Statistics 522: Sampling and Survey Techniques

Topic 10

Topic Overview

This topic will cover

- Two-phase sampling
- Capture-recapture estimation
- Estimation in domains
- Rare events
- Randomized Response

Thomas Jefferson used red ink.

Two-phase sampling

- Also called *double sampling*
- Introduced by Neyman (1938)
- Key idea for application
  - Variable of interest y is expensive to measure
  - A correlated variable x is easy to measure
  - Can't get population-level info for x

Examples

- Sample businesses with probability proportional to sales but sales information is not available in the sampling frame
- Estimate total timber volume in a forest by measuring volume of logs in trucks; weight of trucks is easy to measure

Two Phases

- *Phase I.* Take a probability sample of $n^{(1)}$ units and measure x for every unit; $n^{(1)}$ should be relatively large and should provide a good estimate of the distribution of x.
- *Phase II.* Act as though the phase I sample is the population and select a probability sample; measure y on these.

Phase II

- In phase II you can use any of the information on x (or x's) measured in phase I to determine the probabilities of selection.
- For example, select trucks with probability proportional to weight in the timber problem.
- Can stratify on phase II sample

Example 12.1

- Estimate the percentage of Vietnam-era veterans in US Veterans Administration (VA) hospitals who actually served in Vietnam.
- The 1982 VA annual Patient Census (APC) included a random sample of 20% of the patients in VA hospitals.

APC

- The APC included the question

  "If period of service is 'Vietnam era,' was service in Vietnam?"

- Responses were
  - Yes
  - No
  - Not available

Response

- The responses were taken from medical records.
- Sources of inaccuracies
  - records based on self-reports; recall and definition problems
  - patient may think that response is related to benefits
  - recording errors
- Many responses were "not available".

Phase II

- Military records for a stratified subsample of hospitalized veterans were examined to determine true classification of Vietnam service.
- Strata were the three responses (Y, N, NA) for the 10% of Y and N, 100% of NA sampled.
- APC group

Results

- APC Yes: 67 of 755 sampled, 49 with Vietnam service (73%)
- APC No: 72 of 804 sampled, 11 with Vietnam service (15%)
- APC NA: 505 of 505 sampled, 211 with Vietnam service (42%)

Two-phase theory

- Section 12.1.1 gives some general notation
  - We need the Phase I and Phase II weights.
  - Estimate is an analog of the Horvitz-Thompson estimator (based on weights).

$$\begin{aligned}\hat{t}_x^{(1)} &= \sum_{sam} w_i^{(1)} x_i \\ \hat{t}_y^{(2)} &= \sum_{sam} w_i^{(1)} w_i^{(2)} y_i\end{aligned}$$

  - $\hat{t}_y^{(2)}$ is unbiased.
- Section 12.1.2 give ratio estimation results.

$$\hat{t}_{y,r} = \hat{t}_x^{(1)} \frac{\hat{t}_y^{(2)}}{\hat{t}_x^{(2)}}$$

Two-phase sampling for stratification

- Take an SRS at Phase I
- Stratify at Phase II based on something measured at phase I

$$\hat{t}^{(2)} = N \sum \frac{n_h}{n} \bar{y}_h^{(2)}$$

  where n_h is number of units in stratum h in phase I (n is total number sampled (SRS) in phase I)
- Equation (12.4) gives the variance

Capture-recapture estimation

- A lake has N fish.
- Catch and mark 200 fish; then release them.
- Later catch 100 fish
- Suppose 20 of the fish in the second sample are marked.

The estimate

- We estimate that 20% (20/100) of the fish in the lake are marked.
- 200 fish are marked
- 200 is 20% of N

$$\hat{N} = 200/0.2 = 1000$$

Assumptions

- Population is closed. (N stays the same.)
- Each sample is an SRS. (No units are "hidden".)
- The two samples are independent.
- Marks are not lost or ambiguous.
  - Probability of being in second sample not affected by capture or marking.
  - Not necessarily a mark; may just use identifying characteristics

The estimate

- Let n_1 and n_2 denote the sizes of the two samples.
- Let m denote the number of marked fish in the second sample.

$$\hat{N} = n_1 n_2 / m$$

Ratio estimate

- Let y be 1 for all fish in the lake.
- Let x be 1 for marked fish, 0 otherwise
- Then we estimate $\hat{N} = t_y$ by

  $$t_{y,r} = t_x \hat{B}$$

  where

  $$t_x = \sum x = n_1$$

  and

  $$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{n_2}{m}$$

  So,

  $$\begin{aligned} t_{y,r} &= \frac{n_1 n_2}{m} \\ \hat{\mathrm{V}}(\hat{N}) &\approx \frac{n_1^2 n_2 (n_2 - m)}{m^3} \end{aligned}$$

- Bias can be problem ($m = 0$?)

  $$\tilde{N} = \frac{(n_1 + 1)(n_2 + 1)}{m + 1} - 1$$

- Variance given in eqn (12.8).
- Inverse sampling: sampling continues until n_2 are captured (m is fixed).
- The normal theory CIs are not very good.

A different view

- The ratio estimator can be viewed as a missing data problem in a two-way contingency table.
- Rows are Y/N in sample 1.
- Columns are Y/N in sample 2.
- We observe YY, YN, and NY.
- Estimate NN using the independence assumption and expected counts
- Can't test independence assumption

Confidence Intervals

- In many cases, the distribution of $\hat{N}$ will not be very well-approximated by a normal; it is skewed.
- Sample sizes can be small.

Cormack's method

- Consider the contingency table; we have three of the four entries.
- Consider all possible values of the fourth entry (NN) for which the chi-square test would not reject the hypothesis of independence.
- Use these values (plus the other three) to construct the interval.

Alternatives

- Use the likelihood ratio statistic in place of the chi-square statistic.
- Bootstrap
  - resample from the second sample
  - in the example 20 tagged, 80 not tagged
  - use the 2.5% and 97.5%-tiles

Capture-recapture from lists

- Can use 2 lists (and intersection) to form idea of entire population
- See discussion of assumptions on page 391

Census application

- Called *dual-system estimation*
- Used to estimate undercount
- Take a sample from the census
- Take a sample using a different procedure (the Post-Enumeration Survey)
- Use the contingency table approach

Some details

- 1392 poststrata are used
- based on region, race, ownership of dwelling unit, age, and other variables
- The sample from the census is also used to estimate errors such as nonexistent persons and duplicates.

Multiple-recapture estimation

- More than two samples can be taken.
- Use different markings for different samples.
- Use maximum likelihood estimation.
- $\hat{N}_{ML}$ solves

$$\sum \frac{(n_i - r_i)M_i}{N - M_i} = \sum r_i$$

  where

  - n_i = size of sample i
  - r_i = number recaptured in sample i
  - M_i = number tagged at sample i

Example 12.7

- This methodology was used to estimate prevalence of opiate addiction in Barcelona.
- Samples from
  - emergency room lists
  - people who started treatment
  - overdose deaths

Log-linear models

- We cannot test the assumption that the missing cell follows the same model as the rest of the data.
- We can examine some (hierarchical) models for the data.
- See pages 394-395
- Inverted likelihood ratio test can be used to obtain CI.

Domains

- Sample size within a domain is random.
- We use ratio estimators.
- For large enough sample size, could act as though sample size fixed.
- For complex designs, use weights.
  - Could have empty psu's
- Could use theory of two-phase design
- Overall $deff$ can be used for small-cluster variance estimation.

Small area estimation

Area of research on estimation in small domains with small sample sizes

- *Direct estimation* (unbiased but large variance)

$$\hat{t}_d = \sum_{s_d} w_i y_i$$

- *Synthetic estimation*

$$\hat{t}_d = \frac{\hat{t}_y}{\hat{t}_u} u_d$$

  - u_d auxiliary
  - $\hat{t}_y / \hat{t}_u$ estimated over entire sample.
  - If $\hat{t}_{y,d} / \hat{t}_{u,d}$ varies over different domains, can lead to bias.
- *Composite estimation*

$$\hat{t}_d = \alpha_d \hat{t}_d^{(dir)} + (1 - \alpha_d) \tilde{t}_d^{(tyn)}$$

  α_d related to size of domain
- *Model-based estimators*
  - Use covariates and hypothesized values

Sampling for rare events

- Common problem in epidemiology
- Could take very large sample (non-response issues)
- Stratified sampling with disproportionate allocation
  - Has limited utility
- Two-phase sampling
  - Use phase 1 to estimate prevalence
  - Sample rare characteristic in phase 2
  - Problem if screening not entirely accurate
  - Can be better to gauge probability, not just presence of condition
- Multiple frame surveys
  - Example 1: Detecting Alzheimers

    Frame A: general population
    Frame B: adult care centers
    $B \subset A$
    use one to vary other

  - Example 2: Homeless population in Washington, D.C.

    Frame A: homeless shelters
    Frame B: soup kitchens
    Frame C: encampments
    Frame D: streets
    Need to establish overlap
- Network sampling
  - Ask about other units
  - Weights depend on number of connections
- Snowball sampling
  - Ask to identify other possible sampling units
  - Not random
  - Could be used early in investigation
- Sequential sampling
  - Use previous info to come up with designs/estimators

  - Adaptive design – high prevalence $\Rightarrow$ look at neighbors

- Nonresponse
  - Often nonignorable
  - Illness prevents response

Randomized response

- Used for sensitive questions
- 50 cards in deck
  - for 10, say ‘yes’
  - for 10, say ‘no’
  - for 30, answer the question “have you ever smoked marijuana?”
- Doesn’t *always* work
- Interviewer may have more significant effect.