Biostat 2065
Analysis of Incomplete Data

Gong Tang
Dept of Biostatistics
University of Pittsburgh

September 26, 2005
1. **Estimation of imputation uncertainty**

Methods

1. Apply explicit formulas. Examples: the weighting class estimator for the stratified random sampling and simple hot deck estimator. Difficult to get if the sampling and nonresponse mechanism are complicated.

2. Modify the imputations so that standard errors can be calculated from the *single* imputed dataset.

3. Resample the incomplete data repeatedly and apply the imputation and analysis procedure to each dataset. Then summarize the results.

4. Multiple imputation. Draw from the *predictive distribution* for the missing values conditioned on the observed data for several times. Then compute the appropriate standard errors.
2. Imputation methods that lead to valid inference based on a single imputed data set

**Ultimate cluster (UC)**
Definition: the largest sampling units that are independently sampled at random from the population.

**Lemma.** Let $\hat{\theta}_1, \ldots, \hat{\theta}_k$ be random variables that are (1) uncorrelated and (2) have a common mean $\mu$. Let

$$\bar{\theta} = \frac{1}{k} \sum_{i=1}^{k} \hat{\theta}_i, \quad \hat{\nu}(\bar{\theta}) = \frac{1}{k(k-1)} \sum_{i=1}^{k} (\hat{\theta}_i - \bar{\theta})^2.$$

Then they are unbiased estimators for $\theta$ and $Var(\bar{\theta})$, respectively.

**Example 5.1. Cluster samples with imputed data.** Suppose the population has $K$ UCs and $k$ UCs are randomly sampled. The target is the population total of a variable $Y$, $T$. Horvitz-Thompson estimate:

$$\hat{t}_{HT} = \sum_{j=1}^{k} \frac{\hat{t}_j}{\pi_j},$$

where, $\hat{t}_j$ is an unbiased estimate of $t_j$ (total of $Y$ in UC $j$) and $\pi_j$ is the probability that UC $j$ is selected.
If $\hat{t}_j$s are computed by imputations or weighting adjustments independently within each UC, then the lemma can be applied.

Usually the $\hat{t}_j$s are negatively correlated and the above formula over-estimates the variance.
3. Stratified cluster samples with imputed data

Example 5.2. Suppose there are $H$ strata and stratum $h$ has $K_h$ UCs, $h = 1, \ldots, H$. Let $k_h$ is the number of UCs sampled in stratum $h$. Then the population total can be estimated by

$$
\hat{t} = \sum_{h=1}^{H} \sum_{j=1}^{k_h} \frac{\hat{t}_{hj}}{\pi_{hj}} = \sum_{h=1}^{H} \hat{t}_h,
$$

where $\hat{t}_{hj}$ is unbiased estimate for $t_{hj}$ and $\pi_{hj}$ is the probability of selection of UC $hj$ is stratum $h$. The variance of $\hat{t}$ is

$$
\hat{v}(\hat{t} \mid Y_{obs}) = \sum_{h=1}^{H} \sum_{j=1}^{k_h} \left( \frac{k_h \hat{t}_{hj}}{\pi_{hj}} - \hat{t}_h \right)^2 \frac{k_h}{k_h(k_h - 1)}.
$$
4. **Bootstrap standard errors**

*Example 5.4.* Suppose the data are a random sample \( S = \{i : i = 1, \ldots, n\} \) of independent observations and some observations have incomplete records on variables \( Y \). A consistent estimate \( \hat{\theta} \) of a model parameter is computed by filing the missing values using some imputations method and then estimating \( \theta \) from the filled-in data. The bootstrap estimate of standard error can be obtained by the following procedure: for each \( b = 1, 2, \ldots, B \),

1. Generate a bootstrap sample \( S^{(b)} \) of \( S \), i.e., sampling the subjects or observations with replacement and keeping the same sample size.
2. Fill in the missing data in \( S^{(b)} \) by applying the imputation procedure, say \( \hat{S}^{(b)} \).
3. Compute \( \theta^{(b)} \) on the filled-in data set \( \hat{S}^{(b)} \), say \( \hat{\theta}^{(b)} \).

Then the bootstrap estimate of \( \theta \) is \( \hat{\theta}_{boot} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)} \).

The variance estimate of \( \hat{\theta} \) or \( \hat{\theta}^{(b)} \) is

\[
\hat{V}_{boot} = \frac{1}{B - 1} \sum_{b=1}^{B} (\hat{\theta}^{(b)} - \hat{\theta}_{boot})^2.
\]

This method is computationally intensive. The imputation method should yield consistent estimate for \( \theta \).

Biostat 2065: Analysis of Incomplete Data, 2005
5. Jackknife standard errors

Example 5.6. The data set is the same as in example 5.4. For \( j = 1, 2, \ldots, n \)

1. Delete observation \( j \) from \( S \), yielding the sample \( S^{(-j)} \).

2. Fill in the missing data in \( S^{(-j)} \) by applying the imputation procedure and get a complete data set \( \hat{S}^{(-j)} \).

3. Compute \( \hat{\theta}^{(-j)} \) based on \( \hat{S}^{(-j)} \).

4. Calculate a pseudovalue \( \tilde{\theta}_j = n\hat{\theta} - (n - 1)\hat{\theta}^{(-j)} \).

Then the jackknife estimate of \( \theta \) is

\[
\hat{\theta}_{jack} = \frac{1}{n} \sum_{j=1}^{n} \tilde{\theta}_j = \hat{\theta} + (n - 1)(\hat{\theta} - \bar{\theta}),
\]

where \( \bar{\theta} = \frac{1}{n} \sum_{j=1}^{n} \hat{\theta}^{(-j)} \). The jackknife estimate of the variance of \( \hat{\theta} \) or \( \hat{\theta}_{jack} \) is

\[
\hat{V}_{jack} = \frac{1}{n(n - 1)} \sum_{j=1}^{n} (\tilde{\theta}_j - \hat{\theta}_{jack})^2 = \frac{n - 1}{n} \sum_{j=1}^{n} (\hat{\theta}^{(-j)} - \bar{\theta})^2.
\]

Example 5.7. Stratified cluster samples with imputed data.

Biostat 2065: Analysis of Incomplete Data, 2005
6. Multiple imputation

Simple imputation cannot reflect sampling variability under one model for nonresponse or uncertainty about the correct model for nonresponse.

Under an explicit or implicit model, a predictive distribution is available for imputing missing values and estimate can be derived based on the fill-in data set. The multiple imputation procedure perform the same process $D$ times instead of once: for $d = 1, 2, \ldots, D$, impute missing values by drawing from the predictive distribution and obtain estimate $\hat{\theta}_d$ and its associated variance $W_d$.

The MI estimate is $\bar{\theta}_D = \frac{1}{D} \sum_{d=1}^{D} \hat{\theta}_d$.

The variance estimate of $\theta_D$ is

$$T_D = \bar{W}_D + \frac{D + 1}{D}B_D$$

$$= \frac{1}{D} \sum_{d=1}^{D} W_d + \frac{D + 1}{D} \frac{1}{D - 1} \sum_{d=1}^{D} (\hat{\theta}_d - \bar{\theta}_D)^2.$$  

The test statistic $(\theta - \bar{\theta}_D)T_D^{-1/2} \sim t_\nu$ with

$$\nu = (D - 1)(1 + \frac{1}{D + 1} \frac{\bar{W}_D}{B_D})^2.$$
For small data sets,

\[ \nu^* = (\nu^{-1} + \hat{\nu}_{obs}^{-1})^{-1}, \quad \text{with } \hat{\nu}_{obs} = (1 - \gamma_D)\left(\frac{\nu_{com} + 1}{\nu_{com} + 3}\right)\nu_{com}. \]
7. Maximum likelihood estimation for complete data

Likelihood function: given the data value $Y$, the likelihood function $L(\theta \mid Y)$ is any function of $\theta \in \Omega$ proportional to $f(Y \mid \theta)$. By definition, $L(\theta \mid Y) = 0$ for any $\theta \notin \Omega$.

Example 6.1 Univariate normal sample.
Example 6.2 Exponential sample.
Example 6.3 Multinomial sample.
Example 6.4 Multivariate normal sample.

A maximum likelihood (ML) estimate of $\theta$ is a value of $\theta \in \Omega$ that maximizes the likelihood $L(\theta \mid Y)$.

Score function $D_l(\theta) = \partial \log L(\theta \mid Y)/\partial \theta$. 

Biostat 2065: Analysis of Incomplete Data, 2005