Lab 10
Missing Data

Henian Chen, M.D., Ph.D.

What are missing data?
Data contain various codes to indicate lack of response such as: "Don’t know," "Refused," "Unintelligible," etc.

Why do missing data create difficulty in scientific research?
Because most data analysis procedures were not designed for missing data. Most SAS statistical procedures simply exclude observations with any missing variables from the analysis.

Source of Missing Data
Missing data may occur because of oversights or data recording errors.

Missing data may occur because some subjects either refuse or fail to respond one or more times during the study; some respondents may drop out of the study prematurely.

Some missing data may result from the design of the study (planned missing). For example, some items may only be measured for a subset of the sample (e.g., when alcohol use is measured solely in older children).
Problems Caused by Missing Data

May biased parameter estimates
May inflate Type I and Type II error rates
May degrade the performance of confidence intervals
May reduce statistical power

Three Types of Missing Data

1. Missing Completely At Random (MCAR)

Data are missing because of a random process (e.g., equipment failure).
The missing scores of the participants are not related to other measured or unmeasured variables so estimates of population parameters (e.g., means, variances, correlation) will be unbiased.

Three Types of Missing Data

2. Missing At Random (MAR)

Missing data depend only on the values of other observed variables measured on participants.
Consider a study with income as a key variable of interest. If less educated individuals tend not to report their income, the missing income values may be MAR because whether an individual responds depends on his or her education.
Three Types of Missing Data

3. Missing Not At Random (MNAR)

Missing data depend on the individual's score on the variable of interest, whether measured or unmeasured.

Missing income will be called MNAR if individuals with high or low income tend not to report their income.

Example

"MISSING8266.DBF" is data from a study of the mental health needs of students in the New York City public school system six months after the attack on the World Trade Center. The data we will use in here contains 5 variables for 8266 students.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Missing</th>
<th>Range</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONDUCTD-</td>
<td>5400</td>
<td>2866(54.7%)</td>
<td>0 - 12</td>
<td>1.44 (1.96)</td>
</tr>
<tr>
<td>SANXITY-</td>
<td>5272</td>
<td>2994(36.2%)</td>
<td>0 - 7</td>
<td>1.51 (1.70)</td>
</tr>
<tr>
<td>AGE</td>
<td>8266</td>
<td>0</td>
<td>8 - 21</td>
<td>13.6 (2.57)</td>
</tr>
<tr>
<td>SEX</td>
<td>8218</td>
<td>48(0.6%)</td>
<td>0 - 1</td>
<td></td>
</tr>
</tbody>
</table>

Example cont.

Complete data: 2045 students (24.74%) without any missing values.

Missing data: 6221 students (75.26%) with missing value(s) at least for one variable.

The statistical analysis of interest is the regression analysis of conduct disorder symptoms on the 3 demographic variables (age, sex and mother education).
Traditional Approaches to Missing Data

1. Complete Case Analysis

Only individuals with complete information on all the variables under study are included in the statistical analysis.

This method (also known as listwise deletion) is very convenient but assumes MCAR. If the missing data really are MCAR, this method gives results that are valid, but inefficient, because some information (i.e. the incomplete records) is discarded.

Complete case analysis may lead to inaccurate results. When less educated people are less likely to report their income, complete case analysis will result in an average income that is greater than the population parameter.

2. Single Imputation

a) Mean Substitution:

All the missing values for a variable are replaced by the mean value of that variable.

Use of mean substitution will lead to valid estimates of mean values from the data only if the missing values have the same mean as the rest of the sample. Since the method underestimates the variability among the missing cases as a result of mean substitution, estimates of the variance and covariance parameters will be too small.
### Traditional Approaches to Missing Data
#### 2. Single Imputation cont.

**a) Mean Substitution:**
The mean of CONDUCTD for total sample is 1.44, but the SD decreases from 1.96 to 1.58 because we replaced 2866 missing values by the same mean (1.44). No variance for those 2866 students.

<table>
<thead>
<tr>
<th>Parameter Estimates (N=6447)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>AGE</td>
</tr>
<tr>
<td>SEX</td>
</tr>
<tr>
<td>MEDU</td>
</tr>
</tbody>
</table>

Comparing with the model N=4288, both the parameter estimate, and its standard error are underestimated.

---

**b) Regression Substitution:**
All the missing values of a data set replaced by the predicted value of that variable from a regression analysis based only on the complete cases.

In this model, even if the mean parameters are correctly estimated, the variance parameters are underestimated because this method assumes no residual variance around the regression line predicting the missing variables.

---

**Summary:**
Single estimation treats missing values as if they were known in the complete-data set.

However, this method does not reflect the uncertainty about the predictions of the unknown missing values, and the resulting SEs of parameter estimates will be biased toward zero (underestimated).
Multiple Imputation
MI and MIANALYZE procedures

MI procedure performs multiple imputation (MI) of missing data.

Multiple imputation replaces each missing value with a set of plausible values that include the uncertainty about the right value to impute.

MI and MIANALYZE procedures cont.
In most applications, just three to five imputations are sufficient to obtain excellent results.

MI results in valid statistical inferences that properly reflect the uncertainty due to missing values. MI offers the promise of improving both the accuracy and often the statistical power of results.

MI and MIANALYZE procedures cont.
1. The missing data are filled in \( m \) times (by default, \( m=5 \)) to generate \( m \) complete data sets.
2. The \( m \) complete data sets are analyzed using standard statistical analyses.
3. The results from the \( m \) complete data sets are combined by using the MIANALYZE procedure to generate valid statistical inferences about these parameters.

No matter which complete-data analysis is used, the process of combining results from different data sets is essentially the same.
Statistical Assumptions for MI

1. Missing data are either MCAR or MAR and the data necessary to account for any bias are included in the model.

2. The data are from a multivariate normal distribution. It often makes sense to use a normal model to create multiple information even when the observed data are somewhat non-normal. You can also use a TRANSFORM statement to transform variables to conform to the multivariate normality assumption.

Statistical Assumptions for MI cont.

3. The imputation model is the same as the analysis model. But in practice, the two models may not be the same because it is often unwise to include variables that may account for bias in the missing data in the same analysis as the explanatory equation for some other phenomenon. Thus, one may use education, age, occupation to estimate income, but you would often not want all these variables to compete in a model estimating some disease outcome.

Imputation of MAR variables is usually done by the person who collected the data, the final analysis may be done by many other users who share the data set. Actually the data collector has much better knowledge about the data and is likely the best person to do imputations.

Statistical Assumptions for MI cont.

Generally you should include as many variables as you can when you doing multiple imputation.

The precision you lose when you include unimportant predictors is usually a relatively small price to pay for the general validity of analysis of the resultant multiply imputed data set.
SAS program for MI

```sas
proc mi data=missing8266 nimpute=1 seed=88888
minimum=0 0 0 maximum=12 21 1 round=1
out=imputation1;
mcmc chain=multiple;
var conductd age sex medu;
run;
proc print data=imputation1;
```

nimpute: 1 imputation (the default is 5)
seed: seed to begin random number generator. The default is a value
generated from reading the time of day from the computer's clock. You need
to specify the same seed number in the future to reproduce the results.
minimum: minimum values for imputed variables
maximum: maximum values for imputed variables
round: units to round imputed variable values
out: output data with imputed values (without missing data for n=8266)
mcmc: uses a Markov Chain Monte Carlo method
chain: single / multiple chain

SAS Output

```sas
SAS Output
```

Data Set: WORK.MISSING8266
Method: MC MC
Multiple Imputation Chain: Multiple Chains
Initial Estimates for MC MC: EM Posterior Mode
Start: Starting Value
Prior: Jeffreys
Number of Imputations: 1
Number of Burn-in Iterations: 200
Seed for random number generator: 88888

Missing Data Patterns

<table>
<thead>
<tr>
<th>Group</th>
<th>CONDUCTD</th>
<th>AGE</th>
<th>SEX</th>
<th>MEDU</th>
<th>Freq</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>4288</td>
<td>51.88</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>.</td>
<td>1091</td>
<td>13.20</td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>X</td>
<td>.</td>
<td>X</td>
<td>18</td>
<td>0.22</td>
</tr>
<tr>
<td>4</td>
<td>.</td>
<td>X</td>
<td>X</td>
<td>.</td>
<td>3</td>
<td>0.04</td>
</tr>
<tr>
<td>5</td>
<td>.</td>
<td>.</td>
<td>X</td>
<td>X</td>
<td>2159</td>
<td>26.12</td>
</tr>
<tr>
<td>6</td>
<td>.</td>
<td>X</td>
<td>X</td>
<td>.</td>
<td>680</td>
<td>8.23</td>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>.</td>
<td>X</td>
<td>X</td>
<td>17</td>
<td>0.21</td>
</tr>
<tr>
<td>8</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>X</td>
<td>10</td>
<td>0.12</td>
</tr>
</tbody>
</table>

SAS Output

```sas
proc freq data=imputation1;
tables sex medu;
run;
```

N=8266

<table>
<thead>
<tr>
<th>SEX</th>
<th>Frequency</th>
<th>Percent</th>
<th>SEX</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4334</td>
<td>52.43</td>
<td>0</td>
<td>1583</td>
<td>19.15</td>
</tr>
<tr>
<td>1</td>
<td>3932</td>
<td>47.57</td>
<td>1</td>
<td>6683</td>
<td>80.85</td>
</tr>
</tbody>
</table>

proc means;
var conductd age sex;
run;

proc reg;
model conductd=age sex medu;
run;

Parameter Estimate Standard Error t Value Pr > |t|
Intercept 0.49829 0.12357 4.03 <.0001
AGE 0.07942 0.00801 9.92 <.0001
SEX 0.30059 0.04096 7.34 <.0001
MEDU -0.06385 0.05231 -1.22 0.2223
5 Imputations
proc mi data=missing8266 nimpute=5 seed=88888
minimum=0 8 0 0 maximum=12 21 1 1 round=1
out=imputation5;
mcmc chain=multiple;
var conductd age sex medu;
run;
proc print data=imputation5;
run;
proc means data=imputation5;
var conductd;
by _imputation_;
run;
proc freq data=imputation5;
tables sex medu;
by _imputation_;
run;

SAS Output
Imputation Number = 1

<table>
<thead>
<tr>
<th>CONDUCTD</th>
<th>N</th>
<th>%</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4334</td>
<td>52.43</td>
<td>1.67</td>
<td>1.87</td>
</tr>
<tr>
<td>1</td>
<td>3932</td>
<td>47.57</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Imputation Number = 2

<table>
<thead>
<tr>
<th>CONDUCTD</th>
<th>N</th>
<th>%</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4326</td>
<td>52.33</td>
<td>1.62</td>
<td>1.82</td>
</tr>
<tr>
<td>1</td>
<td>3940</td>
<td>47.67</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Imputation Number = 3

<table>
<thead>
<tr>
<th>CONDUCTD</th>
<th>N</th>
<th>%</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4326</td>
<td>52.33</td>
<td>1.63</td>
<td>1.83</td>
</tr>
<tr>
<td>1</td>
<td>3940</td>
<td>47.67</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Imputation Number = 4

<table>
<thead>
<tr>
<th>CONDUCTD</th>
<th>N</th>
<th>%</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4327</td>
<td>52.35</td>
<td>1.64</td>
<td>1.83</td>
</tr>
<tr>
<td>1</td>
<td>3939</td>
<td>47.65</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**SAS program for MIANALYZE**

To generate regression coefficients for each of 5 imputed data sets:

```sas
proc reg data=imputation5 outest=outreg5 noprint covout;
model conductd=age sex medu;
by _imputation_;
run;
```

To combine the five sets of regression coefficients:

```sas
proc mianalyze data=outreg5;
var intercept age sex medu;
run;
```

**SAS Output**

The MIANALYZE Procedure:  Number of Imputations = 5

**Multiple Imputation Parameter Estimates**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std Error</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>0.359427</td>
<td>0.175940</td>
<td>0.18095 - 0.531883</td>
</tr>
<tr>
<td>age</td>
<td>0.074047</td>
<td>0.010070</td>
<td>0.05338 - 0.094712</td>
</tr>
<tr>
<td>sex</td>
<td>0.267421</td>
<td>0.053243</td>
<td>0.15702 - 0.377825</td>
</tr>
<tr>
<td>medu</td>
<td>-0.069332</td>
<td>0.071753</td>
<td>-0.22073 - 0.082065</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Minimum</th>
<th>Maximum</th>
<th>% Pr &gt; [t]</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>0.390281</td>
<td>0.675177</td>
<td>3.16 0.0003</td>
</tr>
<tr>
<td>age</td>
<td>0.068198</td>
<td>0.080001</td>
<td>7.35 &lt;0.0001</td>
</tr>
<tr>
<td>sex</td>
<td>0.225232</td>
<td>0.300590</td>
<td>5.02 &lt;0.0001</td>
</tr>
<tr>
<td>medu</td>
<td>-0.115245</td>
<td>0.004631</td>
<td>-0.97 0.3475</td>
</tr>
</tbody>
</table>