Statistics for Applications

Chapter 9: Introduction to Survey Sampling

Introduction (1)

- Consider a population $[N] = \{1, \ldots, N\}$ of N individuals.
- Each individual $k \in [N]$ has a qualitative or quantitative characteristic y_k, which is deterministic.
- Examples in sociology/economics: y_k is the salary or individual k, or his/her age, or whether he/she is employed, or the color of his/her eyes, etc...
- Examples in other fields: The individuals are all webpages on the internet and y_k is the number of visits of page k in the past ten days, or the number of pages linked to page k, or the individuals are US American farms and y_k is the production of farm k, etc...

Introduction (2)

- If y_k is qualitative, we transform it into a binary quantity (e.g., $y_k = 1$ if individual k has blue eyes, 0 otherwise).
- We are interested in knowing the total $T = \sum_{k \in [N]} y_k$, the average $\bar{y} = \frac{1}{N} \sum_{k \in [N]} y_k$ or some other quantity $\theta = \theta(y_1, \ldots, y_N)$.
- In practice, N may be too large and even unknown. Hence, it is too costly or impossible to compute θ exactly.
- **Solution:** Sample a smaller proportion of individuals within the population.

Introduction (3)

- If $S \subseteq [N]$, one can define, for instance:

$$\hat{T}_S = \frac{N}{|S|} \sum_{k \in S} y_k, \quad \bar{y}_S = \frac{1}{|S|} \sum_{k \in S} y_k$$

and, in general,

$$\hat{\theta} = \hat{\theta}\left(\{y_k : k \in S\}\right).$$

- **Question:** How to choose S ?

- Choose a random subset $S \subseteq [N]$.

- The probability distribution of $\mathbf{S}$ chosen by the practitioner is called the *design* of the survey.

Sources of error

Running a survey leads to an estimation error. This error has multiple sources:

- Sampling: one does not collect the whole information contained in the population.
- Collection errors: The y_k's may be collected with noise (measurement errors, mistakes by the respondents of the survey, etc...)
- Missing data: Some of the y_k's, for $k \in S$, may be unavailable (e.g., sampled people who may not want to answer).

Goal: Control these errors and find good estimators of the total and/or the average.

Sampling designs (1)

Some designs commonly used:

- Choose a fixed $n < N$ and draw S uniformly in the collection of subsets of $[N]$ of size n:

$$\mathbb{P}[S = s] = \frac{1}{\binom{N}{n}}, \quad \forall s \subset [N] \text{ with } |s| = n.$$

  This is equivalent to sampling n individuals randomly without replacement.

- Choose a fixed $p \in (0,1)$ and let $I_1, \ldots, I_N \overset{i.i.d.}{\sim} Ber(p)$. Take

$$S = \{k \in [N] : I_k = 1\}.$$

  The size of S is random: It is binomial with parameter (N, p). In particular, $\mathbb{E}[|S|] = Np$.

Sampling designs (2)

- A partition $U_1, \ldots, U_d$ of the population $[N]$ may be available and relevant to the problem (e.g., $d = 50$ and U_j is the population in State j, for $j = 1, \ldots, 50$). One can choose

$$S = S_1 \cup \ldots \cup S_d,$$

where each S_j is a random subset of U_j.

- One may want to first partition each of the previous U_j (e.g., into men and women).

- If a partition $U_1, \ldots, U_d$ of $[N]$ is available, one may choose randomly fewer elements of this partition and draw random subsets $S_j \subseteq U_j$, for the selected U_j's.

Inclusion probabilities (1)

- Denote by $p(s) = \mathbb{P}[S = s]$, for $s \subseteq [N]$ (pdf of S).
- For $k \in [N]$, define

$$\pi_k = \mathbb{P}[S \ni k] = \sum_{s \subseteq [N] \,:\, s \ni k} p(s),$$

i.e., the probability that individual k is sampled.

- For $k, l \in [N]$, define

$$\pi_{k,l} = \mathbb{P}[S \supseteq \{k, l\}] = \sum_{s \subseteq [N] \,:\, s \supseteq \{k,l\}} p(s),$$

i.e., the probability that individuals k and l are both sampled.

Inclusion probabilities (2)

- For $k \in [N]$, denote by $\mathbb{I}_k = \mathbb{1}_{S \ni k}$.
- Then, for all $k, l \in [N]$,
  - $\mathbb{E}[\mathbb{I}_k] = \pi_k$,
  - $Var(\mathbb{I}_k) = \pi_k(1 - \pi_k)$,
  - $\Delta_{k,l} := cov(\mathbb{I}_k, \mathbb{I}_l) = \pi_{k,l} - \pi_k \pi_l$.
- $$\sum_{k=1}^{N} \pi_k = \mathbb{E}[|S|], \quad \sum_{k,l=1}^{N} \pi_{k,l} = \mathbb{E}[|S|^2], \quad \sum_{k,l=1}^{N} \Delta_{k,l} = Var(|S|).$$
- E.g., when n individuals are sampled without replacement, then for all $k \neq l \in [N]$,

$$|S| = n \text{ a.s.}, \quad \pi_k = \frac{n}{N}, \quad \pi_{k,l} = \frac{n}{N}\frac{n-1}{N-1}.$$

Estimation (1)

- In the sequel, we are only interested in the estimation of $T = \sum_{k \in [N]} y_k$ and $\bar{y} = \frac{1}{N} \sum_{k \in [N]} y_k$.

- We assume that $\pi_k > 0, \forall k \in [N]$ (i.e., no cut-offs in the population, no unreachable individual, list of individuals not out of date).

- Horvitz-Thompson's estimators of T and $\bar{y}$:

$$\hat{T}_{HT} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in [N]} \frac{y_k}{\pi_k} I_k, \qquad \widehat{\bar{y}}_{HT} = \frac{\hat{T}_{HT}}{N}.$$

  (Note: The y_k's, $k \in S$ are observed and the π_k's, $k \in [N]$ are decided beforehand, they depend on the sampling design.)

Estimation (2)

- $\hat{T}_{HT}$ is unbiased.

- Variance of $\hat{T}_{HT}$:

$$Var(\hat{T}_{HT}) = \sum_{k,l \in [N]} \frac{y_k y_l}{\pi_k \pi_l} \Delta_{k,l}.$$

- If $\pi_{k,l} > 0$, $\forall k, l \in [N]$, there is an unbiased estimator of the variance of $\hat{T}_{HT}$:

$$\hat{V} = \sum_{k,l \in S} \frac{y_k y_l}{\pi_k \pi_l} \frac{\Delta_{k,l}}{\pi_{k,l}}.$$

- In general, this estimator is written in the following way and it is biased:

$$\hat{V} = \sum_{k,l \in S : \pi_{k,l} \neq 0} \frac{y_k y_l}{\pi_k \pi_l} \frac{\Delta_{k,l}}{\pi_{k,l}}.$$

Estimation (3)

- $\mathbb{E}\left[\hat{V}\right] = Var(\hat{T}_{HT}) + \sum_{k,l \in [N]: \pi_{k,l}=0} y_k y_l.$

- If the size of S is fixed, then $Var(\hat{T}_{HT})$ can be written as:

$$Var(\hat{T}_{HT}) = -\frac{1}{2} \sum_{k,l \in [N]} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \Delta_{k,l}.$$

- In that case, another estimator of the variance is then:

$$\tilde{V} = -\frac{1}{2} \sum_{k,l \in S: \pi_{k,l} \neq 0} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \frac{\Delta_{k,l}}{\pi_{k,l}}.$$

- $\mathbb{E}\left[\tilde{V}\right] = Var(\hat{T}_{HT}) - \frac{1}{2} \sum_{k,l \in [N]: \pi_{k,l}=0} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \pi_k \pi_l.$

Confidence intervals

- How to compute confidence intervals for T ?
- In practice, practitioners often use

  $$I = \left[\hat{T}_{HT} - q_{1-\alpha/2}\sqrt{\max(\hat{V}, 0)}, \hat{T}_{HT} + q_{1-\alpha/2}\sqrt{\max(\hat{V}, 0)} \right],$$

  where $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$.
- ⚠ Depending on the design, it is not always the case that $\dfrac{\hat{T}_{HT} - T}{\sqrt{\hat{V}}}$ is approximately standard Gaussian.
- Alternative: bootstrap.

Sampling n individuals without replacement (1)

- $\mathbb{P}[S = s] = \begin{cases} \frac{1}{\binom{N}{n}} & \text{if } |s| = n \\ 0 & \text{otherwise.} \end{cases}$
- $\pi_k = \frac{n}{N}, \qquad \forall k \in [N];$
- $\pi_{k,l} = \frac{n(n-1)}{N(N-1)}, \qquad \forall k, l \in [N]$ with $k \neq l$.
- $\hat{T}_{HT} = \frac{N}{n} \sum_{k \in S} y_k.$
- $\widehat{\bar{y}}_{HT} = \frac{1}{n} \sum_{k \in S} y_k$: Mean value of the y_k's in S.

Sampling n individuals without replacement (2)

- $Var(\hat{T}_{HT}) = N\frac{1-f}{f}\sigma^2$,
- $\tilde{V} = N\frac{1-f}{f}\hat{\sigma}^2$, where:
  - $f = n/N$;
  - $\sigma^2 = \frac{1}{N-1}\sum_{k\in[N]}(y_k - \bar{y})^2$ is the empirical variance of the y_k's in the population;
  - $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{k\in S}(y_k - \bar{y}_S)^2$ is the empirical variance of the y_k's in the random sample S.
- Remark: $\hat{\sigma}^2$ is an unbiased estimator of σ^2.

Sampling n individuals without replacement (3)

Remark: If $y_1, \ldots, y_N$ are binary (i.e., 0 or 1):

- Quadratic risk of $\bar{y}_{HT}$ (bias-variance decomposition):

$$\mathbb{E}\left[(\widehat{\bar{y}}_{HT} - \bar{y})^2\right] = \frac{N-n}{N-1}\frac{\bar{y}(1-\bar{y})}{n}.$$

- When the individuals were sampled with replacement, which corresponded to an i.i.d. Bernoulli statistical model, the MLE $\hat{p}$ satisfied (with $p = \bar{y}$):

$$\mathbb{E}\left[(\hat{p} - p)^2\right] = \frac{p(1-p)}{n}.$$

- Hence, sampling without replacement is more precise than with replacement.

Algorithms for sampling without replacement (1)

Selection draw by draw

For $i = 1, \dots, n$, select randomly an individual among those who have not been selected already.

$\looparrowright$ Complexity $\mathcal{O}(nN)$.

Random sort

- Associated independently a random variable $U_i \sim \mathcal{U}([0,1])$ to individual i, for each $i \in [N]$.
- Sort the individuals by their U_i's.
- Select the n first.

$\looparrowright$ Complexity $\mathcal{O}(N \ln N)$ (to sort N variables).

Algorithms for sampling without replacement (2)

Select-reject

- Initialize $j = 0$.
- For $k = 1, \ldots, N$: With probability $\dfrac{n-j}{N-k+1}$, select individual k and set $j \leftarrow j+1$.

$\looparrowright$ Complexity $\mathcal{O}(N)$.

Reservoir method

- Set $S = \{1, \ldots, n\}$.
- For each $k = n+1, \ldots, N$: With probability $\frac{n}{k}$ choose k, draw randomly (uniformly) an element in S and replace it with k.

$\looparrowright$ Average complexity $\mathcal{O}(n^2 \ln N)$ but does not require knowledge of N from the beginning.

Algorithms for sampling with given inclusion probabilities (1)

- The practitioner may want to design a sample that has given inclusion probabilities $\pi_k, k = 1, \ldots, N$.

- E.g., if the individuals are companies, one may want to assign a larger probability to larger companies.

- If the sizes $e_1, \ldots, e_N$ (numbers of employees) of the companies are known, how to chose a design that satisfies

$$\pi_k \propto e_k, \quad k = 1, \ldots, N \ ?$$

- Remark: any design that satisfies these restrictions will give the same Horwitz-Thompson estimators. The bias will be zero, only the variance will change, according to the values of the $\pi_{k,l}$ that will result from the design choice.

Algorithms for sampling with given inclusion probabilities (2)

Algorithm 1

- Sample $U_1, \dots, U_N \overset{i.i.d.}{\sim} \mathcal{U}([0,1])$.
- For $k = 1, \dots, N$, choose k if $U_k \leq \pi_k$.

$\looparrowright$ Large variance for the HT estimator.
$\looparrowright$ In practice, this is useful when individuals show up one at a time.

Algorithms for sampling with given inclusion probabilities (3)

If $\sum_{i=1}^{N} \pi_i = n$ and we want a random sample of fixed size n:

Algorithm 2 to get a sample of fixed size n

- Set $V_0 = 0$ and $V_k = \sum_{i=1}^{k} \pi_i$, for $k \in [N]$.
- Sample $U \sim \mathcal{U}([0,1])$.
- For $k = 1, \ldots, N$, choose k if $V_{k-1} \leq U + i < V_k$ for some $i \in \{0, \ldots, n-1\}$.

$\looparrowright$ The sample has fixed size n, determined beforehand.
$\looparrowright$ Drawback: This algorithm is very rigid (very little randomness in the choice of S, all depends only on one random variable U).

Conclusions

- A total or an average among a large population is sought.
- A subset of the population is sampled randomly, according to a given sampling design.
- We proposed a few sampling algorithms.
- We proposed unbiased estimators and computed estimators of their variances when the answers of the respondents were collected perfectly.
- What if some answers are incorrect ? If some answers are missing ?