Statistics. HW2. False Discoveries Problem.

Changalidi Anton

(Dated: March 22, 2024)

Part 1 & 2. Problem statement

The problem addressed in the paper (Candès et al., 2018) is a classic statistical problem that we have also mentioned many times in class: the False Discovery Rate (FDR) control.

Modelling

First, let's illustrate it using Figure 1 (same as Figure one in Candès et al., 2018). As in the original article, the author simulated 10^4 independent design matrices (sample size $n = 500$, amount of predictors $p = 200$) and binary responses from a logistic regression for the following two settings:

- one where the predictors $X_1, \ldots, X_p$ follow an $AR(1)$ time series with an autocorrelation coefficient of 0.5 and the response Y is Bernoulli distributed with probability 0.5 (Figure 1A);
- another setting with the same predictor configuration but the response is Bernoulli distributed with a probability determined by a logit function involving a sum of the predictors $X_2, \ldots, X_{21}$ (Figure 1B).

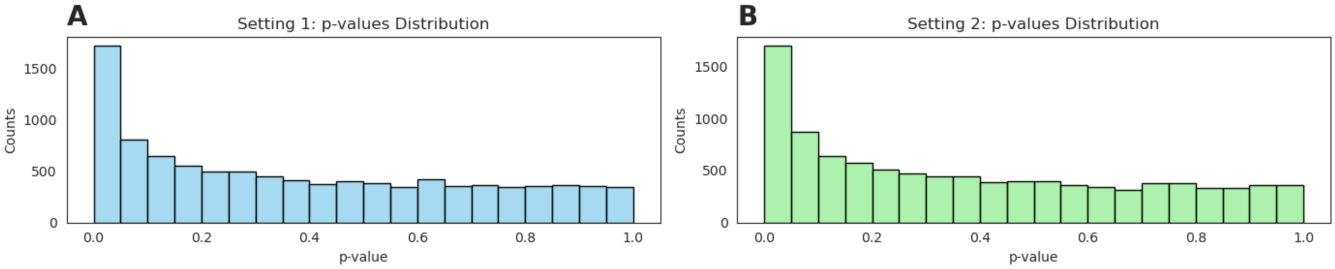

FIG. 1: The distributions of null logistic regression p-values under two simulation settings were obtained by building GLM on these X and Y.

For both cases, p-values were obtained using a Generalized linear model. For the Python code used to generate datasets and images in this and subsequent chapters, please refer to the "Data and Code Availability" section.

Observation

Note, that in both cases Y is not dependent on X_1 (however, due to autoregression in the second case there is an influence of X_1 on all subsequent X's: $X_2, \ldots, X_{21}$). The null hypothesis here states that a particular predictor (X_1) has no effect on the outcome. Under the null hypothesis, p-values should be uniformly distributed between 0 and 1, meaning any small segment of this range should contain a proportionate number of the total p-values.

However, in both settings we observe inflation of probabilities:

- for first setting: for very small p-values (specifically at the 5% percentile), the probability of obtaining such a p-value (or smaller) under the null hypothesis is significantly higher than expected: in this setting, there should NOT be any dependency between X_1 and Y, however, there is. This indicates a high chance of falsely discovering an effect (false positives).
- for the second setting: the exact distribution of null p-values (those that should not indicate any significant effect) is influenced by the actual values of unknown coefficients ($\beta_2, \ldots, \beta_{21}$), however (in reality) they depend on β_0 themselves (broken causation). This means that the likelihood of observing small p-values not only is inflated but also varies significantly depending on the true effects of the predictors.

Therefore the method used to calculate p-values might not be reliable for determining significance in these high-dimensional logistic regression scenarios, leading to many false discoveries (first case). What is more, p-value distribution can depend in general on unknown problem parameters, causing difficulties in interpretation.

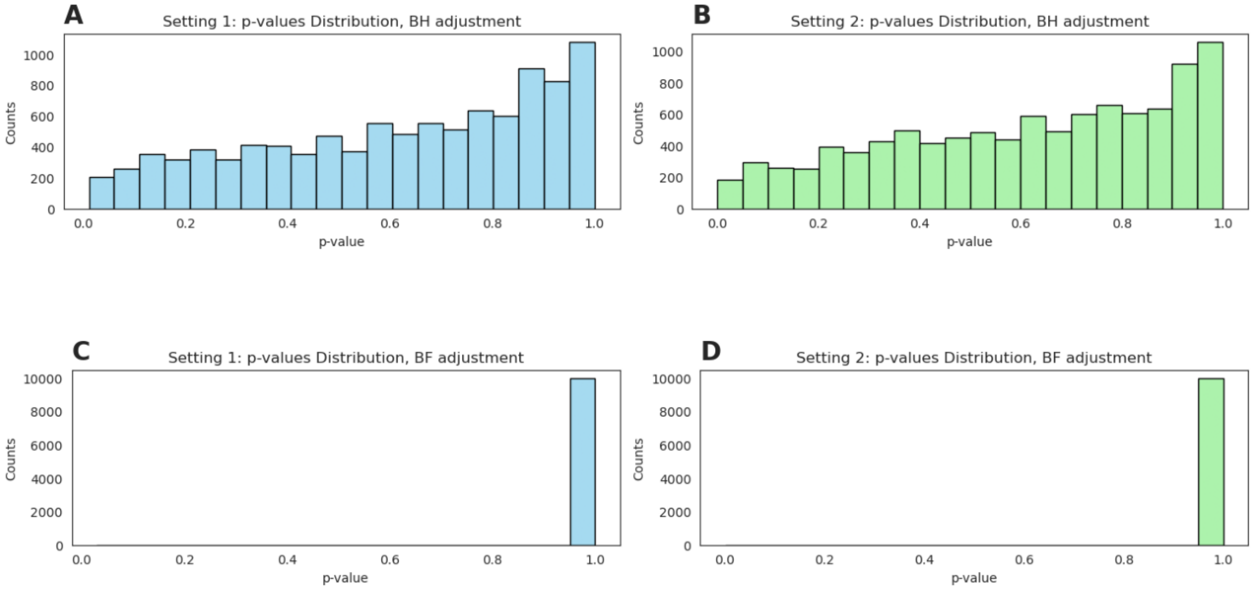

FIG. 2: Distributions of the p-values from the Figure 1 adjusted with BH (A, B) and BF (C, D).

Part 3 & 4. How to fix that?

To be honest, I (the researcher) did not understand the mathematical details of the method implementation in the article well (although I wanted to, but it was a very crazy week). Therefore I decided to try simpler methods: FDR multiple testing adjustment: Bonferroni (BF) (Bonferroni, 1936) and Benjamini-Hochberg (BH) (Benjamini and Hochberg, 1995). They have restrictions: e.g. BF controls the FDR under p-value independence and BF is very conservative, however, it works quite well for most of the cases. Despite Candès et al. , 2018 (and a lot of other works) proposed algorithms that work better and less constraining, for our case these 2 methods should work. The core ideas of these adjustments are:

- The Bonferroni adjustment is a conservative method to control the family-wise error rate when performing multiple comparisons. The idea is to reduce the chance of obtaining false-positive results (Type I errors) by dividing the desired significance level (α) by the number of tests (m). The algorithm simply involves comparing each individual p-value against this adjusted threshold to decide on rejecting the null hypothesis.
- The Benjamini-Hochberg (BH) adjustment aims to control the false discovery rate (FDR), which is the expected proportion of false positives among the rejected hypotheses. The algorithm ranks the individual p-values in ascending order, then finds the largest rank k for which the kth p-value is less than or equal to $(k/m) \cdot \alpha$, where m is the total number of tests and α is the desired FDR level. All hypotheses corresponding to p-values up to the kth are rejected.

Figure 2 shows the histogram for the same p-values, but adjusted using BH (Figure 2A,B) and using BF (Figure 2C,D). Adjusted p-values are not biased towards zero:

- BF adjustment is very conservative and can increase the likelihood of Type II errors (failing to reject a false null hypothesis). p-values here are biased to one, which indicates that the tests are far from showing any significant differences after correcting for the risk of making Type I errors (false positives) due to multiple comparisons.
- BH adjustment controls FDR less strictly: it is not uniform, however, it is much smoother, without strong bias.

The choice ultimately rests with the researcher. Going forward, we will employ the Benjamini-Hochberg (BH) method.

Part 5. Test on microarray dataset

The Kaggle dataset (Crawford, 2017) focuses on gene expression levels from 7129 genes in 72 patients diagnosed with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL), derived from bone marrow and peripheral blood samples. This data comes from a study by Golub et al., 1999, illustrating the potential of gene expression monitoring via DNA microarray for cancer classification. The dataset includes initial and independent datasets with 38 and 34 samples, respectively, aiming to classify AML and ALL patients. In our study, we do not engage in the classification of samples. Instead, our approach focuses on employing the t-test to pinpoint genes whose expression levels significantly vary among different patient groups. The resulting p-values from the t-test will further be considered.

The dataset underwent preprocessing including merging train and test (since we do not need to test something here), and checking for null values. Last but not least, recognizing the importance of normalization for the performance of statistical tests (as many machine learning models), the data underwent standardization scaling.

For the analysis, a straightforward but illustrative methodology is adopted. T-tests were used to identify key genes that significantly contribute to distinguishing between AML and ALL. Since t-test was conducted on every gene, the need to adjust for multiple comparisons arises: BH correction, which was described earlier and that outperformed BF was used. This ensures the reliability of our findings in identifying significant genes. The comparison is essential for refining the selection of gene features that are truly significant in differentiating between AML and ALL. The distribution of both p-values is shown in Figure 3. After adjustment bias towards zero is removed (bias towards one in the case of p-value does not matter). The initial total number of genes identified is 7129. Of these, 1144 are significant at a classical p-value cutoff threshold of 0.05. However, this number is too large to draw meaningful biological conclusions. In contrast, only 34 genes remain significant when applying an adjusted p-value cutoff threshold of 0.05, which appears more realistic and reliable. These genes are potentially of significant importance.

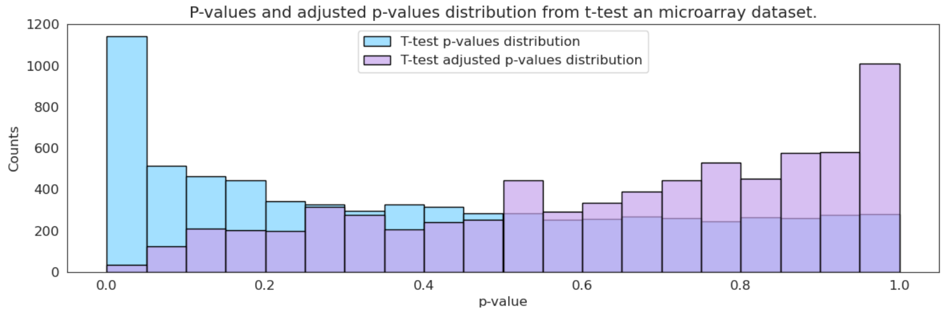

FIG. 3: Distributions of the p-values and adjusted p-values from t-test comparing AML and ALL samples from Crawford, 2017 dataset.

Data and Code availability

All code pertinent to the results presented in this work is available at:
`https://github.com/TohaRhymes/stat_um_24spring/tree/main/hw2`

References

[1] Candès, E., Fan, Y., Janson, L., & Lv, J. (2018). Panning for Gold: Model-X Knockoffs for High-dimensional Controlled Variable Selection. arXiv preprint arXiv:1610.02351. https://arxiv.org/abs/1610.02351

[2] Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289–300. http://www.jstor.org/stable/2346101

[3] Bonferroni, C. E., Teoria statistica delle classi e calcolo delle probabilità, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 1936 https://www.kaggle.com/datasets/quantbruce/real-estate-price-prediction?resource=download

[4] Crawford, C. (2017, August 8). Gene expression dataset (Golub et al.). Kaggle. https://www.kaggle.com/datasets/crawford/gene-expression/

[5] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: Class Discovery and class prediction by Gene Expression Monitoring. Science, 286(5439), 531–537. https://doi.org/10.1126/science.286.5439.531