Looking at the Other Side of Bonferroni

Caitlin McHugh

Department of Biostatistics
University of Washington

10 April 2012

The Paper

CONTROL OF THE MEAN NUMBER OF FALSE DISCOVERIES, BONFERRONI AND STABILITY OF MULTIPLE TESTING

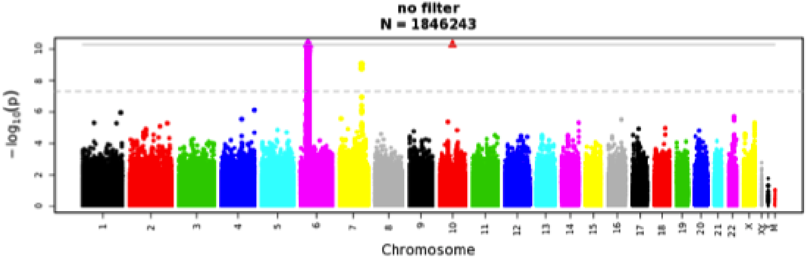
BY ALEXANDER GORDON, GALINA GLAZKO, XING QIU AND ANDREI YAKOVLEV[1]

- Gordon, et al. set out to prove that the Bonferroni testing procedure is not conservative if we simply look at it from a different perspective [2].

Multiple Testing: Control the Type I Error Rate

- When analyzing genetic data, one will commonly perform over 1 million (and growing) hypothesis tests.
- In categorical data analysis, one may want to test all pairwise combinations.
- How do we ensure we are properly controlling for the number of false rejections?

2.5 Million Hypothesis Tests



The Family-Wise Error Rate (FWER)

- For a test, we choose a significance level, α.
- We define the family-wise error rate as

$$\text{FWER} = \mathbb{P}(\#\text{ false rej} \geq 1)$$

- This is the probability of one or more false rejections of the null hypothesis, the probability that there is at least one type I error.

The Per Family Error Rate (PFER)

- We define the per family error rate as

$$\text{PFER} = \mathbb{E}(\#\ \text{false rej})$$

  - This is the expected number of false rejections of the null hypothesis.
  - Can be thought of as the expected number of type I errors, or the expected number of false positives.
- FWER$\leq$PFER.
  - The probability of one or more false rejections is less than or equal to the expected number of false rejections.

Bonferroni Correction

- Say we perform *m* hypothesis tests.
- To adjust our overall significance level, α, simply divide by the number of tests performed, *m*, so our significance level becomes α/m.

$$\alpha^* = \alpha/m$$

- When the p-values are highly correlated, this adjustment is conservative, when thinking about the FWER (the probability of at least one false positive).
  - Note, however, this adjustment is not conservative when thinking about the PFER (the expected number of false positives).
- This controls the probability of one or more false rejections.

Bonferroni does control the PFER!

- Bonferroni correction controls the $\mathbb{E}$(# false rej)=PFER as well as the $\mathbb{P}$(# false rej $\geq$ 1)=FWER [3].
- When controlling the FWER, we have an extremely small threshold (think 0.05/1,000,000).
- When considering the PFER, the threshold is somewhere between 0 and m, the number of tests we are performing.
- Call $Bonf^{\alpha}$ the classical Bonferroni procedure, and denote $Bonf^{\gamma}$ the Bonferroni procedure that controls the PFER, $0 \leq \gamma \leq m$.

What is $Bonf^\gamma$?

- $Bonf^\gamma$ controls the expected number of false positives at level γ.
- PFER$\leq (m_0/m)\gamma$, where m_0 is the number of true null hypotheses in the m hypotheses considered.
- We calculate

$$\begin{aligned}
\text{PFER} &= \mathbb{E}(\#\text{ false rej}) \\
&= \mathbb{E}(\sum_{i\in\mathcal{T}} I_{\{p_i\leq\gamma/m\}}) \\
&= \mathbb{E}(\sum_{i\in\mathcal{T}} \mathbb{P}(p_i \leq \gamma/m)) \\
&\leq (m_0/m)\gamma \\
&\leq \gamma
\end{aligned}$$

where $\mathcal{T}$ is the set of indices of the true null hypotheses and p_i are the observed p-values.

Is *Bonf*$^{\gamma}$ really that cool?

- If the p-values are uniformly distributed on [0,1], *Bonf*$^{\gamma}$ controls the PFER at level $(m_0/m)\gamma$, since PFER$=(m_0/m)\gamma$.
- Since FWER$\leq$PFER, *Bonf*$^{\gamma}$ controls the FWER at level $(m_0/m)\gamma$ also.
- This controls the probability of one or more false rejections.
  - What if we controlled the expected proportion of false rejections, known as the false discovery rate (FDR)?

Benjamin-Hochberg (BH) Correction

- To adjust our overall significance level, α, simply divide by the number of tests performed, and multiply by the ranking of the p-value, so our significance level becomes $\alpha i/m$ for $p_{(i)}$.
- From the ordered p-values, we start at the largest and work down until we find i such that $p_{(i)} \leq \frac{\alpha i}{m}$, call this particular i value k.
- We then reject all $H_{(i)}$, $i \leq k$.
- This controls the expected proportion of false rejections, the FDR [1].

A simple example

- Assume we produce p-values for 5 different tests:

  0.0004 0.0015 0.0095 0.0254 0.0450

- $\alpha = 0.05$
- Using Bonferroni correction, $\alpha^* = \alpha/5 = 0.05/5 = 0.01$.
- Using the BH correction, $\alpha_{(i)} = 0.05i/5 = 0.01i$, so the largest p-value that satisfies the constraint $p_{(i)} \leq 0.01i$ is $p_{(5)} = 0.0450 \leq 0.01i = 0.05$.

A simple example

- Assume we produce p-values for 5 different tests:

  0.0004 0.0015 0.0095 0.0254 0.0450

- No multiple-testing correction: all significant.
- Using Bonferroni correction: 3 smallest p-values are significant.
- Using the BH correction: all significant.

Conclusions

Via simulation, I *will* show that the *Bonf*$^\gamma$ procedure is just as powerful as the BH procedure, when γ is chosen appropriately.

References

Yoav Benjamini and Yosef Hochberg.
Controlling the false discovery rate: a practical and powerful approach to multiple testing.
J. R. Statist. Soc., B(1):289–300, 1995.

Alexander Gordon, Galina Glazko, Xing Qiu, and Andrei Yakovlev.
Control of the mean number of false discoveries, Bonferroni and stability of multiple testing.
The Annals of Applied Statistics, 1(1):179–190, 2007.

Mei-Ling Ting Lee.
Analysis of Microarray gene expression data.
Kluer Academic, Boston, Massachusetts, 2004.