Methods for FDP estimation and confidence

Since both FDR control and FDP estimation and confidence aim to say something about the false discovery proportion Q, we will discuss the difference between FDR control and FDP estimation and confidence briefly before turning to specific methods.

To understand the difference between FDR control and FDP estimation, imagine a researcher repeating the same experiment many times. If the researcher used an FDR controlling method to analyze the experiment, he or she would find a new set of rejected hypotheses every time, obtaining sets $\mathcal{R}_1, \ldots, \mathcal{R}_z$ in z experiments. Each of these sets would have its own unknown false discovery proportion, say $Q_1, \ldots, Q_z$. Control of FDR means that, if z is large, the average of the Q_i's is bounded by α. Since the rejected sets are different upon every repetition of the experiment, FDR control, clearly, is a property of the procedure leading to a rejected set $\mathcal{R}_i$, not of the rejected set $\mathcal{R}_i$ itself. Of course, the fact that $\mathcal{R}_i$ was chosen by a method with FDR control says something about the properties of the set $\mathcal{R}_i$, but only indirectly.

FDP estimation and confidence methods look at the same problem in a different way. A researcher using such a method investigates a set $\mathcal{R}$ and obtains an estimate $\hat{Q}$, or a confidence interval $[L, U]$, for the false discovery proportion Q of that set. Upon repeating the experiment, it is assumed that the researcher investigates the same set $\mathcal{R}$ every time, but obtains new estimates $\hat{Q}_1, \ldots, \hat{Q}_z$ and confidence intervals $[L_1, U_1], \ldots, [L_z, U_z]$. The estimates $\hat{Q}$ are estimates in the classical sense of the fixed parameter Q belonging to the fixed set $\mathcal{R}$, and can have classical properties such as consistency and unbiasedness. The confidence intervals should have the coverage property that, if z is large, the true value of Q is contained in the confidence interval $[L_i, U_i]$ at least a proportion $1 - \alpha$ of the time. Since these FDP estimates and confidence statements are about the same fixed set $\mathcal{R}$, they are directly informative about $\mathcal{R}$ itself.

An important practical difference between FDR control and FDP estimation and confidence is that the routine of the multiple testing procedure is reversed. FDR control starts with the choice of the error rate to be controlled, and the procedure that controls it, and finds a single rejected set that fulfils the criteria. In contrast, FDP estimation and confidence starts with the set of hypotheses the researcher would like to reject, and finds an estimate or confidence interval for the error rate (FDP) of that set.

The potential for FDP estimation, as opposed to FDR or FWER control, is therefore that it allows researchers to approach multiple testing in a different way that reverses the traditional roles of the user and the multiple testing procedure. If reliable estimates of Q are available for every set $\mathcal{R}$ of potential interest, the user can review these sets and their estimates, and select the most promising one. The role of the multiple testing procedure can be to inform the user of the likely FDP in the set by providing an estimate and a confidence interval for Q. Alternatively, methods may find estimates and confidence intervals for the number of type I errors V, rather than for the false discovery proportion Q.

Allowing the user to choose the rejected set $\mathcal{R}$ also opens the possibility of choosing sets that do not necessarily consist of the hypotheses with top p-values. This can be valuable in exploratory genomics research, in which a mixture of statistical and biological considerations may play a role in deciding which hypotheses are candidates for rejection. In such situations it can be important for a researcher to obtain a reliable estimate or confidence statement about the actual set $\mathcal{R}$ that is chosen.

Two issues, however, must be taken into account by FDP estimation methods in order to make this approach viable.

1. The first is selection. Finding an estimate of Q for a fixed set $\mathcal{R}$ is relatively easy, but the rejection set $\mathcal{R}$ of interest is typically chosen on the basis of the same data that are also used for FDP estimation, and this set will probably be selected because it has a small estimate $\hat{Q}$. The value of $\hat{Q}$ for the post hoc selected $\mathcal{R}$ is therefore very likely to be underestimated. Honest estimation of Q should protect against this estimation bias or correct for it.

2. The second difficulty is to provide an assessment of the uncertainty in an estimate $\hat{Q}$. Giving a point estimate of FDP without any assessment of its variability is not very informative. The variance of

the estimate, however, is crucially influenced by the dependence structure of the p-values used for the estimation. Any assessment of this variability is again subject to the above-mentioned problem of selection bias due to the selection of $\mathcal{R}$. Proper confidence statements should take this variability into account.

Storey's approach for FDP estimation

From Goeman and Solari (2014) Multiple Hypothesis Testing in Genomics. Statistics in Medicine 2014 33:1946-78, Section 6.1

Simple and intuitive FDP point estimates for the collection of the top k hypotheses with best p-values can be based on the p-value rank plots and histograms. This was first suggested by Storey, who was motivated by the empirical Bayesian view of FDR. Storey considered only sets of rejected hypotheses of the form $\mathcal{R} = \{H_i : p_i \leq t\}$ for some constant t. In such a rejected set, the expected number of true hypotheses to be rejected is $m_0 t$, because p-values of true hypothesis follow a uniform distribution. This number can be estimated by substituting an estimate $\hat{m}_0$ for m_0. Storey suggests

$$\hat{m}_0 = \frac{\#\{p_i > \lambda\} + 1}{1 - \lambda}, \tag{1}$$

where $0 \leq \lambda < 1$ is some constant. The value of λ is typically taken as $1/2$, although $\lambda = \alpha$ has also been advocated. To understand this estimator, remark that a proportion $1 - \lambda$ of p-values of true hypotheses is expected to be above λ, but a smaller proportion of p-values of false hypotheses, so that $\mathrm{E}(\#\{p_i > \lambda\}) \geq m_0(1 - \lambda)$. Consequently, $\mathrm{E}(\hat{m}_0) \geq m_0$, making $\hat{m}_0$ a conservative estimate of m_0. The estimate (1) is illustrated graphically in the next page.

The addition of 1 to the numerator makes sure that $\hat{m}_0^{-1}$ is always defined. Using this estimate, Storey writes $\mathrm{FDR} \approx m_0 t / \#\mathcal{R}$, so that

$$\hat{Q} \;=\; \frac{\hat{m}_0 t}{\#\mathcal{R}} \;=\; \frac{t(\#\{p_i > \lambda\} + 1)}{(1 - \lambda)\#\{p_i \leq t\}}. \tag{2}$$

Storey's estimate has been derived under the assumption of independence among p-values. At first sight, however, it appears hardly affected by dependence among p-values. The expected number of true hypotheses with p-value smaller than t is $m_0 t$ whatever the joint distribution of these p-values. Dependence, however, does play a large role in the variability of the estimate, since the number of true hypotheses with p-values below t can have high variance especially if p-values are positively correlated, and this may affect the performance of the estimate.

A proof of conservative consistency for the estimate is available, which says that $\lim_{m \to \infty} \hat{Q} \geq Q$. However, this property has been shown to hold only under the assumption of independent p-values or under a form of weak dependence that allows only local correlations, and may fail otherwise. Under more realistic dependence, Storey's estimates are very variable and can underestimate the true FDP by a wide margin. A highly variable estimate does not have to be a problem, as long as we can assess this variability. Some important work has been done in finding the distribution of the estimator, but no definite methodology has emerged. Storey's estimator can have very large variance, large positive skewness, and substantial bias.

Proof that post hoc choice of the threshold t, which defines $\mathcal{R}$, is allowed, is only available under the assumption of independent p-values. In other situations, it is unclear how such a choice biases the final estimate.

In the Van de Vijver data, Storey estimates $\hat{m}_0 = 2230$ ($\hat{\pi}_0 = 0.45$). For the 1340 hypotheses rejected by the Benjamini & Hochberg procedure, this leads to an estimated FDP of only 2.3%.

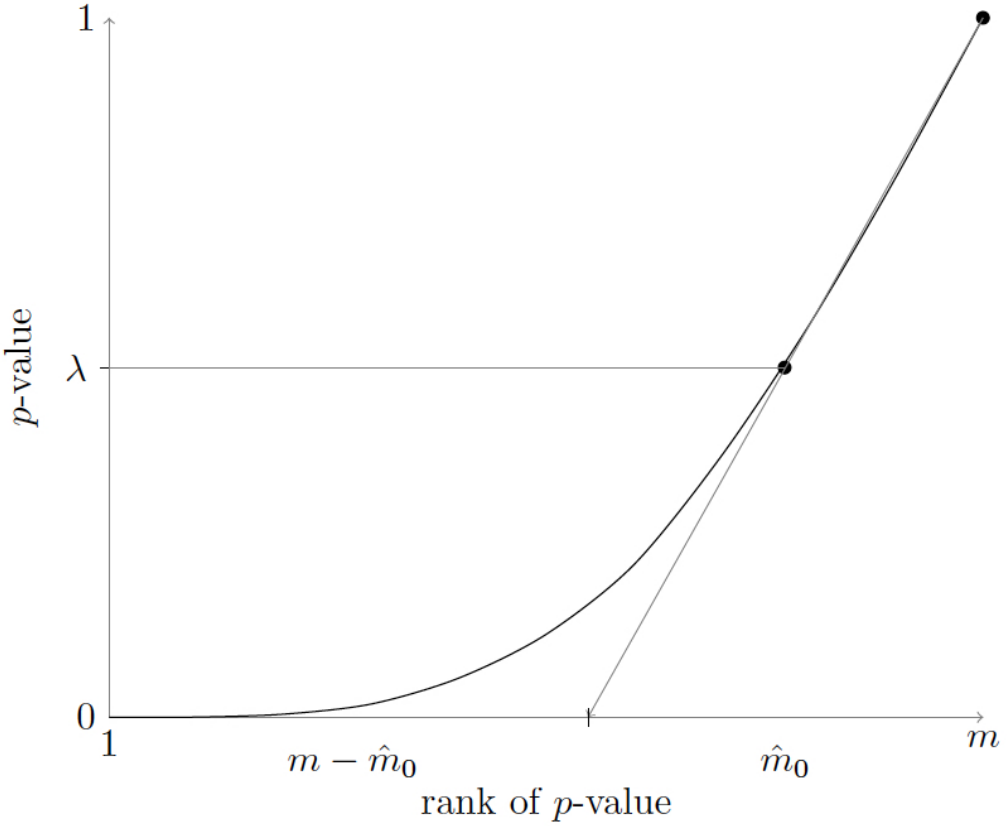


Figure 1: Illustration of Storey's estimate of m_0. The diagonal line connects the point $(m, 1)$ and the point at which the curve is equal to λ. The point at which it splits the x-axis determines the estimated numbers of true and false hypotheses.

```
load("/Users/aldosolari/Documents/mygithub/SL/data/9_VandeVijver.Rdata")
p.raw = p_vandevijver
m = length(p.raw)
lambda = 1/2
hatm0 = ( sum(p.raw > lambda) + 1 )/(1-lambda)
hatm0
```

```
## [1] 2230
```

```
hatpi0 = hatm0/m
hatpi0
```

```
## [1] 0.4533442
```

```
# FDP estimate for the 1340 hypotheses rejected by BH
t_BH = sort(p.raw)[1340]
hatQ = (hatm0*t_BH)/sum(p.raw <= t_BH)
hatQ
```

```
##     pvalue
## 0.02256622
```

Through its incorporation of a π_0 estimate, Storey's method is especially more optimistic than the Benjamini & Hochberg procedure in data sets with strong effects.

Storey's estimate is sometimes used as a way to control FDR, rather than as a way to estimate FDP. This is done by selecting the highest value of t such that the estimate (2) is below α. This actually links quite closely to the Benjamini & Hochberg procedure. If instead of using (1) we estimate m_0 conservatively by m, then we have $\hat{Q} = mt/(\#\{p_i \leq t\})$. Finding the largest t such that this estimate is below α is exactly achieved by the Benjamini & Hochberg procedure. Similarly, if Storey's approach with the estimate of π_0 is used, this results in an adaptive Benjamini & Hochberg procedure.

In the Van de Vijver data control of FDR using Storey's method would lead to 1787 rejections with $\lambda = 1/2$.

```
alpha = 0.05
sum(p.adjust(p.raw,"BH") <= alpha/hatpi0)
```

```
## [1] 1787
```

FDP confidence

Point estimates of FDP are of limited value if no assessment of variability is available. For assessing such variability, a confidence interval is most informative.

In the approach of Goeman and Solari, the user chooses a set $\mathcal{R}$ freely, and uses the method to obtain a confidence interval for the FDP $Q(\mathcal{R})$ of this set. We make the dependence of Q on the set $\mathcal{R}$ explicit in the notation. The confidence intervals are one-sided, using the trivial lower bound $Q(\mathcal{R}) \geq 0$, because only the upper bound to the number of false discoveries is of interest. The resulting confidence interval $[0, \bar{Q}(\mathcal{R})]$ is valid if we have

$$\mathrm{P}(Q(\mathcal{R}) \leq \bar{Q}(\mathcal{R})) \geq 1 - \alpha. \tag{3}$$

Importantly, this confidence bound holds for every possible set $\mathcal{R}$ simultaneously, so that, by the properties of simultaneous confidence bounds, the validity of equation (3) is not compromised by multiple looks at the data, and the confidence statement holds for a post hoc selected set as well as for any other set. The user can therefore look at the confidence bounds $\bar{Q}(\mathcal{R})$ of many sets $\mathcal{R}$ of potential interest and pick the set with the best bound, without destroying the validity of the confidence statement for the selected set. Since the method additionally allows sets $\mathcal{R}$ of any form, not just sets consisting of the hypotheses with the best p-values, the method of Goeman and Solari accommodates the tendency of researchers to cherry-pick among the list of top genes coming from an experiment, composing a meaningful selection of hypotheses to take to the next stage of validation using a mixture of statistical and biological considerations.

All these confidence statements for $Q(\mathcal{R})$ can be alternatively presented as confidence statements on the number of false rejections $V(\mathcal{R}) = \#(\mathcal{R} \cap \mathcal{T})$ by simply multiplying by the fixed quantity $\#\mathcal{R}$ before and after the first inequality sign in (3).

Two variants of the method of Goeman and Solari are relevant for this overview: one that assumes the PDS condition, and one that holds for general dependence structures of p-values. The PDS-based method uses exactly the same combination of the closed testing procedure and the Simes inequality that the methods of Hochberg and Hommel use. Only, instead of using this procedure just to derive FWER-type statements for the individual hypotheses, simultaneous confidence bounds of the form (3) are obtained using the same procedure. Since all confidence bounds are derived from a single application of the closed testing procedure, they depend on the same event for their coverage, making these confidence bounds simultaneous. Because the underlying method is the same, the assumption of PDS on the true hypotheses underlying the Simes-based method of Goeman and Solari is identical to the assumption underlying Hommel's and Hochberg's methods, and almost identical to that underlying the Benjamini & Hochberg procedure.

The result of the method is not a single rejected set, such as the FWER and FDR methods, but rather $2^m - 1$ simultaneous confidence bounds, one for every possible subset $\mathcal{R}$ of $\mathcal{H}$.

In the Van de Vijver data, taking $\mathcal{R} = \mathcal{H}$ first, we find that with 95% confidence there are at least 640 false hypotheses among the 4919.

```r
library(hommel)
hom <- hommel(p.raw, simes = TRUE)
summary(hom)
```

```
## A hommel object for 4919 hypotheses.
## Simes inequality is assumed.
## Use p.adjust(), discoveries() or localtest() to access this object.
##
## With 0.95 confidence: at least 640 discoveries.
## 209 hypotheses with adjusted p-values below 0.05.
```

The smallest set containing at least 640 false hypotheses at this confidence level is the set of 837 hypotheses with smallest p-values. If we would reject this set, the FDP of our findings would be at most (837-640)/837 = 23.5%, with 95% confidence.

```r
ix = sort(p.raw, index.return=TRUE)$ix
fdp(hom, ix = ix[1:837])
```

```
## [1] 0.2353644
```

The connection with Hommel's method becomes obvious if we take $\mathcal{R}$ to be the set of 209 hypotheses rejected by Hommel's method. This set is the largest set for which we find $\bar{Q}(\mathcal{R}) = 0$, which coincides precisely with the FWER statement obtained from Hommel's methods, stating that with 95% confidence each of these 209 rejections is a correct one.

```r
sum(p.adjust(p.raw,"hommel") <= 0.05)
```

```
## [1] 209
```

```r
fdp(hom, ix = ix[1:209])
```

```
## [1] 0
```

The largest set with 95% confidence of an FDP of at most 0.10 is the set of 623 hypotheses with the best p-values.

```r
fdp(hom, ix = ix[1:623])
```

```
## [1] 0.09951846
```

In the Rosenwald data, with 95% confidence at least 14 out of the 38 hypotheses with best p-values are false, yielding an FDP confidence interval ranging from 0 to 63.2% for this set.

```r
p.raw2 = p_rosenwald
hom2 = hommel(p.raw2)
summary(hom2)
```

```
## A hommel object for 7399 hypotheses.
## Simes inequality is assumed.
## Use p.adjust(), discoveries() or localtest() to access this object.
##
## With 0.95 confidence: at least 14 discoveries.
## 4 hypotheses with adjusted p-values below 0.05.
```

So far we only considered sets $\mathcal{R}$ of a type consisting of the k hypotheses with best p-values. We may also, however, take some other set, perhaps partly chosen for biological reasons. For example, in the Rosenwald data we may select a set $\mathcal{R}$ consisting of the hypotheses with p-values ranked $\{2, 5, 6, 7, 8, 9\}$ and find that this set has 95% confidence of an FDP of at most 50%.

```r
ix2 = sort(p.raw2, index.return=TRUE)$ix
fdp(hom2, ix = ix2[c(2,5,6,7,8,9)])
```

```
## [1] 0.5
```

Since all confidence statements obtained from the same data set are simultaneous, they remain valid even if the researcher reviews many possibilities and finally picks a result that stands out.

Comparing the above results to FDR control by the method of Benjamini & Hochberg, which is valid under the same assumptions as the Simes-closed testing combination used by Goeman and Solari, we see that the set of 1340 hypotheses with best p-values rejected in the Van de Vijver data by Benjamini & Hochberg only gets an FDP upper confidence bound of 52.2%. This partly reflects the great variability of FDP: although on average sets selected by the Benjamini & Hochberg have an FDP of at most 5%, the variability of FDP around FDR is large, and the confidence interval for FDP for this particular set ranges from 0 to 52.2%.

```r
fdp(hom, ix = ix[1:1340])
```

```
## [1] 0.5223881
```

In the Rosenwald data, the set of 72 hypotheses selected by Benjamini & Hochberg gets a confidence interval for its FDP of 0 to 80.6%.

```r
fdp(hom2, ix = ix2[1:72])
```

```
## [1] 0.8055556
```

For both data sets, there are smaller rejected sets with much better FDP bounds, as we have seen above.

If desired, point estimates can be calculated for FDP by simply taking a 'midpoint' of the confidence interval, i.e. using $\bar{Q}(\mathcal{R})$ at $\alpha = 0.5$ as an estimate. Calculated by means of a simultaneous confidence interval, this conservative point estimate is robust to selection of $\mathcal{R}$. The estimate has the property that it overestimates the true value at most 50% of the time, even for the final selected $\mathcal{R}$. In practice, the point estimate tends to overestimate FDP.

In the van de Vijver data the set of top 837 hypotheses, which had 95% confidence of an FDP at most 23.5%, gets a point estimate for FDP of only 1.7%. Similarly, the top 38 hypotheses in the Rosenwald data get an

FDP point estimate of 5.3%, with 95% confidence interval [0%, 63.2%].

```
fdp(hom, ix=ix[1:837],alpha=0.5)
```

```
## [1] 0.0167264
```

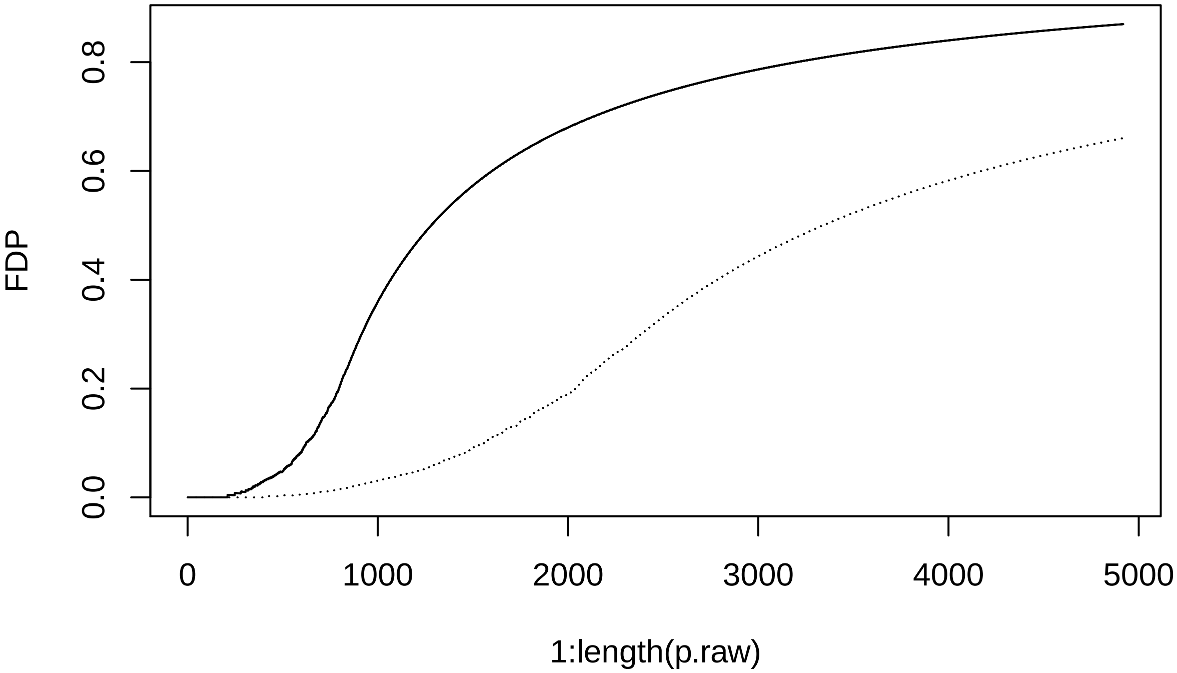
```
fdp(hom2, ix=ix2[1:38],alpha=0.5)
```

```
## [1] 0.05263158
```

The α-midpoint of the confidence interval is often much closer to 0 than the full confidence interval would seem to suggest. Point estimates, however, can augment the FDP confidence intervals, but are not sufficient by themselves as a basis for inference.

The point estimate and the 95% upper confidence bound for FDP for rejected sets consisting of the top k p-values in the Van de Vijver data:

```
upperbound = sapply(1:length(p.raw), function(i) fdp(hom, ix=ix[1:i], alpha=0.05 ) )
pointestimate = sapply(1:length(p.raw), function(i) fdp(hom, ix=ix[1:i], alpha=0.5 ) )
plot(1:length(p.raw), upperbound, type="s", ylab="FDP")
lines(1:length(p.raw), pointestimate, type="s", lty=3)
```



If the PDS assumption cannot be assumed, a variant of the FDP confidence bound method is available based on Hommel's inequality. This reduces all critical values of the Simes inequality by a factor $\sum_{k=1}^{m} 1/k$, but is valid for all dependence structures of the p-values. It relates to the Benjamini & Yekutieli method in the same way that the Simes-based method relates to the Benjamini & Hochberg method. In the Van de Vijver data, the confidence bounds arising from this method say with 95% confidence that 284 false hypotheses are present among the 385 hypotheses with best p-values, with a point estimate of FDP of 2.3%.

```
hom <- hommel(p.raw, simes = FALSE)
summary(hom)
```

```
## A hommel object for 4919 hypotheses.
## Simes inequality is not assumed.
## Use p.adjust(), discoveries() or localtest() to access this object.
##
```

```
## With 0.95 confidence: at least 284 discoveries.
## 107 hypotheses with adjusted p-values below 0.05.
```

Just like the Benjamini & Yekutieli method, the Hommel-based method of calculating FDP confidence bounds can be quite conservative, and is sometimes less powerful than a simple application of Holm's method.

The FDP confidence method takes into account variability in the FDP estimate, by providing confidence intervals rather than a point estimates. It also takes into account post hoc selection of $\mathcal{R}$, by making the FDP confidence intervals simultaneous for all $\mathcal{R}$. These two properties make the method suitable for the reversal of roles described at the beginning. After looking at the data, the user of this method can select the set $\mathcal{R}$ of rejected hypotheses freely, and be informed of the maximal proportion of false rejections made, at the chosen confidence level, when rejecting this set. On the basis of this assessment the user can revise the set, once or many times, to come up with a final set $\mathcal{R}$ with its FDP confidence bound $\bar{Q}$. The validity of the final FDP confidence interval $[0, \bar{Q}]$ is not compromised by the selection process.

FDP confidence bounds are defined for a fixed confidence level $1 - \alpha$, and therefore do not easily admit the use of adjusted p-values. Analogues of adjusted p-values can be given, but since these do not take the form of a single value, but of a whole confidence distribution, they are less straightforward to use.

Use of FDP estimation

FDP estimation methods are tailored to the exploratory setting. They allow the researcher to obtain an estimate or a confidence statement for any set $\mathcal{R}$ that he or she may choose to reject, and the option to pick one or several sets that are biologically meaningful as well as statistically reliable, without paying a price for this cherry-picking in terms of additional false positives. This seems the epitome of exploratory research, empowering researchers and giving them tools to do the selection of interesting findings, rather than taking the selection process out of their hands. FDP statements also relate directly to the actual rejected set, rather than only indirectly through a property of the procedure. Just like with FWER and FDR control methods, however, care must be employed when using these methods.

Point estimates for FDP are available using the methods of Storey or Goeman & Solari with $\alpha = 0.5$. These can be used to examine the proportion of false positives in a rejected set. However, such estimates can be no more than rough indicators without some assessment of variability. To assess this variability, confidence intervals are preferable to standard error estimates because FDP has a strong tendency toward skewness.

It should be realized that all FDP estimates relate to sets, and that FDP statements about sets do not say anything directly about individual hypotheses that are part of these sets. If a set of 100 hypotheses has an FDP of maximally 0.01 with 95% confidence, then each individual hypothesis among these 100 may or may not be a likely candidate for a false positive. Unlike with FDR, however, this problem is alleviated by the fact that FDP confidence methods do simultaneously provide confidence bounds for all subsets of the set $\mathcal{R}$, and a valid confidence statement is always available for the set of hypotheses of real interest.

The methods also allow statements to be made for individual hypotheses, by obtaining confidence statements for subsets of $\mathcal{R}$ of size 1. To completely avoid problems of overinterpretation of FDP statements about sets, however, ideally all $2^m - 1$ simultaneous confidence statements should be reported. This is impossible, of course, and the sets for which the confidence bounds are calculated and reported should be chosen carefully.

Because of the identity between the underlying methods, FDP confidence statements actually come for free with the use of Hochberg's and Hommel's methods. Using Hommel's FWER control and FDP confidence together, the FWER-controlled rejections are augmented by weaker statements for the set of hypotheses that just did not make the FWER threshold, claiming that at least a proportion of these hypotheses is false. These may be forwarded to further validation experiments if the FDP confidence bound for such a set is promising enough.

FDP confidence methods are most fruitfully used in exploratory settings if much freedom is desired in picking and choosing the set of rejected hypotheses of interest, or if more specific and more confident FDP statements are required than those provided by FDR control.