PhD in Economics and Statistics - University of Milano-Bicocca

Lecture 3: Methods for familywise error control

May 7, 2019

Lecturer: Aldo Solari

We first discuss the methods of Bonferroni and Holm, which are valid under any dependence of the p-values, then the method of Hochberg, which is valid if the Simes' inequality holds. For a broader discussion, see [2]. Sections 2 and 3 are taken from [1].

1 Bonferroni

The method of Bonferroni controls FWER at level α by rejecting hypotheses only if they have raw p-value smaller than α/m.

The Bonferroni method is a corollary to Boole's inequality, which says that for any collection of events $E_1, \ldots, E_k$, we have

$$\mathrm{P}\Big(\bigcup_{i=1}^{k} E_i\Big) \leq \sum_{i=1}^{k} \mathrm{P}(E_i).$$

It follows from Boole's inequality that, if $q_1, \ldots, q_{m_0}$ are the p-values of the true null hypotheses, that the probability that there is some i for which $q_i \leq \alpha/m$ is at most α.

Theorem 1.1. *Bonferroni method controls the FWER at level α:*

$$\mathrm{FWER} \leq \pi_0 \alpha \leq \alpha$$

Proof.

$$\mathrm{P}\Big(\bigcup_{i=1}^{m_0} \{q_i \leq \frac{\alpha}{m}\}\Big) \leq \sum_{i=1}^{m_0} \mathrm{P}\Big(q_i \leq \frac{\alpha}{m}\Big) \leq m_0 \frac{\alpha}{m} \leq \alpha. \tag{1}$$

□

The two inequalities in (1) indicate in which cases the Bonferroni method can be conservative.

- The right-hand one shows that Bonferroni does not control the FWER at level α but actually at the stricter level $\pi_0\alpha$, where $\pi_0 = m_0/m$. If there are many false null hypotheses, Bonferroni will be conservative:
- The left-hand inequality is due to Boole's law. This inequality is a strict one in all situations except the one in which all events $\{q_i \leq \alpha/m\}$ are disjoint.

1.1 Independent p-values

With independent p-values, this conservativeness is present but very minor. If $m_0 = m$ and $q_i \sim U(0,1)$, then

$$\mathrm{P}\Big(\bigcup_{i=1}^{m}\{q_i \leq \frac{\alpha}{m}\}\Big) = 1 - \prod_{i=1}^{m}\mathrm{P}\Big(q_i > \frac{\alpha}{m}\Big) = 1 - \Big(1 - \frac{\alpha}{m}\Big)^m \overset{m\to\infty}{\to} 1 - e^{-\alpha}$$

This tells us that if we have many hypotheses, all true, and $\alpha = 0.05$, then $1 - e^{-\alpha} = 0.04877$.

We can also compare the Bonferroni critical value α/m with the corresponding Sidak critical value

$$1 - (1-\alpha)^{1/m}$$

for independent p-values $q_1, \ldots, q_m$ i.i.d. $U(0,1)$:

$$\mathrm{P}\Big(\bigcup_{i=1}^{m}\{q_i \leq c\}\Big) = 1 - \prod_{i=1}^{m}\mathrm{P}\Big(q_i > c\Big) = 1 - (1-c)^m$$

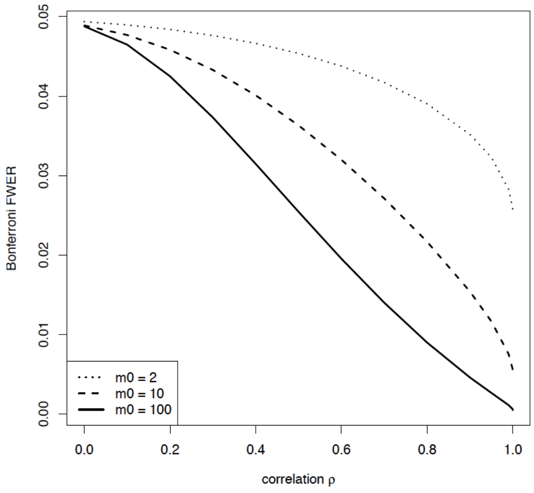
which equals α for $c = 1 - (1-\alpha)^{1/m}$. For $m = 5$ and $\alpha = 0.05$ we find a critical value of 0.01021 for Sidak against 0.01 for Bonferroni. As m increases, the ratio between the two increases to a limit of

$$\frac{\alpha/m}{1-(1-\alpha)^{1/m}} \overset{m\to\infty}{\to} \frac{-\log(1-\alpha)}{\alpha}$$

which evaluates to only 1.026 for $\alpha = 0.05$.

1.2 Positively correlated p-values

Much more serious conservativeness can occur if p-values are positively correlated. For example, in the extreme case that all p-values are perfectly positively correlated, FWER control could already have been achieved with the unadjusted level α, rather than α/m.



1.3 Adjusted p-values

When testing a single hypothesis, we often do not only report whether a hypothesis was rejected, but also the corresponding p-value. By definition, the p-value is the smallest chosen α-level of the test at which the hypothesis would have been rejected. The direct analogue of this in the context of multiple testing is the *adjusted* p-value, defined as the smallest α level at which the multiple testing procedure would reject the hypothesis.

For the Bonferroni procedure, this adjusted p-value is given by

$$\tilde{p}_i = \min(mp_i, 1)$$

where p_i is the raw p-value.

2 Magnitude of the Bonferroni's Threshold

Suppose that $y_1, \ldots, y_m$ are independent with

$$y_i \sim N(\mu_i, 1)$$

and we are interested in the m hypotheses

$$H_i : \mu_i = 0$$

Bonferroni method rejects H_i if

$$y_i \geq |z_{\alpha/m}|$$

in the one-sided case, and if $|y_i| \geq |z_{\alpha/2m}|$ in the two-sided case, where z_α is the α quantile of $N(0,1)$. How large is our threshold $t = |z_{\alpha/m}|$ (one-sided)?

If $\phi(t)$ is the $N(0,1)$ pdf, then we can derive that

$$\mathrm{P}(Z > t) \leq \frac{\phi(t)}{t}$$

where $Z \sim N(0,1)$. To see this, observe that over the interval of integration, we have $z \geq t$

$$\int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz < \frac{1}{t\sqrt{2\pi}} \int_t^\infty z e^{-z^2/2} dz = \frac{1}{t\sqrt{2\pi}} e^{-t^2/2}$$

It follows that

$$\mathrm{P}(Z > \sqrt{2\log m}) \leq \frac{\phi(\sqrt{2\log m})}{\sqrt{2\log m}} = \frac{1}{2m\sqrt{\pi \log m}} < \frac{\alpha}{m}$$

as soon as $\sqrt{\log m} > 1/(2\sqrt{\pi}\alpha)$. Therefore, the $1 - \alpha/m$ quantile of the standard normal distribution must be bounded above by $\sqrt{2\log m}$ as soon as $\sqrt{\log m} > 1/(2\sqrt{\pi}\alpha)$.

It can be proved that $(1-1/t^2)\frac{\phi(t)}{t} < \mathrm{P}(Z > t)$, Then, for any fixed α and $\epsilon > 0$, the following inequalities hold for all large enough m:

$$\sqrt{(1-\epsilon)2\log(m)} \leq |z_{\alpha/m}| \leq \sqrt{2\log(m)}$$

Hence, $|z_{\alpha/m}|$ grows like $\sqrt{2\log m}$, with a small correction factor. For large m, Bonferroni method amounts to reject H_i when

$$y_i \geq \sqrt{2\log m}$$

and there is (asymptotically) no dependence on α.

3 Needle in a haystack problem

The needle in a haystack problem is this: either $\bigcap_{i=1}^m H_i$ is true and $\mu_i = 0$ for all i or just one H_i is false and $\mu_i = \mu > 0$ but we don't know which one. What is the limiting power

$$\lim_{m\to\infty} \mathrm{P}_1(y_{(m)} > |z_{\alpha/m}|)$$

The answer to this question depends on the limiting ratio

$$\frac{\mu^{(m)}}{\sqrt{2\log m}}$$

where $\mu^{(m)} > 0$ is the value of the single non-zero mean, which is a function of m. There are two cases.

- Suppose $\mu^{(m)} > (1+\epsilon)\sqrt{2\log m}$. Then, assuming without loss of generality that $\mu_1 = \mu^{(m)}$,

$$\mathrm{P}_1(y_{(m)} > |z_{\alpha/m}|) \geq \mathrm{P}_1(y_1 > |z_{\alpha/m}|) = \mathrm{P}(Z > |z_{\alpha/m}| - \mu^{(m)}) \to 1$$

- Suppose $\mu^{(m)} < (1+\epsilon)\sqrt{2\log m}$. Then

$$\begin{aligned} \mathrm{P}_1(y_{(m)} > |z_{\alpha/m}|) &\leq \mathrm{P}(y_1 > |z_{\alpha/m}|) + \mathrm{P}(\max_{i>1} y_i > |z_{\alpha/m}|) \\ &= \mathrm{P}(Z > |z_{\alpha/m}| - \mu^{(m)}) + \mathrm{P}(\max_{i>1} y_i > |z_{\alpha/m}|) \\ &\to 0 + (1 - e^{-\alpha}) \approx \alpha \end{aligned}$$

Can we do better than Bonferroni? When $\mu^{(m)} < (1+\epsilon)\sqrt{2\log m}$ we saw that the limiting power is about α, so we are doing no better than flipping a biased coin that disregards the actual data. But this is fact true for any test in this scenario. Even the optimal test given by Neyman-Pearson lemma does no better than flipping a coin. See [1].

4 Holm

Holm's method is a sequential variant of the Bonferroni method that always rejects at least as much as Bonferroni's method, and often a bit more, but still has valid FWER control under the same assumptions. From this perspective, there is no reason, aside from possibly simplicity, to even use Bonferroni's method in preference to Holm's.

Holm remedies part of the conservativeness in the Bonferroni method arising from the right-hand inequality of (1), which makes Bonferroni control FWER at level $\pi_0\alpha$. It does that by iterating the Bonferroni method in the following way.

In the first step, all hypotheses with p-values at most α/h_0 are rejected, with $h_0 = m$ just like in the Bonferroni method. Suppose this leaves h_1 hypotheses unrejected. Then, in the next step, all hypotheses with p-values at most α/h_1 are rejected, which leaves h_2 hypotheses unrejected, which are subsequently tested at level α/h_2. This process is repeated until either all hypotheses are rejected, or until a step fails to result in any additional rejections.

An alternative way of describing Holm's method is via the ordered p-values $p_{(1)}, \ldots, p_{(m)}$. Comparing each p-value $p_{(i)}$ to its corresponding critical value

$$\frac{\alpha}{m - i + 1}$$

Holm's method finds the smallest j such that $p_{(j)}$ exceeds $\alpha/(m - j + 1)$, and subsequently rejects all $j - 1$ hypotheses with a p-value at most $\alpha/(m - j)$. If no such j can be found, all hypotheses are rejected.

5 Hochberg

Bonferroni's and Holm's methods make no assumptions on the dependency structure of the p-values. One such assumption could be that the Simes inequality holds for the subset of true hypotheses. This assumption makes the use of Hochberg's method possible, which is very similar to Holm's method but more powerful.

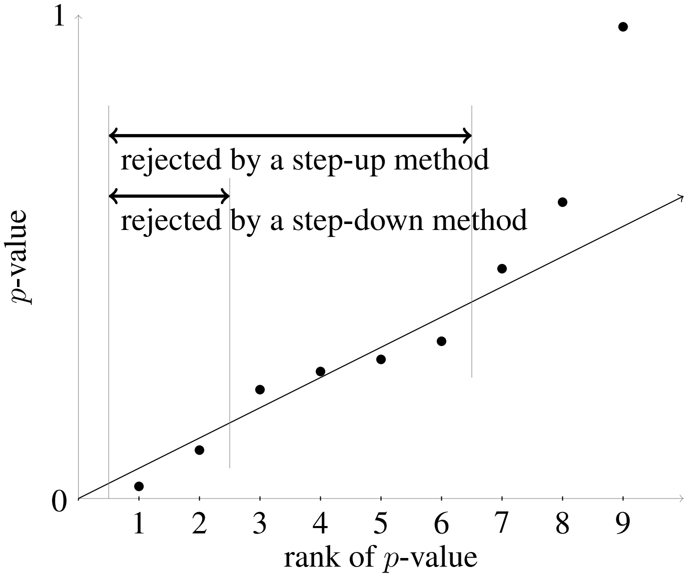
Hochberg's method (not to be confused with Benjamini and Hochberg's method) compares each ordered p-value $p_{(i)}$ to a critical value

$$\frac{\alpha}{m - i + 1}$$

the same as Holm's. It then finds the largest j such that $p_{(j)}$ is smaller than $\alpha/(m - j + 1)$, and subsequently rejects all j hypotheses with p-values at most $\alpha/(m - j + 1)$. Comparing to Holm's method, it is clear that Hochberg's method rejects at least as much as Holm's method, and possibly more.

If the curves $p_{(1)}, \ldots, p_{(m)}$ and $\alpha/m, \alpha/(m - 1), \ldots, \alpha$ never cross or only once, Holm's and Hochberg's methods reject the same number of hypotheses. If the same curves cross multiple times, Holm's method uses the first crossing point as the final critical value, while Hochberg's uses the last crossing point. Holm's method is known as a *step-down* method and Hochberg's as its *step-up* equivalent.

Figure 1: Comparison of rejections by step-up and step-down methods with the same critical values. The dots are observed ranked p-values. The line represents the critical values. Step-down methods reject all hypotheses up to, but not including, the first p-value that is larger than its critical value. Step-up methods reject all hypotheses up to and including the last p-value that is smaller than its critical value.



References

[1] E. Candes et al. Stats 300c: Theory of statistics. *Lecture notes*, 2018.

[2] J. J. Goeman and A. Solari. Multiple hypothesis testing in genomics. *Statistics in medicine*, 33(11):1946–1978, 2014.