HDS 2. False discovery rate

Matti Pirinen, University of Helsinki

20.10.2021

In the previous lecture, we considered a multiple testing problem where thousands of tests were done while only a few of them were expected to be non-null. There the family-wise error control (FWER) seemed a reasonable way to filter a few most prominent candidates for true positives from the vast set of null variables.

Let's now consider situations where we may have a considerable amount of true positives $m = p - p_0$ among the p tested variable, e.g., $m/p = 10\%$. Let's start by writing up easy ways to generate P-values for such a situation.

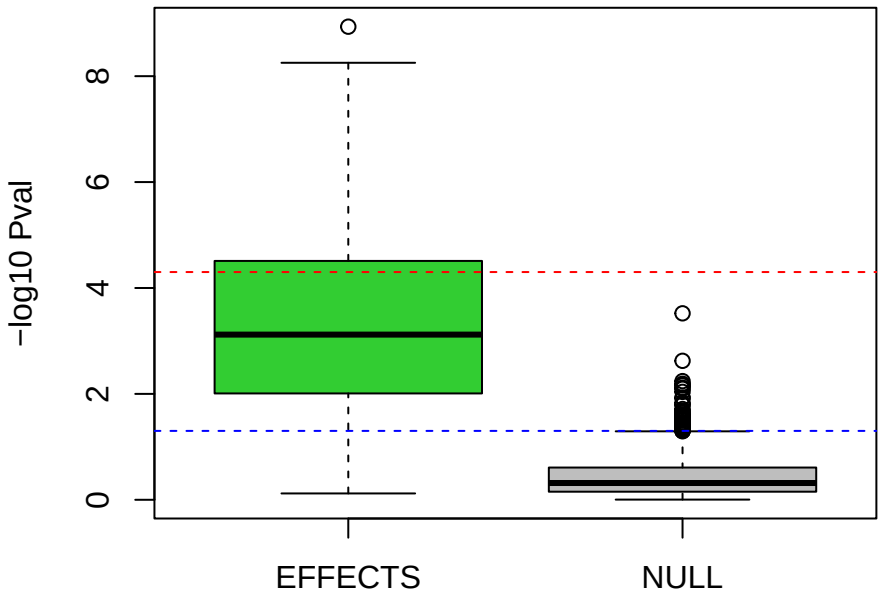
Distribution of z-scores Suppose that in linear regression $y = \mu + x\beta + \varepsilon$ the true slope is β. If $\text{Var}(x) = v_x$ and error variance is σ^2, then sampling variance of $\widehat{\beta}$ is $v_\beta = \sigma^2/(nv_x)$ where n is the sample size, and z-score for testing the slope is

$$z = \frac{\widehat{\beta}}{\sqrt{v_\beta}} \sim \mathcal{N}\left(\frac{\beta}{\sqrt{v_\beta}}, 1\right).$$

We can compute P-values for such z-scores using `pchisq(z^2, df = 1, lower = F)`, because under the null $z \sim \mathcal{N}(0,1)$ and hence $z^2 \sim \chi^2_1$, i.e., z^2 follows the central chi-square distribution with one degree of freedom. With this method, we can easily generate P-values for p_0 null variables ($\beta = 0$) and m non null variables $\beta \neq 0$ and test various inference methods on those P-values. Note that the P-value distribution of the non-null variables will depend on sample size n, true value β and the ratio of predictor and noise variances v_x/σ^2, but the P-value distribution of the null variables is always the same, namely, Uniform(0,1).

Let's generate P-values for $p = 1000$ variables that we assume each have been tested independently for a linear association with its own outcome variable. Let's also assume that $m = 100$ of them actually do have an effect each explaining 1% of the variance of the corresponding outcome. We follow the procedure from Lecture 0 to generate such data, and set $\sigma^2 = 1$ and $v_x = 1$.

```
set.seed(11102017)
n = 1000 #sample size
p = 1000 #number of variables to be tested
m = 100 #number of non-null tests
b = sqrt( 0.01 / (1 - 0.01) ) #This means each predictor explains 1%, See Lecture 0.
#Generating P-values: 1,...,m are true effects, m+1,...,p are null.
eff = c(rep(T, m), rep(F, p-m)) #indicator for non-null effects
pval = pchisq( rnorm(p, b*sqrt(n)*as.numeric(eff), 1)^2, df = 1, lower = F)
boxplot(-log10(pval)[eff], -log10(pval)[!eff], col = c("limegreen","gray"),
        names = c("EFFECTS","NULL"), ylab = "-log10 Pval")
abline(h = -log10(0.05), col = "blue", lty = 2) #significance threshold 0.05
abline(h = -log10(0.05 / p), col = "red", lty = 2) #Bonferroni corrected threshold of 0.05
```



We see that true effects tend to have smaller P-values (i.e. larger -log10 P-values) than null effects, but there is some overlap between the distributions and therefore, from P-values alone, we cannot have a rule that would detect all true positives but would not report any false positives.

Remember the characterization of test results according to this table:

Test result	not null	null	Total
negative (not significant)	FN	TN	p-D
positive (significant, discovery)	TD	FD	D
Total	p-p0	p0	p

Let's print such a table when we first do inference based on P-value threshold (say 0.05) and then use Bonferroni correction.

```
alpha = 0.05
dis = (pval < alpha) #logical indicating discoveries
table(dis, !eff, dnn = c("discov","null"))
```

```
##        null
## discov  FALSE TRUE
##   FALSE    10  851
##   TRUE     90   49
```

```
dis = (p.adjust(pval, method = "bonferroni") < alpha)
table(dis, !eff, dnn = c("discov","null"))
```

```
##        null
## discov  FALSE TRUE
##   FALSE    73  900
##   TRUE     27    0
```

The problem with the raw P-value threshold is that there are many false discoveries (over one third of all discoveries are false). The problem with the Bonferroni correction is that only a third of all true effects are discovered. We want something in between. We want to directly control the **proportion** of false discoveries made out of all discoveries.

Definition of false discovery rate

Let's define **False Discovery Proportion (FDP)** as a random variable

$$\text{FDP} = \frac{\text{FD}}{\max\{1, D\}} = \begin{cases} \frac{\text{FD}}{D}, \text{ if } D > 0. \\ 0, \text{ if } D = 0. \end{cases}$$

False Discovery Rate (FDR) is the expectation of FDP:

$$\text{FDR} = \text{E}(\text{FDP}).$$

Thus, it is the (expected) proportion of false discoveries among all discoveries. By controlling FDR at a given level α_F, we will tolerate more false discoveries as the number of tests increases as long as we will also keep on doing more true discoveries. If a method guarantees to keep $\alpha_F = 0.1$, then, in expectation, when we make 10 discoveries we allow 1 of them to be a false discovery, whereas when we make 100 discoveries we allow 10 false discoveries among them. Note the difference to FWER control where we always allow at most 1 false discovery in the experiment, no matter whether we are doing 1, 10, 1000 or 100,000 discoveries altogether.

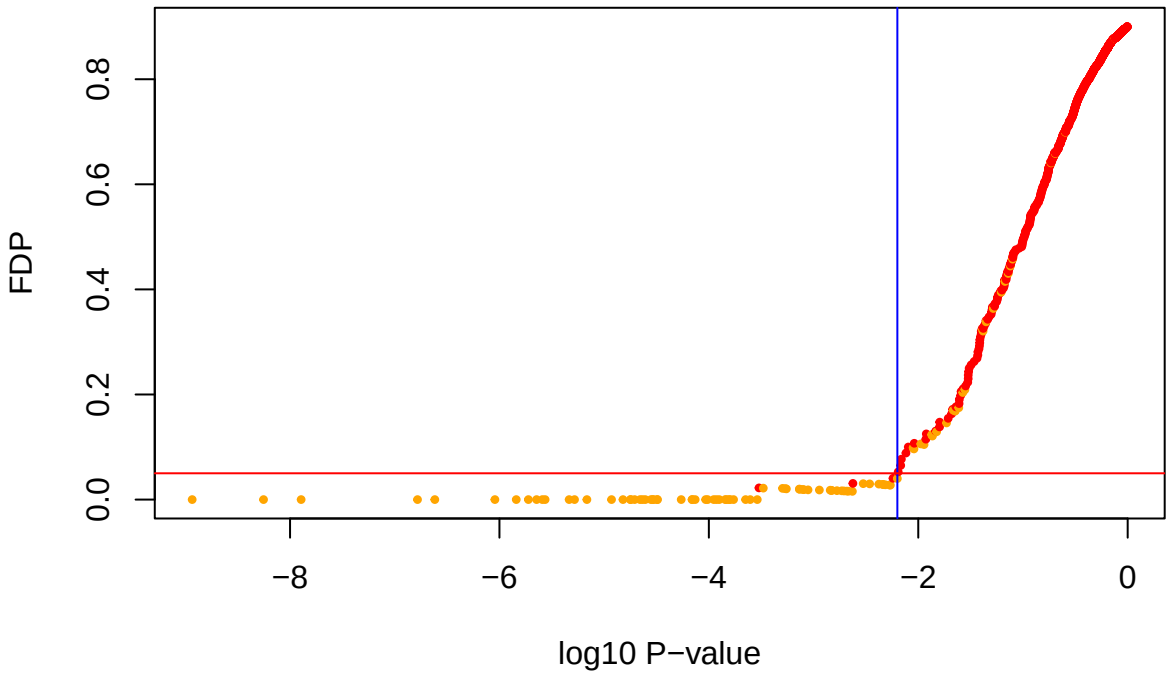
Given a rule for calling variables significant, the false positive rate is the rate that truly null variables are called significant and is measured by P-value. The FDR, instead, is the rate at which significant variables are truly null. For example, a false positive rate of 5% means that, on average, 5% of the truly null variables in the study will be called significant. A FDR of 5% means that, on average, among all variables called significant, 5% of them are truly null.

Let's see which P-value threshold would yield a false discovery proportion (FDP) of $\alpha_F = 0.05$ in our ongoing example. (Note FDP is the realized value of FD/D whereas FDR is its expectation. In real life we can't know FDP, but in these simulations we can!)

```
sort.pval = sort(pval) #sorted from smallest to largest
sort.eff = eff[order(pval)] #whether each sorted pval is from a true positive
fdp = cumsum(!sort.eff)/(1:p) #which proportion of discoveries are false
cols = rep("red",p); cols[sort.eff] = "orange" #true pos. in orange
plot(log10(sort.pval), fdp, xlab = "log10 P-value", ylab = "FDP",
     col = cols, cex = 0.7, pch = 20)
alpha = 0.05
i = max( which(fdp < alpha) )
print( paste("fdp <",alpha,"when P-value is <",signif(sort.pval[i],3)) )
```

```
## [1] "fdp < 0.05 when P-value is < 0.00632"
```

```
abline(v = log10(sort.pval[i]), col = "blue")
abline(h = alpha, col = "red")
```




```
#shows the step where fdp < alpha breaks
cbind(D=i:(i+1), FD=fdp[i:(i+1)]*c(i,i+1), fdp=fdp[i:(i+1)], pval=sort.pval[i:(i+1)])
```

```
##       D FD        fdp        pval
## [1,] 75  3 0.04000000 0.006315223
## [2,] 76  4 0.05263158 0.006431357
```

So if we had a method that controlled FDR at 0.05, we expect it to give roughly D=75 discoveries of which FD=3 are false. Compare this to D=139 and FD=49 with unadjusted P-value threshold of 0.05 and to D=27 and FD=0 with Bonferroni adjusted threshold of 0.05/p. But how can we in general control FDR given the set of P-values? Such a method was formulated by Yoav Benjamini and Yosef Hochberg in 1995.

Benjamini-Hochberg procedure (1995) Let H_j be the null hypothesis for test j and let P_j be the corresponding P-value. Denote the ordered sequence of P-values as $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(p)}$ and let $H_{(j)}$ be the hypothesis corresponding to the jth P-value. Benjamini-Hochberg procedure at level α_F (BH(α_F)) is to

$$\text{reject the null hypotheses } H_{(1)}, \ldots, H_{(k)}, \text{ where } k \text{ is the largest index } j \text{ for which } P_{(j)} \leq \frac{j}{p}\alpha_F.$$

Theorem (BH). For independent tests and for any configuration of false null hypotheses, BH(α_F) controls the FDR at level α_F.

Original proof in Benjamini and Hochberg (1995, J. R. Statist. Soc. B. 57(1):289-300).

Intuitive explanation why BH works. Consider P-value $P_{(j)}$. Since $P_{(j)}$ is the jth smallest P-value, if we draw a significance threshold at $P_{(j)}$ we have made exactly j discoveries. On the other hand, we expect that out of all p_0 null effects about $p_0 \cdot P_{(j)}$ give a P-value $\leq P_{(j)}$. Thus we have an approximation for false discovery proportion at threshold $P_{(j)}$

$$\text{FDP}(P_{(j)}) \approx \frac{p_0 \cdot P_{(j)}}{j} \leq \frac{p \cdot P_{(j)}}{j}.$$

If we simply ask that for which j is this estimated $\text{FDP}(P_{(j)}) \leq \alpha_F$, we get

$$\frac{p \cdot P_{(j)}}{j} \leq \alpha_F \Leftrightarrow P_{(j)} \leq \frac{j}{p}\alpha_F,$$

which is the condition of BH procedure.

Proof of BH. (Adapted from Aditya Guntuboyina.)

Assume we have p hypotheses and corresponding P-values $\boldsymbol{P} = (P_1, \ldots, P_p)$ and that $\text{BH}(\alpha_F)$ method rejects k of these hypotheses. This means that when the P-values are in the ascending order $(P_{(1)}, \ldots, P_{(p)})$, then k is the largest index that satisfies the BH-condition

$$P_{(k)} \leq \frac{k}{p}\alpha_F. \qquad (1)$$

We can write FDP as $\frac{1}{\max\{k,1\}} \sum_{j \in I_0} \text{I}\{P_j \leq \alpha_F \cdot k/p\}$, where I_0 is the set of p_0 true null hypotheses, and I is an indicator function.

$$\text{FDR} = \text{E(FDP)} = \text{E}\left(\sum_{j \in I_0} \frac{\text{I}\{P_j \leq \alpha_F \cdot k/p\}}{\max\{k,1\}} \right) \qquad (2)$$

The problem with trying to simplify this further is that P_j and k are not independent. The idea in this proof is to replace k with a closely related quantity $\tilde{k}_j$ that is independent of P_j.

For any j, denote by $\tilde{\boldsymbol{P}}_j = (P_1, \ldots, P_{j-1}, 0, P_{j+1}, \ldots P_p)$ the sequence of P-values where P_j has been changed to 0. Denote by $\tilde{k}_j$ the number of rejections that $\text{BH}(\alpha_F)$ method does when it is applied to $\tilde{\boldsymbol{P}}_j$. Importantly, $\tilde{k}_j$ only depends on $(P_i)_{i \neq j}$ but not on P_j, that, by an assumption of the BH theorem, is independent of $(P_i)_{i \neq j}$.

Clearly $k \leq \tilde{k}_j$ since the difference between the two sequences of P-values is that the element j in $\tilde{\boldsymbol{P}}_j$ is $0 \leq P_j$ while all other elements are the same between the two sequences. It is also clear that $\tilde{k}_j \geq 1$ since $\text{BH}(\alpha_F)$ always rejects the hypothesis corresponding to P-value 0.

We next show that

$$\frac{\text{I}\{P_j \leq \alpha_F \cdot k/p\}}{\max\{k,1\}} \leq \frac{\text{I}\{P_j \leq \alpha_F \cdot \tilde{k}_j/p\}}{\tilde{k}_j}, \qquad \text{for all } j = 1, \ldots, p. \qquad (3)$$

Since the right-hand side is non-negative, the inequality (3) holds when $\text{I}\{P_j \leq \alpha_F \cdot k/p\} = 0$.

Consider then the remaining case $\text{I}\{P_j \leq \alpha_F \cdot k/p\} = 1$. Since k is the largest index for which the BH-condition (1) holds, $P_j \leq P_{(k)}$, and hence the ordered sequences of P-values $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}_j$ agree on every element that is $\geq P_j$. In particular, they agree on every element that is $\geq P_{(k)}$. Since k was the largest index in ordered sequence $\boldsymbol{P}$ for which the BH-condition (1) holds, it follows that $\text{BH}(\alpha_F)$ rejects exactly the same hypotheses from both sequences. Thus $\tilde{k}_j = k$ and $\text{I}\{P_j \leq \alpha_F \cdot \tilde{k}_j/p\} = \text{I}\{P_j \leq \alpha_F \cdot k/p\} = 1$, which proves the inequality (3) also in this case.

Let's get back to formula (2) and apply (3).

$$\begin{aligned}
\text{FDR} &= \text{E(FDP)} = \text{E}\left(\sum_{j \in I_0} \frac{\text{I}\{P_j \leq \alpha_F \cdot k/p)\}}{\max\{k,1\}} \right) \leq \text{E}\left(\sum_{j \in I_0} \frac{\text{I}\{P_j \leq \alpha_F \cdot \tilde{k}_j/p)\}}{\tilde{k}_j} \right) \\
&= \text{E}\left(\text{E}\left(\sum_{j \in I_0} \frac{\text{I}\{P_j \leq \alpha_F \cdot \tilde{k}_j/p\}}{\tilde{k}_j} \,\middle|\, \tilde{k}_j \right) \right) = \text{E}\left(\sum_{j \in I_0} \frac{\text{E}(\text{I}\{P_j \leq \alpha_F \cdot \tilde{k}_j/p\} \,|\, \tilde{k}_j)}{\tilde{k}_j} \right) \\
&= \text{E}\left(\sum_{j \in I_0} \frac{\alpha_F \tilde{k}_j}{p\tilde{k}_j} \right) = \frac{p_0}{p}\alpha_F \leq \alpha_F.
\end{aligned}$$

The last line follows since the sum is over p_0 true null hypotheses whose P-values are uniformly distributed in (0,1) and because P_j is independent of $\tilde{k}_j$. This proves that BH(α_F) controls FDR at level α_F (actually even more strictly at level $\alpha_F \cdot p_0/p$). $\square$

Like in Holm procedure, in BH the rejection of the tested hypotheses depends not only on their P-values but also on their rank among all the P-values. We see that the critical threshold increases from the Bonferroni threshold α_F/p for the smallest P-value to the unadjusted threshold α_F for the largest P-value. Crucially, if any P-value $P_{(j)}$ is below its own threshold $j \cdot \alpha_F/p$, it means that also ALL the hypotheses corresponding to smaller P-values will be rejected, no matter whether they are below their own thresholds.

Let's apply BH to our current data, first manually and then through `p.adjust`.

```
alpha = 0.05
i.BH = max(which(sort.pval <= ((1:p)*alpha)/p))
print(paste("Reject 1,...,",i.BH," i.e, if P-value <=",signif(sort.pval[i.BH],3)))
```

```
## [1] "Reject 1,..., 66  i.e, if P-value <= 0.00298"
```

```
print(paste0("Discoveries:",i.BH,
             "; False Discoveries:",sum(!sort.eff[1:i.BH]),
             "; fdp=",signif(sum(!sort.eff[1:i.BH])/i.BH,2)))
```

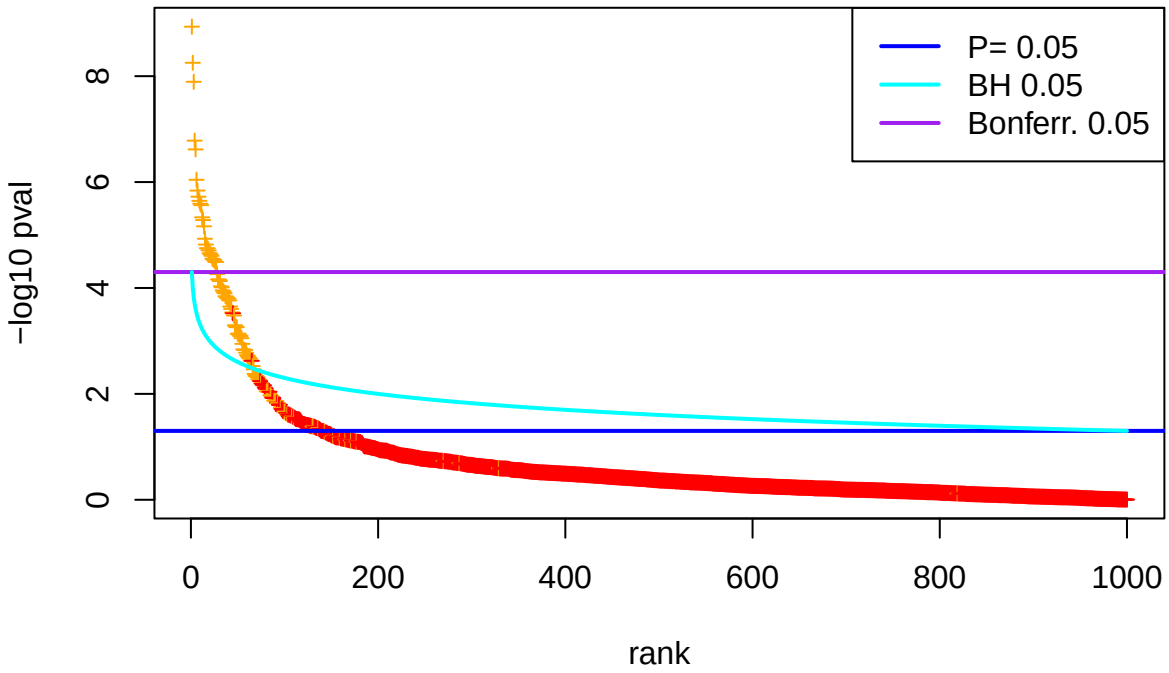
```
## [1] "Discoveries:66; False Discoveries:2; fdp=0.03"
```

```
pval.BH = p.adjust(pval,method = "BH")
#These are pvals adjusted by a factor p/rank[i]
#AND by the fact that no adjusted P-value can be larger than any of other adjusted
#P-values that come later in the ranking of the original P-values (See home exercises)
sum(pval.BH < alpha) #should be D given above
```

```
## [1] 66
```

Let's draw a picture with P-value threshold (blue), Bonferroni threshold (purple) and BH threshold (cyan). Let's sort the P-values from smallest to largest and draw P-values on -log10 scale.

```
plot(1:p, -log10(sort.pval), pch = 3, xlab = "rank", ylab = "-log10 pval", col = cols, cex = 0.7)
abline(h = -log10(alpha), col = "blue", lwd = 2)
abline(h = -log10(alpha/p), col = "purple", lwd = 2)
lines(1:p, -log10((1:p)*alpha/p), col = "cyan", lwd = 2)
legend("topright", legend = paste(c("P=","BH","Bonferr."), alpha),
       col = c("blue","cyan","purple"), lwd = 2)
```



The Figure tells that BH quite nicely draws the threshold around the point where true effects start to mix with null effects, whereas Bonferroni is too stringent and raw P-value based significance level is too liberal when the goal is to separate orange and red.

Exercise. Suppose that all your $p = 1000$ P-values were between $(0.01, 0.05)$, so they are not very small but still they are at the left tail of the whole Uniform(0,1) distribution. How would the three methods (Bonferroni, BH, "P<0.05") label variables significant in this case, when we control each at level 0.05 as in the Figure above? Earlier we saw that by using significance level $P \leq 0.05$ we tend to get a lot of false positives when p is large. How can our FDR method, that should keep false discovery rate small, then agree with the traditional significance threshold in this case? Why does it think that there are not a lot of false positives among these 1000 variables?

Benjamini-Yekutieli procedure It is important to remember that BH procedure was proven for **independent** test statistics and therefore it is not as generally applicable as Bonferroni and Holm methods. (In practice, however, BH has been shown to be quite robust to violations of independence assumption.) An extension of BH has been proven to control FDR in all cases by Benjamini and Yekutieli (The Annals of Statistics 2001, Vol. 29, No. 4, 1165-1188).

Theorem. When the Benjamini-Hochberg procedure is conducted with $\alpha_F / \left(\sum_{j=1}^{p} \frac{1}{j}\right)$ in place of α_F, it always controls the FDR at level $\leq \alpha_F$. (Proof skipped.)

This Benjamini-Yekutieli (BY) procedure can also be done by `p.adjust`.

```
alpha = 0.05
i.BY = max(which(sort.pval <= ((1:p)*alpha/sum(1/(1:p))/p)))
print(paste("Reject 1,...,",i.BY," i.e, if P-value <",signif(sort.pval[i.BY],2)))
```

```
## [1] "Reject 1,..., 44  i.e, if P-value < 0.00029"
```

```
print(paste0("Discoveries:",i.BY,
             "; False Discoveries:",sum(!sort.eff[1:i.BY]),
             "; fdp=",signif(sum(!sort.eff[1:i.BY])/i.BY,2)))
```

```
## [1] "Discoveries:44; False Discoveries:0; fdp=0"
```

```
pval.BY = p.adjust(pval,method = "BY") #these are pvals adjusted by factor p/rank[i]*sum(1/(1:p))
sum(pval.BY < alpha) #should be D given above
```

```
## [1] 44
```

As we see, BY is more conservative than BH, which is the price to pay for proven guarantees of control of FDR in case of all possible dependency structures.

Relationship between FWER and FDR We have

$$\mathrm{FDR} = \mathrm{E}(\mathrm{FDP}) = \Pr(D = 0) \cdot 0 + \Pr(D > 0) \cdot \mathrm{E}\left(\frac{\mathrm{FD}}{D} \,\middle|\, D > 0\right) = \Pr(D > 0) \cdot \mathrm{E}\left(\frac{\mathrm{FD}}{D} \,\middle|\, D > 0\right).$$

FDR weakly controls FWER. If all null hypotheses are true, then the concept of FDR is the same as FWER. To see this note that here $\mathrm{FD} = D$ and therefore if $\mathrm{FD} = 0$ then $\mathrm{FDP} = 0$ and if $\mathrm{FD} > 0$ then $\mathrm{FDP} = 1$. Thus, $\mathrm{FDR} = \mathrm{E}(\mathrm{FDP}) = \Pr(\mathrm{FD} > 0) = \mathrm{FWER}$. This means that FDR *weakly controls FWER*: If all null hypotheses are true, then any method that controls FDR at level α also controls FWER at level α. However, if some null hypotheses are false, then FDR doesn't typically control FWER.

FWER controls FDR. Because $\mathrm{FDP} \leq \mathrm{I}(\mathrm{FD} > 0)$, where I is the indicator function, by taking expectation,

$$\mathrm{FDR} = \mathrm{E}(\mathrm{FDP}) \leq \Pr(\mathrm{FD} > 0) = \mathrm{FWER}.$$

Thus, any method that controls FWER at level α also controls FDR at level α.

Examples **Example 2.1.** Let's see how BH controls FWER under the global null hypothesis.

```
p = 1000 #variables for each data set
alpha = 0.1 #target FWER
R = 1000 #replications of data set
res = matrix(NA, ncol = 3, nrow = R) #collect the number of discoveries by BH, Holm, Bonferr.
for(rr in 1:R){
#Generate P-values that are null.
  pval = runif(p)
  res[rr,] = c(sum(p.adjust(pval, method = "BH") < alpha),
               sum(p.adjust(pval, method = "holm") < alpha),
               sum(p.adjust(pval, method = "bonferroni") < alpha))
}
apply(res > 0, 2, mean) #which proportion report at least one discovery?
```

```
## [1] 0.102 0.097 0.097
```

All are close to the target FWER value $\alpha = 0.1$.

Example 2.2. BH method has a property that status of being a discovery at level α can change if a new variable is included among the tested variables.

```
pval = c(0.014, 0.09, 0.05, 0.16)
p.adjust(pval, method = "BH")
```

```
## [1] 0.056 0.120 0.100 0.160
```

```
p.adjust(c(pval, 0.001), method = "BH")
```

```
## [1] 0.03500000 0.11250000 0.08333333 0.16000000 0.00500000
```

Now the status at FDR level $\alpha = 0.05$ of P-value 0.014 has changed when we added one more variable with lower P-value. This is because this new variable is taken as a discovery and, intuitively, this inclusion then allows some more false discoveries to be included among the discoveries without breaking the idea of controlling FDR.

Example 2.3. Bonferroni, Holm and BH behave differently in whether the P-values close to each other can become the same after the adjustment. Let's demonstrate this with one tie in the middle of 4 P-values.

```
pval = c(0.01, 0.02, 0.02, 0.03)
rbind(pval, p.adjust(pval, "bonferroni"))
```

```
##      [,1] [,2] [,3] [,4]
## pval 0.01 0.02 0.02 0.03
##      0.04 0.08 0.08 0.12
```

```
rbind(pval, p.adjust(pval, "holm"))
```

```
##      [,1] [,2] [,3] [,4]
## pval 0.01 0.02 0.02 0.03
##      0.04 0.06 0.06 0.06
```

```
rbind(pval, p.adjust(pval, "BH"))
```

```
##            [,1]       [,2]       [,3] [,4]
## pval 0.01000000 0.02000000 0.02000000 0.03
##      0.02666667 0.02666667 0.02666667 0.03
```

Bonferroni multiplies each P-value by same value (here 4) and hence tie will stay as it is but not unite with its neighbors.

Holm first multiplies by $(p - j + 1)$, i.e., the 2nd value by 3 (to get 0.06) and 3rd value by 2 (to get 0.04) and the 4th value stays as it is (0.03). But then it needs to make sure that the adjusted P-values are in ascending order and it does this, for each P-value, by taking the maximum with P-values on the left. Hence all last 3 P-values become 0.06.

BH first multiplies by (p/j), i.e., the 1st value by 4 (to get 0.04), the 2nd value by 2 (to get 0.04) and 3rd value by 4/3 (to get 0.0266667) and the 4th value stays as it is (0.03). It also makes sure that the adjusted P-values become in ascending order but it makes this by taking, for each P-value, the minimum with all other P-values on the right hand side. Hence also 1st and 2nd P-values become 0.0266667.

We say that Holm method is a step-down procedure and BH method is a step-up procedure. Note that in this terminology, "up" and "down" refer to the ordering of the test statistics and not to the ordering of the P-values.

Step-down procedures start from the hypothesis with the largest test statistic (and lowest P-value) and step down through the sequence of hypotheses in descending order of their test statistics while rejecting the hypotheses. The procedure stops at the first non-rejection and labels all remaining hypotheses as not-rejected.

Step-up procedures start from the hypothesis with the lowest test statistic value (and highest P-value) and step up through the sequence of hypotheses in ascending order of their test statistics while retaining hypotheses. The procedure stops at the first rejection and labels all remaining hypotheses as rejected.

Single step procedures are the ones where the same criterion is used for each hypothesis, independent of its rank among the test statistics. Bonferroni correction and fixed significance level testing are examples of single step procedures.

Questions.

1. If you use BH(α_F) method to choose the significant variables, are you guaranteed to have at most $\alpha_F D$ false discoveries among the chosen variables, where D is the total number of discoveries made by the procedure?
2. If you have $p = 1000$ variables to test, when should you use FDR control and when FWER control?
3. What could be a situation when you would rather use BY procedure than BH procedure to control for FDR?