note0801

Daniel Sorensen

August 19, 2024

This note offers first, a visual representation of the algorithm that computes false discovery rates (Benjamini and Hochberg, 1995) and secondly, an accessible proof of the basic result:

$$\begin{aligned} \text{FDR} &= \text{E}\left(\frac{V}{R} I(R>0)\right) \\ &= \frac{m_0}{m} q \leq q. \end{aligned}$$

Visual insight of the workings of the FDR-BH algorithm

Visual insight into the workings of the false discovery rate (FDR-BH) algorithm can be gained as follows. Consider an experiment that outputs m $p-$values; unknown to the analyst, a proportion π_0 are associated with the null hypothesis and a proportion $\pi_1 = (1-\pi_0)$ with the non-null hypothesis. The distribution of $p-$values originating from the true null hypothesis is uniform $\mathcal{U}(0,1)$, so that for $p-$value i, $\Pr(P_i \leq p_i | H_i = \text{true}) = \int_0^{p_i} du = p_i$. The distribution of $p-$values originating from the false null hypothesis tends to concentrate closer to 0 and is non-uniform; for $p-$value i, $\Pr(P_i \leq p_i | H_i = \text{false}) = H\left(p_i\right)$. Marginally, the $p-$values are draws from the mixture distribution

$$\begin{aligned} \Pr\left(P_i \leq p_i\right) &= F\left(p_i\right) \\ &= \pi_1 \Pr(P_i \leq p_i | H_i = \text{false}) + \pi_0 \Pr(P_i \leq p_i | H_i = \text{true}) \\ &= \pi_1 H\left(p_i\right) + \pi_0 p_i. \end{aligned}$$

The sorted $p-$values are an approximation to the inverse distribution function $\left(F\left(p_i\right)\right)^{-1}$. One can plot these sorted $p-$values against the indices (i/m), $i = 1, 2, \ldots, m$, and overlay a straight line with intercept equal to zero and slope equal to q. This represents a rejection line. Starting in the top-right corner (highest $p-$value) move towards the bottom-left corner and identify the first $p-$value that falls below the rejection line. Then reject all hypotheses whose $p-$values are smaller than or equal to this $p-$value (regardless of whether any of those $p-$values happen to be above the rejection line). The expected proportion of false positives among the rejected hypotheses (the discovery set) is smaller than or equal to q.

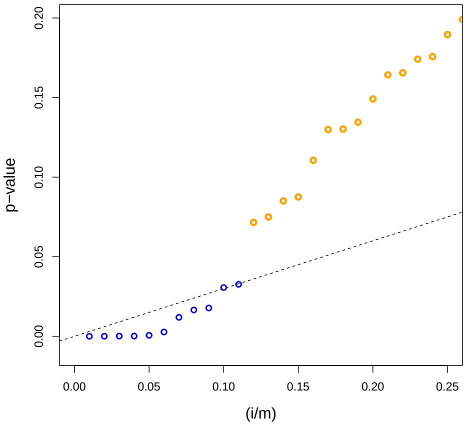

Figure 1: Plot of a subset of ordered $p-$values from an experiment involving 100 identical hypothesis tests against the scaled index (i/m) , $m = 100$. There are 90 true null hypotheses and 10 non-null. The dashed line represents the line of rejection (intercept zero and slope $q = 0.3$). Orange circles represent ordered $p-$values from the non-rejected hypotheses; blue circles represent ordered $p-$values from the 11 rejected hypotheses and constitute the discovery set.

Figure 1 provides the visual representation of the FDR-BH algorithm for an experiment involving $i = 1, 2, \ldots, m$, identical hypothesis tests, $m = 100$, $q = 0.30$. The first blue circle, starting from the right of the figure, corresponds to the first $p-$value that is below the rejection line $(i/m)q$. All hypotheses beyond and including this point are rejected and constitute the discovery set, despite the fact that the second blue circle starting from the right, corresponding to $i = 10$, happens to be larger than $(i/m)q$: $0.031 \,\text{vs}\, 0.030$.

Proof of Benjamini and Hochberg's Theorem

Benjamini and Hochberg (1995) prove the following theorem: for continuously distributed test statistics, the false discovery rate, defined as the expectation of the false discovery proportion, is given by

$$\begin{aligned} \text{FDR} &= \text{E}\left(\frac{V}{R} I(R > 0)\right) \\ &= \frac{m_0}{m} q \leq q \end{aligned}$$

where the expectation is taken over the true distribution of the data. This section provides a proof of this result that differs from the original proof of Benjamini and Hochberg (1995) and is more aligned with the proof in Benjamini and Yekutieli (2001), for the case of jointly independent test statistics. Other sources of the proof that follows are a set of unpublished notes by Jens Ledet Jensen (2023): ”Statistical Inference for High Dimensional data”, Mathematical Institute, Aarhus University, and Ewens and Grant (2005), where more mathematical rigour can be found.

Consider the set of m $p-$values associated with the m hypotheses. An observed $p-$value is p_i and the random variable is P_i. The observed ordered $p-$values are here denoted $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$. Let $(i/m)\,q = q_i$. Moving towards smaller $p-$values, $i = m, m-1, \ldots, 1$; the first i where $p_{(i)} \leq q_i$ results in a discovery set $R = i$, i hypotheses are rejected and $m - i$ hypotheses with $p_{(j)} > q_j$, $j = i+1, \ldots, m$ are accepted.

Let the set M_0 contain the m_0 true null hypotheses, and let the set M_1 contain the $m - m_0$ false null hypotheses. Thus H_i, $i \in M_0$, denotes the ith true null hypothesis and H_i, $i \in M_1$, denotes the ith false null.

The number of true null hypotheses rejected in error (number of false discoveries) is V, and the total number of rejections (the size of the discovery set) is R.

With this notation in place, the FDR can be written

$$\begin{aligned}
\mathrm{FDR} &= \mathrm{E}\left(\frac{V}{R}\, I(R > 0)\right) = \sum_{j=1}^{m} \frac{1}{j}\, \mathrm{E}\left[V\, I(R = j)\right] \\
&= \sum_{j=1}^{m} \frac{1}{j}\, \mathrm{E}\left[\sum_{i \in M_0} I(P_i \leq q_j)\, I(R = j)\right] \\
&= \sum_{i \in M_0} \sum_{j=1}^{m} \frac{1}{j}\, \mathrm{E}\left[I(P_i \leq q_j)\, I(R = j)\right], \qquad (1)
\end{aligned}$$

where $(P_i \leq q_j, R = j)$ is the joint event that H_i is rejected and that in total, j null hypotheses are rejected. This joint event can be rewritten as $\left(P_i \leq q_j, C_{j-1}^{(i)}\right)$, where $C_{j-1}^{(i)}$ is the event that, of the remaining $m - 1$ hypotheses other than H_i, exactly $j - 1$ are rejected. In the case of independent test statistics, $(P_i \leq q_j)$ and $C_{j-1}^{(i)}$ are independent events. Therefore

$$\begin{aligned}
\mathrm{E}\left[I(P_i \leq q_j)\, I\left(C_{j-1}^{(i)}\right)\right] &= \Pr\left(P_i \leq q_j\right) \Pr\left(C_{j-1}^{(i)}\right), \quad i \in M_0, \\
&= q_j \Pr\left(C_{j-1}^{(i)}\right) \\
&= \frac{j}{m} q \Pr\left(C_{j-1}^{(i)}\right). \qquad (2)
\end{aligned}$$

The second line on the right hand side follows because P_i, $i \in M_0$, is assumed to have a uniform distribution in $[0, 1]$, and the third uses the definition $(j/m)\,q = q_j$. Substituting in (1) yields the final result:

$$\begin{aligned} \text{FDR} &= \sum_{i \in M_0} \sum_{j=1}^{m} \frac{1}{j} \frac{j}{m} q \Pr\left(C_{j-1}^{(i)}\right) \\ &= \sum_{i \in M_0} \frac{q}{m} \sum_{j=1}^{m} \Pr\left(C_{j-1}^{(i)}\right) \\ &= \sum_{i \in M_0} \frac{q}{m} = \frac{m_0}{m} q. \end{aligned} \tag{3}$$

The third line uses the law of total probability $\sum_{j=1}^{m} \Pr\left(C_{j-1}^{(i)}\right) = 1$.

Benjamini and Yekutieli (2001) show that result (3) holds in cases where there is a certain type of correlation among the test statistics (positive regression dependence). This requires replacing the constants $(i/m)\,q$ by

$$\frac{iq}{m \sum_{j=1}^{m} \frac{1}{j}}.$$

NOTE A connection between the proof of the theorem and the implementation of the FDR algorithm calls for taking a closer look into the event $C_{j-1}^{(i)}$. A little extra notation is needed. First, recall that when $R = j$, j hypotheses with ordered $p-$values $p_{(j)} \leq q_j$ are rejected. Consider removing the ith $p-$value from the ordered $p-$values $P_{(1)}, P_{(2)}, \ldots, P_{(m)}$. Denote the remaining $m-1$ ordered $p-$values as $P_{(1)}^{-i}, P_{(2)}^{-i}, \ldots, P_{(m-1)}^{-i}$. For example, if $i = 3$, $P_{(1)}^{-3} = P_{(1)}, P_{(2)}^{-3} = P_{(2)}, P_{(3)}^{-3} = P_{(4)}, \ldots, P_{(m-1)}^{-3} = P_{(m)}$. With this notation the joint event $(P_i \leq q_j, R = j)$ can be expressed as $\left(P_i \leq q_j, C_{j-1}^{(i)}\right)$, equal to

$$\begin{aligned} & \left(P_i \leq q_j, P_{(j)} \leq q_j, P_{(j+1)} > q_{j+1}, \ldots, P_{(m)} > q_m\right) \\ = & \left(P_i \leq q_j, P_{(j-1)}^{-i} \leq q_j, P_{(j)}^{-i} > q_{j+1}, \ldots, P_{(m-1)}^{-i} > q_m\right), \end{aligned} \tag{4}$$

that holds for $i < j$ and also for $i = j$, in which case $P_{(j-1)}^{-i} = P_{(j-1)} \leq P_{(j)} \leq q_j$. Expression (4) reveals that at the cut-off j with the ordered $p-$value $P_{(j-1)}^{-i} \leq q_j$, $j-1$ hypotheses are rejected and $H_{j+1}, H_{j+2} \ldots, H_m$ are accepted; this is precisely how the algorithm is implemented. One can also note that

$$C_{j-1}^{(i)} = \left(P_{(j-1)}^{-i} \leq q_j, P_{(j)}^{-i} > q_{j+1}, \ldots, P_{(m-1)}^{-i} > q_m\right)$$

depends only on $P_l, l \neq i$; therefore with independent test statistics $(P_i \leq q_j)$ and $C_{j-1}^{(i)}$ are independent events.

References

Benjamini, Y. and Y. Hochberg (1995). Controlling false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, Methodological 57*, 289–300.

Benjamini, Y. and Y. Yekutieli (2001). The control of false discovery rate in multiple testing under dependency. *Annals of Statistics 29*, 1165–1188.

Ewens, W. J. and G. R. Grant (2005). *Statistical Methods in Bioinformatics*. Springer.