- The number of beds with and without nets were recorded (along with other information).

Three stages

- Stage 1
  - select districts stratified by region
- Stage 2
  - select villages stratified by PHC or not
- Stage 3
  - select compounds

Data – compound

- Record the total number of nets for each compound.
- Estimate the total number of nets for each village (number of compounds times the average number of nets per compound).
- Find the estimated variance of the total for each village.

PHC/non-PHC villages

- Estimate the total nets for PHC villages in each district.
- Sampling was proportional to population so use Chapter 6 methods for estimate of total and its variance.
- Do same for non-PHC villages.

Districts

- Add the estimates for the two strata (PHC and non-PHC) to get estimates for each district sampled.
- Variances add.

Region

- We have estimated total nets and variance for each district.
- Use two-stage cluster methods to estimate total nets for each region.

The Gambia

- Add estimated totals for each region to estimate the total for the country.
- Variances add for stratification.

Ratio estimation

- If we are interested in the proportion of beds with a net, we would use a ratio estimator, incorporating number of beds x against number of nets y.
- Could be done at different levels
  - Compound
  - Village
  - District

  - Region
  - Country

- Combine across strata

$$\begin{aligned}\hat{B} &= \sum_h \left(\frac{N_h}{N}\right)\frac{\hat{t}_{y,h}}{\hat{t}_{x,h}} \\ \hat{t}_y &= \sum_h \frac{t_{x,h}\hat{t}_{y,h}}{\hat{t}_{x,h}}\end{aligned}$$

- Works well if sample size large, $\frac{\hat{t}_{y,h}}{\hat{t}_{x,h}}$ varies across strata.

Weights

- Weight is the reciprocal of the probability that the observation unit is selected to be in the sample.
- Weights determine the estimates.
- Variances can depend on more knowledge of the sampling design (probabilities of pairs).
- For stratified sampling $w_{h,j} = N_h/n_h$.
- For cluster sampling with equal probabilities $w_{i,j} = \frac{NM_i}{nm_i} = \frac{1}{j\text{th ssu in } i\text{th psu in sample}}$.
- Basic idea: $\hat{t}_y = \sum w_i y_i$, $\hat{\bar{y}} = \frac{\hat{t}_y}{\sum w_i}$.

Self-weighting samples

- Weights for each observation unit is the same.
  - if clusters have different sizes, this means pps.
  - often yield smaller SE's.
- Standard methods for histograms, means, medians, quantiles are valid.
- Standard methods for standard errors are NOT
  - we do not have iid observations

Non-self-weighting samples

- Disproportionate sampling probabilities often occur with stratification.
  - sample a higher proportion of large businesses
  - National Health and Nutrition Examination Survey (NHANES) oversamples blacks and Mexican-Americans.
  - has more to do with optimal allocation.

Estimating a distribution function

- Often interested in more complicated statistics than means, totals
  - Example: 95th quantile
  - Weights help this process
- Probability mass function

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

- Distribution function

$$F(y) = \frac{\text{number of units } \leq y}{N}$$

- Means, quantiles, measures of variability, etc. can all be computed from these quantities.
  - Example: mean – $\bar{y}_u = \sum y f(y)$

Example 7.3 (page 230)

- Consider an artificial population of heights of 1000 men and 1000 women.
- There are in the data set `htpop.dat`.
- (There are also files `htsrs.dat` and `htstrat.dat`.)
- These are comma-delimited files with the variable names in the first record.

Import the data and check it (`SLL230.sas`)

- In SAS, use file; import data; delimited file; browse to find file; options to specify comma as the delimiter; specify data set name (`a1`).
- Use `proc print` to check the data.
- Variables are *height* (cm) and *gender*.

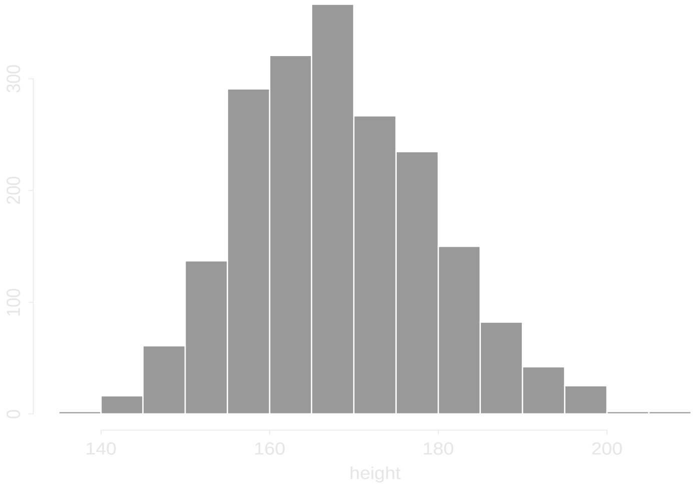
Output

```
Obs           HEIGHT    GENDER
 1              173      F
 2              163      F
 3              160      F
 4              148      F
 5              160      F
```

Generate a histogram

```
proc univariate data=a1;
  var height;
  histogram height/normal;
run;
```

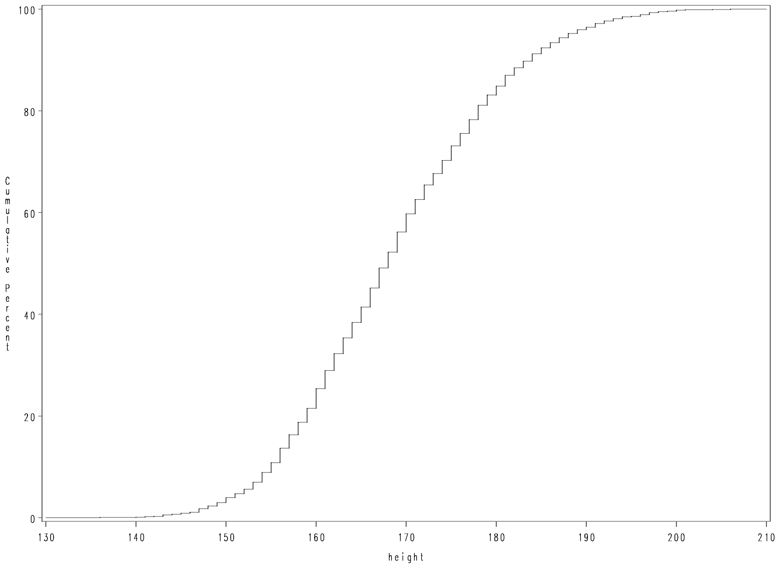
Output



Plot the CDF

```
proc capability data=a1;
   var height;
   cdfplot height;
run;
```

Output



An SRS

- Suppose we take an SRS of size 200 from our artificial population ($N = 2000$)
- The sample is self-weighting.
- Each person in the sample represents 10 people in the population.
- The SRS is in the data set `htsrs.dat`.

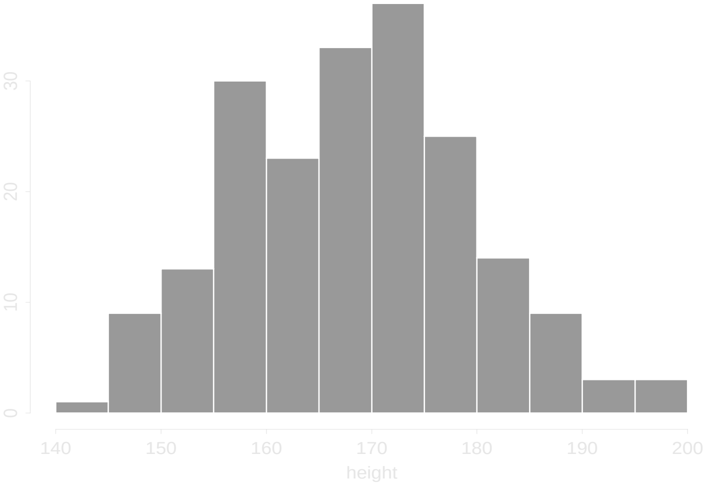
A stratified sample

- Suppose we took a stratified sample of size 200 with 160 women and 40 men.
- The sample is not self-weighting.
- In the sample each woman represents 1000/160=6.25 women and each man represents 1000/40=25 men.
- The stratified sample is in the data set `htstrat.dat`.

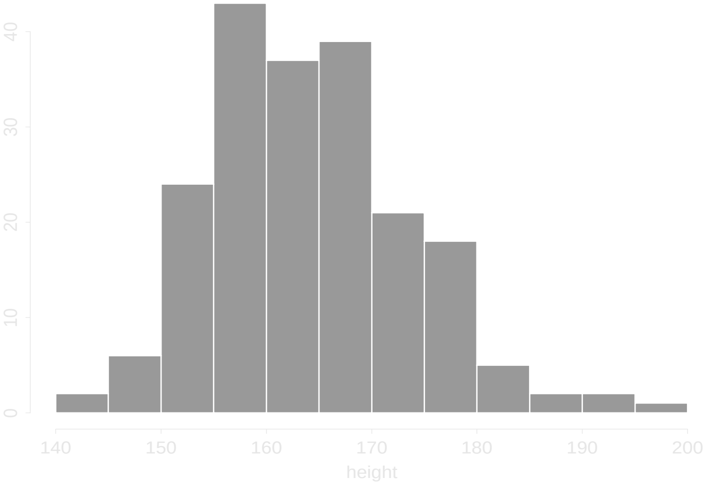
Estimates

- For the SRS, the usual estimates of mean, median and the distribution will be valid.
- For the stratified sample, the usual estimates will be biased because males are underrepresented in the sample (and presumably they tend to be taller).

Histogram for the SRS



Histogram for the stratified sample



Weights

- We can use weights to adjust a non-self-weighting sample to obtain estimates of population characteristics that are based on the distribution.
- This method does *not* work to do inference: for example, to compute standard errors.

Method

- Use the weights to estimate $f(y)$ and $F(y)$.

$$\hat{F}(x) = \frac{\sum_{y_i \le x} w_i}{\sum w_i}$$

- Use these estimates to construct estimates of any population characteristic.
  - Median
  - Quantiles
  - Variance
  - Standard deviation
- Can used smoothed versions
- See pages 230-234

Plots

- There is no one representative plot.
- Try everything. (For example: with weights; without weights)
- For stratified samples, plot the distributions for the strata using side by side boxplots (page 237).
  - You can use `proc univariate` with the `plot` option and a `by` statement.
- See Section 7.4 for some other plots.

Design effects

- Consider two plans
  - an SRS
  - an alternative
- V_1 = the variance of the SRS
- V_2 = the variance of the alternative
- The design effect is $deff = \frac{V_2}{V_1}$.

For an estimate of the mean

- SRS: $V_1 = \left(1 - \frac{n}{N}\right) \frac{S^2}{n}$
- The alternative: $V_2 = \mathrm{V}(\hat{\bar{y}})$

Stratified sampling with proportional allocation

- $Num = \sum_h \frac{N_h}{N} S_h^2$
- $Den = \sum_h \frac{N_h}{N}(S_h^2 + (\bar{y}_{h,U} - \bar{y}_U)^2)$
- The extent to which stratified sampling is better depends on the size of the terms $(\bar{y}_{h,U} - \bar{y}_U)^2$.
- Generally, $deff \leq 1$; stratified is better than SRS.

Cluster sampling

- For single stage cluster sampling when all psus have M ssus.
- $deff$ is approximately $1 + (M-1)ICC$, where ICC is the intraclass correlation coefficient.
- Generally, ICC is positive, so cluster designs have less precision (per observation) than an SRS.

Bed net survey

- For the bed net survey in The Gambia, the design effect is approximately 5.89.
- This is due to the use of clusters.
- Villages tend to be homogeneous with respect to bed net use.

Use in the MOE

- First calculate the MOE for the SRS.
- Then multiply this quantity by the $\sqrt{deff}$ to obtain the MOE for the alternative design.
- Interpretation as ratio of sample sizes.
- $deft =$ proportion of SE's
- See Section 7.5.1, page 241.

Uses of $deff$

- Should be reported for a survey design.
- Useful for planning sample sizes for future studies.
  - Estimate the sample size needed for an SRS.
  - Multiply by the design effect.

The National Crime Victimization Survey (NCVS)

- Most US crime statistics come from the FBI Uniform Crime Reports.
- These reports underestimate crime because not all crimes are reported to the police.
- The NCVS is a large national survey administered by the Bureau of Justice Statistics with interviews conducted by the Bureau of the Census.
- It uses a stratified multistage cluster design.

Frame

- Household members 12 years of age and older are asked about their experiences as victims of crime within the last 6 months.
- The psus are counties, groups of adjacent counties, or metropolitan statistical areas (MSAs).

Examples of psu's

- Montgomery Alabama MSA includes Autauga, Elmore, and Montgomery counties.
- Columbus Ohio MSA includes Delaware, Fairfield, Franklin, Madison, Pickaway and Union counties.

Large psu's

- Any psu with 550,000 (use 1980 census data for 1990 survey) or more ssus is automatically included in the survey.
- These are psus are called *self-representing* with selection probability 1.

Other psu's

- Other psus are grouped into strata so that each strata group has a population of about 650,000.
- The stratification is based on
  - geographic location
  - demographic information from the 1980 census
  - Uniform Crime Report crime rates.

Selection of psu's

- One psu is selected from each stratum with probability proportional to population size (1980 census).
- These psus are called *non-self-representing*.
- They represent themselves and all other psus in their stratum.

The 1990 NCVS

There were

- 84 self-representing psu's
- 153 non-self-representing psu's, one from each of the 153 strata where sampling is pps.

Second stage sampling

Use enumeration districts (EDs)

- geographic areas used in the 1980 census
- each contains about 300 to 400 households
- they vary considerably in population and land area

ED's

ED's are selected

- with probability proportional to (1980 census) size
- number of EDs selected within a psu is determined so that the sample of EDs is approximately self-weighting

Selection of ED's

- In the census listing, EDs are arranged by geographic location.
- Systematic sampling (every xth unit is selected) is used.

Stages

- First stage
  - sample all of the 84 self-representing psus
  - select one psu from each of the 153 non-self-representing strata
- Second stage
  - select EDs with probability proportional to size
- Third stage
  - Each selected ED is divided into clusters of approximately four housing units each
  - Census lists these in geographic order
  - Select a sample of these clusters.
- Fourth and fifth stages
  - Then sample all (approx 4) housing units in the selected clusters
  - Select all persons aged 12 or over in the selected housing units

Summary of stages

1. psu (county, counties, MSA)
2. enumeration district
3. cluster of four housing units
4. household
5. person with household

Interviews

- Interviews with persons aged 12 and over are taken every month for a 6-month period.
- Interviews are also done every 6 months over a 3-year period.
- The first interview is used for *bounding*, to establish a time frame for the reports.

1990 Survey

- 62,600 housing units
- 56,800 were given the main questionnaire, others were given a new one being phased in.
- 8,200 of the 56,800 were ineligible.
  - vacant
  - demolished
  - no longer used as residences
- Of the remaining 48,600 housing units, no interviews were conducted in 1600 (3.3% nonresponse rate)
  - residents could not be reached
  - residents refused to participate
- 95,000 persons gave responses

Weights

- The survey was designed to be self-weighting.
- Weight is 1/(probability of housing unit selection).
- This is the base weight.
- Each person represents 1658 other persons in the US.

Adjustments to weights

A cluster may have more housing units than expected.

- sample a subset
- assign a weight-control factor (WCF)
- if 1/3 of units sampled then $WCF = 3$.

Other adjustments

- $weight =$ base weight $\times WCF \times WHHNAF \times HHNAF \times FSF \times SSF$
- See pages 245-246.
- weight is (an estimate of) the number of persons in the population represented.

Variance

- In NSR strata, one psu is selected, so we have between-psu variance.
- Within an ED, the clusters of four housing units are likely to be positively correlated.
- Persons within households are clusters (positively correlated).
- Systematic sampling used to choose EDs (because of list ordering, we hope to do better than SRS).

Sampling and experimental design (DOE)

Randomization

- SRS randomly selects a subset from a population.
- DOE design where subjects are randomly assigned to treatments
    - Fisher's permutation test

Stratification and blocking

- In sample surveys, we increase precision by grouping similar items.
- In DOE, we often use blocking to reduce the MSE.

Clustering

- With clustering we have groups of items that are usually similar.
- Split-plot designs are a DOE analog.