p-value correction and False Discovery Rate (FDR)

In a statistical test, based on the value of statistic computed from data, we make try to prove a hypothesis. Take the example of the study of drug effect on control and treatment groups. Assuming that the data sets X and Y of control and treatment groups are Gaussian, we can perform a two sample independent t test between the groups to detect any significant difference in the means of two populations.

Any statistic is assumed to follow a distribution the under null hypothesis. In a two sample t-test, the computed t statistics is assumed to follow a t-distribution under null hypothesis.

The probability of getting a value equal to or more of the observed statistic in the appropriate distribution is called the "p-value", which is a measure of statistical significnce of the data set under null hypothesis.

If the computed p-value of sample data is less than a pre-decided value α, we reject the null hypothesis and call the result statistically significant.

[for two sided test, reject null if $p < \alpha/2$. For one sided test, reject null if $p < 1 - \alpha$].

We should remember one important thing about the p-value. if the p-value is 0.04, it means that under the null hypothesis, there is a 4% chance of getting our observed result. It does not mean there is a 4% chance that the null hypothesis is true.

Multiple testing and false positives

There are occasions when we conduct many statistical tests simulataneously. We look at two examples here.

Example 1 : Consider a microarray or RNA seq experiment for differential gene expression analysis. Assume that the array has 10000 genes, each with n control and m treatment samples.

We will thus perform 10000 t tests, one test per gene between its control and treatment values.

Let $\alpha = 0.05$ be the significa level we fix for rejecting null hypothesis.

In each one of the 10000 tests, p-value is computed on the same t distribution, t(n+m-2).

The 10000 t tests have 10000 t statistic values, from same t(n+m-2) distribution. Correspondingly, we get 10000 p-values from the t distribution table. We compare these p-values with a significance $\alpha = 0.05$ (for example) to reject or accept null.

Therefore, for an arbitrary cutoff value $\alpha = 0.05$, there will be $10000 * 0.05 = 500$ genes that will be rejecting null hypothesis. We will select them for further analysis. If we take $\alpha = 0.02$, we will select 200. *This is true even when when all the 10000 tests are accepting the null hypothesis!.*

Example 2 : In order to test the effect of 40 chemical substances on a certain human cell lines, following experiment was performed. For each drug, n tissue samples were treated with placebos, n samples with the drug and some quantity was measured as a response. A statistical analysis, say two sample independent t test was performed on them to get a p-value for testing the null hypothesis. This was done for all the drugs to get 40 p-values from 40 multiple tests.

Even if we assume that all the 40 drugs accepted the null hypothesis (which is the equality of population means of placebo and treatment groups), for a significance level of $\alpha = 0.05$, we will end up with $40 * 0.05 = 2$ drugs rejecting the null hypothesis!

> We can summarize as follows : If N statistical tests are conducted simultaneously (multiple testing) each with a significance level of α, then we will end up with $N\alpha$ tests rejecting the null hypothesis even if every one of them accept the null hypothesis in reality!.

Some important terminologies

In a multiple testing procedure, the $N\alpha$ tests that rejected the null hypothes (called **discoveries**) are made up of two components :
(i) The component consisting of tests that genuinely rejected the null hypothesis due to the effect under study. They are called **true positives**.
(ii) The component consisting of tests that actually support the null hypothesis, were a part of distribution under null hypothesis, but were rejected because the value of their statistic is greater than the cutoff value arbitrary chosen through a significance α. These are called **false positives**.

False positives are also called as **Type-I erros**, which occur when a null hypothesis is rejected by mistake.

When testing multiple hypothesis, the more multiple hypothesis are tested, the more false positives ("false discoveries") occur. The probability of making one or more false discoveries in a multiple hypothesis testing is called **Family-Wise Error Rate (FWER)** .

The proportion of discoveries ("significant results") that are actually false positives is named as **False Discovery Rate (FDR)**.

The Family-Wise Error Rate(FWER) and the False Discovery rate(FDR) can be corrected by applying a algorithmic procedure on the list of p-values from the multiple test. In the following section, we will learn about three such algorithms: the *Bonferroni adjustment* and *Home-Bonferroni method* for controlling FWER and *Benjamini-Hochberg procedure* for controlling FDR.

1. Procedures for controlling family-wise error rate

These algorithms try to minimize the family-wise error rate. For a multiple test procedure with a decided significance level α, these algorithms compute a new significance level α' which is lower than α. This is done i such a way that even if null hypothesis is true for all the tests, the probability of getting one result that is significant at this new valueα' is α.

Even if all the N tests reject null hypothesis, we end up in selecting $n\alpha$ significant events for a threshold α. The probability of getting one or more false positive among these $n\alpha$ (family-wise error rate) is not known to us. These procedures give us a new threshold α' so that the probability of getting one FWER is 0.05.

1.1 Bonferroni adjustment for controlling the family-wise error rate

Let Let $p_1, p_2, p_3, \ldots\ldots, p_m$ be the p-values from m comparisons from a signle experiment [eg. m genes in a RNAseq experiment]. In Bonferroni method, the significance level $alpha$ is considered to be the significance level of all the m statistical tests. For an individual test i, either its p-value p_i is divided by m or p_i is kept same and the significance level α is divided by m, which will be compared with p_i to reject the null hypothesis. Thus,

Corrected p-value the for the i^{th} test $= \ p_{ci} \ = \ min(p_i \times m, 1) \qquad \text{(Reject Null if } p_{ci} \le \alpha)$

or, Corrected significance level $= \alpha_c = \dfrac{\alpha}{m}$ (Reject Null if $p_i < \alpha_c$)

The Bonferroni adjustment for multiple testing is very stringent (conservative) especially when the number of comparisons are high. For example, whem we test $m = 5000$ null hypothesis with $\alpha = 0.05$ we get $= \alpha_c = \dfrac{\alpha}{m} = \dfrac{0.05}{5000} = 0.00001$ as the maximum p-value that can reject null hypothesis. In a typical experiment, we will rarely get this much smaller p-value.

1.2 Holm-Bonferroni method of controlling family-wise error rate

In Holme's method, the significance level of each gene is computed using its rank in an ascending order of p-values for testing the null hypothesis.

Let $p_1, p_2, p_3, \ldots\ldots, p_m$ be the p-values from m comparisons from a signle experiment [eg. m genes in a RNAseq experiment].

Arrange these m p-values in **descending order** to get their rank k.

Then, rank k=1 for gene with lowest p-value, rank k=m for highest p-value.

For any gene i with p-value p_i whose rank is k in an ascending order of p-values, the null is rejected if

$$p_i \leq \frac{\alpha}{m - k + 1}$$

Thus the Bonferroni-Homes procedure sorts the p-values in ascending order and successively compares them to $\dfrac{\alpha}{m}, \dfrac{\alpha}{m-1}, \dfrac{\alpha}{m-2}, \ldots\ldots \dfrac{\alpha}{3}, \dfrac{\alpha}{2}, \dfrac{\alpha}{1}$

Alternately, we can also compute the corrected p_{ci} through the multiplication:

$$p_{ci} = p_i \times (m - k + 1)$$

and reject the Null if $p_{ci} \leq \alpha$

Holm-Bonferroni correction is less stringent than the Bonferroni correction.

<u>Example</u> : . Let us consider the following set of p-values from 7 comparisons in an experiment:

$$p1 = 0.022, p2 = 0.005, p3 = 0.085, p4 = 0.041, p5 = 0.035, p6 = 0.029, p7 = 0.001$$

We arrange them in a table in ascending order pf p-values and rank them to compute the corrected α values:

$$m = 7, \quad \alpha = 0.05$$

name	p_i	*rank* k	$\alpha' = \dfrac{\alpha}{m-k+1}$	[*reject if* $p_i < \alpha'$]	$p_{ci} = p_i(m-k+1)$
$p7$	0.001	1	0.00714	*rejected*	0.007
$p2$	0.005	2	0.00833	*rejected*	0.03
$p1$	0.022	3	0.01	*accepted*	0.11
$p6$	0.029	4	0.0125	*accepted*	0.116
$p5$	0.035	5	0.0166	*accepted*	0.105
$p4$	0.041	6	0.025	*accepted*	0.082

*p*3	0.085	7	0.05	*accepted*	0.085

In the above table note that eventhough the probabilities p1,p6 and p5 of statistical tests 1,6 and 5 by themselves are below 0.05, these tests fail to get accepted after correction. If our intention is to select the rejected cases for a given α, we can stop the test when we encounter the first acceptance, since all the subsequent tests will accept the null hypothesis.

2. Benjamini-Hochberg procedure for controlling False Discovery Rate (FDR)

This procedure controls the False Discovery rate (FDR), which is the proportion of "discoveries" (significat results) that are false positives.

In experiments like microarray, RNAseq etc, we discover few hundred genes out of thousands and ready to accept, say, a nominal 10% false positives among them.

The Benjamini-Hochberg procedure is as follows:

Choose a False Discovery Rate Q you are ready to tolerate.

Sort the m p-vlues, from smallers to the largest.

The smallest p-values has rank of k=1, next larger has rank k=2 etc, and the highest has a rank of k=m.

Compare each p-value to its Benjamini-Hochberg critical value given by $(\frac{k}{m})Q$.

The largest p-value $< (\frac{k}{m})Q$ is significant, and all the p-values smaller than it are also significant, even the ones that aren't less than their Benjamini-Hochberg critical value.

Another way of doing this is to compute p_c, the corrected p-value with rank k as $p_c = \text{p-value} \times \frac{m}{k}$ and reject null if $p_c < Q$.

<u>Example</u>:

The table below applies the Benjamini_Hochberg procedure for 16 multiple tests names T1, T2,...,T16, with FDR=0.1

```
               m=16, FDR=0.1
---------------------------------------------------------------------------------------------------

  Test      p-value   k    (k/m)*Q     corrected p-value        Decision on null
                                       ( p-value*(m/k) )    (significant if p-value < (k/m)*Q)
                                                                          or
                                                            (significant if corrected p-value < Q

---------------------------------------------------------------------------------------------------
   T4         0.0010   1    0.00625      0.0160                 Significant
   T7         0.0080   2    0.01250      0.0640                 Significant
   T11        0.0122   3    0.01875      0.0650                 Significant
   T8         0.0192   4    0.02500      0.0768                 Significant
   T10        0.0295   5    0.03125      0.0944                 Significant
   T16        0.0322   6    0.03750      0.0858                 Significant
   T15        0.0458   7    0.04375      0.1047                 Significant *****
   T1         0.0490   8    0.05000      0.0980                 Significant
   T3         0.0528   9    0.05625      0.0939                 Significant (highest significant val
   T2         0.0650   10   0.06250      0.1040                 Not significant
   T13        0.0722   11   0.06875      0.1050                 Not significant
   T9         0.0856   12   0.07500      0.1141                 Not significant
   T12        0.0939   13   0.08125      0.1155                 Not significant
   T5         0.1260   14   0.08750      0.1440                 Not significant
   T6         0.1770   15   0.09375      0.1888                 Not significant
   T14        0.2670   16   0.10000      0.2670                 Not significant
---------------------------------------------------------------------------------------------------
```

In the above computation, note that the highest significant p-value is 0.0528. All p-values less than this are significant, even if Benjamini-Hochberg procedure has not selected it. This has happenend to p-value=0.0458, marked by multiple stars. Eventhough 0.0458 is not less than its (k/m)*Q value of 0.4375, it is still considered significant since the highest significant value is after this.

Meaning of this computation is this: Out of the 16 comparisons, T4,T7,T11,T8,T10,T16,T15,T1 and T3 are significant, and the False Discovery Rate (ie., the fraction of false positives among the significant results) is controlled to 0.1 (10%) level.

R-scripts

R statistics library has an inbuilt function p.adjust() for controlling Family-wise Error Rate as well as False Discovery Rate. The function always returns a vector of corrected p values. The function calland the choices are as follows: The basic function call is, p.adjust(p, method = p.adjust.methods, n = length(p)) where, p ---> a vector of p values to be corrected p.adjust.methods ----> name of algorithm to be applied for p value correction. Choices are,
p.ajust.methods="holm" is Homes's method for controlling FWER
p.adjust.mthods="hochberg" applies Hochberg procedure for controlling FWER.
p.adjust.methods = "hommel" applies Hommel method for controlling FWER.
p.adjuat.methods = "bonferroni" allplies Bonferroni method for controlling FWER.
p.adjust.methods = "BH" or "fdr" applies Benjamini-Hochberg FDR control p.adjust.methods = "BY" applies Benjamini-Yekutaili FDR control procedure

The R script below demonstrates their call.

```
###################################################

##Perform FWER control using Bonferroni procedure

pvalue = c(0.022, 0.005, 0.085, 0.041, 0.035, 0.029, 0.001)

## compare this p.corrected with statistical significance alpha (eg. alpha=0.05)
##     to reject each null hypothesis.

p.corrected = p.adjust(p=pvalue, method = "holm")

print("Adjuted p values for Holm FWER control:")

print(p.corrected)

print("--------------------------------------------------------")


## Control FDR using Benjamini-Hochberg method
pvalue = c(0.049, 0.065, 0.0528, 0.001, 0.126, 0.177, 0.008, 0.0192, 0.0856,
           0.0295, 0.0122, 0.0939, 0.0722, 0.267, 0.458, 0.0322)

### compare the p.corrected with your level of FDR to reject the null for each test
p.corrected = p.adjust(p=pvalue, method="BH")

print("FDR control using Benjamini-Hochberg : ")

print(p.corrected)

###############-------------------------------------------------
```

Executing the above script in R prints the following results and figures of probability distribution on the screen:

```
[1] "Adjuted p values for Holm FWER control:"
[1] 0.110 0.030 0.116 0.116 0.116 0.116 0.007
[1] "--------------------------------------------------------"
[1] "FDR control using Benjamini-Hochberg : "
 [1] 0.10560000 0.11552000 0.10560000 0.01600000 0.15507692 0.20228571
 [7] 0.06400000 0.07680000 0.12450909 0.08586667 0.06506667 0.12520000
[13] 0.11552000 0.28480000 0.45800000 0.08586667
```