

## NOTES ON DEALING WITH MISSING DATA

### 0. Introduction

This set of notes is an introduction to the topic of missing data. In practice, missing data arises in many, probably most, studies, and so it is important to know how to deal with it.

We will focus mainly on the simple one-sample problem where it is desired to estimate a population mean  $\mu$  based on a sample of i.i.d. observations  $Y_1, \dots, Y_n$ , some of which may be missing. We will write  $R_i = 1$  if  $Y_i$  is observed and  $R_i = 0$  if  $Y_i$  is missing. Certain more general situations, such as the two-sample problem of estimating a difference in means, can be handled by similar techniques. For other problems, such as estimating correlations, a more complex development is needed, although some of the basic concepts are the same.

Missing data is a difficult problem because  $R_i$  can depend on  $Y_i$  in ways that are not necessarily well understood. For example, in a salary survey, it is likely that people who are very rich or very poor will tend not to reveal their salary, but the exact form of the dependence may be very unclear. It is difficult to build statistical models for the dependence because, by definition, we are trying to model a process that cannot be directly observed (i.e. when  $R_i = 0$ , we do not get to see  $Y_i$ ). Also, in certain situations, missing data (even if completely random) can lead to technical complications making the statistical analysis more difficult than for complete data. This is not the case for the one-sample setting we focus on here, but it is the case in more complex settings, such as regression problems with scattered missing values of the explanatory variables and repeated measures studies with dropout or intermittent missing measurements.

In general, if a study has a high percentage of missing data, it will be impossible to draw definite conclusions from the study, unless the mechanism behind the missingness is extremely well understood.

There is a lot that can be said on the topic of missing data. Indeed, it is possible to devote a whole course to this topic. We are just going to scratch the surface of the area. The following are basic references on missing data for further reading.

Little, R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

### 1. Missing Completely At Random (MCAR)

The simplest case of missing data is what is called “Missing Completely At Random” (MCAR). This means that there is no dependence whatsoever between  $R_i$  and  $Y_i$ .

In the MCAR situation, life is easy: we can just ignore the missing data and analyze the observed data in the usual way. That is, we discard the units with missing data, and regard the sample of units with observed values as if it were the original sample.

However, if there is dependence between  $R_i$  and  $Y_i$ , this simple approach will give biased answers. The reason is that, when such dependence exists, the sample of units with observed data is a biased subsample of the original sample. Hence it is important not to be too quick to follow the simple approach of just disregarding the missing data. One must consider carefully whether the MCAR assumption is justified.

The MCAR assumption may be reasonable in certain situations, such as missingness due to dropped test tubes or other technical problems in a laboratory experiment. It is clearly not reasonable in a salary survey.

In some cases, there may be multiple types of missing data, some of which can be assumed to be MCAR and some of which cannot. For example, in a political survey, missingness due to the fact that the person we are trying to reach is in the hospital is probably MCAR (or nearly so), while missingness due to the fact that the person we contacted does not want to discuss politics is probably not MCAR. Special theory is needed for this case. The ideas presented in the following subsections can be used to develop the necessary techniques. Some recent work on the multiple type problem has been done by Ofer Harel (2003, Ph.D. thesis, Penn State University).

## 2. Explanatory Variables

In certain cases, we may have a vector  $\mathbf{X}_i$  of background variables (observed on *all* sample units) that explain to some degree the relationship between  $R_i$  and  $Y_i$ . The most extreme case is where  $\mathbf{X}_i$  *completely* explains the relationship between  $R_i$  and  $Y_i$ . That is,  $R_i$  is independent of  $Y_i$  given  $\mathbf{X}_i$ :

$$\Pr(R_i = 1|Y_i, \mathbf{X}_i) = \Pr(R_i = 1|\mathbf{X}_i). \quad (1)$$

When this condition holds, we can use the explanatory variables  $\mathbf{X}_i$  to get an (approximately) unbiased estimate of the population mean  $\mu$ .

In real life, the condition (1) will usually not hold exactly, but it may hold to a good enough approximation to yield a reasonable answer. Even when the condition does not hold even approximately (i.e. there are other factors that significantly effect the relationship between  $R_i$  and  $Y_i$ ), adjusting for known explanatory variables will usually at least reduce the missing data bias problem.

In the following, we will present theory under the assumption that the condition (1) holds exactly. The setup is that  $(Y_i, R_i, \mathbf{X}_i)$  are i.i.d. random vectors (note well that the vector  $\mathbf{X}_i$  of explanatory variables is regarded as a random vector).

### 2.1. A Single Binary Explanatory Variable

The simplest case is where there is single explanatory variable  $X_i$  with just two possible values, say 1 or 2. For concreteness, we will use the example of gender (say 1

for male and 2 for female). Let  $n$  denote the total number of subjects in the original sample, and  $n_k$  the number of subjects in gender group  $k$  for  $k = 1, 2$ . Further, let  $n'_k$  denote the number of subjects in gender group  $k$  with an observed value of  $Y_i$  (i.e.  $R_i = 1$ ).

If we use the overall observed sample mean  $\bar{Y}$  to estimate  $\mu$ , we will get a biased answer. Because of the dependence between the missingness and gender, the subsample of subjects with observed data is not a representative subsample of the original sample.

We can construct a correct estimate as follows. Let  $\bar{Y}_k, k = 1, 2$ , be the observed sample means for the two gender groups. Because the missingness is independent of the response variable  $Y$  given gender (i.e. within each gender group, the missingness is MCAR), these sample means are unbiased estimates of the corresponding population means  $\mu_k = E[Y_i | X_i = k]$ . In parallel, we can unbiasedly estimate the probabilities  $\psi_k = \Pr(X_i = k)$  by the corresponding proportions in the original sample:  $\hat{\psi}_k = n_k/n$ . Thus, we can unbiasedly estimate the overall population mean  $\mu$  by

$$\hat{\mu} = \hat{\psi}_1 \bar{Y}_1 + \hat{\psi}_2 \bar{Y}_2. \quad (2)$$

We can compare the above estimator with the naive overall observed sample mean  $\bar{Y}$ , which can be written as

$$\bar{Y} = \tilde{\psi}_1 \bar{Y}_1 + \tilde{\psi}_2 \bar{Y}_2,$$

where  $\tilde{\psi}_k$  are the proportions of subjects in gender group  $k$  in the subsample with observed data:  $\tilde{\psi}_k = n'_k/n$ . Thus, we move from a biased estimate of  $\mu$  to an unbiased one by appropriately reweighting the results in the respective gender groups.

It should be obvious that the above discussion extends directly to the case of a single categorical explanatory variable with more than two levels, e.g. region of origin (Israel, Europe, North America, South America, Africa, Asia), or, more generally, to the case where the overall population can be broken up into a number of distinct subpopulations, with MCAR missingness within each subpopulation. We get an estimate of the form (2), with a term for each of the subpopulations.

## 2.2. The General Case

We now proceed to the general case where  $\mathbf{X}_i$  may include continuous explanatory variables (or a mix of discrete and continuous explanatory variables), such that the method in the preceding subsection does not apply. We will describe two approaches to handling this situation. The first approach, the *regression imputation approach*, involves modeling the relationship between  $Y_i$  and  $\mathbf{X}_i$ . The second approach, the *inverse propensity weighting approach*, involves modeling the relationship between  $R_i$  and  $\mathbf{X}_i$ . The choice of approach will be driven by which of these two relationships the researchers feel more confident modeling. It is easily shown that for the case of a binary (or categorical) explanatory variable discussed in the preceding subsection, the two approaches are equivalent, and both lead to the procedure described in the preceding subsection. We also describe a combination method known as the “double robust” method.

The validity of the estimate of  $\mu$  that is obtained depends on two conditions: (a) the conditional MCAR condition (1), and (b) the validity of the model that is assumed to describe the relationship between  $\mathbf{X}_i$  and  $Y_i$  or between  $\mathbf{X}_i$  and  $R_i$ . In real life, the condition (1) and the assumed model for the relevant relationship will not hold exactly, but if they hold to a reasonable approximation then we will get a reasonable estimate of  $\mu$ .

### 2.2.1. Regression Imputation Approach

The idea here is to fill in the missing  $Y$  values with imputed values based on a prediction model for  $Y_i$  as a function of  $\mathbf{X}_i$ . If  $Y_i$  is missing, we replace it by a model-based prediction  $\hat{Y}_i$ . We thus define  $Y_i^*$  to be equal to  $Y_i$  when  $Y_i$  is observed and  $\hat{Y}_i$  when  $Y_i$  is missing, and we estimate  $\mu$  by the sample mean of the  $Y_i^*$  values.

We will focus here on the case where  $Y_i$  is continuous and a classical linear regression model is used. A similar, though more complex, development can be given for more general regression models. Also, for the case where  $Y_i$  is a 0–1 binary variable, a similar approach could be followed based on (for example) the logistic regression model (here we would take  $\hat{Y}_i$  to be the predicted probability that  $Y_i = 1$ ).

Thus, we assume that the relationship between  $Y_i$  and  $\mathbf{X}_i$  can be described by a classical linear regression model

$$Y_i = \beta_0 + \sum_{j=1}^p \beta_j X_{ij} + \epsilon_i, \quad (3)$$

with the  $\epsilon_i$  i.i.d.  $N(0, \sigma_\epsilon^2)$  (or at least i.i.d. with mean 0 and common variance  $\sigma_\epsilon^2$ ). We estimate the vector  $\boldsymbol{\beta}$  of regression parameters  $(\beta_0, \beta_1, \dots, \beta_p)$  using the standard least squares estimator

$$\hat{\boldsymbol{\beta}} = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathbf{Y}.$$

Here  $\mathbf{Y}$  is the vector of observed  $Y$  values and  $\mathcal{X}$  is a matrix with rows of the form  $[1 X_{i1} \dots X_{ip}]$ , where only the units  $i$  with an observed  $Y$  value are included (in the same order as they appear in the vector  $\mathbf{Y}$ ). We then define (for all units  $i$ , whether  $Y_i$  was observed or not)

$$\hat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j X_{ij} = \mathbf{X}(i)^T \hat{\boldsymbol{\beta}},$$

where  $\mathbf{X}(i)^T = [1 X_{i1} \dots X_{ip}]$ . For units with missing  $Y$  values, we replace the missing value  $Y_i$  with the predicted value  $\hat{Y}_i$ , and estimate  $\mu$  using the sample average of the resulting completed sample of observed and imputed values, as described above.

We now present an estimator of the variance of the resulting estimator  $\hat{\mu}$ . This development depends on the assumed model (3). Note first that, from classical regression theory, we have

$$\sum_{i:R_i=1} Y_i = \sum_{i:R_i=1} \hat{Y}_i$$

(this is a consequence of the fact that the sum of the residuals in a regression model is equal to zero). So we have

$$\begin{aligned}
\hat{\mu} &= \frac{1}{n} \left[ \sum_{i:R_i=1} Y_i + \sum_{i:R_i=0} \hat{Y}_i \right] = \frac{1}{n} \left[ \sum_{i:R_i=1} \hat{Y}_i + \sum_{i:R_i=0} \hat{Y}_i \right] \\
&= \frac{1}{n} \sum_{i=1}^n \hat{Y}_i = \frac{1}{n} \sum_{i=1}^n \mathbf{X}(i)^T \hat{\boldsymbol{\beta}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}(i)^T \boldsymbol{\beta} + \frac{1}{n} \sum_{i=1}^n \mathbf{X}(i)^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= \frac{1}{n} \sum_{i=1}^n W_i + \bar{\mathbf{X}}^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \tag{4}
\end{aligned}$$

where  $W_i = \boldsymbol{\beta}^T \mathbf{X}(i)$  and  $\bar{\mathbf{X}} = n^{-1} \sum_{i=1}^n \mathbf{X}(i)$ .

Now the two terms in (4) are uncorrelated because the conditional expectation of  $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$  given all the  $X_{ij}$  values is equal to zero. The variance of the first term is  $\sigma_W^2/n$ , where  $\sigma_W^2$  is the variance of  $W$ . If  $\boldsymbol{\beta}$  were known, we would estimate  $\sigma_W^2$  by the classical estimate

$$s_W^2 = \frac{1}{n-1} \sum_{i=1}^n (W_i - \bar{W})^2.$$

Since  $\boldsymbol{\beta}$  is unknown, we replace it with the estimate  $\hat{\boldsymbol{\beta}}$ , and estimate  $\sigma_W^2$  by

$$\hat{s}_W^2 = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{X}(i)^T \hat{\boldsymbol{\beta}} - \bar{\mathbf{X}}^T \hat{\boldsymbol{\beta}})^2. \tag{5}$$

As for the second term in (4), by classical regression theory we have  $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma_\epsilon^2 (\mathcal{X}^T \mathcal{X})^{-1}$ , and so  $\text{Var}(\bar{\mathbf{X}}^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) = \sigma_\epsilon^2 \bar{\mathbf{X}}^T (\mathcal{X}^T \mathcal{X})^{-1} \bar{\mathbf{X}}$ . We estimate  $\sigma_\epsilon^2$  by the classical regression mean square error (MSE) (based on the  $n'$  units with observed  $Y$  values)

$$s_\epsilon^2 = \frac{1}{n' - p - 1} \sum_{i:R_i=1} (Y_i - \hat{Y}_i)^2.$$

Thus, overall we have

$$\widehat{\text{Var}}(\hat{\mu}) = \frac{\hat{s}_W^2}{n} + s_\epsilon^2 \bar{\mathbf{X}}^T (\mathcal{X}^T \mathcal{X})^{-1} \bar{\mathbf{X}}. \tag{6}$$

### 2.2.2. Inverse Propensity Weighting Approach

This approach is based on a model for

$$\pi(\mathbf{x}) = \Pr(R_i = 1 | \mathbf{X}_i = \mathbf{x}).$$

A typical model is the logistic regression model

$$\pi(\mathbf{x}) = \frac{\exp(\boldsymbol{\gamma}^T \mathbf{x})}{1 + \exp(\boldsymbol{\gamma}^T \mathbf{x})}, \tag{7}$$

and we will assume this model in our development. The quantity  $\pi(\mathbf{x})$  is called the *propensity*, because it reflects the tendency of a unit with  $\mathbf{X} = \mathbf{x}$  to yield an observed

$Y$  value (“propensity” is a fancy word for “tendency”). The quantity  $\boldsymbol{\gamma}^T \mathbf{x}$  is called the *propensity score*.

Let us assume momentarily that  $\boldsymbol{\gamma}$  is known. Our estimator of  $\mu$  will then be

$$\hat{\mu} = \frac{1}{n} \sum_{i:R_i=1} \pi(\mathbf{X}_i)^{-1} Y_i = \frac{1}{n} \sum_{i=1}^n \pi(\mathbf{X}_i)^{-1} R_i Y_i. \quad (8)$$

Intuitively, the idea behind this is as follows. Suppose, for example, that  $\pi(\mathbf{x}) = \frac{1}{4}$ . This means that, on the average, for every unit with  $\mathbf{X} = \mathbf{x}$  that yields an observed value, there are three other units with  $\mathbf{X} = \mathbf{x}$  that yield missing values. So, given a unit with  $\pi(\mathbf{x}) = \frac{1}{4}$  that yielded an observed value, we make up for the missing values by introducing three additional virtual values each equal to that unit’s observed value. In other words, we multiply the observed unit’s value by 4, which is  $\pi(\mathbf{x})^{-1}$ . A similar argument applies for value of  $\pi(\mathbf{x})$  other than  $\frac{1}{4}$ .

We show that the above estimator is an unbiased estimator for  $\mu$  by showing that  $E[\pi(\mathbf{X}_i)^{-1} R_i Y_i] = \mu$ . This is accomplished by the following argument:

$$\begin{aligned} E[\pi(\mathbf{X}_i)^{-1} R_i Y_i] &= E[E[\pi(\mathbf{X}_i)^{-1} R_i Y_i | \mathbf{X}_i]] = E[\pi(\mathbf{X}_i)^{-1} E[R_i Y_i | \mathbf{X}_i]] \\ &= E[\pi(\mathbf{X}_i)^{-1} E[R_i | \mathbf{X}_i] E[Y_i | \mathbf{X}_i]] = E[\pi(\mathbf{X}_i)^{-1} \pi(\mathbf{X}_i) E[Y_i | \mathbf{X}_i]] \\ &= E[E[Y_i | \mathbf{X}_i]] = E[Y_i] = \mu. \end{aligned}$$

In the second line above, we used the conditional independence condition (1) and the fact that  $E[R_i | \mathbf{X}_i] = \Pr(R_i = 1 | \mathbf{X}_i) = \pi(\mathbf{X}_i)$ .

In practice,  $\boldsymbol{\gamma}$  will have to be estimated. The standard maximum likelihood estimate  $\hat{\boldsymbol{\gamma}}$  for the logistic regression model can be used. We then replace  $\boldsymbol{\gamma}$  by  $\hat{\boldsymbol{\gamma}}$  in the expression for  $\hat{\mu}$  (i.e.  $\pi(\mathbf{X}_i)$  is replaced by an estimate  $\hat{\pi}(\mathbf{X}_i)$  based on the estimate  $\hat{\boldsymbol{\gamma}}$  of  $\boldsymbol{\gamma}$ ). The resulting estimate of  $\mu$  is no longer exactly unbiased, but since  $\hat{\boldsymbol{\gamma}}$  is (under mild conditions) a consistent estimator of  $\boldsymbol{\gamma}$ , the estimate  $\hat{\mu}$  with  $\boldsymbol{\gamma}$  replaced by  $\hat{\boldsymbol{\gamma}}$  is a consistent estimator of  $\mu$ .

We now develop an estimator of the variance of the estimator  $\hat{\mu}$  (based on  $\hat{\boldsymbol{\gamma}}$ ). This development is specific to the logistic regression model (7), but a similar argument can be used for other models. Let us write

$$\lambda_i = \pi(\mathbf{X}_i)^{-1} = 1 + \exp(-\boldsymbol{\gamma}^T \mathbf{X}(i)),$$

and define  $\hat{\lambda}_i$  by the corresponding expression with  $\boldsymbol{\gamma}$  replaced by  $\hat{\boldsymbol{\gamma}}$ . By Taylor approximation we have

$$\hat{\lambda}_i - \lambda_i \doteq -\exp(\boldsymbol{\gamma}^T \mathbf{X}(i)) \mathbf{X}(i)^T (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}).$$

Thus,

$$\hat{\mu} \doteq \frac{1}{n} \sum_{i=1}^n U_i - \bar{\mathbf{X}}_*^T (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}), \quad (9)$$

where  $U_i = \lambda_i R_i Y_i$  and

$$\bar{\mathbf{X}}_* = \frac{1}{n} \sum_{i=1}^n \exp(-\boldsymbol{\gamma}^T \mathbf{X}(i)) \mathbf{X}(i).$$

Similarly to what we saw for the regression imputation method, the two terms in (9) are asymptotically uncorrelated because the conditional expectation of  $\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}$  given all the  $X_{ij}$  values tends to zero. The variance of the first term is  $\sigma_U^2/n$ , where  $\sigma_U^2$  is the variance of  $U$ . If  $\boldsymbol{\gamma}$  were known, we would estimate  $\sigma_U^2$  by the classical estimate

$$s_U^2 = \frac{1}{n-1} \sum_{i=1}^n (U_i - \bar{U}_i)^2.$$

Since  $\boldsymbol{\gamma}$  is unknown, we replace it with the estimate  $\hat{\boldsymbol{\gamma}}$ . We denote the resulting estimator by  $\hat{s}_U^2$ .

We turn now to the second term in (9). By standard theory for the logistic regression model, the covariance matrix of  $\hat{\boldsymbol{\gamma}}$  may be estimated by  $\hat{\mathbf{V}} = (\mathcal{X}^T \boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}}) \mathcal{X})^{-1}$ , where  $\boldsymbol{\Omega}(\boldsymbol{\gamma})$  is a diagonal matrix with diagonal entries  $\pi(\mathbf{X}_i)(1 - \pi(\mathbf{X}_i))$ . Hence the variance of the second term in (9) can be estimated by  $\tilde{\mathbf{X}}^T \hat{\mathbf{V}} \tilde{\mathbf{X}}$ , where

$$\tilde{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \exp(-\hat{\boldsymbol{\gamma}}^T \mathbf{X}(i)) \mathbf{X}(i).$$

Thus, overall we have

$$\widehat{\text{Var}}(\hat{\mu}) = \frac{\hat{s}_U^2}{n} + \tilde{\mathbf{X}}^T \hat{\mathbf{V}} \tilde{\mathbf{X}}. \quad (10)$$

### 2.2.3. Double Robust Method

Robins, Rotnitzky, and Zhao (1994, JASA) described a “double robust” method for obtaining an estimate of  $\mu$  which is valid if *either* the relationship between  $Y_i$  and  $\mathbf{X}_i$  *or* the relationship between  $R_i$  and  $\mathbf{X}_i$  is modelled correctly. Suppose we model  $E[Y_i|\mathbf{X}_i]$  as  $\mathbf{X}(i)^T \boldsymbol{\beta}$  and  $\Pr(R_i = 1|\mathbf{X}_i)$  as  $\pi(\mathbf{X}_i)$ . Then the estimator of  $\mu$  is given by

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \left[ \frac{R_i}{\pi(\mathbf{X}_i)} Y_i - \left( \frac{R_i - \pi(\mathbf{X}_i)}{\pi(\mathbf{X}_i)} \right) \mathbf{X}(i)^T \boldsymbol{\beta} \right]. \quad (11)$$

Note that the first term inside the sum drops out if  $R_i = 0$ , so that there is no dependence on unobserved  $Y_i$  values.

We have

$$\begin{aligned} E[\hat{\mu}] &= \frac{1}{n} \sum_{i=1}^n E \left[ \frac{\Pr(R_i = 1|\mathbf{X}_i)}{\pi(\mathbf{X}_i)} E[Y_i|\mathbf{X}_i] - \left( \frac{\Pr(R_i = 1|\mathbf{X}_i) - \pi(\mathbf{X}_i)}{\pi(\mathbf{X}_i)} \right) \mathbf{X}(i)^T \boldsymbol{\beta} \right] \\ &= \frac{1}{n} \sum_{i=1}^n E \left[ E[Y_i|\mathbf{X}_i] - \left( \frac{\Pr(R_i = 1|\mathbf{X}_i) - \pi(\mathbf{X}_i)}{\pi(\mathbf{X}_i)} \right) (\mathbf{X}(i)^T \boldsymbol{\beta} - E[Y_i|\mathbf{X}_i]) \right]. \end{aligned}$$

The second term inside the sum in the second line is equal to zero if either the model  $E[Y_i|\mathbf{X}_i] = \mathbf{X}_i^T\boldsymbol{\beta}$  or the model  $\Pr(R_i = 1|\mathbf{X}_i) = \pi(\mathbf{X}_i)$  is correct. In this case we get  $E[\hat{\mu}] = \mu$ . In practice, we substitute estimates for  $\boldsymbol{\beta}$  and  $\pi(\mathbf{X}_i)$ , and we get a consistent estimator of  $\mu$ . An expression for the variance of this estimator can be worked out and written down, but because the development is complicated we will not present the variance expression here.

### 3. Simple Methods for Informatively Missing Data

Often we are faced with a dataset with missing data where we suspect the missingness is related to the response, but the exact relationship is not known, and the relationship cannot be explained in terms of known background variables. In this section, we present some simple methods for dealing with this situation. It is possible to generalize the discussion to the case where the relationship can be partially explained in terms of background variables with a remaining element of the relationship that cannot be explained. However, for simplicity, we will focus here on the case where no adjustments for explanatory variables are being made.

#### 3.1. Worst Case Analysis

The “worst-case analysis” approach involves presenting an extreme “low-ball” estimate  $\hat{\mu}_L$  of  $\mu$  and an extreme “high-ball” estimate  $\hat{\mu}_U$  of  $\mu$ , with the idea that the truth is somewhere in the middle. We define  $Y_L^*$  to be a  $Y$  value sufficiently low so that  $Y$  values lower than  $Y_L^*$  are implausible, and  $Y_U^*$  to be a  $Y$  value sufficiently high so that  $Y$  values higher than  $Y_U^*$  are implausible. We then form the estimate  $\hat{\mu}_L$  by replacing all missing  $Y$  values with  $Y_L^*$ , and the estimate  $\hat{\mu}_U$  by replacing all missing  $Y$  values with  $Y_U^*$ . Thus

$$\hat{\mu}_L = \hat{\pi}\bar{Y} + (1 - \hat{\pi})Y_L^*, \quad (12)$$

$$\hat{\mu}_U = \hat{\pi}\bar{Y} + (1 - \hat{\pi})Y_U^*, \quad (13)$$

where  $\bar{Y}$  is the sample mean of the observed  $Y$  values and  $\hat{\pi}$  is the proportion of units in the original sample having an observed  $Y$  value (i.e.  $\hat{\pi} = n'/n$ , where  $n'$  is the number of units with observed data).

This approach is a very conservative, play-it-safe approach. We can have a high degree of confidence in the conclusion that we state, at the cost that the conclusion may be very imprecise. If the percentage of missing data is high, the range spanned by  $\hat{\mu}_L$  and  $\hat{\mu}_U$  will be so wide as to be essentially useless. This simply reflects the fact that with a high percentage of missing data, it is impossible to draw firm conclusions.

The specification of  $Y_L^*$  and  $Y_U^*$  calls for some attention. In some situations, there will be natural bounds on the range of  $Y$ , which can be used to define  $Y_L^*$  and  $Y_U^*$ . A common case of this is where the response variable  $Y$  is a 0–1 binary variable (e.g. the answer to a yes/no survey question), in which case we can take  $Y_L^* = 0$  and  $Y_U^* = 1$ . In other situations, the choice of  $Y_L^*$  and  $Y_U^*$  will not be so obvious, and will involve some element of judgment, based on knowledge of the subject matter and results of past studies in the area.

Note that  $\hat{\mu}_L$  and  $\hat{\mu}_U$  are NOT the endpoints of a confidence interval for  $\mu$ . They are two alternate point estimates based on two different assumptions about the missing data. To construct a confidence interval, we need to take into account the variability in  $\bar{Y}$  and  $\hat{\pi}$ .

The variance of  $\bar{Y}$  as an estimator of  $\mu_1 = E[Y_i|R_i = 1]$  can be estimated by  $s^2/n'$ , where  $s^2$  is the usual sample variance for the observed data. The variance of  $\hat{\pi}$  can be estimated by the usual variance formula for proportions:  $\hat{\pi}(1 - \hat{\pi})/n$ . The random variables  $\bar{Y}$  and  $\hat{\pi}$  are uncorrelated because  $E[(\bar{Y} - \mu_1)|R_1, \dots, R_n] = 0$ , while  $\hat{\pi}$  is a function of the  $R_i$ 's only (see if you can fill in the missing steps in this argument). By the multivariate central limit theorem,  $(\bar{Y}, \hat{\pi})$  is asymptotically bivariate normal.

Now we develop a variance formula for  $\hat{\mu}_L$  and  $\hat{\mu}_U$ . We can write

$$\hat{\mu}_L = q(\bar{Y}, \hat{\pi}),$$

where

$$q(u, w) = uw + (1 - w)Y_L^*.$$

We have

$$\frac{\partial q}{\partial u} = w, \quad \frac{\partial q}{\partial w} = u - Y_L^*.$$

Hence, applying the delta method, we can estimate the variance of  $\hat{\mu}_L$  by

$$\widehat{\text{Var}}(\hat{\mu}_L) = \left(\frac{\partial q}{\partial u}\right)^2 \widehat{\text{Var}}(\bar{Y}) + \left(\frac{\partial q}{\partial w}\right)^2 \widehat{\text{Var}}(\hat{\pi}) + 2 \left(\frac{\partial q}{\partial u}\right) \left(\frac{\partial q}{\partial w}\right) \widehat{\text{Cov}}(\bar{Y}, \hat{\pi}) \quad (14)$$

$$= \frac{1}{n'} \hat{\pi}^2 s^2 + \frac{1}{n} \hat{\pi}(1 - \hat{\pi})(\bar{Y} - Y_L^*)^2, \quad (15)$$

where the partial derivatives in the first line above are evaluated at  $(u, w) = (\bar{Y}, \hat{\pi})$ . A corresponding result holds for  $\hat{\mu}_U$  (just replace “L” by “U” in all the above).

The above results allow us to build a  $100(1 - \alpha)\%$  confidence interval (for any given  $\alpha$ , e.g.  $\alpha = 0.05$ ) around each of the point estimates  $\hat{\mu}_L$  and  $\hat{\mu}_U$ . In the end, we can construct an interval estimate for the the population mean  $\mu$  that takes into account both the different assumptions about the missing data and the variability in  $\bar{Y}$  and  $\hat{\pi}$ . The lower and upper limits of this interval are given by

$$\tilde{\mu}_L = \hat{\mu}_L - z_{1-\frac{1}{2}\alpha} \widehat{\text{Var}}(\hat{\mu}_L)^{\frac{1}{2}}, \quad (16)$$

$$\tilde{\mu}_U = \hat{\mu}_U + z_{1-\frac{1}{2}\alpha} \widehat{\text{Var}}(\hat{\mu}_U)^{\frac{1}{2}}, \quad (17)$$

where the variance estimates are as given above and  $z_{1-\frac{1}{2}\alpha}$  is the  $100(1 - \frac{1}{2}\alpha)$ -th percentile of the standard normal distribution (e.g. 1.96 for  $\alpha=0.05$ ).

### 3.2. Simple Sensitivity Analysis

The idea here is to make some simple assumption about the behavior of the missing data as a function of some adjustable parameter, and then examine how the results vary over a “reasonable” range of the parameter. For example, we can assume

$$E[Y_i|R_i = 0] = aE[Y_i|R_i = 1], \quad (18)$$

where  $a$  is a factor that can be varied. We then estimate the population mean by

$$\hat{\mu}(a) = \hat{\pi}\bar{Y} + (1 - \hat{\pi})[a\bar{Y}] = \bar{Y}(\hat{\pi} + (1 - \hat{\pi})a). \quad (19)$$

We then can vary  $a$  over a “reasonable” range and see how the estimate changes. What constitutes a “reasonable range” is a matter of judgment, which will depend on intuition, knowledge of the subject matter area, and results from relevant past studies.

Using a delta method argument as in the preceding subsection, we can show that  $\hat{\mu}$  has an asymptotic normal distribution with variance that can be estimated by

$$\widehat{\text{Var}}(\hat{\mu}) = \frac{1}{n'}(\hat{\pi} + (1 - \hat{\pi})a)^2 + \frac{1}{n}\hat{\pi}(1 - \hat{\pi})(1 - a)^2\bar{Y}^2.$$

This result can be used to form confidence intervals around the estimate  $\hat{\mu}(a)$ .

If there is a high percentage of missing data, then typically  $\hat{\mu}(a)$  will vary widely over any reasonable range of  $a$ . In this case, it will be impossible to draw definite conclusions from the data.

The model (18) is only one possible approach to simple sensitivity analysis in the presence of missing data. Another possible approach is to assume that a certain fraction  $c$  of the missing data are missing completely at random, and apply a “worst-case” analysis to the remaining fraction of the missing data. This would give lower and upper estimates of the following form:

$$\hat{\mu}_L(c) = [\hat{\pi} + c(1 - \hat{\pi})]\bar{Y} + (1 - c)(1 - \hat{\pi})Y_L^*, \quad (20)$$

$$\hat{\mu}_U(c) = [\hat{\pi} + c(1 - \hat{\pi})]\bar{Y} + (1 - c)(1 - \hat{\pi})Y_U^*, \quad (21)$$

The value of  $c$  then could be varied over a “reasonable” range and we can examine how the results change as  $c$  is varied.

There is an endless range of other possible approaches along similar lines that can be used. The statistician can use whatever approach seems best for the situation at hand.

### 3.3. Recontacting (a Sample of) the Nonresponders

In certain situations it is possible to make a second attempt to get the data that initially was missing. For example, in a mail survey, we can mail out the questionnaire a second time to those who did not respond the first time, or contact these nonrespondents by phone. Often, instead of recontacting all the the nonresponders, we will recontact a random sample of them, say  $100\omega\%$  of them. We then estimate the population mean  $\mu$  by

$$\hat{\mu} = \hat{\pi}\bar{Y}_O + (1 - \hat{\pi})\bar{Y}_M, \quad (22)$$

where  $\bar{Y}_O$  is the sample mean among units with observed data on the first attempt,  $\bar{Y}_M$  is the sample mean among the units for which a second attempt was made, and  $\hat{\pi}$  is the proportion of units that yielded an observed value on the first attempt. We may

note that  $\bar{Y}_O$  is an unbiased estimate of  $\mu_1 = E[Y_i|R_i = 1]$ , while  $\bar{Y}_M$  is an estimate of  $\mu_0 = E[Y_i|R_i = 0]$ .

If we are fortunate enough to get responses on all the units for which a second attempt was made, then  $\bar{Y}_M$  is in fact an unbiased estimate of  $\mu_0$ . Hence, in this case,  $\hat{\mu}$  is an unbiased estimator of  $\mu$ . Further, using delta method arguments, it can be shown that  $\hat{\mu}$  is asymptotically normal with variance that can be estimated by the following formula:

$$\widehat{\text{Var}}(\hat{\mu}) = \frac{\hat{\pi}s_O^2}{n} + \frac{(1-\hat{\pi})s_M^2}{\omega n} + \frac{1}{n}[\hat{\pi}(\bar{Y}_R - \hat{\mu})^2 + (1-\hat{\pi})(\bar{Y}_M - \hat{\mu})^2], \quad (23)$$

where  $s_O^2$  and  $s_M^2$  are the sample variances, respectively, for those with observed data on the first attempt and those sampled for a second attempt. Given some initial guesses of the variances  $\sigma_O^2 = \text{Var}(Y_i|R_i = 1)$  and  $\sigma_M^2 = \text{Var}(Y_i|R_i = 0)$  based on past studies, we can decide what sampling fraction  $\omega$  is appropriate for sampling the units with missing data on the first attempt.

In real life, usually some fraction of the units on which a second attempt was made will fail to yield a response. In this case, we will not have solved the missing data problem completely, but at least we will have reduced it. The approaches described in the preceding two subsections can be used to explore the impact of the missing data in the second-attempt sample on the estimate of  $\mu_0$ . The results of this analysis can then be used to explore the impact of these missing data on the estimate of  $\mu$ .

Nearly all professional surveys use this second-attempt technique wherever possible in order to address the missing data problem. It is possible to extend the technique to allow for a number of additional attempts; for details, one may consult a text on survey methods.

On the other hand, there are situations where the second-attempt technique cannot be applied. For example, in a medical study, if a patient fails to show up for the 6-month visit, the missing data cannot be recovered. In this case, we remain with a missing value.

#### 4. Missing Data Modeling

In this section we will describe model-based methods for handling missing data that are more sophisticated than the simple approaches presented in Sections 3.1 and 3.2 above. The discussion here is based on the paper of Greenless et al. (1982, JASA). We first present the technical development of the method, and then make some remarks about its application. For concreteness, we will assume here that  $Y$  is a continuous variable, but a parallel development can be given for the discrete case (just replace the density function with a probability mass function and integrals with sums).

The method is built upon the following assumptions.

*Assumption 1:* The response variable  $Y_i$  is distributed according to a specified parametric density function  $f(y; \boldsymbol{\theta})$ , such as  $N(\mu, \sigma^2)$  or some other parametric distribution.

*Assumption 2:* The conditional probability  $\Pr(R_i = 1|Y_i = y)$  is given by a function  $\pi^*(y, \gamma)$  of specified form, depending on some unknown parameters  $\gamma$ . For instance, if the response is number of years of education, and we suspect that well-educated people will be more likely to respond than poorly-educated people, then we could use a function  $\pi^*(y, \gamma)$  that is monotone in  $y$ . A typical choice is the logistic model

$$\pi^*(y, \gamma) = \frac{e^{\gamma_0 + \gamma_1 y}}{1 + e^{\gamma_0 + \gamma_1 y}}. \quad (24)$$

Alternately, we might suspect that the probability of response is higher in the middle range of the distribution of  $Y$  than at either extreme. This may be the case, for instance, in a survey of salary. In this case, we could use a logistic model involving a quadratic function of  $y$ , that is

$$\pi^*(y, \gamma) = \frac{e^{\gamma_0 + \gamma_1 y + \gamma_2 y^2}}{1 + e^{\gamma_0 + \gamma_1 y + \gamma_2 y^2}}. \quad (25)$$

Given the above assumptions, we can construct the likelihood function of the data. The likelihood construction here differs a bit from the standard case because of the missing data. Let us define  $\pi(\boldsymbol{\theta}, \gamma) = \Pr_{\boldsymbol{\theta}, \gamma}(R_i = 1)$ . This is given by

$$\pi(\boldsymbol{\theta}, \gamma) = \int f(y; \boldsymbol{\theta}) \pi^*(y, \gamma) dy. \quad (26)$$

Then the story is as follows. For a unit  $i$  with observed data, the available information is the fact that  $Y_i$  was observed along with the value of  $Y_i$  itself. Thus, this unit's contribution to the likelihood function is given by  $\pi^*(Y_i, \gamma) f(Y_i; \boldsymbol{\theta})$ . For a unit  $i$  with missing data, the only information available is the fact that the  $Y_i$  value was missing. Thus, this unit's contribution to the likelihood function is  $1 - \pi(\boldsymbol{\theta}, \gamma)$ . Hence the overall likelihood function is given by

$$L(\boldsymbol{\theta}, \gamma) = \left[ \prod_{i: R_i=1} \pi^*(Y_i, \gamma) f(Y_i; \boldsymbol{\theta}) \right] (1 - \pi(\boldsymbol{\theta}, \gamma))^{n_{miss}}, \quad (27)$$

where  $n_{miss}$  is the number of units with missing data. From here, maximum likelihood estimates (MLE) of  $\boldsymbol{\theta}$  and  $\gamma$  can be obtained in the standard manner (i.e. we compute the partial derivatives of the log likelihood function and set them to zero). From standard maximum likelihood theory, the MLE vector  $(\hat{\boldsymbol{\theta}}, \hat{\gamma})$  has an approximate multivariate normal distribution with expectation  $(\boldsymbol{\theta}, \gamma)$  and covariance matrix given by the inverse of the Fisher information matrix  $\mathbf{I}(\boldsymbol{\theta}, \gamma)$ . In practice, this matrix will be estimated by  $\mathbf{J}(\hat{\boldsymbol{\theta}}, \hat{\gamma})$ , where  $\mathbf{J}$  is  $-1$  times the matrix of second derivatives of the log likelihood function.

Given the estimates of  $\boldsymbol{\theta}$  and  $\gamma$ , we can construct an estimate of the population mean  $\mu$  in one of the following two ways.

- a. Fully model-based estimator: We define

$$\mu(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[Y] = \int y f(y; \boldsymbol{\theta}) dy. \quad (28)$$

We then take  $\hat{\mu} = \mu(\hat{\boldsymbol{\theta}})$ . This estimator relies completely on the model specified by Assumptions 1 and 2.

b. Imputation-based estimator: Here we instead use observed data wherever we have it, and use the model only to fill in the missing data. We thus define  $Y_i^*$  to be equal to  $Y_i$  when the value  $Y_i$  is observed, and equal to  $\mu_0(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$  otherwise, where

$$\mu_0(\boldsymbol{\theta}, \boldsymbol{\gamma}) = E_{\boldsymbol{\theta}, \boldsymbol{\gamma}}[Y_i | R_i = 0] = [1 - \pi(\boldsymbol{\theta}, \boldsymbol{\gamma})]^{-1} \int y f(y; \boldsymbol{\theta}) (1 - \pi^*(y, \boldsymbol{\gamma})) dy. \quad (29)$$

This quantity can be computed either by numerical integration (consult a calculus or numerical analysis book for a discussion of this) or by simulation. For the simulation approach, see the discussion of multiple imputation in the following section. We then take our estimator  $\hat{\mu}$  to be the mean of the  $Y^*$  values. In other words,

$$\hat{\mu} = \hat{\pi} \bar{Y} + (1 - \hat{\pi}) \mu_0(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}), \quad (30)$$

where, as earlier,  $\bar{Y}$  denotes the sample mean of the observed data and  $\hat{\pi}$  is the proportion of units in the sample that yielded an observed value.

Let us now consider how to estimate the variances of the above estimators of  $\mu$ . For Method (a), the delta method can be used. Define  $\nabla \mu(\boldsymbol{\theta})$  to be the gradient of  $\mu(\boldsymbol{\theta})$ , i.e. the vector whose  $r$ -th component is  $\partial \mu(\boldsymbol{\theta}) / \partial \theta_r$ . Then, according to the multivariate delta method, and under the condition that Assumptions 1 and 2 hold, the estimator  $\mu(\hat{\boldsymbol{\theta}})$  has an asymptotic normal distribution with expectation  $\mu(\boldsymbol{\theta})$  and variance that can be estimated by

$$\widehat{\text{Var}}(\mu(\hat{\boldsymbol{\theta}})) = \nabla \mu(\hat{\boldsymbol{\theta}})^T \mathbf{J}(\hat{\boldsymbol{\theta}})^{-1} \nabla \mu(\hat{\boldsymbol{\theta}}) = \sum_r \sum_s \left( \frac{\partial \mu(\hat{\boldsymbol{\theta}})}{\partial \theta_r} \right) \left( \frac{\partial \mu(\hat{\boldsymbol{\theta}})}{\partial \theta_s} \right) [\mathbf{J}(\hat{\boldsymbol{\theta}})^{-1}]_{rs}. \quad (31)$$

For Method (b), in principle a theoretical formula for estimating the variance of  $\hat{\mu}$  can be worked out, but the development is complex. An alternate, simpler, way to arrive at an estimate of the variance is through the multiple imputation approach described in the following section.

Now for a couple of remarks about this approach.

The key point to note is that the approach depends critically on the model defined by Assumptions 1 and 2. Now since we only get to observe  $Y_i$  when  $R_i = 1$ , it is impossible to check these model assumptions directly. Hence it is proper to carry out the analysis under an number of different plausible models and see how the results change depending on the model. The specification of what models are plausible is a matter of judgment, guided by intuition, knowledge of the subject matter, and results of past studies. As an example, if on logical grounds we expect the distribution of  $Y$  in the overall population to be symmetric about its expectation, then we can focus mainly on symmetric density functions  $f(y; \boldsymbol{\theta})$  (though it would be wise to consider some models with a modest degree of skewness as well). As usual, if the percentage of missing data is high, we probably will not be able to draw clear conclusions: the

range of answers we get under different plausible scenarios will be too wide to say anything definite.

Another point to mention is that the entire discussion above can be generalized to the case where there are explanatory variables. We can allow for one set of explanatory variables  $\mathbf{X}_i$  affecting the distribution of  $Y_i$  and another set of explanatory variables  $\mathbf{Z}_i$  affecting the probability that  $R_i = 1$  (there can be some overlap between  $\mathbf{X}_i$  and  $\mathbf{Z}_i$ ).

. Under this extended model,  $Y_i$  will have density function  $f(y; \boldsymbol{\theta}|\mathbf{X}_i)$  (this could be, for instance, a classical linear regression model for  $Y_i$  as a function of  $\mathbf{X}_i$ ), while  $\Pr(R_i = 1|Y_i = y, \mathbf{Z}_i)$  would be represented by some function  $\pi^*(y, \boldsymbol{\gamma}|\mathbf{Z}_i)$  (for instance, a logistic model incorporating the components of  $\mathbf{Z}_i$  as explanatory variables along with  $y$  and possibly  $y^2$ ). In parallel with (26), we define

$$\pi(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{X}_i, \mathbf{Z}_i) = \int f(y; \boldsymbol{\theta}|\mathbf{X}_i)\pi^*(y, \boldsymbol{\gamma}|\mathbf{Z}_i)dy. \quad (32)$$

We then construct the likelihood function as

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \left[ \prod_{i:R_i=1} \pi^*(Y_i, \boldsymbol{\gamma}|\mathbf{Z}_i)f(Y_i; \boldsymbol{\theta}|\mathbf{X}_i) \right] \left[ \prod_{i:R_i=0} (1 - \pi(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{X}_i, \mathbf{Z}_i)) \right]. \quad (33)$$

Given  $L(\boldsymbol{\theta}, \boldsymbol{\gamma})$ , we estimate  $\boldsymbol{\theta}$  and  $\boldsymbol{\gamma}$  by maximum likelihood as before. The population mean  $\mu$  then can be estimated as follows. Define

$$\mu(\boldsymbol{\theta}|\mathbf{X}_i) = E_{\boldsymbol{\theta}}[Y_i|\mathbf{X}_i] = \int yf(y; \boldsymbol{\theta}|\mathbf{X}_i)dy, \quad (34)$$

$$\mu_0(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{X}_i, \mathbf{Z}_i) = E_{\boldsymbol{\theta}, \boldsymbol{\gamma}}[Y_i|R_i = 0, \mathbf{X}_i, \mathbf{Z}_i] \quad (35)$$

$$= [1 - \pi(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{X}_i, \mathbf{Z}_i)]^{-1} \int yf(y; \boldsymbol{\theta}|\mathbf{X}_i)(1 - \pi^*(y, \boldsymbol{\gamma}|\mathbf{Z}_i))dy. \quad (36)$$

The Method (a) estimate of  $\mu$  then becomes

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \mu(\hat{\boldsymbol{\theta}}|\mathbf{X}_i). \quad (37)$$

A formula for estimating the variance of this estimator can be derived using an argument similar to that used in Section 2.2 and 2.3. The Method (b) estimate of  $\mu$  becomes

$$\hat{\mu} = \hat{\pi}\bar{Y} + (1 - \hat{\pi}) \left[ \frac{\sum_{i:R_i=0} \mu_0(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}|\mathbf{X}_i, \mathbf{Z}_i)}{n'} \right]. \quad (38)$$

The variance of this estimator can be estimated by the multiple imputation method described in the following section.

## 5. Multiple Imputation

The term ‘‘imputation’’ refers to filling in missing values with guessed values. If the filled-in sample is analyzed in a naive way, as if all the values were real observations, the results can be misleading. The naive analysis fails to take into account the uncertainty

in the guessed values, and thus the variance of the parameter estimator of interest is underestimated. Multiple imputation is a general-purpose method for dealing with this problem. The idea is to carry out the imputation a number of times according to a random mechanism that reflects the uncertainty in the guessed values. We then factor in the variability of the parameter estimate values over the different imputations.

We illustrate the application of this idea in the context of the Greenlees et al. method. The procedure is as follows.

A. Begin with the model  $Y_i \sim f(y; \boldsymbol{\theta})$  and  $\Pr(R_i = 1|Y_i = y) = \pi^*(y, \boldsymbol{\gamma})$ . As explained in the preceding section, it is wise to try a number of different models and see how the answers vary across the models. Also, as above, it is possible to incorporate explanatory variables  $\mathbf{X}_i$  and  $\mathbf{Z}_i$  into the analysis, but for simplicity we will not do this here.

B. Estimate  $\boldsymbol{\theta}$  and  $\boldsymbol{\gamma}$  by maximum likelihood, as in the Greenlees method.

C. Draw values  $\tilde{\boldsymbol{\theta}}$  and  $\tilde{\boldsymbol{\gamma}}$  at random from the multivariate normal distribution with mean  $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$  and covariance matrix given by the estimated covariance matrix of  $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$ . This step is designed to factor in the uncertainty in the values of  $\boldsymbol{\theta}$  and  $\boldsymbol{\gamma}$ .

D. If  $Y_i$  is observed, then set  $\tilde{Y}_i = Y_i$ . If  $Y_i$  is missing, then fill in an imputed value  $\tilde{Y}_i$  for  $Y_i$  according to the conditional distribution of  $Y_i$  given  $R_i = 0$ . This can be accomplished through the following scheme.

- i. Draw a candidate value  $\tilde{Y}_i$  from the distribution  $f(y; \tilde{\boldsymbol{\theta}})$ .
- ii. Draw a  $U(0, 1)$  random number  $\tilde{U}_i$ .
- iii. If  $\tilde{U}_i \leq 1 - \pi^*(\tilde{Y}_i, \tilde{\boldsymbol{\gamma}})$ , then accept  $\tilde{Y}_i$ . Otherwise, reject  $\tilde{Y}_i$ .
- iv. Repeat steps (i)-(iii) until a  $\tilde{Y}_i$  value is accepted.

E. Set

$$\tilde{\mu} = \frac{1}{n} \sum_{i=1}^n \tilde{Y}_i, \quad (39)$$

$$\tilde{v} = \frac{1}{n-1} \sum_{i=1}^n (\tilde{Y}_i - \tilde{\mu})^2. \quad (40)$$

F. Repeat Steps C-E  $m$  times, where  $m$  is some positive integer (in many applications,  $m$  of 5 or 10 will be enough). This gives  $\tilde{\mu}_1, \dots, \tilde{\mu}_m$  and  $\tilde{v}_1, \dots, \tilde{v}_m$ .

G. Compute

$$\mu^* = \frac{1}{m} \sum_{k=1}^m \tilde{\mu}_k, \quad (41)$$

$$v^* = \frac{1}{m} \sum_{k=1}^m \tilde{v}_k + \left( \frac{m+1}{m} \right) \left[ \frac{1}{m-1} \sum_{k=1}^m (\tilde{\mu}_k - \mu^*)^2 \right]. \quad (42)$$

(The factor  $(m + 1)/m$  is a correction factor to make the method more accurate for small  $m$ .)

H. Take  $\mu^*$  as the estimate of  $\mu$  and regard  $\mu^*$  as approximately normal with mean  $\mu$  and variance  $v^*$ .

## 6. Preventing Missing Data

In this section, we discuss a number of common-sense techniques to reduce missing data in studies involving people. We begin with some general points, and then present some points that are specific to surveys.

### General Points

a. Explain to the person what you are trying to accomplish. If you explain to the person in a clear and appealing manner what the goals of the research are and what you are asking from him, the person will be more likely to cooperate.

b. Make the research demands reasonable. In a survey, if the questionnaire is too long or the questions take too much thought, many people will just throw it away. In a medical study, if too many clinic visits are involved, the person may refuse to participate. Or worse, the person may agree initially but lose interest later and drop out or miss visits. In general, if the research requires excessive time and effort on the participants, people will not want to participate.

c. Incentives. In certain cases (but not all), it may be appropriate and useful to offer a monetary or other incentive to encourage participation.

### Points Relating Mainly to Surveys

a. Time the survey appropriately. Avoid carrying out a survey during times of the year (e.g. holiday or vacation periods) when people will be difficult to reach.

b. Interviewers. The interviewers should be able to interact with the interviewees in a pleasant way. In some cases, special attention may be needed in the choice of interviewers because of the target population.

c. Questionnaire design. The questionnaire should be easy to understand. In particular, the questions should be clear. The questions should be phrased so as to have a unequivocal meaning: there should not be more than one way to interpret the question. (This affects both the response rate and the validity of the responses that are obtained.) If the questionnaire is too confusing, many people will just give up.

d. Follow-up. With people who do not respond to the initial contact, it is a good idea to follow up with additional attempts. This idea was already discussed in Section 3.3.

e. The type of survey can influence the response rate. There are three main types of survey: mail, phone, and in-person interview. Typically the personal interview is

best for maximizing the response rate and the phone survey is next best, although this could depend on the situation. Obviously cost is a factor: mail surveys are cheap, in-person interviews are expensive, and phone interviews are somewhere in between. One should pick the method that is most suitable for the specific situation, in view of cost and response rate considerations. Often it will be effective to conduct a survey by mail with phone follow-up on the nonrespondents.

In conducting a mail survey, it usually will help to include a pre-addressed, pre-stamped envelope for returning the completed questionnaire. It also will usually help to send reminders to the nonrespondents. The reminders can be either postcard reminders or resends of the original questionnaire. In line with General Point (a), it is useful to prepare a letter explaining the purpose of the survey which can be included along with the survey itself (in some cases it is useful to send the explanatory letter before the survey itself, to give people advance notice about the survey).

f. Sensitive subject matter. Certain topics, because of their sensitive nature, are inherently prone to nonresponse. People tend not to like to answer questions such as “Have you ever seen a psychiatrist?” or “Have you ever used drugs?”

One device that has been proposed in the survey research literature for attempting to get around this problem is what is called the method of *randomized response*. This method works as follows. Assume we are interested in getting answers to a sensitive yes/no question like the ones presented above. We tell the interviewee, in effect, to flip a coin with known probability of heads  $\phi$  (in practice this can be accomplished by asking the person to roll dice, or pick a card out of a given deck, etc.) The result of the coin flip is observed by the interviewee but not by the interviewer. The interviewee is told that if the coin comes up heads, he should answer the question with the truth, but if it comes up tails he should answer the opposite. The result of this process is that, for any individual person, it is impossible to tell with certainty from the person’s answer what the truth is. This can make the interviewees more willing to give a response (and to respond in accordance with the instructions rather than just answering “no”).

At the same time, the aggregate data on the sample as a whole can be used to estimate the population proportion  $\pi$  of people for whom the true answer is “yes.” Under the foregoing scheme, the probability that a subject will answer “yes” (assuming everyone follows the rules) is  $p = \phi\pi + (1 - \phi)(1 - \pi)$ . This can be estimated by the sample proportion  $\hat{p}$  of “yes” answers in the survey. The desired proportion  $\pi$  then can be estimated by  $\hat{\pi} = [\hat{p} - (1 - \phi)] / (2\phi - 1)$ . As regards the choice of  $\phi$ , as  $\phi$  gets closer to  $\frac{1}{2}$ , the individual responses become better hidden, but the variance of  $\hat{\pi}$  increases.