Lecture 8. Multiple Hypothesis Testing

[1] *School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen)*

1. Global Testing Problem: One Sample Multiple Hypothesis Testing Problem

For a global testing problem (simultaneous multiple hypothesis testing problem) in the form

$$H_0 = \bigcap_{i=1}^{m} H_{0i} \ \ v.s. \ \ H_1 = \bigcup_{i=1}^{m} H_{1i}. \tag{1.1}$$

We require m to a non-infinite number when given a fixed number sample. But when allow the sample size, denoted as n, goes to infinite, we allow $m = m(n)$ to depend on n as well. Furthermore, we don't specifically require m to be in low-dimensional, high-dimensional, or ultra-high-dimensional range.

In the global testing problem, the global null we are interested states that all of the individual nulls are true. One may often encounter such problems in bioinformatics field. For example, suppose we have n genes and data about expression levels for each gene among healthy individuals and those with breast cancer,

	Expression level of Gene i
k Healthy patients:	$x_{i1}^{(0)}, x_{i2}^{(0)}, \cdots, x_{ik}^{(0)}$
ℓ breast cancer patients:	$x_{i1}^{(1)}, x_{i2}^{(1)}, \cdots, x_{i\ell}^{(1)}$

Naturally, the $i-$th null hypothesis, denoted as H_{0i}, states that the mean expression level of the $i-$th gene is the same in both groups of patients, i.e.,

$$H_{0i}: \ \mathbb{E}\big(x_{ij}^{(0)}\big) = \mathbb{E}\big(x_{ij}^{(1)}\big) \ \ v.s. \ \ H_{1i}: \ \mathbb{E}\big(x_{ij}^{(0)}\big) \neq \mathbb{E}\big(x_{ij}^{(1)}\big).$$

For simplicity, we assume for each hypothesis H_{0i} $v.s.$ H_{1i}, we already have a test statistic T_i associated with a rejection region $R_i(\alpha') = \{T_i > c(\alpha')\}$, hence, a p-value p_i that can be defined as

$$p_i = \mathbb{P}\big(T_i \geq t_i | H_0\big) = 1 - F_{T_i}(t_i)$$

where t_i is the observed value of the statistic T_i. When the distribution function F_{T_i} is continuous under the null hypothesis, we have $F_{T_i}(T_i) \sim Uniform[0,1]$, and hence the $p_i \sim Uniform[0,1]$. But in general, when we have rather irregular rejection region or when the distribution function F_{T_i} is discontinuous under the null hypothesis, the distribution of p_i can still be hard to track.

To avoid being buried under tedious discussion, we assume $p_i \sim Uniform[0,1]$ for each $i = 1, \cdots, m$ in the later paragraph.

2. Approaches in Global Hypothesis Testing

2.1. Approaches in setting the significant level

We introduce some approaches in setting the significant level in a global testing problem:

- ♣ Bonferroni Approach:

  Procedure: Given a significant level α for the global testing problem (1.1), the Bonferroni approach tests the global null H_0 by simply testing each H_{0i} at level α/m and rejecting the global null whenever any of the H_{0i} is rejected, i.e., the Bonferroni approach rejects the global null H_0 if

  $$\min_{1\le i\le m} p_i < \frac{\alpha}{m}.$$

  Apparently,

  $$\begin{aligned}\mathbb{P}_{H_0}\Big(\text{Type-I error}\Big) &= \mathbb{P}_{H_0}\Big(\bigcup_{i=1}^{m}\Big\{p_i < \frac{\alpha}{m}\Big\}\Big)\\ &\le \sum_{i=1}^{m}\mathbb{P}_{H_0}\Big(p_i < \frac{\alpha}{m}\Big) = \sum_{i=1}^{m}\frac{\alpha}{m} = \alpha,\end{aligned}$$

  So Bonferroni approach is valid. Even if one think the union bound might seem crude, at least in the case when the hypotheses are independent,the size of Bonferroni's test is very near α. Indeed, under independence, the size of the test is

  $$\begin{aligned}\mathbb{P}_{H_0}\Big(\text{Type-I error}\Big) &= 1 - \mathbb{P}_{H_0}\Big(\bigcap_{i=1}^{m}\Big\{p_i \ge \frac{\alpha}{m}\Big\}\Big)\\ &= 1 - \Big(1 - \frac{\alpha}{m}\Big)^m \to 1 - e^{-\alpha},\end{aligned}$$

This tells us that if we have many hypotheses (e.g. n $= 10^4$ genes in the biological example),then Bonferroni's test has size approximately $1-e^{-\alpha}$, which for small α is approximately α.

Remark 2.1. To gain some more intuition about this test, Bonferroni's test looks only at the smallest p-value, and checks if this value is below $\alpha = n$. Hence, the test is most suited for situations where we expect at least one of the p-values to be very significant. Thus, we expect Bonferroni's test to be powerful against alternatives with this property. In the biological example, we might apply this test if we expect one (or a few) of the genes to be very significantly linked to prostate cancer.

- ♣ Šidák Approach:

  Procedure: Given a significant level α for the global testing problem (1.1), further assume that the sub-tests are mutually independent or positively quadrant dependent, i.e., $\{p_i\}_{1\le i\le m}$ are mutually independent or positively quadrant dependent, then Šidák approach tests the global null H_0 by simply testing each H_{0i} at level $\alpha' = 1-(1-\alpha)^{1/m}$ and rejecting the global null whenever any of the H_{0i} is rejected, i.e., Šidák approach rejects the global null H_0 if

$$\min_{1\le i\le m} p_i < 1-(1-\alpha)^{1/m}.$$

Remark 2.2. In general, unlike the Bonferroni approach works in all cases, Šidák approach requires the sub-tests to be mutually independent or positively quadrant dependent in order to get an valid control on type-I error, then we have

$$\begin{aligned}
\mathbb{P}_{H_0}\Big(\text{Type-I error}\Big) &= \mathbb{P}_{H_0}\Big(\bigcup_{i=1}^{m}\Big\{p_i < 1-(1-\alpha)^{1/m}\Big\}\Big)\\
&= 1-\mathbb{P}_{H_0}\Big(\bigcap_{i=1}^{m}\Big\{p_i \ge 1-(1-\alpha)^{1/m}\Big\}\Big)\\
&\le 1-\prod_{i=1}^{m}\mathbb{P}_{H_0}\Big(p_i \ge 1-(1-\alpha)^{1/m}\Big)\\
&= 1-\prod_{i=1}^{m}(1-\alpha)^{1/m} = \alpha
\end{aligned}$$

so Šidák approach is valid under these conditions and we can see it controls the type-I error at level α exactly when the sub-tests are mutually independent, and is conservative when the sub-tests are positively quadrant dependent.

Definition 2.3 (**Negative Quadrant Dependent**). For a sequence of random variables $\{p_i\}_{1\le i\le m}$ (assume they have continuous distribution functions for simplicity), we say they are positively quadrant dependent if

for arbitrary $\mu_1, \cdots, \mu_m \in \mathbb{R}$, we have

$$PQD(\mu_1, \cdots, \mu_m) \triangleq \mathbb{P}\left(\bigcap_{i=1}^{m} \{p_i \leq \mu_i\}\right) - \prod_{i=1}^{m} \mathbb{P}\left(p_i \leq \mu_i\right) \geq 0,$$

or equivalently,

$$\widetilde{PQD}(\mu_1, \cdots, \mu_m) \triangleq \mathbb{P}\left(\bigcap_{i=1}^{m} \{p_i \geq \mu_i\}\right) - \prod_{i=1}^{m} \mathbb{P}\left(p_i \geq \mu_i\right) \geq 0.$$

Similarly, we define $\{p_i\}_{1\leq i\leq m}$ being negatively quadrant dependent if $PQD(\mu_1, \cdots, \mu_m) \leq 0$, or equivalently, $\widetilde{PQD}(\mu_1, \cdots, \mu_m) \leq 0$.

2.2. Approaches in combination type test

We introduce some combination type test approaches in a global testing problem:

- ♣ Fisher's Combination Test:

  Procedure: Given a significant level α for the global testing problem (1.1), further assume that the sub-tests are mutually independent, i.e., $\{p_i\}_{1\leq i\leq m}$ are mutually independent, then Fisher's combination test rejects the null hypothesis H_0 if

  $$T = -\sum_{i=1}^{m} 2\log p_i > \chi^2_{2m}(1-\alpha).$$

  Notice that the test statistic T increases when each p_i approach to zero, so it makes sense that smaller p-values will push up the value of T and the Fisher's combination test rejects large value of T. When the sub-tests are mutually independent, note that Fisher's Combination Test aggregates all of the p-values in log scale, rather than aggregates p-values directly. Hence, we expect this test to be powerful when there are many small effects but less powerful when there are a few strong effects. In other words, Fisher's combination test tends to be more powerful in dense data and as an opposite, Bonferroni's approach tends to be more powerful in sparse data cases.

  As a matter of fact, when sub-tests are mutually independent, we have

  Proposition 2.4. *Suppose $\{p_i\}_{1\leq i\leq m}$ are mutually independent and follows uniform distribution under the null hypothesis. Then under the null hypothesis, $T \sim \chi^2_{2m}$.*

  Proof. By solve the equation

  $$F^{-1}(x) = -2\log x,$$

we conclude that

$$F(x) = e^{-x/2},$$

which is the distribution function of Exponential distribution with rate parameter $\lambda = 2$. Since $\{p_i\}_{1\le i\le m}$ are mutually independent and follows uniform distribution under the null hypothesis, so we conclude $\{-2\log p_i, i = 1, \cdots, m\}$ are i.i.d follow Exp(2). Further, we have

$$T = -\sum_{i=1}^{m} 2\log p_i \sim \text{Gamma}(m, 2) \overset{d}{=} \chi^2_{2m}.$$

□

So Fisher's combination test is valid.

Remark 2.5. Note that Fisher's Combination Test aggregates all of the p-values (in log scale), rather than just looking at the minimum p-value. Hence, we expect this test to be powerful when there are many small effects and less powerful when there are only a few strong effects. In this sense, Fisher's Combination Test is the opposite of Bonferroni's test. We emphasize that Fisher's Combination Test requires the p_i's to be independent. It is only with this assumption that we were able to conclude that T is distributed as χ^2_{2m}.

- ♣ Simes's Approach:

  Procedure: Given a significant level α for the global testing problem (1.1), further assume that the sub-tests are mutually independent, i.e., $\{p_i\}_{1\le i\le m}$ are mutually independent, then Simes's approach uses the Simes Statistic to tests the global null H_0,

$$T_{Simes} = \min_{1\le i\le m} \left\{ p_{(i)} \cdot \frac{m}{i} \right\},$$

  where

$$p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)}$$

  is the order statistics of $\{p_i\}_{1\le i\le m}$. Simes's approach rejects H_0 if $T_{Simes} \le \alpha$.

  As a matter of fact, when sub-tests are mutually independent, we have

  Proposition 2.6. *Suppose $\{p_i\}_{1\le i\le m}$ are mutually independent and follows uniform distribution under the null hypothesis. Then under the null hypothesis, $T_{Simes} \sim Uniform[0, 1]$.*

  So Fisher's combination test is valid.

3. Post-hoc Multiple Comparison Procedures

Multiple comparison problem is a special case of the multiple hypothesis testing, and often links to ANOVA and its analysis results, therefore, sometimes been

called the Post-hoc multiple comparison problem. Compare to the classical multiple hypothesis testing problem with form

$$\text{Multiple hypothesis problem:} \quad H_{0i} \ \ v.s. \ \ H_{1i}, \ \ i = 1, 2, \cdots, m, \cdots$$

multiple comparison problem specifies each H_{0i} to be a comparison problem, i.e.,

$$\text{Multiple comparison problem:} \tag{3.1}$$
$$H_{0ij} : \mu_i = \mu_j \ \ v.s. \ \ H_{1ij} : \mu_i \neq \mu_j, \ \ i, j = 1, 2, \cdots, m, \cdots, \ \text{ and } i \neq j$$

Those multiple comparison problem can be further separated into two types, pairwise multiple comparison problem and the simultaneous multiple comparison problem. Where the pairwise multiple comparison problem tests each

$$H_{0ij} \ v.s. \ H_{1ij}$$

separately with significance α, and the simultaneous multiple comparison problem is a special case of global testing problem that tests

$$\text{Simultaneous multiple comparison problem:} \tag{3.2}$$
$$H_0 = \bigcap \{H_{0ij} : \mu_i = \mu_j\} \ \ v.s. \ \ H_1 = \bigcup \{H_{1ij} : \mu_i \neq \mu_j\},$$
$$i, j = 1, 2, \cdots, m, \cdots, \ \text{ and } i \neq j,$$

with a single significant level α for this H_0 $v.s.$ H_1. Notice that (3.2) can also be written as

$$H_0 : \mu_1 = \cdots = \mu_m, \ \ v.s. \ \ H_1 : \mu_i \neq \mu_j \text{ for some } i \text{ and } j \text{ where } i \neq j$$

with a single significant level α for this H_0 $v.s.$ H_1.

One of the most commonly seen case of the multiple comparison problem is in post-hoc ANOVA, where μ_i represents the mean of $i-$th group and we are interested in testing whether there are differences between different groups. To avoid complicity, we only consider data meets ANOVA condition in the later content if no further explanation is made. In other words, consider we have data generated from classical normal distributed fixed effect model, i.e.,

$$\left\{Y_{ij}, j = 1, \cdots, n_i, i = 1, \cdots, m\right\}$$

was generated from homoscedastic normal distribution

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \text{where } \ j = 1, \cdots, n_i, \text{ and } i = 1, \cdots, m, \tag{3.3}$$

where $\epsilon_{ij} \sim_{i.i.d} N(0, \sigma^2)$, i.e., the data are generated from m groups with $N = \sum_{i=1}^{m} n_i$. Denote

$$\text{Group mean: } \bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}, \quad i = 1, \cdots, m,$$

$$\text{Mean square error: } S_E^2 = \frac{1}{N-m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2, \quad i = 1, \cdots, m,$$

In the following content, we will introduce several approaches that are specifically designed for post-hoc multiple comparison problems. But as a preview, we emphasize that those post-hoc multiple comparison procedures are difficult to compare as some of them are designed for pairwise comparison(E.g., one Step-down procedure called SNK approach), some of them are designed for simultaneous comparison (E.g., Dunnett's approach), while some of them are designed for pairwise comparison but are adaptive to simultaneous comparison if we combine them with approaches introduced in Section 2 (E.g., combine Fisher's LSD with Bonferroni's approach).

3.1. Approaches for simultaneous multiple comparison problem

We introduce some approaches that are designed for simultaneous multiple comparison problem (3.2). Apparently, they are not as general as approaches introduced in Section 2 as they only focus on (3.2) though they share similar ideas:

- ♣ Dunnett's Approach:
  Unlike taking intersection with respect to both i and j in the H_0 of (3.2), Dunnett's approach is usually invoked for an slightly different simultaneous multiple comparison problem:

$$H_0 = \bigcap_{j=2}^{m} \{H_{0j} : \mu_1 = \mu_j\} \quad v.s. \quad H_1 = \bigcup_{j=2}^{m} \{H_{1j} : \mu_1 \neq \mu_j\}, \\ \text{for some } j = 2, \cdots, m. \tag{3.4}$$

  In other words, instead of interested in testing all pairwise differences, Dunnett"s approach focus on the comparison between other groups (group $2, 3, \cdots, m$, one may think them as various treatment groups) to a specific group (denoted as group 1, and one may think it as the control group). This restriction can of course been extend to (3.2) but with some complicity in computing the corresponding critical value (investigate corresponding limiting distribution).
  Procedure: Given a significant level α for the simultaneous multiple comparison problem (3.4), for $j = 2, \cdots, m$, define the test statistic

$$T_j \triangleq \frac{\bar{Y}_j - \bar{Y}_1}{S_E \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_j}}}, \quad \text{and } T = \max_{2 \leq j \leq m} |T_j|$$

In Dunnett's original paper Dunnett (1955), the procedures then follows by try to find the vector of critical values $(d_2, \cdots, d_m)$, such that

$$\mathbb{P}\Big(|T_j| \le d_j, j = 2, \cdots, m\Big) = 1 - \alpha$$

and we reject the global null hypothesis H_0 in (3.4) whenever

$$\max_{2 \le j \le m} \Big(|T_j| - d_j\Big) > 0$$

Remark 3.1. Generally, the critical value $(d_2, \cdots, d_m)$ is hard to find, but Dunnett and Sobel (1955) gives the following proposition,

Proposition 3.2. *Assume the data is generated from homoscedastic normal distribution (3.3), then statistics* $\{T_j\}_{2 \le j \le m}$ *are positively quadrant dependent.*

Therefore we may pick a critical value slightly conservative, i.e., $d_j = t_{N-m}(1 - \alpha/2)$, since

$$\prod_{j=2}^{m} \mathbb{P}\Big(|T_j| \le d_j\Big) \le \mathbb{P}\Big(|T_j| \le d_j, j = 2, \cdots, m\Big) = 1 - \alpha,$$

and each T_j follows $t-$distribution with degree of freedom $N - m$.

So Dunnett's approach is valid.

4. Optimality of Bonferroni's Global Test

Despite being a simple procedure, we show here that Bonferroni'smethod is somehow optimal for testing against sparse alternatives. This claim relies on power calculations, which require us to specify alternatives.

Consider an independent Gaussian sequence model:

$$Y_i \sim \mathcal{N}(\mu_i, 1), \quad i = 1, \ldots, n,$$

where $Y = \{Y_1, \cdots, Y_n\}$ are mutually independent with each other. We are interested in the n hypotheses

$$H_{0i} : \mu_i = 0 \ \ v.s. \ \ H_{1i} : \mu_i \neq 0,$$

so that in this case, the global null asserts that all the means μ_i vanish, while under the alternative H_1, at least some means $\mu_i \neq 0$.

When use Bonferroni's approach, we rejects the overall H_0 if

$$\max_{1\le i\le n} |Y_i| > \mu(1-\alpha/(2n))$$

with $\mu(1-\alpha/(2n))$ being the $(1-\alpha/(2n))-$th quantile of the standard normal distribution. Put another way, Bonferroni rejects the global null hypothesis if the largest Y_i is large enough. For the special case where the n tests are mutually independent, we may calculated that

$$\mathbb{P}_{H_0}(\text{Type I Error}) := q(\alpha) \approx 1 - e^{-\alpha} \approx \alpha.$$

To see this, notice that according to Markov's inequality, for $Z \sim N(0,1)$, we have

$$\begin{aligned}\mathbb{P}(Z>t) &= \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \cdot \int_0^\infty e^{-tz-z^2/2} dz \\ &\le \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \cdot \int_0^\infty e^{-tz} dz = \frac{e^{-\frac{t^2}{2}}}{t\cdot\sqrt{2\pi}} = \frac{\phi(t)}{t}\end{aligned}$$

where $\phi(\cdot)$ is the density function of the standard normal distribution. On the other hand, we have

$$\begin{aligned}\mathbb{P}(Z>t) &= \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \ge \int_t^\infty (1-3x^{-4}) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\ &= \left(\frac{1}{t} - \frac{1}{t^3}\right) \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} = \left(\frac{1}{t} - \frac{1}{t^3}\right) \phi(t).\end{aligned}$$

Hence, overall we have for $\forall t > 0$,

$$\frac{\phi(t)}{t}\left(1 - \frac{1}{t^2}\right) \le \mathbb{P}(Z>t) \le \frac{\phi(t)}{t},$$

That is, for large t, $\phi(t)/t$ is a good approximation to the normal tail probability and we may use it to calculate t, which is the Gaussian quantile. Roughly speaking, for large n

$$\mathbb{P}(Z>t) = \alpha/n \ \text{“}\Leftrightarrow\text{”}\ \frac{\phi(t)}{t} \approx \alpha/n.$$

Holding α fixed, then, we can show that for large n,

$$|\mu(1-\alpha/(2n))| \approx \sqrt{2\log(2n)}\left[1 - \frac{1}{4}\frac{\log\log(2n)}{\log(2n)}\right] \approx \sqrt{2\log(2n)}.$$

Hence, the quantiles grow like $\sqrt{2\log(2n)}$, with a small correction factor. Figure.1 plots $\mu(1-\alpha/n)$ and $\sqrt{2\log n}$. Notice that Bonferroni then basically amounts to rejecting when $\max |Y_i| > \sqrt{2\log n}$.

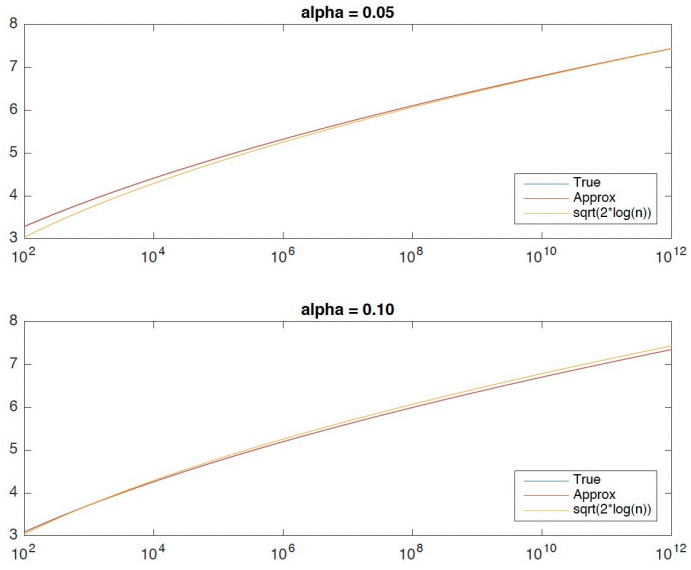

Figure 1: Bonferroni Approach

One remarkable fact about all of this is that there is (asymptotically) no dependence on α. That is, whatever α we use, our rejection threshold for $\max |Y_i|$ is asymptotic to $\sqrt{2 \log n}$. This is a consequence of the fact that, under H_0,

$$\frac{\max |y_i|}{\sqrt{2 \log n}} \xrightarrow{p} 1.$$

In other words, the first order term $\sqrt{2 \log n}$ asymptotically dominates the terms containing α. If we set

$$B = 2 \log(n/\alpha) - \log(2\pi) = 2 \log(n/\alpha) - 1.8379.$$

Then

$$|\mu(1 - \alpha/n)| \approx \sqrt{B \left(1 - \frac{\log B}{B}\right)}.$$

Figure 1 shows that this approximation is nearly indistiguishable from $\mu(1 - \alpha/n)$, even for modest values of n.

5. Sharp Detection Threshold for the "Needle in a Haystack"

Definition 5.1 (**Asymptotic Power**). For a sequence of problems with $n \to \infty$, the asymptotic power is the limiting power

$$\mathbb{P}_{H_1}(\text{Reject } H_0).$$

A natural question arise that we want to see how powerful is Bonferroni,or, put it another way, what is the limiting power $\mathbb{P}_{H_1}(\max Y_i > \mu(1-\alpha/(2n)))$?

To answer the above question, we first look at a simple problem called the **Needle in a Haystack Problem**, for which we specify the alternative hypotheses. The needle in a haystack problem is this: under the alternative, we have one and only one $\mu_i = \mu(n) > 0$. But we don't know which one. For this problem, we shall see that the answer to the power question depends very sensitively on the limiting ratio

$$\frac{\mu(n)}{\sqrt{2\log n}},$$

where the n in $\mu(n)$ are emphasizing the dependence between this non-zero mean and the sample size. There are two cases:

(i) Asymptotic full power above threshold: Suppose $\mu(n) > (1+\epsilon)\sqrt{2\log n}$. Then, without loss of generality, assume that $\mu_1 = \mu(n)$, we have

$$\begin{aligned}\mathbb{P}_{H_1}(\text{Reject } H_0) &= \mathbb{P}_{H_1}\Big(\max|Y_i| > \mu(1-\alpha/(2n))\Big)\\ &\geq \mathbb{P}_{H_1}\Big(Y_1 > \mu(1-\alpha/(2n))\Big)\\ &= \Phi\Big(\mu(1-\alpha/(2n)) - \mu(n)\Big) \to 1.\end{aligned}$$

(ii) Asymptotic powerlessness below threshold: Suppose $\mu(n) < (1-\epsilon)\sqrt{2\log n}$. Then, without loss of generality, assume that $\mu_1 = \mu(n)$, we have

$$\begin{aligned}\mathbb{P}_{H_1}(\text{Reject } H_0) &= \mathbb{P}_{H_1}\Big(\max|Y_i| > \mu(1-\alpha/(2n))\Big)\\ &\leq \mathbb{P}_{H_1}\Big(Y_1 > \mu(1-\alpha/(2n))\Big) + \mathbb{P}\Big(\max_{i>1}|Y_i| > \mu(1-\alpha/(2n))\Big)\\ &= \Phi\Big(\mu(1-\alpha/(2n)) - \mu(n)\Big) + \mathbb{P}\Big(\max_{i>1}|Y_i| > \mu(1-\alpha/(2n))\Big)\\ &\to 0 + \alpha = \alpha.\end{aligned}$$

meaning this is a bad test because we can obtain the same level and power by flipping a biased coin that rejects α of the time.

Therefore, we may effectively see that $\sqrt{2\log n}$ constitutes a sharp detection threshold. When $\mu(n) = (1+\epsilon)\sqrt{2\log n}$, we can always detect the needle that $\mu_1 > 0$. We can even achieve

$$\mathbb{P}_{H_0}\Big(\text{Type I Error}\Big) \to 0 \text{ and } \mathbb{P}_{H_1}\Big(\text{Type II Error}\Big) \to 0.$$

if we use $\sqrt{2\log n}$ instead of $\mu(1-\alpha/(2n))$ as our threshold. In other words, asymptotically we make no mistakes.

However, when $\mu(n) = (1-\epsilon)\sqrt{2\log n}$, with $q(\alpha) = 1 - e^{-\alpha} \approx \alpha$ being the asymptotic size, Bonferroni's global test gives

$$\mathbb{P}_{H_0}\Big(\text{Type I Error}\Big) \to q(\alpha) \;\text{ and }\; \mathbb{P}_{H_1}\Big(\text{Type II Error}\Big) \to 1 - q(\alpha).$$

that is, it does no better than flipping a coin.

References

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50(272), 1096-1121.

Dunnett, C. W., & Sobel, M. (1955). Approximations to the probability integral and certain percentage points of a multivariate analogue of Student's t-distribution. Biometrika, 42(1/2), 258-260.