Chapter 9. Non-responses and Randomized Responses

1 Non-responses

The following two types (among others) of errors are very common in survey sampling:

(1) Non-response errors (e.g. Respondents will throw away the questionnaires),

(2) Errors of reporting (e.g. the person may lie about his/her true income),

1.1 Effect of non-response on sample mean

Divide the population of size N into "response group" of and "non-response group". Denote

N_1 = the size of "response group", μ_1 = population mean of "response group"

N_2 = the size of "non-response group", μ_2 = population mean of "non-response group".

So $N = N_1 + N_2$. Further define

$$W_1 = \frac{N_1}{N}, \qquad W_2 = \frac{N_2}{N}.$$

If we draw a random sample of size n, only n_1 of n respond, the bias due to non-response is

$$bias = E(\bar{y}_1) - \mu = \mu_1 - (W_1\mu_1 + W_2\mu_2) = W_2(\mu_1 - \mu_2).$$

If both W_2 and $|\mu_1 - \mu_2|$ are small, then the bias is not serious. Otherwise, one must be cautious.

1.2 Proportion estimation in the presence of non-response

If we estimate the proportions, then the formula $\mu = W_1\mu_1 + W_2\mu_2$ reduces to $p = W_1p_1 + W_2p_2$, then a conservative $(1-\alpha)$ level C.I. for p (ignoring f.p.c.) is given by $(\hat{p}_L, \hat{p}_U)$, where

$$\hat{p}_L = W_1\left(\hat{p}_1 - z_{\alpha/2}\sqrt{\hat{p}_1\hat{q}_1/n_1}\right) + W_2 \times 0,$$

$$\hat{p}_U = W_1 \left(\hat{p}_1 + z_{\alpha/2} \sqrt{\hat{p}_1 \hat{q}_1 / n_1} \right) + W_2 \times 1$$

if we know W_2.

If W_2 is unknown, Cochran (1977, p362) proposed the following conservative C.I.:

$$\hat{p}_L = \hat{p}'_1 - z_{\alpha/2} \sqrt{\hat{p}'_1 \hat{q}'_1 / n},$$

where $\hat{p}'_1$ is the proportion of positive response by assuming all sample non-respondents would have given a negative response (i.e. $p_2 = 0$), and

$$\hat{p}_U = \hat{p}''_1 + z_{\alpha/2} \sqrt{\hat{p}''_1 \hat{q}''_1 / n}$$

where $\hat{p}''_1$ is the proportion of positive response by assuming all sample non-respondents would have given a positive response (i.e. $p_2 = 1$).

Example. Let $n = 500$, $n_1 = 400$, and $\hat{p}_1 = 15\%$, so that $n_1 \hat{p}_1 = 60$ members in the sample respond positively and the non-response rate is 20%. Then

$$\hat{p}'_1 = \frac{60}{500} = 12\%, \qquad \hat{p}''_1 = \frac{60 + 100}{500} = 32\%.$$

Therefore,

$$\hat{p}_L = 12\% - 1.96 \times \sqrt{12\% \times 88\% / 500} = 0.0915$$
$$\hat{p}_U = 32\% + 1.96 \times \sqrt{32\% \times 68\% / 500} = 0.3609$$

1.3 Ways to deal with non-response

1.3.1 Call-backs

In order to avoid a large number of non-responses in the sample, it is customary to call back the non-respondents (especially those who are not home) a fixed number of times.

If a call is made in the daytime, most likely, one would miss families like "double income no kids" (DINK), whose family incomes are most likely above the average family income, thus creating bias.

1.3.2 Imputation method

Missing items may occur in surveys for several reasons: An interviewer may fail to ask a question: a respondent may refuse to answer the question or can not provide the information.

In a questionnaire, many questions are usually asked. If the respondent fails to answer one question (such as his/her age), it seems a huge waste to throw the whole data set away. **Imputation** is commonly used to assign values to the missing items. A replacement value, often from another person (or an average of several persons) in the survey who is similar to the item non-respondent on other variables, is imputed for the missing value. Details of this will be omitted.

2 Randomized Responses to Sensitive Questions.

2.1 Why?

Sometimes you want to conduct a survey asking some sensitive questions such as

Do you have AIDS?

Have you had any extramarital affairs?

Do you use cocaine?

Have you ever shoplifted?

Have you ever cheated in the exams?

Did you understate your income on your tax return?

In these circumstances, you might expect that many people will find very uncomfortable to answer such questions and so may choose not answer them, or if they do, the answers may be very evasive. In these cases, we may resort to **randomized responses**.

2.2 How?

We shall illustrate how to do this by an example.

Example 1. The respondent throws a coin with $P(Head) = P$. But the interviewer does not know the outcome of coin tossing.

(a) If it is a head, he answers the original (sensitive) question.

(e.g., have you ever cheated on an exam?)

(b) If it is a tail, he answers a totally unrelated (innocent) question.

(e.g., is the last digit of your home number odd?)

Let

$$\begin{aligned} P &= P(\text{asked sensitive question}) \\ p_S &= P(\text{say "Yes" } | \text{ asked sensitive question}) \\ p_I &= P(\text{say "Yes" } | \text{ asked innocent question}) \\ n &= \text{the total number of people being asked, i.e. sample size} \end{aligned}$$

Therefore,

$$\begin{aligned} \psi &= P(\text{respondent replies "Yes"}) \\ &= P(\text{Yes}|\text{asked sensitive question})P(\text{asked sensitive question}) \\ &\quad +P(\text{Yes}|\text{asked innocent question})P(\text{asked innocent question}) \\ &= p_S P + p_I(1-P). \end{aligned}$$

Hence,

$$p_S = \frac{\psi - (1-P)p_I}{P}.$$

Here, P and p_I are known, but ψ is unknown, but it can be estimated from the sample, denoted by $\hat{\psi}$ = estimated proportions of "Yes" from the sample. Thus,

$$\hat{p}_S = \frac{\hat{\psi} - (1-P)p_I}{P}.$$

Clearly,

$$\begin{aligned} E\hat{p}_S &= p_S, \\ Var(\hat{p}_S) &= \frac{Var(\hat{\psi})}{P^2} = \frac{1}{P^2} \times \frac{\psi(1-\psi)}{n} \left(\frac{N-n}{N-1}\right), \\ \widehat{Var}(\hat{p}_S) &= \frac{1}{P^2} \times \frac{\hat{\psi}(1-\hat{\psi})}{(n-1)}(1-f), \qquad with \quad f = \frac{n}{N}. \end{aligned}$$

If N is unknown, we can treat it as $N = \infty$.

Suppose that out of the 800 people surveyed, 275 says "Yes". Then

$$\hat{\psi} = \frac{275}{800}, \qquad \widehat{Var}(\hat{\psi}) = \frac{\hat{\psi}(1-\hat{\psi})}{(n-1)} = 2.82 \times 10^{-4}$$

$$\hat{p}_S = \frac{\hat{\psi} - (1-P)p_I}{P} = \frac{\frac{275}{800} - \left(1 - \frac{1}{2}\right) \times \frac{1}{2}}{\frac{1}{2}} = 0.1875$$

(Here $p_I = 1/2$. Note that if the innocent question is: "is the last digit of your home number odd?", the probability of yes is 1/2.)

$$\widehat{Var}(\hat{p}_S) = \frac{2.82 \times 10^{-4}}{\frac{1}{4}} = 1.129 \times 10^{-3}$$

So a 95% confidence interval for p is

$$0.1875 \pm 2\sqrt{1.129 \times 10^{-3}} = (0.1203, 0.2547)$$

Remarks.

(a) The "penalty" for randomized response appears in the factor $1/P^2$ in the variance expression. If $P = 1/3$, e.g., the standard deviation is 3 times as great as it would have been, had everyone in the sample been asked the sensitive question and responded truthfully.

(b) One need to think before choose P: the large the P is, the smaller the variance of $\hat{p}_S$. But if P is too large, respondents may think that the interviewer will know which question is being answered. Some respondents may think that $P = 0.5$ is "fair".

(c) Instead of throwing a coin, respondent could also be asked to draw a card from a deck of 52 cards. If it is a "heart" or a "spade", he/she answers the sensitive question. Otherwise, he/she answers the innocent question.