STA 250: Statistics

Notes 16. Large scale testing and multiplicity correction

Book chapters:

1 Background: t-tests in microarray association studies

The advent of microarray technology has made it possible to simultaneously measure the expression levels of all genes of an individual (sample extracted from blood or a tissue). The expression level of a gene is a quantitative measure of how active it is in the gene-to-mRNA-to-protein synthesis procedure. By studying differences in the expression level of a gene between tissues of healthy individuals and tissues of individuals affected with a certain disease, say with colon cancer, we can learn whether the gene is associated with the condition. This often a first step in understanding connections between a disease and our genetic architecture which may eventually lead to effective treatments of the disease. Here is a typical dataset of expression levels (standardized).

<table>
<thead>
<tr>
<th>Gene#</th>
<th>Tumor1</th>
<th>Tumor2</th>
<th>...</th>
<th>Tumor18</th>
<th>Normal1</th>
<th>Normal2</th>
<th>...</th>
<th>Normal18</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.39</td>
<td>6.19</td>
<td>...</td>
<td>0.03</td>
<td>6.53</td>
<td>-0.22</td>
<td>...</td>
<td>-1.53</td>
</tr>
<tr>
<td>2</td>
<td>-10.82</td>
<td>0.117</td>
<td>...</td>
<td>-1.11</td>
<td>24.72</td>
<td>1.22</td>
<td>...</td>
<td>14.06</td>
</tr>
<tr>
<td>3</td>
<td>-0.78</td>
<td>-6.47</td>
<td>...</td>
<td>0.03</td>
<td>15.69</td>
<td>-1.39</td>
<td>...</td>
<td>-1.53</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>11.40</td>
<td>...</td>
<td>27.14</td>
<td>57.25</td>
<td>44.71</td>
<td>...</td>
<td>60.50</td>
</tr>
<tr>
<td>5</td>
<td>4.17</td>
<td>1.33</td>
<td>...</td>
<td>-1.11</td>
<td>-3.78</td>
<td>-0.22</td>
<td>...</td>
<td>-8.86</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>7457</td>
<td>-7.17</td>
<td>-1.06</td>
<td>...</td>
<td>-2.28</td>
<td>0.14</td>
<td>-4.25</td>
<td>...</td>
<td>-3.97</td>
</tr>
</tbody>
</table>

The association of any gene with the disease could be tested by testing for the gene’s expression level difference between the normal and the affected individuals. If $X^g = (X^g_1, \ldots, X^g_n)$ are the observed expression levels of gene $g$ from affected individuals and $Y^g = (Y^g_1, \ldots, Y^g_m)$ are the same from the normal individuals, we could set up the model: $X^g_i \overset{iid}{\sim} \text{Normal}(\mu^g, (\sigma^g)^2)$ and $Y^g_j \overset{iid}{\sim} \text{Normal}(\mu^g, (\sigma^g)^2)$ and test $H^g_0 : \mu^g_1 = \mu^g_2$ against $H^g_1 : \mu^g_1 \neq \mu^g_2$ and run a size-$\alpha$ t-test, with some small $\alpha$, to accept/reject $H^g_0$. Rejecting $H^g_0$ will be taken as “discovering” an association between the gene $g$ and the disease.

2 The issue of multiplicity

However, naively carrying out these t-tests for all genes results in a major problem. Empirically, the tests will produce a large number of discoveries. In most cases this is scientifically untenable, as at most a handful of genes are assumed to be associated with the disease. It is also statistically invalid. Indeed, if we used size-5% tests for a total $G = 10,000$ genes,
then even if none of them was associated with the disease (i.e., all $H_0^g$'s are true) we will still make approximately $10000 \times 0.05 = 500$ discoveries by chance.

3 Old classical approach to multiplicity correction: Familywise error

From the above discussion it’s clear that controlling for type-I error probability of each individual test may lead to absurd results, especially when $G = \text{the total number of tests to be performed, is large.}$ There have been many proposals to correct for this multiplicity effect. An early favorite is the Bonferroni correction in which one tries to control the familywise error (FWE) of “making one or more discoveries when all null hypotheses are true”. Operationally we choose a small familywise significance level $\alpha_F$, usually 5% or 1% and carry out each individual test at a level $\alpha = \alpha_F/G$. This works because

$$P(\text{one or more discoveries } | \text{ all } H_0^g \text{ true}) \leq \sum_{g} P(H_0^g \text{ is rejected} | H_0^g \text{ true}) = G \cdot \alpha = \alpha_F.$$ 

However, the Bonferroni correction is deemed inadequate for most modern scientific studies. Many scientists complain it’s too stringent and do not make any discoveries unless in special cases of a very strong signal. Most scientists agree that controlling for the familywise error is unappealing and the familywise null hypotheses: “all $H_0^g$ are true” is simply implausible and impractical.

4 Modern classical approach: False discovery rate

In modern statistics, considerations of “false discovery rate” have replaced the Bonferroni approach. Here the concern is “among all the discoveries made by our rule, how many are false?”, and the goal is to control this false discovery rate. If we knew the truth about each $H_0^g$, we could set the following table of error counts:

<table>
<thead>
<tr>
<th></th>
<th>$H_0^g$ true</th>
<th>$H_0^g$ false</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery</td>
<td>$U$</td>
<td>$T$</td>
<td>$G - R$</td>
</tr>
<tr>
<td>Non-discovery</td>
<td>$V$</td>
<td>$S$</td>
<td>$R$</td>
</tr>
<tr>
<td>Total</td>
<td>$G_0$</td>
<td>$G - G_0$</td>
<td>$G$</td>
</tr>
</tbody>
</table>

and define the false discover rate as $\text{FDR} = \text{E}\{\frac{V}{G} \cdot I(R > 0)\}$. In reality we will never know $V$. But we can work with multiple testing rules that guarantee $\text{FDR} < \delta$ for a chosen false discovery level $\delta > 0$ (usually, 5%, 10% or 20%).

One such rule is known as the Benjamini-Hochberg procedure which is extremely easy to implement. We first gather the p-values from all $G$ individual tests, call these $p_g, g = 1, \cdots, G$. Next, rearrange the p-values from the smallest to the largest: $p_{[1]} \leq p_{[2]} \leq \cdots \leq p_{[G]}$ and find the maximum rank $r$ such that $p_{[r]} \leq \delta \cdot r/G$. Declare the $r$ genes associated with the smallest $r$ p-values as “discoveries”. Under very mild conditions Benjamini and Hochberg (1995) proved that the resulting multiple discovery rule has $\text{FDR} \leq \delta$.

Benjamini and Hochberg’s proof is involved, and we will not get into the details. One interesting thing in the proof is the simple observation that “p-values are uniformly distributed under the null”. Formally, let $p_g(x^g, y^g)$ denote the p-value for the $g$-th test based on data $X^g = x^g, Y^g = y^g$. If indeed $H_0^g$ was true, then $p_g(X^g, Y^g) \sim \text{Uniform}(0, 1)$.
5 A simulation example

To illustrate the procedure we will work with a simplified version of the gene association study. The scalar test statistic for every gene $g$ will be denoted by $Z^g$ with $Z^g \sim \text{Normal}(\theta_g, 1)$ and we will test $H^g_0 : \theta_g = 0$ against $H^g_1 : \theta_g \neq 0$ [the t-test statistics can be reduced to such z-statistics by some simple transformations]. A size-\(\alpha\) ML test for testing $H^g_0$ is “reject $H_0$ if $|Z^g| > z(\alpha)$” and so the p-value gathered from gene $g$ is $2\{1 - \Phi(|Z^g|)\}$.

In the following we simulate $Z^g$ for $g = 1, \ldots, G = 10,000$ genes. For every gene $g$, we first make coin toss to decide between “null” or “non-null” with probability 99% and 1% respectively. If “null” then we simulate $Z^g \sim \text{Normal}(0, 1)$, otherwise we simulate $Z^g \sim \text{Normal}(3, 1)$. Roughly 99% of $H^g_0$ should be true under such a simulation.

```
G <- 1e4
p0 <- 0.99
theta <- c(0, 3)
ix <- sample(2, G, replace = TRUE, prob = c(p0, 1 - p0))
z <- rnorm(G, theta[ix])
```

One instance of this simulation is shown in the above histogram. For this simulation we had 9897 cases of true $H^g_0$ and 103 cases of false $H^g_0$. The red dots in the graph shows the $Z^g$ values of the non-null cases.

Using a simple size-5% test on each gene results in 581 discoveries, most of which (> 80%) are false discoveries. A level 5% Bonferroni correction carries out a size 0.0005% test on each gene and make only 3 discoveries. The code below applies B-H discovery rule with FDR control level $\delta = 10\%$ and makes 22 discoveries, only one of which is a false discovery.

```
p.vals <- 2 * (1 - pnorm(abs(z)))
p.order <- order(p.vals)
r <- max(which(p.vals[p.order] < 0.1 * 1:G / G))
cat("#Discoveries =", r, "\n")
cat("#False Discoveries =", sum(ix[p.order[1:r]] == 1), "\n")
```
6 Bayesian approach

There are many excellent Bayesian solutions available to the problem multiplicity. We will not discuss them here. I refer you to the article by Scott and Berger (2006) [http://dx.doi.org/10.1016/j.jspi.2005.08.031].