NOTES FOR CAUSAL INFERENCE ON JAN 13TH
BY: JIN XU
18/01/2010

MAR/MNAR

There was some confusion about MAR/MNAR. The only relevance for this class is that there is an analogy with unfounding.

(defined as \( \text{Pr}(W | X, Y(0), Y(1)) = \text{unfounding Pr}(W | X) \))

Independence vs Exchangeability

\[ X_1, X_2, \ldots, X_n | \pi \sim_{iid} \text{Bern}(\pi) \text{ trials} \]

\[ p(X_1, X_2, \ldots, X_n | \pi) = \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i} = \pi^{\sum_{i=1}^{n} x_i}(1-\pi)^{n-\sum_{i=1}^{n} x_i} \]

Add a prior distribution on \( \pi, g(\pi) = 1, 0 < \pi < 1 \)

\[ p(X_1, X_2, \ldots, X_n) = \int p(X_1, X_2, \ldots, X_n | \pi)g(\pi)d\pi = \frac{\Gamma(\sum x_i + 1)\Gamma(n-\sum x_i + 1)}{\Gamma(n + 1)} \]

\[ p(X_1, X_2, \ldots, X_n) = p(X_{i_1}, X_{i_2}, \ldots, X_{i_n}) \]

Where \( i_1, i_2, \ldots, i_n \) is a permutation of 1, 2, \ldots, n.

So, given \( \pi, X_1, X_2, \ldots, X_n \) are independent of each other, while the marginal distributions of \( X_1, X_2, \ldots, X_n \) (after averaging over \( g(\pi) \)) are not independent of each other anymore. However, \( X_1, X_2, \ldots, X_n \) are still exchangeable to each other, since \( p(X_1, X_2, \ldots, X_n) = p(X_{i_1}, X_{i_2}, \ldots, X_{i_n}) \).

Fisher’s Approach

\( H_0: Y_i(0) = Y_i(1) \) for all \( i \), this is a “sharp null hypothesis”, completely specifies outcomes.
**Test Statistic**  
Definition of a test statistic, \( T = T(W, Y_{\text{obs}}, X) \)

\[
T = \frac{1}{N_T} \sum_{i=1, w_i=1}^{N} y_i(1) - \frac{1}{N_C} \sum_{i=1, w_i=0}^{N} y_i(0) = \frac{1}{N_T} \sum_{i=1, w_i=1}^{N} y_{i,\text{obs}} - \frac{1}{N_C} \sum_{i=1, w_i=0}^{N} y_{i,\text{obs}}
\]

Test statistic should be shown to detect "expected" alternative. Evaluating observed test statistic \( T_{\text{obs}} \): \( p \)-value = \( pr_w(T \text{ is as or more extreme than } T_{\text{obs}}) \). Compute \( T \) for every possible realization of \( W \), and use the distribution of those \( T \)'s to find \( p \)-value.

**Examples**  
\( N=6 \) units, \( Y \) is score for cough from 1 to 6, 1 is worst and 6 is best.

<table>
<thead>
<tr>
<th>( i )</th>
<th>( y_i(0) )</th>
<th>( y_i(1) )</th>
<th>( y_{i,\text{obs}}(0) )</th>
<th>( y_{i,\text{obs}}(1) )</th>
<th>( w_i )</th>
<th>( y_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>?</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>?</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>?</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>?</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>?</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>?</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

There is \( \binom{6}{3} = 20 \) possible \( w \) vectors

<table>
<thead>
<tr>
<th>( w )</th>
<th>( T )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1,1,0,0,0)</td>
<td>8/3-5/3=1</td>
</tr>
<tr>
<td>(1,1,0,1,0,0)</td>
<td>12/3-1/3=3.67</td>
</tr>
<tr>
<td>(1,1,0,0,1,0)</td>
<td>8/3-5/3=1</td>
</tr>
<tr>
<td>( \cdots )</td>
<td>( \cdots )</td>
</tr>
<tr>
<td>(0,0,0,1,1,1)</td>
<td>5/3-8/3=-1</td>
</tr>
</tbody>
</table>

\( \Rightarrow \) Prob Distn on \( T=Pr_w(T) \)  
\( \Rightarrow \) 2-side p-values=16/20=0.8  
\( \Rightarrow \) observed \( T \) is not unusual under \( H_o \), no reason to reject \( H_o \)

**Confidence Intervals**  
can use relationship of tests + CIs to drive for \( C \), if hypothesis constant addictive trt effect \( C \)

\( H_o : Y_i(1) = Y_i(0) + C \)

Test for a grid of \( C \) values. If don’t reject \( H_o \) at 0.1, so \( C \) is in the 90% CI.

**Use of covariates**  
create more powerful test statistic
**Example** redefine \( \hat{y}_i = y_i - x_i \) (x=pre-test), notice that causal effect:

\[
\hat{y}_i(1) - \hat{y}_i(0) = (y_i(1) - x_i) - (y_i(0) - x_i) = y_i(1) - y_i(0)
\]

\( H_0 \) is the same in terms of \( \hat{y}'s \)

\[
\hat{T} = \frac{1}{N_t} \sum_{i=1}^{N_t} \hat{y}_{i}^{obs} - \frac{1}{N_c} \sum_{i=0}^{N_c} \hat{y}_{i}^{obs} = \frac{1}{N_t} \sum_{i=1}^{N_t} (y_{i}^{obs} - x_i) - \frac{1}{N_c} \sum_{i=0}^{N_c} (y_{i}^{obs} - x_i) = T - (\bar{x}_t - \bar{x}_c)
\]

**Neyman Approach**

Statistical Science 1990, translated version of Neyman 1923, paper by Rubin

\[
T = \frac{1}{N} \sum_{i=1}^{N} (y_i(1) - y_i(0)) = \bar{y}(1) - \bar{y}(0)
\]

Population average treatment effect (ATE)
Inference for \( \tau \): test \( H_0 : \tau = 0 \) (this is not a “sharp” null hypothesis); CIs for \( \tau \)

**Estimator** \( \hat{\tau} \) Examine properties of \( \hat{\tau} \) (estimate of variance of \( \hat{\tau} \))

\[
\hat{\tau} = \bar{y}_{i}^{obs} - \bar{y}_{c}^{obs} = \frac{1}{N} \sum_{i=1}^{N} \left( w_i y_i(1) \frac{N_t}{N} - (1 - w_i) y_i(0) \frac{N_c}{N} \right)
\]

\[
E_w(\hat{\tau}) = \frac{1}{N} \sum_{i=1}^{N} \left( E(w_i) y_i(1) \frac{N_t}{N} - (1 - E(w_i)) y_i(0) \frac{N_c}{N} \right)
\]

If completely randomized experiment with fixed \( N_t, N_c \)

\[
E(w_i) = N_t/N \Rightarrow E_w(\hat{\tau}) = \frac{1}{N} \sum_{i=1}^{N} (y_i(1) - y_i(0)) = \tau
\]

\[
Var_w(\hat{\tau}) = \frac{S_{\tau}^2}{N_t} + \frac{S_{c}^2}{N_c} - \frac{S_{tc}^2}{N}
\]

\[
S_{\tau}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i(1) - \bar{y}(1))^2
\]

\[
S_{c}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i(0) - \bar{y}(0))^2
\]

\[
S_{tc}^2 = \frac{1}{N-1} \sum_{i=1}^{N} ((y_i(1) - y_i(0)) - (\bar{y}(1) - \bar{y}(0)))^2
\]

We don’t and can’t know \( S_{tc}^2 \).
Example 2 units

\[ \tau = \frac{1}{2} ((y_1(1) - y_1(0)) + (y_2(1) - y_2(0))) \]

\[ \hat{\tau} = \begin{cases} y_1(1) - y_2(0), & \text{with prob 1/2,} \\ y_2(1) - y_1(1), & \text{with prob 1/2.} \end{cases} \]  

(1a)

(1b)

\[ \text{Var}_w(\hat{\tau}) = \frac{1}{2} (y_1(1) - y_2(0) - \tau)^2 + \frac{1}{2} (y_2(1) - y_1(0) - \tau)^2 \]

\[ = \frac{1}{4} (y_1(1) - y_2(1))^2 + \frac{1}{4} (y_1(0) - y_2(0))^2 + \frac{2}{4} (y_1(1)y_1(0) + y_2(0)y_2(1) - y_1(0)y_2(1) - y_1(1)y_2(0))) \]

3 Approaches to deal with \( \text{Var}_w(\hat{\tau}) \)

1. If assume constant trt effect, then \( S^2_{tc} = 0 \) \( \Rightarrow \text{Var}_w(\hat{\tau}) \approx \frac{S^2_c}{N_c} + \frac{S^2_t}{N_t} \), conservative

2. \( S^2_{tc} = S^2_c + S^2_t - 2\rho_{tc}S_cS_t \), \( \rho_{tc} = \text{correl of potential outcomes}. \)

Smallest value of \( S^2_{tc} \) corresp to \( \rho = 1 \) \( \Rightarrow \) largest possible value of \( \text{Var}(\hat{\tau}) \)

\( \Rightarrow \) assume \( \rho = 1 \)

\( \Rightarrow \text{Var}_w(\hat{\tau}) = \frac{S^2_c}{N_c} + \frac{S^2_t}{N_t} - \frac{(S_c - S_t)^2}{N} \), conservative!

3. A pooled approach (see next time).