677 Final

Zengqi Chen

2024-05-05

Main Points

Chapter 15 of "Computer Age Statistical Inference" by Bradley Efron and Trevor Hastie talks about the topics of large-scale hypothesis testing and the control of false discovery rates (FDRs), which are particularly pertinent in the era of big data. This chapter is crucial for understanding the statistical tools needed to handle complex datasets where multiple hypotheses are tested simultaneously.

Large-Scale Hypothesis Testing

In the context of modern statistical analysis, particularly in fields like genomics, neuroimaging, and other areas where large datasets are common, the need to perform multiple hypothesis tests simultaneously is a significant challenge. The traditional methods of hypothesis testing, which control for the probability of making a Type I error (false positive) in a single test, do not scale well when thousands or even millions of tests are conducted because they lead to a high overall chance of making one or more false discoveries.

Efron and Hastie discuss the adjustment of these methods to handle multiple comparisons without inflating the overall error rate. They describe the application of techniques such as the Bonferroni correction and its more sophisticated alternatives, which adjust the threshold for statistical significance according to the number of tests performed.

False Discovery Rates (FDRs)

The concept of False Discovery Rate (FDR), introduced by Benjamini and Hochberg (1995), represents a paradigm shift in handling multiple comparisons. FDR is defined as the expected proportion of false discoveries among the rejected hypotheses. This is mathematically represented as:

$$\text{FDR} = \mathbb{E}\left(\frac{V}{R}\right)$$

where V is the number of false positives, and R is the total number of rejections. If no hypotheses are rejected, R is set to 1 to ensure the formula is well-defined.

Empirical Bayes Methods for Large-Scale Testing

Empirical Bayes methods are particularly highlighted in this chapter for their effectiveness in estimating the parameters used in computing the FDR. These methods use the observed data to estimate the distribution of the test statistics under the null hypothesis and the alternative hypothesis, allowing for a more accurate computation of FDR. The Empirical Bayes approach is often more powerful and flexible compared to traditional methods, especially when the assumptions about the distribution under the null hypothesis may not hold exactly.

Local False Discovery Rates (Local FDRs)

While FDR gives an overall measure of error among multiple tests, local false discovery rates (local FDRs) provide a measure at the individual test level. The local FDR is defined as the probability that a given

hypothesis is null given the observed test statistic:

$$\text{local FDR}(t) = \frac{\Pr(\text{Null}|T = t)}{\Pr(\text{Alternative}|T = t)}$$

where T is the test statistic. This measure helps researchers understand the reliability of individual tests within the context of a large dataset.

Choice of Null Distribution

A crucial aspect discussed is the selection of an appropriate null distribution. This selection is fundamental in calculating both FDR and local FDRs. The authors discuss various strategies to estimate this distribution from the data, which is often a critical step in ensuring the robustness of hypothesis testing results.

Computational Methods

Estimating False Discovery Rates (FDR)

One of the primary computational challenges in large-scale testing is the estimation of the False Discovery Rate (FDR), which provides a way to control the expected proportion of incorrect rejections among all rejections. The fundamental FDR computation can be formulated as follows:

Given a set of p-values $P_1, P_2, ..., P_m$ from m independent tests, the FDR can be controlled using the Benjamini-Hochberg (BH) procedure, which is designed to ensure that the expected proportion of false positives (incorrect rejections) among the rejected hypotheses does not exceed a pre-specified level α.

The BH procedure operates as follows:

1. Sort the p-values in ascending order.
2. For each p-value P_i, compute its rank i in the ordered list.
3. Find the largest k such that $P_k \leq \frac{k}{m}\alpha$, where m is the total number of hypotheses tested.

This procedure can be efficiently implemented in R using the `p.adjust` function.

```
# Generate random p-values for demonstration
set.seed(123)
p_values <- runif(100)

# Adjust p-values using the Benjamini-Hochberg method
adjusted_p_values <- p.adjust(p_values, method = "BH")

# Determine which hypotheses to reject
alpha <- 0.05

adjusted_p_values <= alpha
```

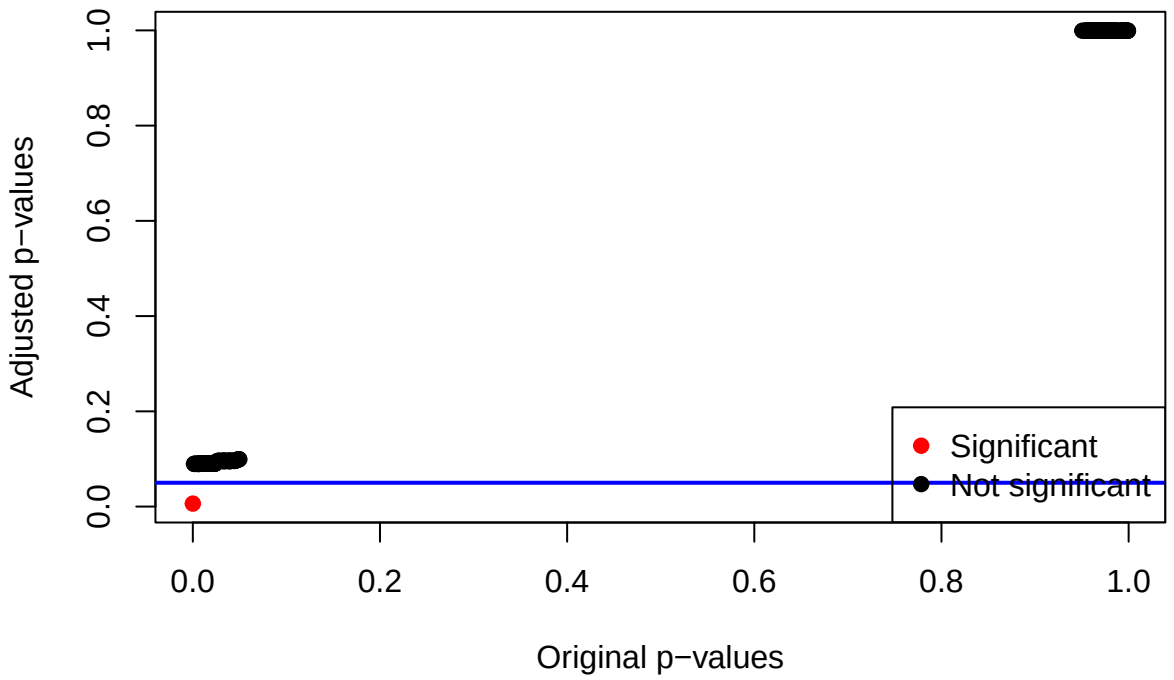
```
##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [97] FALSE FALSE FALSE FALSE
```

Visualizing FDR and p-values

```r
set.seed(123)
p_values <- c(runif(100, min = 0, max = 0.05), runif(100, min = 0.95, max = 1))

# Apply FDR correction
adjusted_p_values <- p.adjust(p_values, method = "fdr")

# Plotting
plot(p_values, adjusted_p_values, pch = 19, col = ifelse(adjusted_p_values <= 0.05, "red", "black"),
     xlab = "Original p-values", ylab = "Adjusted p-values")
abline(h = 0.05, col = "blue", lwd = 2)
legend("bottomright", legend = c("Significant", "Not significant"), pch = 19, col = c("red", "black"))
```



Simulating Data to Study FDR

```r
# Simulate some data where some hypotheses are true
true_effects <- rnorm(50, mean = 3, sd = 1)
null_effects <- rnorm(950, mean = 0, sd = 1)
effects <- c(true_effects, null_effects)
p_values <- pnorm(abs(effects), lower.tail = FALSE) * 2  # Two-tailed test

# Apply FDR correction
adjusted_p_values <- p.adjust(p_values, method = "fdr")

# Determine which hypotheses to reject
```

```r
significant_results <- adjusted_p_values <= 0.05
cat("Number of significant results:", sum(significant_results), "\n")
```

```
## Number of significant results: 12
```

Comparison of Different FDR Methods

```r
# Comparing different methods
methods <- c("holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none")
sapply(methods, function(m) {
    adj_p <- p.adjust(p_values, method = m)
    sum(adj_p <= 0.05)
})
```

```
##       holm   hochberg     hommel bonferroni         BH         BY        fdr
##          5          5          5          5         12          4         12
##       none
##         91
```

Empirical Bayes for Large-Scale Testing

Empirical Bayes methods are extensively used for estimating parameters that are subsequently used in FDR calculations. These methods involve estimating the distribution of the test statistics under the null hypothesis and combining this with the prior distribution of the parameters.

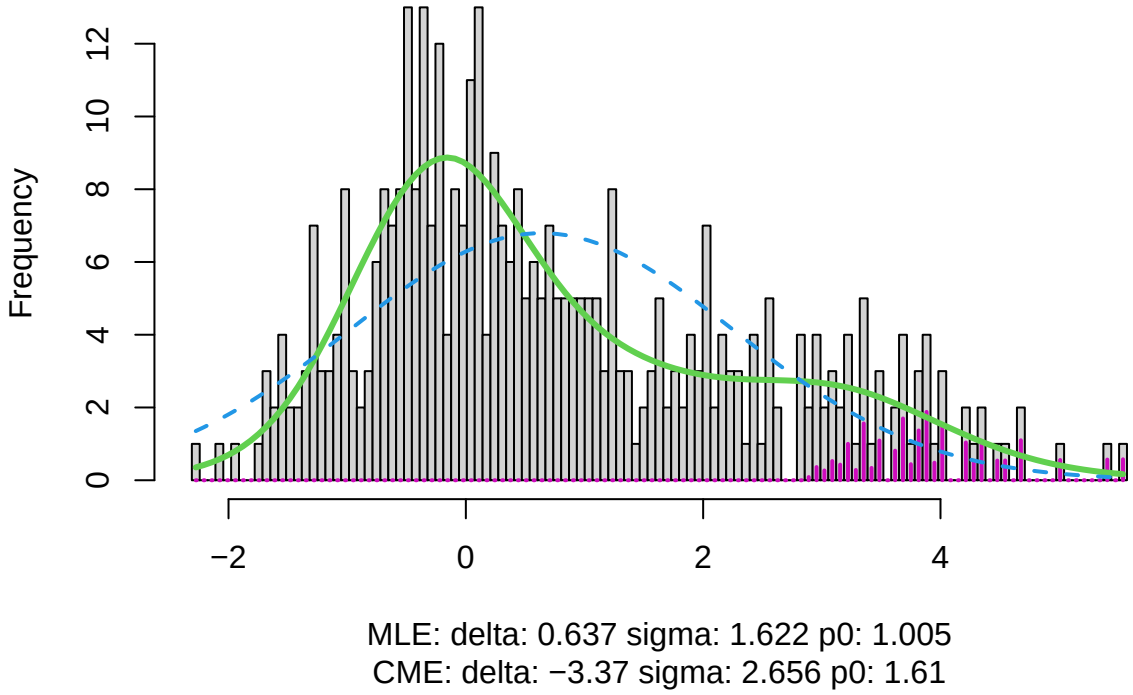
The key computational aspect here involves modeling the distribution of p-values under the null hypothesis, which is often assumed to be uniform. However, deviations from this uniform distribution in observed p-values can indicate the presence of true effects. Empirical Bayes methods adjust for such deviations by estimating the prior distribution of the test statistics.

using the `locfdr` package to computes local FDR values:

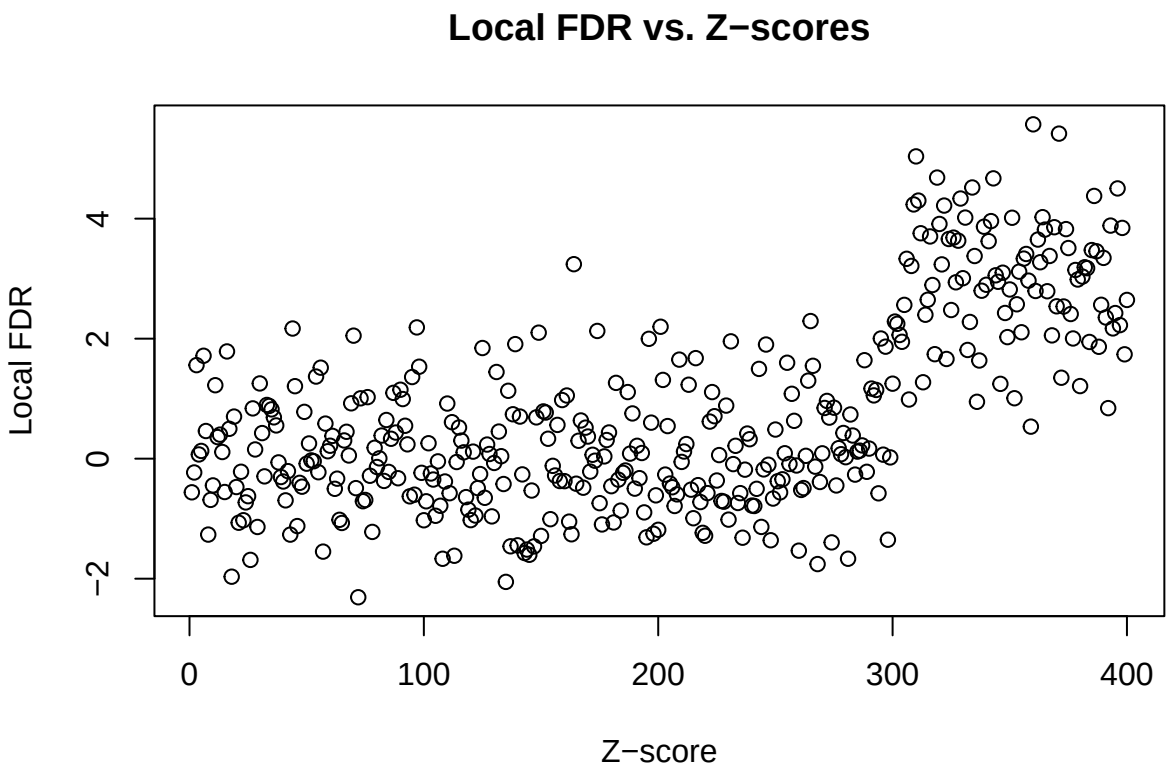
```r
library(locfdr)


# Simulate some test statistics under null and alternative hypotheses
set.seed(123)
z_scores <- c(rnorm(300), rnorm(100, mean=3))

# Compute local FDR values
fdr_values <- locfdr(z_scores)
```




```
# Plot local FDR values
plot(z_scores, fdr_values$lfdr, main="Local FDR vs. Z-scores", xlab="Z-score", ylab="Local FDR")
```



Selection of Null Distribution

Selecting an appropriate null distribution is crucial for accurately computing FDR and local FDRs. Computational methods involve using techniques like bootstrap or smoothing methods to estimate the null distribution from the data. The choice of the null distribution impacts the entire process of hypothesis testing, especially in how the p-values are interpreted.

For example, if the null distribution is incorrectly specified as too broad, then even significant deviations might be considered as noise, leading to a high number of false negatives. On the other hand, a too-narrow null distribution might lead to too many false positives.

Mathematical foundation

Probability Framework for Multiple Testing

Large-scale hypothesis testing involves handling multiple hypotheses simultaneously, which fundamentally depends on understanding the behavior of Type I (false positive) and Type II (false negative) errors across many tests. The key mathematical challenge is to control the overall error rates, which can dramatically increase with the number of hypotheses tested.

Bonferroni Correction

A basic approach to control the family-wise error rate (FWER) in multiple testing is the Bonferroni correction. The correction is based on the union bound from probability theory:

$$\text{FWER} = \Pr(\text{at least one false positive}) \leq \sum_{i=1}^{m} \Pr(\text{false positive in test } i)$$

Given individual tests are performed at significance level α, for m independent tests, the Bonferroni correction suggests testing each hypothesis at a significance level of α/m to ensure that the FWER does not exceed α. This is overly conservative, especially as m becomes large, but it provides a simple and rigorous control of Type I error.

False Discovery Rates (FDR)

The concept of FDR is a more sophisticated approach that allows a certain proportion of false positives among the rejected hypotheses, which is more appropriate for large-scale testing scenarios. Introduced by Benjamini and Hochberg, the FDR is mathematically defined and controlled as follows:

$$\text{FDR} = \mathbb{E}\left(\frac{V}{\max(R, 1)}\right)$$

where V is the number of false positives and R is the total number of rejected hypotheses. The expectation is over the random selection of hypotheses being true or false under the null hypothesis.

The FDR controlling procedure by Benjamini and Hochberg involves sorting the p-values and selecting a cutoff point. The largest k where:

$$P_k \leq \frac{k}{m}\alpha$$

This threshold ensures that the expected proportion of incorrect rejections is controlled at the level α.

Empirical Bayes and Local FDR

Empirical Bayes methods are used for estimating the distribution parameters necessary for FDR calculations. The local FDR, an extension of the FDR concept, gives the probability that a specific null hypothesis is true given the observed data:

$$\text{local FDR}(t) = \Pr(\text{Null} \mid T = t)$$

This is typically estimated using the ratio of the density of the test statistic under the null hypothesis to its density under the alternative hypothesis. These densities are often estimated using smoothing techniques or parametric models fitted to the data.

Mathematical Modeling of Null and Alternative Hypotheses

The accurate modeling of the null and alternative hypotheses is critical. The null hypothesis typically assumes no effect (e.g., zero mean difference), while the alternative suggests some non-zero effect. The distributions of test statistics under both hypotheses need to be well approximated to accurately compute p-values and, subsequently, FDRs.

Historical Context of Chapter 15: Large-Scale Hypothesis Testing and FDRs

The historical evolution of methods discussed in this Chapter primarily revolves around addressing the challenges posed by large-scale hypothesis testing. This context sets the stage for understanding the

significant shifts in statistical practices driven by the need to control errors in multiple testing scenarios, particularly as datasets grew in size and complexity.

Evolution of Multiple Testing

Early 20th Century

The statistical community initially tackled multiple testing issues as isolated statistical anomalies, primarily through individual hypothesis tests. As scientific experiments grew in complexity, especially with advancements in biotechnology and social sciences, the need to address multiple hypotheses simultaneously became evident.

Mid 20th Century

The introduction of the family-wise error rate (FWER) marked a foundational approach to controlling Type I errors across multiple hypothesis tests. Techniques like the Bonferroni correction, derived from Boole's inequality, provided a stringent but overly conservative method for ensuring that the probability of making even one Type I error was controlled. This method was practical for a small number of tests but quickly became inadequate as the number of hypotheses grew, particularly highlighted by the explosion of data in genetics and other fields.

Pivotal Developments

Late 20th Century

The seminal work by Yoav Benjamini and Yosef Hochberg in 1995 introduced the concept of the False Discovery Rate (FDR), which was a critical turning point. They proposed a less conservative, more powerful method that was better suited for the burgeoning field of genomics, where researchers dealt with thousands of tests simultaneously. The FDR was designed to control the expected proportion of false discoveries, thereby providing a more balanced approach to Type I error control.

Their method, now known as the Benjamini-Hochberg procedure, was revolutionary because it allowed scientists to make discoveries at a faster pace without increasing the risk of false findings disproportionately. This method quickly became a staple in statistical analyses across multiple disciplines, from genomics to machine learning.

Integration of Bayesian Approaches

Early 21st Century

The integration of Bayesian statistics through Empirical Bayes methods further refined the approach to multiple testing. These methods allowed for the estimation of parameters from the data itself, enhancing the adaptability and accuracy of FDR calculations. The concept of local FDRs, which provided the probability that a specific hypothesis was null given the observed data, offered even finer resolution in hypothesis testing.

Statistical Practice Implications of Chapter 15: Large-Scale Hypothesis Testing and FDRs

This Chapter introduces advanced statistical methods tailored for the analysis of large-scale data, emphasizing the practical implications of False Discovery Rates (FDR) and Empirical Bayes methods in hypothesis testing. These methods address the critical challenges in modern statistical practices, especially in areas dealing with voluminous data and multiple comparisons.

Enhancing Decision-Making in Complex Datasets

The adoption of FDR methods allows statisticians to navigate the complex waters of large datasets more effectively. By focusing on the proportion of false discoveries rather than the traditional family-wise error

rate (FWER), researchers can make more informed decisions without being overly conservative, which is particularly crucial in fields like genomic research and big data analytics where decisions directly impact further research directions and funding.

Data-Driven Insights and Model Refinement

Empirical Bayes methods bring a data-driven approach to statistical inference, allowing more nuanced insights and model refinement. By estimating parameters directly from the data rather than relying on strict assumptions or fixed thresholds, these methods provide a flexible framework that adapts to the underlying data structure. This adaptability is invaluable for sectors like finance and healthcare, where predictive accuracy and response to changing patterns can significantly influence outcomes.

Implications for Research and Development

- **Genomics**: Efficient handling of multiple hypothesis tests enables more accurate identification of genetic markers linked to diseases.
- **Drug Development**: Improved statistical methods enhance the reliability of clinical trial results, influencing decision-making in pipeline development.
- **Technology and E-commerce**: Large-scale A/B testing frameworks benefit from refined FDR controls, optimizing user experience and product features based on robust statistical analysis.

Challenges and Considerations

While the advancements in statistical methodologies provide powerful tools for handling complex datasets, they also introduce challenges:

- **Skill Gap**: There is a growing need for training and knowledge dissemination to ensure that practitioners can effectively apply these advanced methods.
- **Computational Demand**: Implementing these methods requires substantial computational resources, especially in real-time data analysis scenarios.
- **Ethical Considerations**: The decision thresholds set by FDR and Empirical Bayes methods can have significant implications, particularly in sensitive areas such as personalized medicine and criminal justice.

Summary

The integration of False Discovery Rates and Empirical Bayes methods into statistical practice marks a significant evolution in the field. These methods not only enhance the precision and reliability of hypothesis testing in large-scale studies but also push the boundaries of what can be achieved with modern data analysis. As datasets continue to grow in size and complexity, the importance of these methodologies will only increase, underscoring their role as essential tools.

References

- **Benjamini, Y., & Hochberg, Y. (1995).** "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing."
- **Efron, B., & Hastie, T. (2016).** "Computer Age Statistical Inference."
- **Storey, J.D. (2002).** "A direct approach to false discovery rates."