

BH FDR control simulation

2022-04-15

Data simulation

(example adapted from a youtube video by Josh Starmer)

We construct a simple data that is comprised of two groups of subjects, each of size 25. Subjects have data from $m = 10,000$ variables, for example gene expression levels and we want to compare expression levels of all genes between the groups.

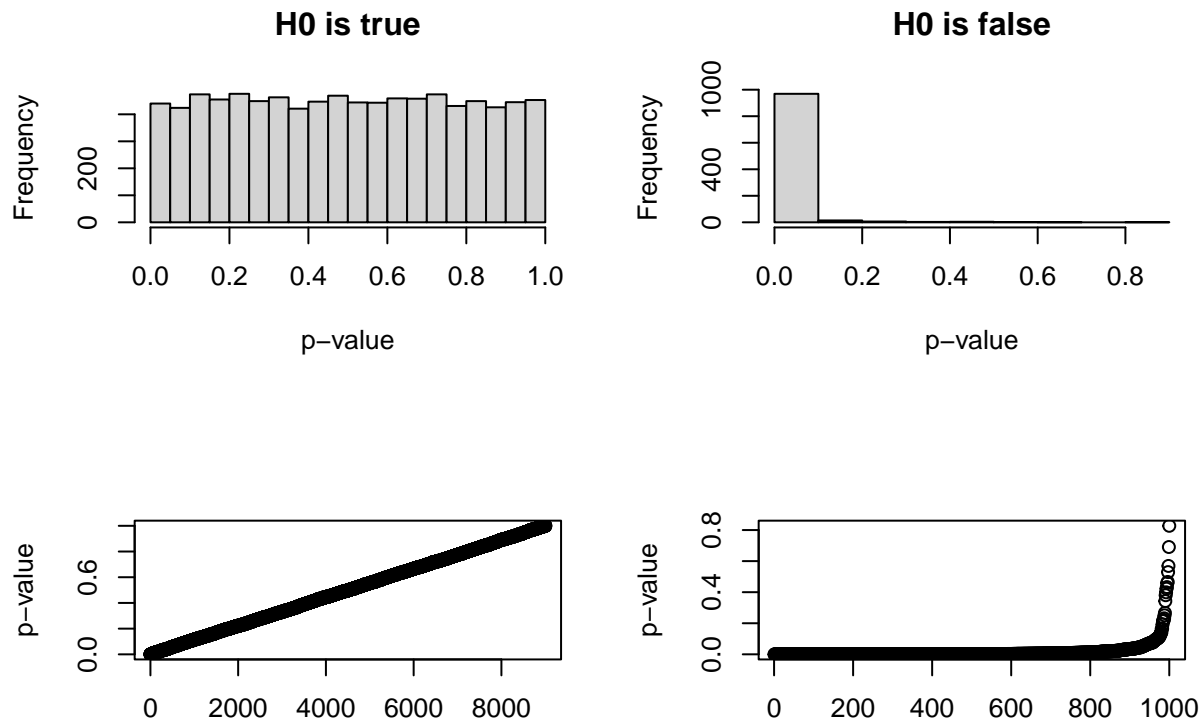
To demonstrate how BH controls the FDR, we simulate the gene expression data as follows:

For $M_0 = 9000$ genes, the expression levels are sampled from $N(0, 1)$ for the entire set of 50 subjects.

For $M - M_0 = 1000$ genes, the expression levels for one group are sampled from $N(0, 1)$ and for the second group from $N(1, 1)$

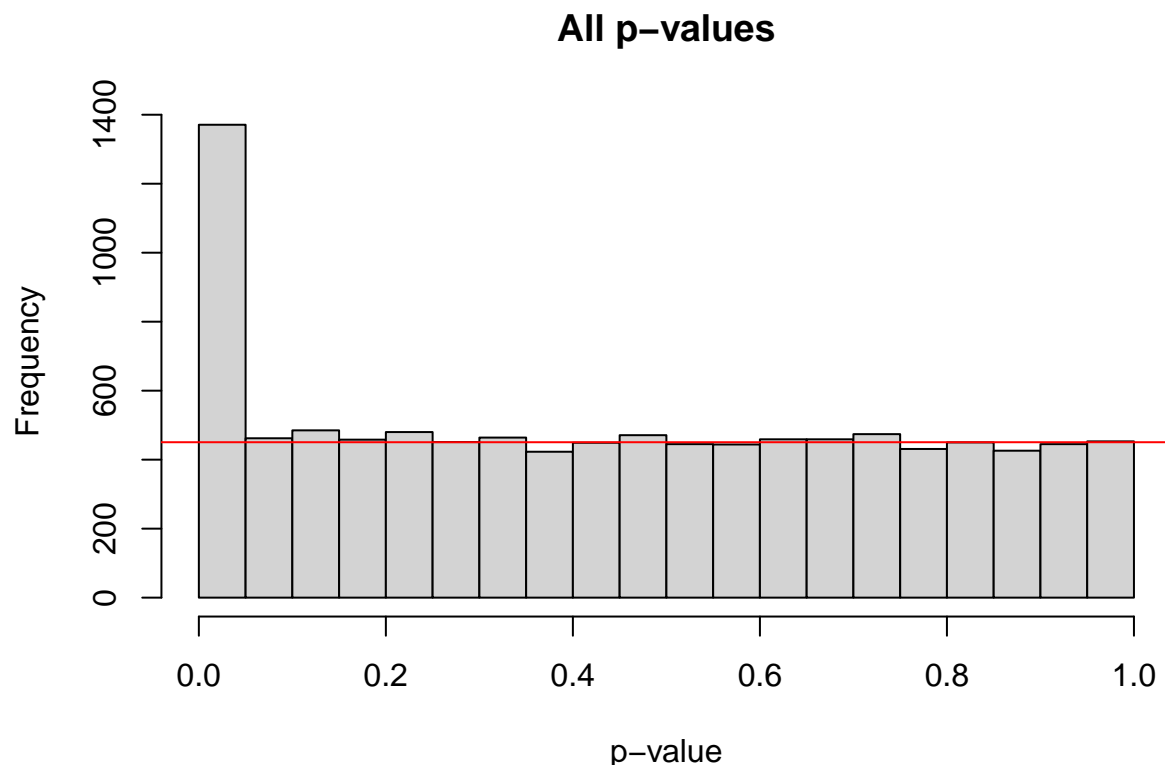
The p-values are then computed using two-sample t-tests.

To demonstrate the distributions of the test results for true nulls and false nulls we can draw histograms of the p-values



The histogram on the right shows only the p-values for which H_0 is true. The p-values are uniformly distributed between 0 and 1.

The histogram on the left shows only the p-values for which H_0 is false. As expected most of the 1000 cases are significant (which is how we designed them to be)

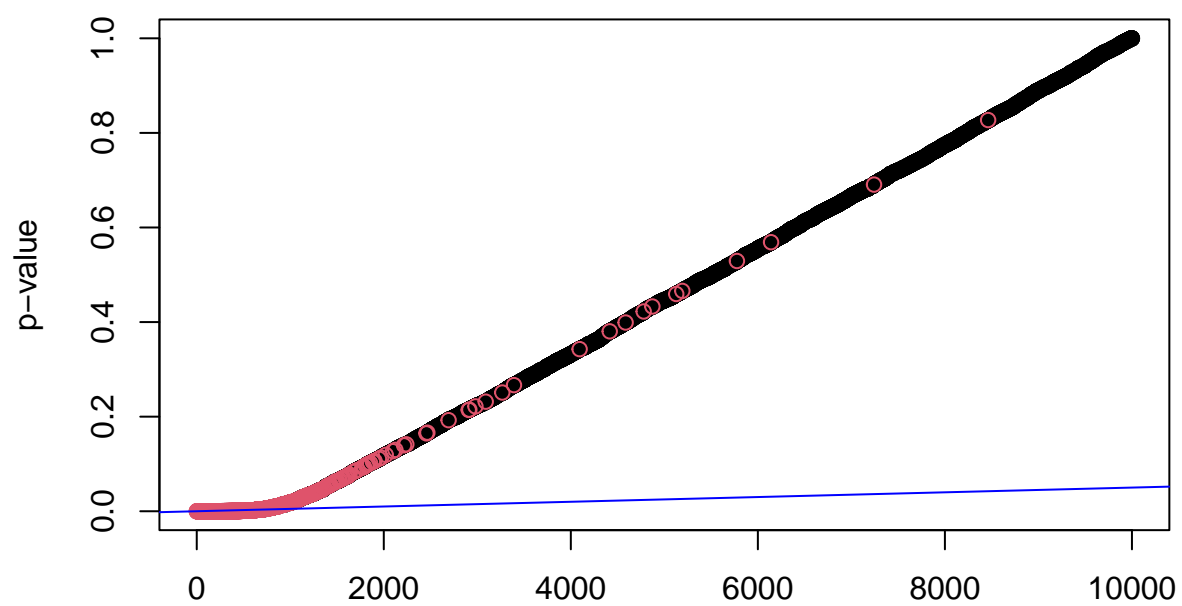


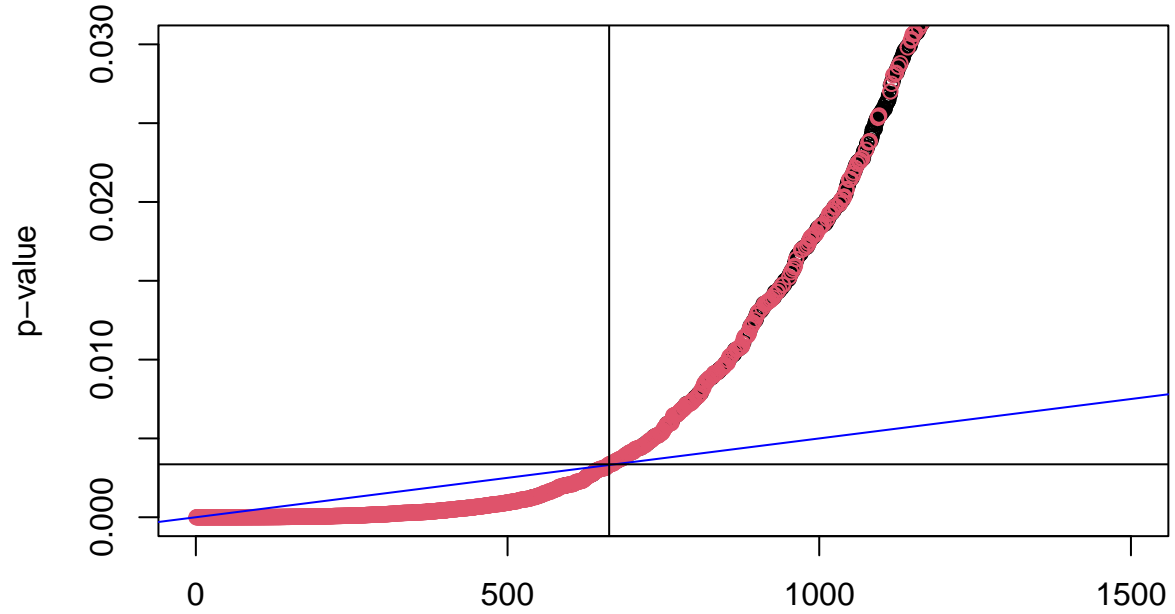
If we look at the distribution of all p-values, we can “eyeball” the total number of true positive tests (tests that should be rejected) as the frequency count on the highest bar minus the value of the red line (which is the average count of all the p-values from the second bar to the last (i.e. all p-values that are greater than 0.05))

Using the simulated data we can calculate the exact value for each of the cells in the 2x2 table from the paper and calculate the FDR.

	Non-significant	Significant
Null True	8560	440
Null False	69	931

The BH method is intended to control the FDR at a level of $\frac{\alpha}{m}k$. If we plot all p-values and draw the line $y = \frac{\alpha}{m}k$ we are essentially visualizing the procedure. The point where this line intersects with the p-values is where the FDR is controlled at the desired level.





We demonstrated that the distribution of p-values for which H_0 is true is $U[0, 1]$. Therefore the expected number of p-values lying in any interval $[0, k]$ is $m_0 k$. From here, with simple algebra we can demonstrate that the horizontal line is equal to $\frac{\alpha k}{m}$ and the expected number of p-values among these for which H_0 is true (i.e. false discoveries) is less than or equal to $m_0 h = m_0 \frac{\alpha k}{m}$, therefore:

$$\text{FDR} \leq m_0 \frac{\alpha k}{m} = \frac{\alpha m_0}{m} \leq \alpha$$

In this instance the k corresponding to the intersection point is 662 and the total false positive in this case are 29 and the FDR is therefore 0.044

To show that the Expected value is indeed control at that level, we run the simulate data above 100 times and the results for the FDR controlled p-values are presented in the histogram below:

Simulated Distribution of FDRs

