1. Complete-case analysis (I)

Complete-case analysis refers to analysis based on the cases with all variables observed. 

Advantages: (a) Simplicity; (b) Comparability.

Drawbacks: (a) Bias; (b) Loss of precision.

Suppose $\theta$ is a scalar parameter of interest. Denote $\hat{\theta}_{CC}$, $\hat{\theta}_{NM}$ and $\hat{\theta}_{EFF}$ the complete-case estimator, estimator based on the hypothetical complete data and the efficient estimator on the observed data, respectively. From

$$Var(\hat{\theta}_{CC}) = Var(\hat{\theta}_{NM})(1 + \Delta^*_CC),$$
$$Var(\hat{\theta}_{CC}) = Var(\hat{\theta}_{EFF})(1 + \Delta_{CC}),$$

$\Delta^*_CC$ and $\Delta_{CC}$ can be used to evaluate the relative efficiency of $\hat{\theta}_{CC}$.

Example 3.1. Bivariate normal monotone data $\{x_i, y_i\}_{i=1}^n$, where $y$ is missing for $i = r + 1, \ldots, n$ and data are MCAR. Then

1. For the mean of $X$, $\Delta^*_CC = \Delta_{CC} = \frac{n-r}{r}$;

2. For the mean of $Y$, $\Delta^*_CC = \frac{n-r}{r}$, but $\Delta_{CC} \approx \frac{(n-r)\rho^2}{n(1-\rho^2)+r\rho^2}$, where $\rho$ is the correlation coefficient between $X$ and $Y$.

3. For the regression coefficient of $Y$ on $X$, $\Delta^*_CC = \frac{n-r}{r}$ and $\Delta_{CC} = 0$.

Biostat 2065: Analysis of Incomplete Data, 2005
2. Complete-case analysis (II)

Example 3.2. In general, bias on mean estimator: \( \mu_{CC} - \mu = (1 - \pi_{CC}) \mu_{CC} - \mu \), where \( \pi_{CC} = pr(m = 0) \).

Example 3.3. Estimation of regression parameters.
When the probability of being a complete case only depends on \( X \), whether observed or missing values of \( X \), then the complete cases are a random sample for regression (or conditional distribution) of \( Y \) on \( X \), noted as \([Y \mid X]\). Then the complete-case estimator is consistent, though may not be efficient.

Example 3.4. Inference on odds ratio.
Suppose count data \((X, Y)\) form a \(2 \times 2\) contingency table, the complete-case analysis estimator is unbiased if the logarithm of the probability of response is an additive function of \( X \) and \( Y \).
3. Survey sampling & inference

Example 3.5. Randomization inference in surveys with complete response. Suppose in a size-$N$ population, the variable of interest is $Y$ (survey variable) and the variables for design are $Z$ (design variables). For each subject $i$, denote $\pi_i = pr(I_i = 1 \mid Y_i, Z_i)$. Sampling mechanisms:

1. Simple random sampling. $\pi_i = c$, where $c$ is a constant.

2. Stratified or unconfounded random sampling. $\pi_i = pr(I_i = 1 \mid Z_i)$.

Let $T$ denote a population quantity of interest. For example, $T$ is the population mean of $Y$, then an estimator is stratified mean:

$$t = \bar{y}_{st} = \frac{1}{N} \sum_{j=1}^{J} N_j \bar{y}_j,$$

where $\bar{y}_j$ is the sample mean in stratum $J$. The estimator $\bar{y}_{st}$ has variance

$$Var(\bar{y}_{st}) = \frac{1}{N^2} \sum_{j=1}^{J} N_j^2 \left( \frac{1}{n_j} - \frac{1}{N_j} \right) S^2_{yj},$$

where $S^2_{yj}$ is the population variance of $Y$ in stratum $j$. In variance estimation, $S^2_{yj}$ is replaced by the corresponding sample variance, i.e., $s^2_{yj}$.

Biostat 2065: Analysis of Incomplete Data, 2005.
4. Weighted estimates

Horvitz-Thompson estimator
Suppose in the sample, unit $i$ has probability $\pi_i$ being selected from the population, $i = 1, 2, \ldots, n$. Then the H-T estimator of the population mean is

$$y_{HT} = \frac{1}{N} \sum_{i=1}^{n} \pi_i^{-1} y_i.$$

The stratified mean estimator can be written as

$$\bar{y}_{st} = \bar{y}_w = \frac{1}{n} \sum_{i=1}^{n} w_i y_i,$$

where the weights, $w_i = \frac{n\pi_i^{-1}}{\sum_{k=1}^{n} \pi_k^{-1}}$, are scaled to sum to the sample size $n$. 

Biostat 2065: Analysis of Incomplete Data, 2005.
5. Weighted estimates for data with nonresponse

Suppose some of the sampled units do not respond, then for a unit $i$ the probability of
response is

$$pr(\text{selection and response}) = pr(\text{selection}) \times pr(\text{response } | \text{ selection}) = \pi_i \phi_i.$$  

Then a weighted estimator for the population mean is

$$\bar{y}_w = \frac{1}{r} \sum_{i=1}^{r} w_i y_i,$$

where $w_i = \frac{\pi_i \phi_i}{\sum_{k=1}^{r} (\pi_k \phi_k)^{-1}}$.

Note: in practice, $\pi_i$ is determined by the sampling mechanism; $\phi_i$ is usually unknown
and needs to be estimated.

**Example 3.6.** In stratified random sampling, suppose within each stratum (or weighting
class) $j$ the number of respondents is $r_j$ and the probability of response is uniform. Then
a natural estimate of $\phi_i$ is $\hat{\phi}_i = r_j/n_j$, for unit $i$ in stratum $j$. If the sampling weights
are uniform or constant, then the estimator can be written as:

$$\bar{y}_{wc} = n^{-1} \sum_{j=1}^{J} n_j \bar{y}_{jR},$$

where $\bar{y}_{jR}$ is the mean value of the respondents in stratum $j$. 
6. Propensity weighting

Let $X$ denote the set of variables observed for all subjects; variable $Y$ is only partially observed and we are interested in certain characteristics of $Y$, for example, the mean of $Y$. When the number of $X$ is limited, the weighting class estimator can be applied if data are MAR:

$$pr(M | X, Y) = pr(M | X).$$

The weighting class is $X$ and the response probability (or propensity) as $p(x_i) = pr(m_i = 0 | x_i)$, for unit $i$. From

$$pr\{m_i = 0 | y_i, p(x_i)\} = E[pr\{m_i = 0 | y_i, x_i\} | y_i, p(x_i)]$$

$$= E[pr\{m_i = 0 | x_i\} | y_i, p(x_i)] \quad \text{(by MAR)}$$

$$= E\{p(x_i) | y_i, p(x_i)\} = p(x_i),$$

in fact the weighting class can be defined by levels of $p(X)$ and within each level the respondents are a random subsample.
In practice, $p(x)$ is unknown,

1. we may estimate $p(X)$ by logistic regression of $M$ on $X$.

2. Then coarsen the estimates $\hat{p}(X)$ into five or six values (or intervals).

3. Regard the above classification as weighting class and apply the weighting class estimate with assuming that units in each class has the same value of response propensity; or simply weight each respondent $i$ by the $\hat{p}(x_i)^{-1}$.

4. Sensitive to model specification.
7. Weighted Generalized Estimating Equations

Let $Y = (Y_1, \ldots, Y_K)$ be a vector of variables and the observed data of $Y$, $\{y_i\}_{i=1}^n$, are subject to missingness. Assume the covariates $x_i = (x_{i1}, \ldots, x_{ip})$ are fully observed for all subjects, $i = 1, \ldots, n$. The regression of $Y$ on $X$ has mean structure $g(x, \beta)$ and $\beta$ is the set of parameters of interest, with size $d$.

1. When there are no missing values, the inference on $\beta$ can be done by sloving the following GEE:

$$\sum_{i=1}^n D_i(x_i, \beta)\{y_i - g(x_i, \beta)\} = 0,$$

where $D_i(x_i, \beta)$ is a $d \times K$ matrix and chosen according to the format or distribution of $y_i$.

2. When there are missing values, for example, the first $r$ cases are complete cases. The complete-case estimation is based on the following GEE:

$$\sum_{i=1}^r D_i(x_i, \beta)\{y_i - g(x_i, \beta)\} = 0,$$
3. When there are missing values and the missing-data mechanism is determined by

\[ pr(m_i = 0 \mid x_i, y_i, z_i) = pr(m_i = 0 \mid x_i, z_i; \alpha) = p(x_i, z_i; \alpha), \]

where \( z_i \) is a set of fully observed auxiliary variables. Then the inference on \( \theta \) can be done by solving the following weighted GEE:

\[ \sum_{i=1}^{r} p(x_i, z_i; \hat{\alpha})^{-1} D_i(x_i, \beta) \{ y_i - g(x_i, \beta) \} = 0, \]

where \( \hat{\alpha} \) is usually estimated by regression of \( M \) on \( X \) and \( Z \).

(a) The weighted GEE can correct bias of unweighted GEE when the mechanism depends on \( z \).
(b) Estimating \( \alpha \) leads to more efficient estimates than simply using the true value of \( \alpha \).
(c) This method can be extended to more general missing-data mechanisms with care.
8. Post-stratification and raking to known margins

Example 3.9. Suppose in the stratified sampling data, the proportion $N_j/N$ is known from external sources. The post-stratified mean is

$$\bar{y}_{ps} = \frac{1}{N} \sum_{j=1}^{J} N_j \bar{y}_j R.$$ 

Example 3.10. For a $J \times L$ contingency table classified by variables $X_1$ and $X_2$, in each cell $(j, l)$ suppose $Y$ is recorded for $r_{jl}$ units out of $n_{jl}$ sampled units among a $N_{jl}$ sub-population. The post-stratified and weighting class estimates are:

$$\bar{y}_{ps} = \frac{1}{N} \sum_{j=1}^{J} \sum_{l=1}^{L} N_{jl} \bar{y}_{jl} R, \quad \bar{y}_{ps} = \frac{1}{n} \sum_{j=1}^{J} \sum_{l=1}^{L} n_{jl} \bar{y}_{jl} R.$$ 

When only the marginal distribution of $N_{jl}$ is available, a modified estimate can be obtained by a raking method, i.e., finding estimates $\{N_{jl}^*\}$ of $\{N_{jl}\}$ satisfying:

$$N_{jl}^* = \sum_{l=1}^{L} N_{jl}^* = N_{j+}, \quad N_{jl}^* = \sum_{j=1}^{J} N_{jl}^* = N_{+l}, \quad j = 1, \ldots, J; l = 1, \ldots, L,$$

and keep the same relationship between $X_1$ and $X_2$ from the sample:

$$N_{jl}^* = a_j b_l n_{jl}.$$
9. **Available-case analysis**

1. When there are many missing-data patterns and the data are MCAR, the available-case analysis use all cases where the variable of interest is present and usually more efficient than the complete-case analysis.

2. Some pitfalls.

**Homework 2 (due Sep. 22, 2005)**