Session 09

Applied Problems for Survey Sampling

Double Sampling for Nonresponse

Non-sampling Error

Definition of non-sampling error: the differences between estimates and population quantities that do not arise solely from the fact that only a sample, instead of the whole population, is observed.

Here is an example of when the sampling frame does not match up perfectly with the target population. For telephone surveys, if we are interested in sampling the entire population in a given city, telephone directories are inadequate because of telephone numbers that are unlisted or the homeless that do not have telephones.

What would be another example where an important part of the target population would be missed? Serious non-sampling error might have occurred for non-response.

Non-response: the self selection of respondents may produce bias. For example, only people with certain opinions will respond to some questions.

In a well designed research study, handling non-responses is important. How do we handle this common phenomena?

An important use of double sampling:

Double sampling can be used to adjust for non-response in the form of call backs.

Non-response is an important problem to consider in any survey. We can consider the two groups: response and non-response in two strata.

The two steps of double sampling for non-response:

Step 1:

n' initial simple random samples are selected from a population of N units. These units are classified into two strata: response and non-response.

n_1' of these respond - stratum 1
n_2' of these do not respond - stratum 2

Step 2:

Call back n_2 samples by simple random sampling from the n_2' non-respondents (by giving more incentives, etc.)

Thus, we are in a double sampling setting where $n_1 = n_1'$, n_2 is the number of call backs.

Example

In a college with 1000 students, a questionnaire is mailed to a simple random sample of 106 students asking them about the amount of time they spend per week studying. Out of these students, 46 respond. From the 60 non-respondents, a simple random sample of 20 is selected and intensive efforts are made by telephone and personal visit to obtain responses. The data obtained are as follows:

	Students responding to questionnaire	Students contacted and responded to telephone and visit
Sample mean	20.5 hours	10.9 hours
Sample st. dev.	6.2 hours	5.1 hours
Sample size	46	20

Now, let's estimate the mean and also the variance of the estimate.

Solution

To estimate the average hours students spend per week studying, they use a double sampling:

Step 1: 106 students are randomly sampled; 46 respond, 60 non-respondents.

$n' = 106$, $n_1' = 46$, $n_2' = 60$

$$w_1 = \frac{n_1'}{n'} = \frac{46}{106} = 0.434$$
$$w_2 = \frac{n_2'}{n'} = \frac{60}{106} = 0.566$$

Step 2: From the 60 non-respondents, a simple random sample of 20 students were sampled and the following responses were obtained:

$n_1 = 46$, $n_2 = 20$, $\bar{y}_1 = 20.5$, $\bar{y}_2 = 10.9$

Selecting the Number of Call Backs

- c_0 : the initial cost of sampling each respondent (the set-up cost for each respondent)
- c_1 : the cost of a standard response (cost of producing the response)
- c_2 : the cost of a call back response

Total cost = ($n' \times c_0$) + ($n'_1 \times c_1$) + ($n_2 \times c_2$)

We want to determine the value $k (k > 1)$ where: $n_2 = \frac{n'_2}{k}$.

As $\bar{y}_d = \sum_{h=1}^{2} w_h \bar{y}_h$, its variance can be derived and one can find the value of k and n' that minimize the expected cost of sampling for a desired fixed value of $\hat{V}ar(\bar{y}_d)$, which we denote as V_0 .

When N is large, the optimal value of k and n' are:

$$k = \sqrt{\frac{c_2(\sigma^2 - w_2\sigma_2^2)}{\sigma_2^2(c_0 + c_1 w_1)}}$$

$$n' = \frac{N(\sigma^2 + (k-1)w_2\sigma_2^2)}{NV_0 + \sigma^2}$$

where σ^2 is the variance of the entire population and σ_2^2 is the variance of the non-response group.

Example

In a college of 1000 students we want to find out students' average weekly living expenditure. The response rate is anticipated to be about 60%. It is thought that the response group has a higher variance than the non-response group. The overall variance $\sigma^2 \sim 120$ the variance of the non-response group $\sigma_2^2 \sim 80$, $c_0 = 0$, $c_1 = 1$, $c_2 = 4$.

Variance of a Mixture Distribution

$X \sim (\mu_X , \sigma_X^2)$
$Y \sim (\mu_Y , \sigma_Y^2)$

W is a mixture of hX and $(1 - h)Y$

$$W = hX \oplus (1-h)Y$$

Then:

$\mu_w = h\mu_X + (1 - h)\mu_Y$

What is the variance of the overall population?

$\sigma_W^2 = h\sigma_X^2 + (1 - h)\sigma_Y^2 + h(1 - h)(\mu_X - \mu_Y)^2$

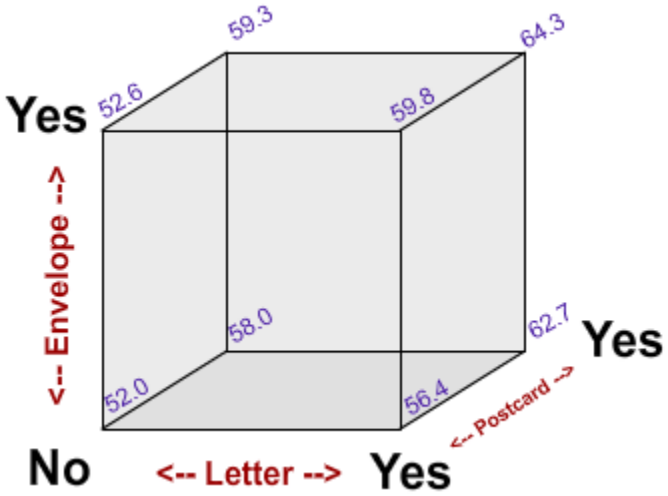
Designing surveys to reduce non-response

Good survey practice is to discover why the non-response occurs and resolve as many of the problems as possible before commencing the survey.

Example: To design an experiment to find out how to best improve the response rate.

A factorial experiment employed in the 1992 Census Implementation Test to explore the individual effect and the interactions for 3 factors on the response rate:

1. pre-notice letter
2. stamped return envelope
3. reminder postcard



- Letter, Postcard, Envelope >>> 64.3
- Letter, Postcard, no envelope >>> 62.7 (not bad!)

Some factors that may influence the response rate and data accuracy:

- Survey content: sensitive questions will have high non-response rate, Try randomizing the response technique
- Time of survey - select wisely
- Data collection method such as Computer Assisted Telephone Interviewing (CATI) has been shown to improve data accuracy. CATI: interview questions are stored in a computer, and recalled in programmable sequences and displayed for each interviewer on a video display terminal. And, interviewers enter answers received via telephone directly into computer right away.
- Incentives and penalties

Remark: The quality of the survey data is largely determined at the design stage.

Interpenetrating Subsample

There are *k* interviewers and they are each different in their manner of interviewing and hence may obtain slightly different responses. To make notation simple, we assume that each interviewer conducts the same number of interviews. Let *n* denote the total sample size and *n* = *k** *m*. There are *k* subsamples and each interviewer will be assigned *m* subjects.

Objective: to use simple random sampling to estimate μ

Interviewer 1 - $y_{11}, y_{12}, y_{13}, \ldots, y_{1m}$
Interviewer 2 - $y_{21}, y_{22}, y_{23}, \ldots, y_{2m}$
Interviewer 3 - $y_{31}, y_{32}, y_{33}, \ldots, y_{3m}$

Interviewer *k* - $y_{k1}, y_{k2}, y_{k3}, \ldots, y_{km}$

The average for the *i*th interviewer is denoted as:

$$\overline{y}_i = \frac{1}{m}\sum_{j=1}^{m} y_{ij}$$

The grand average is denoted as:

$$\overline{y} = \frac{1}{k}\sum_{i=1}^{k} \overline{y}_i$$

The grand average $\overline{y}$ is unbiased for μ and the estimated variance of $\overline{y}$ is:

$$\hat{Var}(\overline{y}) = \frac{N-n}{N} \cdot \frac{s_k^2}{k}$$

$$where \quad s_k^2 = \frac{\sum_{i=1}^{k} (\overline{y}_i - \overline{y})^2}{k-1}$$

The technique of interpenetreting the subsample gives an estimate of the variance of ybar that accounts for interviewer biases. In practice, the estimated variance given in the above formula is usually larger than the standard estimate of the variance by using simple random sampling.

Example for interpenetreting subsample

A researcher has 10 research assistants, each with his/her own equipment that they use to measure the time (in seconds) it take for people to respond to a command. A simple random sample of 80 people are taken. Since the researcher believes the assistants will produce slightly biased measurements, he decides to randomly divide the 80 people into 10 subsamples of 8 persons each. Each assistant is then assigned to one subsample. The measurements are given in the following table.

assistants	**time it takes to respond**							
1	52	73	62	75	71	68	55	65
2	62	65	73	67	78	71	67	59
3	43	54	52	48	56	51	62	57
4	73	64	63	59	71	78	67	76
5	88	76	69	83	85	66	74	73
6	55	71	63	75	68	72	69	60
7	72	65	77	69	74	82	73	67
8	55	43	58	62	42	61	53	61
9	62	52	59	63	69	72	64	58

10	77	65	79	69	72	68	71	67

Minitab output:

	mean
Subsample 1	65.125
Subsample 2	67.750
Subsample 3	52.875
Subsample 4	68.875
Subsample 5	76.750
Subsample 6	66.625
Subsample 7	72.375
Subsample 8	54.375
Subsample 9	62.375
Subsample 10	71.000

If one neglects the interviewer effect, then $\hat{SD}(\overline{y}) \approx 1$, thus it is important to take into consideration the interviewer effect. Otherwise, one underestimates $\hat{SD}(\overline{y})$.

Estimation of means and totals over subpopulation

Quite often, obtaining a frame that lists only those elements of the population that one is interested in is impossible. For example, perhaps you want to sample households with children, however, the best frame available is a list of all households. Therefore, we wish to estimate the parameters of a subpopulation of the population represented in the frame.

Main Issue: You do not know the size of the subpopulation.

Notation:

- N - the number of elements in the population
- N_1 - the number of elements in the subpopulation
- n - sample size from the population
- n_1 - the number of sampled elements from the subpopulation
- y_{1j} - the jth sampled observation that falls in the subpopulation

An unbiased estimator of μ_1, the subpopulation mean is:

$$\overline{y}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} y_{ij}$$

Its variance is estimated by:

$$\hat{Var}(\overline{y}_1) = \left(\frac{N_1 - n_1}{N_1}\right)\frac{s_1^2}{n_1}$$

$$where \quad s_1^2 = \frac{\sum_{j=1}^{n_1}(y_{ij} - \overline{y}_1)^2}{n_1 - 1}$$

Usually we do not know N_1, so we will estimate it using:

$$\frac{N_1 - n_1}{N_1} by \frac{N - n}{N}$$

Example: Let's say we want to estimate the average weekly amount spent on food by married graduate students in a certain college at Penn State. There are 80 graduate students in the college. 15 are sampled and 10 are married. A summary of the data follows:

Descriptive Statistics: food cost

Variable	marital status	N	Mean	SE Mean	StDev
food cost	m	10	135.3	14.1	44.4
	s	5	87.60	9.73	21.76