Hypothesis Testing for Personalizing Treatment

Huitian Lei and Susan Murphy

Department of Statistics
University of Michigan

April 12, 2013
Two pre-defined subgroups: subgroup 1 and subgroup 2
Two active treatments: treatment A and treatment B.
The mean treatment response in subgroup \( i \) under treatment \( TR \):
\[ \mu_{i,TR}, \text{ where } i \in \{1, 2\}, \; TR \in \{A, B\} \]
Subgroup treatment effects:
\[ \theta = (\theta_1, \theta_2), \text{ where } \theta_1 = \mu_{1,A} - \mu_{1,B} \]
and \[ \theta_2 = \mu_{2,A} - \mu_{2,B}. \]
[Motivation] Current practice: Testing qualitative interaction

- Gail and Simon (1985): likelihood ratio test of qualitative interaction: $\theta_1 \theta_2 < 0$
- Their hypothesis:
  
  $$H_{GS} : \theta_1 \theta_2 \geq 0$$

- Qualitative interaction is not the only treatment-subgroup interaction informative in personalizing treatment (e.g, Simon 2002).
  - $\theta_1 \neq 0, \theta_2 = 0$
  - $\theta_1 = 0, \theta_2 \neq 0$

- Mismatch between the test and the scientific goal.
[Motivation] Current practice: Testing subgroup hypotheses separately

- \( H_1 : \theta_1 = 0 \) and \( H_2 : \theta_2 = 0 \)
- \( H_0 : \theta_1 = \theta_2, H_1 : \theta_1 = 0 \) and \( H_2 : \theta_2 = 0 \)
- Usually control the familywise error rate.
- Mismatch between the controlled errors and the scientific goal.
Objectives

- Match the decision procedure and controlled errors with the scientific goal
- Distinguish different types of interaction informative in personalizing treatment
- Same power to detect qualitative interactions as Gail and Simon test
[Procedure] Hypothesis

The hypothesis that the subgroup indicator is not useful for personalizing treatment:

\[ H : \theta_1 = \theta_2 = 0, \text{ or } \theta_1 \theta_2 > 0 \]

The complement of \( H \) consists of:

\( \theta_1 \neq 0, \theta_2 = 0 \)
\( \theta_1 = 0, \theta_2 \neq 0 \)
\( \theta_1 \theta_2 < 0 \)
### Decision Space

<table>
<thead>
<tr>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
</tr>
<tr>
<td><strong>2</strong></td>
</tr>
<tr>
<td><strong>3</strong></td>
</tr>
<tr>
<td><strong>4</strong></td>
</tr>
</tbody>
</table>

**Table:** The decision space
Controlled errors over $\theta \in H : \theta_1 = \theta_2 = 0$, or $\theta_1 \theta_2 > 0$

<table>
<thead>
<tr>
<th>Decision 2 $\theta_1 \neq 0, \theta_2 = 0$</th>
<th>Decision 3 $\theta_1 = 0, \theta_2 \neq 0$</th>
<th>Decision 4 $\theta_1 \theta_2 &lt; 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1 = \theta_2$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\theta_1 &gt; \theta_2 &gt; 0$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$\theta_2 &gt; \theta_1 &gt; 0$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$\theta_1 &lt; \theta_2 &lt; 0$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$\theta_2 &lt; \theta_1 &lt; 0$</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table: The controlled errors
[Procedure] Standard test statistics

- Data from two-arm randomized clinical trials
- Assume equal unknown variance

\[ T_0 = \frac{\bar{X}_{1A} - \bar{X}_{1B} - \bar{X}_{2A} + \bar{X}_{2B}}{\hat{\sigma}\sqrt{\frac{4}{n_1} + \frac{4}{n_2}}} = \sqrt{p_2} T_1 - \sqrt{p_1} T_2 \]  \hspace{1cm} (1)

\[ T_1 = \frac{\bar{X}_{1A} - \bar{X}_{1B}}{\hat{\sigma}\sqrt{\frac{4}{n_1}}} \]  \hspace{1cm} (2)

\[ T_2 = \frac{\bar{X}_{2A} - \bar{X}_{2B}}{\hat{\sigma}\sqrt{\frac{4}{n_2}}} \]  \hspace{1cm} (3)
[Procedure] The two-stage decision making procedure

c₀ and c₁ are the critical values in stage I and stage II.

- In stage I, utilize test statistic T₀ and compare it with critical value ±c₀. If T₀ > c₀ or T₀ < −c₀, proceed to stage II. Otherwise if |T₀| ≤ c₀, stop the testing procedure and make decision 1.

- In stage II, utilize test statistics T₁ and T₂ and compare them with critical value ±c₁. Decisions are made according to the decision rule specified in the following table.
[Procedure] The decision rule

<table>
<thead>
<tr>
<th>Decision Rule</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>T_0</td>
</tr>
<tr>
<td>$</td>
<td>T_0</td>
</tr>
<tr>
<td>$</td>
<td>T_0</td>
</tr>
<tr>
<td>$</td>
<td>T_0</td>
</tr>
<tr>
<td>$T_0 &gt; c_0, T_1 &gt; c_1,</td>
<td>T_2</td>
</tr>
<tr>
<td>$T_0 &lt; -c_0, T_1 &lt; -c_1,</td>
<td>T_2</td>
</tr>
<tr>
<td>$T_0 &lt; -c_0,</td>
<td>T_1</td>
</tr>
<tr>
<td>$T_0 &gt; c_0,</td>
<td>T_1</td>
</tr>
<tr>
<td>$T_0 &gt; c_0, T_1 &gt; c_1, T_2 &lt; -c_1$</td>
<td>4</td>
</tr>
<tr>
<td>$T_0 &lt; -c_0, T_1 &lt; -c_1, T_2 &gt; c_1$</td>
<td>4</td>
</tr>
</tbody>
</table>

Table: The decision Rule
The decision rule

Decision Rule at p=0.5

T1
T2
dec 1
dec 2
dec 3
dec 4

(decision rule)

(University of Michigan)
April 12, 2013 12 / 21
Choose \((c_0, c_1)\) so that the probability of controlled errors for \(\theta \in H\) is at most \(\alpha\).

Fix \(c_1 = z_\alpha\) (Gail and Simon critical value) and vary \(c_0\).

\(p_1\) is the sample proportion of subgroup 1.

<table>
<thead>
<tr>
<th>(p_1)</th>
<th>(c_0)</th>
<th>(c_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10 or 0.90</td>
<td>2.06</td>
<td>1.64</td>
</tr>
<tr>
<td>0.20 or 0.80</td>
<td>2.05</td>
<td>1.64</td>
</tr>
<tr>
<td>0.30 or 0.70</td>
<td>1.96</td>
<td>1.64</td>
</tr>
<tr>
<td>0.40 or 0.60</td>
<td>1.96</td>
<td>1.64</td>
</tr>
<tr>
<td>0.50</td>
<td>1.96</td>
<td>1.64</td>
</tr>
</tbody>
</table>

*Table*: The critical values at \(\alpha = 0.05\)
Same power to detect qualitative interactions as Gail and Simon test when $0.10 \leq p_1 \leq 0.90$.

$p_1, p_2$ are the sample proportions of subgroup 1 and subgroup 2.

$$\{(T_1, T_2): \sqrt{p_2}T_1 - \sqrt{p_1}T_2 > c_0, T_1 > c_1, T_2 < -c_1\}$$

$$= \{(T_1, T_2): T_1 > c_1, T_2 < -c_1\}$$

$$\{(T_1, T_2): \sqrt{p_2}T_1 - \sqrt{p_1}T_2 < -c_0, T_1 < -c_1, T_2 > c_1\}$$

$$= \{(T_1, T_2): T_1 < -c_1, T_2 > c_1\}$$
Better power to detect $\theta_1 \neq 0, \theta_2 = 0$ or $\theta_1 = 0, \theta_2 \neq 0$, compared to a likelihood ratio test of $H : \theta_1 = \theta_2 = 0$, or $\theta_1 \theta_2 > 0$. 

![Power Comparison Diagram](attachment:power_diagram.png)
[Pros and Cons] Compared to Gail and Simon test

- less powerful to detect qualitative interactions when one subpopulation is rare (e.g., \( p_1 = 0.05 \)).
- Gail and Simon test is d-admissible. The proposed procedure has larger risk over \( H : \theta_1 = \theta_2 = 0 \), or \( \theta_1 \theta_2 > 0 \) in exchange for better power at the complement of \( H \).
[Pros and Cons] Compared to Gail and Simon test

Power of detect qualitative interaction

n = 200, p1 = 0.05
Pros and Cons: Compared to testing subgroup hypotheses separately

- **Pros**: The controlled errors matches the scientific goal.
- **Inflated error probability for testing** $H_1 : \theta_1 = 0$ and $H_2 : \theta_2 = 0$
  - Controlling FWER at 0.05, probability of making decision 2 or 3 may exceed 0.25 when $\theta_1 = \theta_2$
- **Overly conservative error control for testing** $H_0 : \theta_1 = \theta_2$, $H_1 : \theta_1 = 0$ and $H_2 : \theta_2 = 0$. 
Thank you.

ehlei@umich.edu
[Back up slides] Total error rate

total error rate, n=200, p1=0.5

(theta1, theta2)
Total error rate

$\theta_1$

$\theta_2$

0.05

0.1

0.1

0.2

0.2

0.3

0.3

0.4

0.4

0.5

0.5

0.6

0.6

0.7

0.7

0.8

0.8

0.9

0.9

0.0 0.5 1.0 1.5 2.0 2.5 3.0

0.0 0.5 1.0 1.5 2.0 2.5 3.0