

All or None procedure: An outline

Nanayaw Gyadu-Ankama, Shoubhik Mondal, Steven Cheng

1 Motivation

The All or none method was first evolved in the context of quality control. Berger mentioned about this method in 1982 (see Berger, 1982, Technometrics(1)). Some time the quality of a product is determined by several parameters. Berger gave the method to test each parameter individually and deciding the product is acceptable only if each parameter passes its test. Likewise most human diseases are characterized by more than one symptoms, signs, patient-reported outcome etc. For example Arthritis patients experience not only pain, but also swelling and stiffness in their joints, Migraine is characterized by moderate-to-severe headache pain which is frequently accomapied by nausea, photophobia, and phonophobia. As a result when a treatment is assesed for its effect on these disoreders in a clinical trial setting, the effect is typically examined via multiple endpoints that describe the change in the multiple measures. Generally endpoints are classified in two categories, primary endpoints which define most important aspects of the disease affected by the treatment and secondary endpoints. Again for clarity Offen et al(2007) differentiate between two types of multiplicity of the primary endpoints. The first case is when treatment is effective if it improves at least one of the multiple endpoints. The second case is when a treatment is effective when it improves on all the multiple primary endpoints. Multiple primary endpoints in second case is called co-primary endpoints where simultaneous improvement is required to declare a treatment effective. Most of the time these co-primary endpoints are correalted. Offen et al.(2007(2)) gave examples of 20 such diseases where more than two co-primary endpoints exist.
2 Formulation of test

Let there are \( m \) co-primary endpoints in a randomized, two-treatment group trial. Let \( (X_j, j = 1, ..., m) \) represent the observations on the \( m \) endpoints of an individual in the treatment group with a mean vector \( \mu_X \) and \( (Y_j, j = 1, ..., m) \) the corresponding observations of an individual in the group receiving the placebo or the standard treatment with mean vector \( \mu_Y \). Let \( \Delta = \mu_X - \mu_Y = (\Delta_1, ..., \Delta_m)^t \).

Testing that a new drug is better than the placebo (or standard treatment) on all \( m \) co-primary endpoints is equivalent to testing the following superiority hypothesis:

\[
H_0: \bigcup_{j=1}^{m}(\Delta_j \leq 0) \quad \text{i.e. } \Delta_j \leq 0 \text{ for at least one } j \text{ against}
\]

\[
H_1: \bigcap_{j=1}^{m}(\Delta_j > 0) \quad \text{i.e. } \Delta_j > 0 \text{ for all } j
\]

3 Different test procedures

3.1 Min test

The all or none procedure has the following form: Reject all hypotheses if \( t_{\min} = (\min_{1 \leq i \leq m} t_i) \geq t_\alpha(\nu) \), where \( t_\alpha(\nu) \) is the \((1 - \alpha)\) quantile of the \( t \)-distribution with \( \nu = n_1 + n_2 - 2 \) df. This procedure is popularly known as the min test (Laska and Meisner, 1989(3)).

Here we are testing each individual sub-hypotheses at \( \alpha \) significance level. The regulatory position is to test each primary endpoint at the 2.5\% level (one-sided) if 5\% is allowable studywise false-positive rate regardless of how many endpoints are on the co-primary list. Since this procedure does not use a multiplicity adjustment (each hypothesis \( H_i \) is tested at level \( \alpha \)), it may appear that it must be highly powerful as a test of the individual test sub-hypothesis. But in reality, the min test is very conservative.

A simulation study was done to show that overall power of the min test is less than the power of the test for individual hypothesis. \((X_1, ..., X_m)\) follows multivariate normal distri-
bution with mean vector \( \mu_x = (0.5604, \ldots, 0.5604) \) and \( \Sigma \), \((Y_1, \ldots, Y_m)\) follows multivariate normal distribution with mean distribution \( \mu_y = (0, \ldots, 0) \) and \( \Sigma \). For convenience \( \Sigma \) has been taken such that all the diagonals are equal to one. Simulation has been done for 2 and 3 endpoints for different value of \( \rho \). Below table give a comparison of the power

<table>
<thead>
<tr>
<th>endpoints</th>
<th>( \rho )</th>
<th>sub-hypothesis</th>
<th>Min test</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>0.7982</td>
<td>0.6408</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.7982</td>
<td>0.6555</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
<td>0.7982</td>
<td>0.6752</td>
</tr>
<tr>
<td>2</td>
<td>0.8</td>
<td>0.7982</td>
<td>0.7277</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0.8017</td>
<td>0.5165</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>0.8017</td>
<td>0.5538</td>
</tr>
<tr>
<td>3</td>
<td>0.4</td>
<td>0.8017</td>
<td>0.5929</td>
</tr>
<tr>
<td>3</td>
<td>0.8</td>
<td>0.8017</td>
<td>0.6869</td>
</tr>
</tbody>
</table>

when there are 3 co-primary endpoints and the correlation among the endpoints is 0.2, the overall power is 55% while the power at the individual subhypothesis is 80%. As \( \rho \) increases the power of min test increases and if the endpoints are perfectly correlated, then the power for detecting the same effect size on all endpoints is the same as the common power for detecting the same effect size at the individual subhypothesis level. To maintain the power of the min test one need to increase the sample size. Table based on Offen et al. (2007) gave the percentage rise in sample size to maintain the overall power to 80%.

Figure 1: samplesize increment
Secondly when there are two co-primary endpoints, it can be shown that the probability rejecting $H_0$ over the complete null space as displayed in figure 2 is maximized at either $(\Delta_1, \Delta_2) = (0, \infty)$ or $(\Delta_1, \Delta_2) = (\infty, 0)$. In general with $m$ co-primary endpoints, the type I error takes it maximum on the boundaries of the $m$-dimensional null space, where $(m-1)$ of the co-ordinates takes the value $\infty$, while the remaining one takes the value 0. This configuration leads to marginal $\alpha$ level t-tests. But this extreme cases are not realistic.

A more reasonable approach is to include in the null hypothesis only those cases that are realistic from a clinical perspective. One way to make the null space more closely mimic real life solutions is to restrict the null space(Chuang et al. 2007(4)). For two co-primary endpoints, one option is to look at only the shaded areas in quadrants II, III, and IV in figure 3. These shaded areas consist of points whose vertical and horizontal distances from the X and Y axis are no more than $M$ units. For any $(\Delta_1, \Delta_2)$ in the 2nd or the 4th quadrant in the restricted null space, it can be shown that the maximum of the power function at $(\Delta_1, 0)$ or $(0, \Delta_2)$ is greater than that at $(\Delta_1, \Delta_2)$. This means that for all practical purposes, we only need to consider the power function over the portion of the X-axis between (0,0) and (M,0) and over the Y-axis that is between (0,0) and (0,M). Generally $M$ represents effect size $(\Delta_j/\sigma_j)$, (where $\sigma_j$ is the diagonal element of the $\Sigma$) that could be realistically be expected of the new drug on the endpoint when the drug has no benefit on the other. Now under the construction of simulation studey to detect power of the min test, the power function can be derived as follows (Chuang et al. 2007(4) for two endpoints. Let $T_1$ and $T_2$ be the two test statistics for comparing an investigational treatment against placebo in a superiority trial,
where $T_j = \frac{X_j - Y_j}{\sqrt{2/n}}$. Assuming that the decision rule takes the form of $T_1 > c, T_2 > c$. The power function is given by

$$Pr(T_1 > c, T_2 > c|\Delta_1, \Delta_2) = Pr\left(\frac{\bar{X}_1 - \bar{Y}_1}{\sqrt{2/n}} > c, \frac{\bar{X}_2 - \bar{Y}_2}{\sqrt{2/n}} > c|\Delta_1, \Delta_2\right)$$

$$= Pr\left(\frac{\bar{X}_1 - \bar{Y}_1 - \Delta_1}{\sqrt{2/n}} > c - \Delta_1\sqrt{\frac{n}{2}}, \frac{\bar{X}_2 - \bar{Y}_2 - \Delta_2}{\sqrt{2/n}} > c - \Delta_2\sqrt{\frac{n}{2}}|\Delta_1, \Delta_2\right)$$

$$= Pr\left(Z_1 > c - \Delta_1\sqrt{\frac{n}{2}}, Z_2 > c - \Delta_2\sqrt{\frac{n}{2}}\right)$$

where $Z_1$ and $Z_2$ has a bivariate normal distribution with mean vector $(0,0)$, a standard deviation of 1 and correlation $\rho$. Now as the null space is restricted instead of the entire null space indicated in figure 3, Chuang et al. (2007) showed that it is possible to use higher significance level (called the adjusted significance level).

Figure 4 gives the adjusted significance levels to get overall 0.025 significance level, based on a simulation study (Chuang et al. (2007)) for 3 choices of sample size per treatment group (50, 100, 150) and 4 choices of correlation (0, 0.2, 0.4, 0.6) for M = 0.5 and 0.8. For the case of M = 0.5, n = 50, 3 co-primary endpoints and a zero correlation coefficient, the adjusted significance level is 0.041. The shaded cells indicates negligible adjustment. Whe the restricted null space is defined by M = 0.8, adjustment is only possible when n = 50 and the correlation is equal to 0. Considering the most registration trials enrol more than 100 subjects per treatment group, the impact of this approach in terms of adjusting the significance level upwards for testing individual endpoints is minimal.

The restricted null space does not have to be square within quadrants II-IV. For example, one can consider rectangle area and different critical values for different endpoints. The central idea is to replace the full null space by a reasonable restricted null space and control the false positive rate over the restricted space. This can be done by choosing suitable value of M based on the underlying disease and factors such as available treatment options for the disease.

Unfortunately this method offers very limited relief from the requirement of the IU test for the kind of sample size typically seen in phase III trials, so also in terms of power.
3.2 Average Type I Error Approach

Chuang et al. (2007) proposed this method which is based on looking at the average power function over the restricted null space with weight given by a function similar to the concept of a prior distribution over the restricted null space. They call it the average type I error. Here instead of looking at the maximum false positive rate over a restricted null space this approach looks at the average false positive rate over that space. They describe it with a uniform prior.

Let $\alpha(\Delta_1, \Delta_2)$ denote the power function, where $(\Delta_1, \Delta_2)$ are the effect size described previously. Let $U(\Delta_1, \Delta_2)$ represent the uniform distribution of $(\Delta_1, \Delta_2)$ in the restricted null space, i.e. $U(\Delta_1, \Delta_2) = 1/3M^2$, as we are considering three quadrants. It is clear that $\alpha(\Delta_1, \Delta_2)$ is smaller than $\alpha(0, 0)$, $\alpha(0, M)$, $\alpha(M, 0)$ in third, second and fourth quadrant respectively. $\alpha(M, 0)$ and $\alpha(0, M)$ are smaller than $\alpha(\infty, 0)$ and $\alpha(0, \infty)$ respectively. Even though points $\alpha(0, \infty)$ and $\alpha(\infty, 0)$ are not in the restricted null space, they are the value of $\alpha(0, M)$ and $\alpha(M, 0)$ as $M$ goes to $\infty$. The upper bound for the average type I error is given by

$$\frac{1}{3M^2} \int \int \alpha(\Delta_1, \Delta_2) d\Delta_1 d\Delta_2 \leq \frac{\alpha(0, 0) + \alpha(0, \infty) + \alpha(\infty, 0)}{3}$$
For the simple case when the two endpoints are independent and each endpoint is tested at the $\alpha^*$ level, the upper bound becomes

$$\frac{\alpha^*^2 + 2\alpha^*}{3}$$

One can set this quantity to 0.025 and solve for $\alpha$. Similarly when the number of endpoints are 3, then null space has seven subparts consist of points that are $(-, -, -), (-, -, +), (-, +, -), (+, -, -), (-, +, +), (+, -, +), (-, +, +)$. Uniform prior will be $\alpha(\Delta_1, \Delta_2, \Delta_3) = \frac{1}{7M^2}$. When endpoints are independent then under the uniform prior the upper bound for the type I error will be

$$\frac{\alpha(0,0,0) + \alpha(0,0,\infty) + \alpha(0,\infty,\infty)}{7} = \frac{\alpha^*^3 + 3\alpha^*^2 + 3\alpha^*}{7}$$

Figure 5 gives the adjusted significance level for four choices of $m$ and five choices of correlation (0, 0.2, 0.4, 0.6, 0.8). When correlation is 0.4, the approach gives the adjusted significance level is 0.035 for 2 co-primary points, 0.048 for 3 and so on. The adjustment is independent of choice of M. But if we compare with the Figure 4 adjustment for "Average Type I Error Approach" is more than the test based on min test on restricted null space.

But this method has more power than the previous one. Sample size adjustment to get a specified power is less for the this approach. The formula for comparison of the sample size between two method is given by

$$\frac{n_{FB}}{n_S} = \frac{(c_{FB} - z_{FB}^2)^2}{(1.96 - z_{FB}^S)^2}$$

7
Where $z_{\beta}^S$ is the 100(1-\(\beta\)) percent power to detect an effect size \(\Delta\) at one-sided 2.5 percent level in a single endpoint. $z_{\beta}^{FB} = c_{FB} - \Delta\sqrt{n_{FB}/2}$ denote the value so that the region $((x_1, ..., x_m), x_i > z_{\beta}^{FB})$ has $(1 - \beta)$ probability i.e. power of the average type I error test. $c_{FB}$ is the critical value of the average type I error approach.

In Figure 6 a comparison of sample size increment for average type I error approach and IU test (in bracket) is given. For example when correlation 0.4 and the required power is 0.8 then the sample size increment is 14 and 26 for average type I error approach and IU test respectively. In all the cases this average type I error approach has the lower increment in the sample size for a fixed power.

In this method the author put more concern about the sample size determination, while neglecting the adjustment of significance level. They mentioned that the restricted null space may not be square, it can be rectangle but in avargae type I error process they are assuming it a square to come up with uniform prior. In real scenerio treatment may not have the same highest effect on all the co-primary end points. In those case prior distribution will be different. More investigation is required in this perspective.

<table>
<thead>
<tr>
<th>Number of co-primary endpoints</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>80% Power (%)</td>
<td>90% Power (%)</td>
<td>80% Power (%)</td>
<td>90% Power (%)</td>
</tr>
<tr>
<td>0.0</td>
<td>18 (31)</td>
<td>12 (23)</td>
<td>19 (49)</td>
<td>11 (36)</td>
</tr>
<tr>
<td>0.2</td>
<td>16 (29)</td>
<td>11 (22)</td>
<td>18 (46)</td>
<td>11 (34)</td>
</tr>
<tr>
<td>0.4</td>
<td>14 (26)</td>
<td>10 (22)</td>
<td>17 (46)</td>
<td>11 (34)</td>
</tr>
<tr>
<td>0.6</td>
<td>14 (22)</td>
<td>11 (18)</td>
<td>15 (35)</td>
<td>11 (27)</td>
</tr>
<tr>
<td>0.8</td>
<td>11 (16)</td>
<td>8 (13)</td>
<td>12 (25)</td>
<td>9 (21)</td>
</tr>
</tbody>
</table>

Figure 6: Sample size increment to maintain a fixed power
4 Appendix

References


