Notation

Consider a collection of m hypotheses $\{H_1, \cdots, H_m\}$. Let p_i denote the p-value corresponding to hypothesis H_i when the hypotheses are unordered. Let $H_{[i]}$ and $p_{[i]}$ denote the ith hypothesis and p-value when the hypotheses are ordered *a priori*. Let $H_{(i)}$ and $p_{(i)}$ denote the ith hypothesis and p-value when the hypotheses are ordered *a posteriori*, in ascending order by the p-values: $p_{(1)} \leq \cdots \leq p_{(m)}$.

Methods for Family Wise Error Rate Control

2.1 Family-wise Error Rate

Definition 2.1.1. Across m tests, the **family-wise error rate** (FWER) is the probability of incorrectly rejecting at least 1 null hypothesis. *Multiple testing correction* refers to adjusting the observed p-values such that, across all m tests, the FWER is $\leq \alpha$. ■

2.2 Bonferroni's Method

In **Bonferroni's method**, hypothesis H_i is *rejected* if $p_i \leq \alpha/m$. Using sub-additivity, the probability of a type I error is:

$$\mathbb{P}\left(\exists\, p_i \leq \frac{\alpha}{m}\right) = \mathbb{P}\left\{\bigcup_{i=1}^{m}\left(p_i \leq \alpha/m\right)\right\} \leq \sum_{i=1}^{m}\mathbb{P}\left(p_i \leq \frac{\alpha}{m}\right) \leq \sum_{i=1}^{m}\frac{\alpha}{m} = \alpha.$$

2.3 Holm's Method

In **Holm's method** [7] the p-values are first sorted in ascending order. The smallest p-value $p_{(1)}$ is compared to α/m. If $p_{(1)} > \alpha/m$, then $H_{(1)}$ is not rejected and the procedure stops. Otherwise, $H_{(1)}$ is rejected, and the second smallest p-value $p_{(2)}$ is compared with $\alpha/(m-1)$. If $p_{(2)} > \alpha/(m-1)$, then $H_{(2)}$ is not rejected and the procedure stops. Else, $H_{(2)}$ is rejected, and the third smallest p-value $p_{(3)}$ is compared with $\alpha/(m-2)$. The procedure continues until one of the ordered p-values exceeds its threshold, or all null hypotheses have been rejected. If the procedure stops at $p_{(k)} > \alpha/(m-k+1)$, then hypotheses $H_{(1)}, \cdots, H_{(k-1)}$ are *rejected*, and hypotheses $H_{(k)}, \cdots, H_{(m)}$ are *retained*.

Holm's method is described as a **step-down** procedure because the p-values are visited in decreasing order of significance. Holm's method is uniformly more powerful than Bonferroni's, yet applicable under the same circumstances. Therefore, unless Bonferroni's method is favored for simplicity, Holm's method is preferred.

2.4 Hochberg's Method

Hochberg's method [6] is the reflection of Holm's method. The p-values first are sorted in ascending order. Now, however, the largest p-value $p_{(m)}$ is compared to α. If $p_{(m)} \leq \alpha$, then $H_{(m)}$ is rejected and the procedure stops. Otherwise, $H_{(m)}$ is retained, and the next largest p-value $p_{(m-1)}$ is compared with $\alpha/2$. If $p_{(m-1)} \leq \alpha/2$, then $H_{(m-1)}$ is rejected and the procedure stops. Else, $H_{(m-1)}$ is retained, and $p_{(m-2)}$ is compared with $\alpha/3$. The procedure continues until one of the ordered p-values falls below its threshold, or all null hypotheses have been retained. If the procedure stops at $p_{(k)} \leq \alpha/(m-k+1)$, then hypotheses $H_{(1)}, \cdots, H_{(k)}$ are *rejected*, while hypotheses $H_{(k+1)}, \cdots, H_{(m)}$ are *retained*.

Hochberg's method is described as a **step-up** procedure because the p-values are visited in increasing order of significance. Hochberg's method is uniformly more powerful than Holm's. However, it assumes non-negative dependence among the p-values.

2.5 Hommel's Method

Hommel's method [8] is based on the **closure principle**: if H_i is false, then every intersection of hypotheses including H_i is necessarily false. By way of example, consider a collection of $m = 3$ hypotheses $\mathcal{H} = \{H_1, H_2, H_3\}$. The *closure* of $\mathcal{H}$ is:

$$\overline{\mathcal{H}} = \big\{H_1, H_2, H_3, H_1 \cap H_2, H_1 \cap H_3, H_2 \cap H_3, H_1 \cap H_2 \cap H_3\big\}.$$

Elementary hypothesis H_1 is rejected at level α if the 4 intersection hypotheses including H_1 are each rejected at level α. These are:

$$\big\{H_1, H_1 \cap H_2, H_1 \cap H_3, H_1 \cap H_2 \cap H_3\big\}.$$

Analogous criteria apply to both H_2 and H_3. Hommel specifically proposed using **Simes' method** [10]: to test the *intersection hypothesis* $H_1 \cap \cdots \cap H_k$, order the corresponding p-values $p_{(1)} \leq \cdots \leq p_{(k)}$, and reject the intersection $\cap_{i=1}^{k} H_i$ if there exists $i \in \{1, \cdots, k\}$ such that $p_{(i)} \leq \alpha \cdot i/k$.

Hochberg's method is uniformly more powerful than Hommel's, and is valid under the same conditions. Namely, non-negative dependence among the p-values.

2.6 Fallback Method

The **fallback method** [11] is useful when the hypotheses may be prospectively ordered. For example, consider testing $H_{[1]}$ and $H_{[2]}$. Suppose $\alpha_{[1]} = 0.01$ is allocated to $H_{[1]}$ and $\alpha_{[2]} = 0.04$ is allocated to $H_{[2]}$. If $p_{[1]} \leq \alpha_{[1]} = 0.01$, then $H_{[1]}$ is rejected, and $p_{[2]}$ is

compared with $\alpha_{[1]} + \alpha_{[2]} = 0.05$. If $p_{[1]} > 0.01$, then $H_{[1]}$ is not rejected, and $p_{[2]}$ is compared with $\alpha_{[2]} = 0.04$.

If the hypotheses are correctly ordered, meaning in order of decreasing power to reject, then the fallback method is more powerful than any procedure discussed so far, and requires no assumptions beyond those of Bonferroni's method. Any ordering of the hypotheses and any allocation of the α is admissible. However, the hypothesis ordering and α allocation must be specified *a priori*.

2.7 Gatekeeping Method

Like the fallback method, **gatekeeping** [3] is applicable when hypotheses exhibit a hierarchical ordering. Suppose $H_{[1]}$ is the primary hypothesis, and $\{H_{[21]}, H_{[22]}, H_{[23]}\}$ is a collection of three secondary hypotheses. In gatekeeping, $p_{[1]}$ is first compared with α. If $p_{[1]} > \alpha$, then $H_{[1]}$ is **retained**, and the procedure stops; the secondary hypotheses are never considered. However, if $p_{[1]} \leq \alpha$, then $H_{[1]}$ is rejected, and the set $\{H_{[21]}, H_{[22]}, H_{[23]}\}$ is tested at level α. The secondary hypotheses may be tested in *series*, for example using gatekeeping, or in *parallel*, with appropriate adjustment for $m = 3$ hypothesis tests.

Relative to fallback, an advantage of gatekeeping is that the full α is available for testing the primary hypotheses. The procedure extends to multiple primary hypotheses, tested in series or in parallel. When the primary hypotheses are tested in parallel, the α of any rejected primary hypothesis becomes available for testing the secondary hypotheses. The potential disadvantage of gatekeeping is that, if no primary hypotheses are rejected, then no α remains for testing the secondary hypotheses. In contrast, the fallback method reserves α for testing the secondary hypotheses, even if no primary hypotheses are rejected. Like the fallback method, gatekeeping requires no assumptions beyond those of Bonferroni's method, however the hypothesis ordering and α allocation must be specified *a priori*.

False Discovery Rate Control

3.1 Definition

Definition 3.1.1. Suppose that, after examining m hypotheses, $R \leq m$ have been rejected, but $V \leq R$ were in fact null. The *false discovery proportion* is defined as:

$$Q = \begin{cases} V/R, & R > 0, \\ 0, & R = 0. \end{cases}$$

Thus, Q is the proportion of hypothesis incorrectly rejected, if at least 1 hypothesis was rejected, and is zero otherwise. The **false discovery rate** (FDR) is the expected value of the false discovery proportion:

$$\text{FDR} \equiv \mathbb{E}(Q) = \mathbb{E}\left(\frac{V}{R}\Big| R > 0\right)\mathbb{P}(R > 0) \tag{3.1.1}$$

■

In contrast, FWER is the probability that at least 1 hypothesis was incorrectly rejected:

$$\text{FWER} = \mathbb{P}(V > 0). \tag{3.1.2}$$

Since the false discovery proportion is bounded above by an indicator that some hypothesis was incorrectly rejected $Q \leq \mathbb{I}(V > 0)$, the FDR $\leq$ FWER. Consequently, any procedure that controls the FWER automatically controls the FDR, but not conversely.

3.2 Benjamini-Hochberg

Benjamini-Hochberg (BH) [1] controls the FDR (3.1.1) rather than the FWER (3.1.2). Like Hochberg's method, BH is a step-up procedure in which the p-values are sorted in ascending order, then examined in increasing order of significance. Like Simes' method, the threshold with which $p_{(k)}$ is compared is $\alpha \cdot k/m$. However, in contrast to Simes' method, BH does **not** assess the intersection hypothesis.

If $p_{(m)} \leq \alpha$, $H_{(m)}$ is rejected and the procedure stops. Otherwise, $H_{(m)}$ is retained and the next largest p-value $p_{(m-1)}$ is compared with $\alpha(m-1)/m$. If $p_{(m-1)} \leq \alpha(m-1)/m$, then $H_{(m-1)}$ is rejected and the procedure stops. Else, $H_{(m-1)}$ is retained, and $p_{(m-2)}$ is compared with $\alpha(m-2)/m$. The procedure continues until one of the ordered p-values falls below its threshold, or all null hypotheses have been retained. If the procedure stops at $p_{(k)} \leq \alpha \cdot k/m$, then the hypotheses $H_{(1)}, \cdots, H_{(k)}$ are *rejected*, and hypotheses $H_{(k+1)}, \cdots, H_{(m)}$ are *retained*.

Although often described as more powerful than methods that control the FWER, BH is not truly comparable since it controls the FDR, which is a less stringent criterion.BH assumes the p-values exhibit non-negative dependence. The Benjamini–Yekutieli [2] (BY) procedure extends BH to allow for arbitrary dependence among the p-values. BY proceeds as in BH, save $p_{(k)}$ is compared with $\alpha \cdot k / \{m \cdot \gamma(m)\}$, where $\gamma(m) = \sum_{j=1}^{m} 1/j$, which is a tighter threshold. For additional discussion of the FDR, see [4]

References

[1] B Benjamini and Y Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to MultipleTesting". In: *JRSSB* 57.1 (1995), pp. 289–300. URL: https://www.jstor.org/stable/2346101.

[2] Y Benjamini and D Yekutieli. "The Control of the False Discovery Rate in Multiple Testing under Dependency". In: *Annals of Statistics* 29.4 (2001), pp. 1165–1188. URL: https://projecteuclid.org/euclid.aos/1013699998.

[3] A Dmitrienko, WW Offen, and PH Westfall. "Gatekeeping strategies for clinical trials that do not require all primary effects to be significant". In: *Statistics in Medicine* 22 (2003), pp. 2387–2400. URL: https://doi.org/10.1002/sim.1526.

[4] B Efron. *Large-Scale Inference*. Cambridge University Press, 2010. URL: https://doi.org/10.1017/CBO9780511761362.

[5] U.S. Food and Drug Administration. *Multiple Endpoints in Clinical Trials Guidance for Industry*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials-guidance-industry.

[6] Y Hochberg. "A sharper Bonferroni procedure for multiple tests of significance". In: *Biometrika* 75.4 (1988), pp. 800–802. URL: https://www.jstor.org/stable/2336325.

[7] S Holm. "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scandinavian Journal of Statistics* 6 (1979), pp. 65–70. URL: https://www.jstor.org/stable/4615733.

[8] G Hommel. "A Stagewise Rejective Multiple Test Procedure Based on a Modified Bonferroni Test". In: *Biometrika,* 75.2 (1988), pp. 383–386. URL: https://www.jstor.org/stable/2336190.

[9] T Sellke, MJ Bayarri, and JO Berger. "Calibration of p Values for Testing Precise Null Hypotheses". In: *The American Statistician* 55.1 (2001), pp. 62–71. URL: https://doi.org/10.1198/000313001300339950.

[10] RJ Simes. "An Improved Bonferroni Procedure for Multiple Tests of Significance". In: *Biometrika* 73.3 (1986), pp. 751–754. URL: https://doi.org/10.1093/biomet/73.3.751.

[11] BL Wiens. "A Fixed Sequence Bonferroni Procedure for Testing Multiple Endpoints". In: *Pharmaceutical Statistics* 2 (2003), pp. 211–215. URL: https://doi.org/10.1002/pst.64.