Using pwrFDR in design and analysis of a multiple testing experiment (Version 3.2.2)

Grant Izmirlian

2024-12-18

The package, `pwrFDR`, is for computing Average and TPX Power under various sequential multiple testing procedures such as the Benjamini-Hochberg False Discovery Rate (BH-FDR) procedure. Before we begin, some review of multiple testing and sequential procedures is in order. Consider a multiple testing experiment with m simultaneous tests of hypotheses. The most widely used multiple testing procedure is Bonferroni's procedure which guarantees control of the family-wise error rate (FWER) which is the probability of one or more false positives. It is applied by referring all p-values to the common threshold α/m. The Benjamini-Hochberg procedure guarantees control of the false discovery rate (FDR). Since it gained widespread use in the early 2000's, most practitioners are at least vaguely familiar with the notion that the target of protected inference is different for the Bonferroni (FWER) and the BH-FDR (FDR) procedures. The domain of application and in particular the cost of a false positive guides the choice of the target for protected inference, with higher costs (drug development) requiring a more conservative target of control, and lower costs (thresholding in –omics studies) allowing for a less conservative target of control. Application of a sequential procedure in a multiple testing experiment (MTE) usually begins with ordering the m p-values from smallest to largest and then comparing each sorted p-value with a corresponding member of a sequnce of criterion values. This sequence of criterion values, also a non-decreasing sequence and specific to the particular procedure, is the product of α and a multiple testing penalty. All procedures begin with marking rows for which the sorted p-value is less than its corresponding criterion value.

Sequential procedures differ in two main features. First is the choice of the sequence of criterion values, and secondly, by whether the procedure is step-up or step-down. This latter distinction provides a recipe for calling tests significant based upon marked/unmarked rows of p-value and criterion pairs. A step-up procedure calls significant all tests up until the last marked row. A step-down procedure calls significant tests belonging to a block of contiguous marked rows beginning with the first. If the first row is not marked,

a step-down procedure calls nothing significant.

We now discuss the number of significant calls and of these which are true positives and which are false positives. Let R denote the number of tests called significant by the procedure. This partitions into the unobserved false positive count, V, e.g. the number of tests called significant which are distributed as the null, and unobserved true positive count, T, e.g. the number of tests called significant which are distributed as the alternative, $V + T = R$. The ratio, FDP $= V/R$ is called the false discovery proportion and the ratio, TPP $= T/M$ is called true positive proportion. Here M is the number of statistics distributed as the alternative (more on this below). Within the fairly broad scope of sequential procedures considered here the goal of protected multiple inference will be to control some summary of the false discovery proportion distribution: $\mathbb{P}\{\text{FDP} > x\} = \mathbb{P}\{V/R > x\}$. Protected inference must be done within the context of some definition of multiple test or aggregate power so that multiple testing experiments can be sized and so that we have some idea of the probability of success as defined appropriately for the application. We will consider definitions of aggregate power based upon some summary of the true positive proportion distribution: $\mathbb{P}\{\text{TPP} > x\} = \mathbb{P}\{T/M > x\}$.

The BH-FDR procedure is a step-up procedure with criterion sequence $\alpha j/m$. It guarantees control of the FDR, which is the expected FDP:

$$\text{FDR} = \mathbb{E}[\text{FDP}] = \mathbb{E}[V/R]$$

The type of aggregate power usually used in conjunction with the BH-FDR procedure is the average power. It is the expected TPP:

$$\text{AvgPwr} = \mathbb{E}[\text{TPP}] = \mathbb{E}[T/M]$$

Let's begin by computing the sample size required for 80% average power under the BH-FDR procedure at FDR = 15% when the effect size is 0.79. There is one more parameter required for calculation of sample size for multiple test power besides the usual required power, type I error and effect size which are sufficient to calculate the sample size in the single testing case. Whereas in the single test case, we condition upon the statistic being drawn from the null or the alternative, in the multiple testing case we must somehow make a specification regarding the number of tests distributed as the alternative. We handle this by posing the mixture model as the common distribution of the test statistics. Under the mixture model, the population from which each test statistic is drawn is determined via an a priori density r_1 coin flip per test statistic, the value 1 signifying the alternative distribution. This is the additional parameter which must be specified.

In applications, a reasonable working value is drawn from substance experts. Let us assume this is 5%, the value typically used in larger –omics studies like mRNA profiling and RNAseq ([1, 2]).

In order to do this we load the `pwrFDR` library as well as the `ggplot2` and `TableMonster` libraries. The latter two libraries are for plotting and for easy generation of nice looking latex tables.

```
> library(pwrFDR)
> library(ggplot2)
> library(TableMonster)
```

You can use this vignette file to follow along or if you prefer, open the companion script file (all supporting text removed) at `/usr/local/lib/R/site-library/pwrFDR/doc/pwrFDR-vignette.R`.

We are now ready to call `pwrFDR` to calculate sample size required for 80% average power under the BH-FDR procedure at $\alpha = 0.15$ and above mentioned effect size and prior probability:

```
> avgpwr.fdr.r05 <- pwrFDR(effect.size=0.79, alpha=0.15, r.1=0.05, average.power=0.80)
```

Notice that we did not specify the number of tests. This calculates the infinite tests limit which exists for procedures controlling the FDR and for procedures controlling the FDX, but not for procedures controlling the family-wise error rate (FWER). While we're at it, in order to see how much the alternative hypothesis prior probability, r_1 affects the required sample size, let's calculate sample size required for 80% average power under BH-FDR at $\alpha = 0.15$ under the above settings ammended to incorporate a higher prior probability, $r.1 = 0.10$.

```
> avgpwr.fdr.r10 <- update(avgpwr.fdr.r05, r.1=0.10)
```

The following line generates a publication ready table.

```
> print(avgpwr.fdr.r05, label="tbl:minf", result="tex", cptn="$m=\\infty$")
```

or we can join the two tables into one, also adding a caption

```
> print(join.tbl(avgpwr.fdr.r05, avgpwr.fdr.r10), label="tbl:minf-r05-r10",
+                result="tex", cptn="$m=\\infty, r_1=0.05, 0.10$")
```

From the third and seventh lines in table 1 we can see that sample sizes of 42 and 37 are required for 80% average power under BH-FDR at $\alpha = 0.15$ when the effect size is 0.79 and the prior probabilities are $r_1 = 5\%$

Parameter	result a	result b
N.tests	Inf	Inf
r.1	0.05	0.1
n.sample	42	37
effect.size	0.79	0.79
alpha	0.15	0.15
FDP.cnt	BHFDR	BHFDR
average.power	0.8	0.7999
gamma	0.0466	0.0925
sigma.rtm.Rom	0.2806	0.3863
sigma.rtm.VoR	1.742	1.201
sigma.rtm.ToM	2.111	1.513

Table 1: $m = \infty, r_1 = 0.05, 0.10$

and 10% respectively. The meaning of the other entries in the table are as follows. The first line shows that the sample size was calculated using the infinite tests limit since no value of m (`N.tests` in the routine) was specified. The second, forth, fifth, and seventh rows show values of the user specified parameters, r_1, the effect size, α and the average power. The sixth row indicates that the default method of FDP control, "BHFDR" was used as there was no user specified value. The eighth row shows the value of the rejection rate or positive rate, which is the inifinite tests consistent limit of the proportion of positive calls, R/m. The bottom three rows show the asymptotic standard deviations for the rejection proportion, R/m, the false discovery proportion, V/R and the true positive proportion, T/M. We shall see why it is useful to know these below.

In any case we can always use simulation. In this case we must specify the number of tests. The simulation method will not find sample size required for specifed power, so we must also specify the sample size instead and compute the power (average power in this case). The simulation routine generates replicate data-sets, each containing m full data records, each consisting of a population indicator (bernouli, probability r_1), test statistic distributed under the null or alternative corresponding to the value of the population indicator, and corresponding p-values. For each simulation replicate the requested procedure is applied to the m test statistics, and then the numbers of rejected tests, R, and true positives, T, are recorded. The number of statistics distributed as the alternative, M, is also recorded. Of course the number of false positives isn't recorded because it can be found via subtraction: $V = R - T$. These per simulation replicate statistics are in the `reps` component of the `detail` attribute which is obtained in this setting via the expression `detail(avgpwr.fdr.sim.r05.m1e5)$reps`. In the following code-block, we call the simulation option with 10,000 tests at a sample size of 42. The first line of code calls `pwrFDR` at the previous parameter settings in simulation mode. The second line calculates the empirical FDR as the mean of the FDP divided by

$(1 - r_1) = 0.95$. Note that it is actually an slight abuse of nomenclature that we refer to *both* the expected value of V/R *and* α as the false discovery rate, even though the former is in fact $(1-r_1)\alpha$. Notice that we use the operator `%over%` instead of the ordinary division operator, `/`, since, when it is applied component-wise, any occurrences of 0/0 are treated as 0.

```
> avgpwr.fdr.sim.r05.m1e5 <- pwrFDR(effect.size=0.79, alpha=0.15, r.1=0.05,
+                                   n.sample=avgpwr.fdr.r05$n.sample, N.tests=10000,
+                                     meth="sim")
> avgpwr.fdr.sim.r05.m1e3 <- update(avgpwr.fdr.sim.r05.m1e5, N.tests=1000)
> avgpwr.fdr.sim.r05.m100 <- update(avgpwr.fdr.sim.r05.m1e5, N.tests=100)
> avgpwr.fdr.r10.sim.m100 <- update(avgpwr.fdr.r10, n.sample=avgpwr.fdr.r10$n.sample,
+                                     average.power=NULL, method="sim", N.tests=100)
```

Parameter	result a	result b	result c	result d
N.tests	10000	1000	100	100
r.1	0.05	0.05	0.05	0.1
n.sample	42	42	42	37
effect.size	0.79	0.79	0.79	0.79
alpha	0.15	0.15	0.15	0.15
emp.FDR	0.1416	0.141	0.1378	0.1275
FDP.cnt	BHFDR	BHFDR	BHFDR	BHFDR
average.power	0.8008	0.7983	0.7865	0.8039
gamma	0.0467	0.0472	0.048	0.0938
se.Rom	0.0028	0.0091	0.0277	0.0361
se.VoR	0.017	0.0532	0.1754	0.1141
se.ToM	0.0209	0.0677	0.243	0.1569

Table 2: Results of simulation calls with varying 'm' and 'r.1'.

Next, looking at the first three columns of table 2, we see that passing from 10000, to 1000, to 100 simultaneous tests changes nothing regarding the sample size required for average power 80.08%. This is because the average power is independent of the number of tests and is in fact the infinite tests limit of the true positive proportion. The empirical FDR also is the same, at least to within simulation error. The only values which change are the standard errors of the positive proportion, false discovery proportion and true positive proportion, as these are of order one over the square root of number of tests. Note that these empirical standard errors times $\sqrt{m}$, e.g. the square root of `N.tests` as shown in the table, aggree well with their asymptotic values shown in the table 2. The final column which was run with identical parameters except for $r_1 = 0.10$ for $m = 100$ simultaneous tests shows a smaller sample size required for 80% average power, smaller emprical FDR and nearly twice as large rejection fraction, γ. This makes sense because there twice

as many statistics are expected to be distributed as the null.

So judging by the results shown in table 2 alone, it seems that the BH-FDR procedure controls the FDR, no matter the number of test statistics, just as stated in the results proved by Benjamini and Hochberg. Lets pay special attention to what is being controlled. As we mentioned previously, the FDR is the expected proportion of false discoveries, $\mathbb{E}[\text{FDP}] = \mathbb{E}[V/R]$. And above, in the table we corroborate that the empirical false discovery rate (eFDR) is indeed less than or equal to the nominal value. The eFDR is the average over 1000 multiple testing experiments defined by the parameters in the calling sequence. What we are in fact guaranteed of controlling is an average value over many identical multiple testing experiments. This average itself is only meaningful if the distribution of the FDP is tightly distributed above its mean, as is the case with several thousand simultaneous tests, $m = 10000$. If the FDP is not tightly distributed above its mean, the FDR says little to nothing about the FDP for any one given multiple testing experiment.

References

[1] Alizadeh AA and Eisen MB and Davis RE and Ma C and Lossos IS and Rosenwald A and Boldrick JC and Sabet H and Tran T and Yu X and Powell JI and Yang L and Marti GE and Moore T and Hudson J Jr and Lu L and Lewis DB and Tibshirani R and Sherlock G and Chan WC and Greiner TC and Weisenburger DD and Armitage JO and Warnke R and Levy R and Wilson W and Grever MR and Byrd JC and Botstein D and Brown PO and Staudt LM. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

[2] Mortazavi A and Williams BA and McCue K and Schaeffer L and Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*, 5:1–8, 2008.

List of Figures

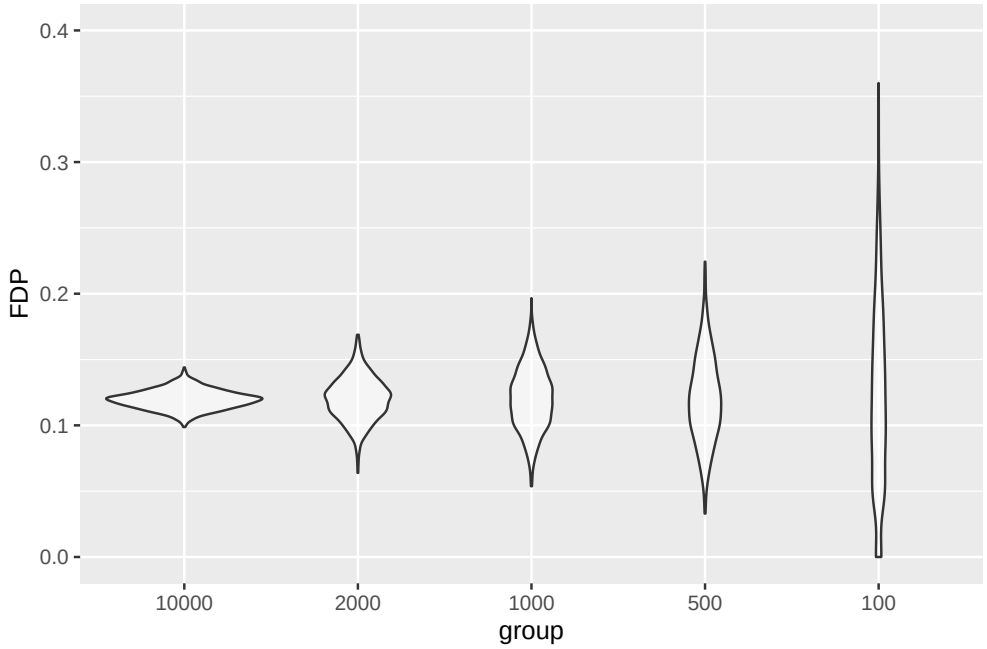

Figure 1: Violin plots of FDP distribution for numbers of simultaneous tests varying from 10,000 down to 100, effect.size=0.79, n.sample=47, r.1=0.20, alpha=0.15

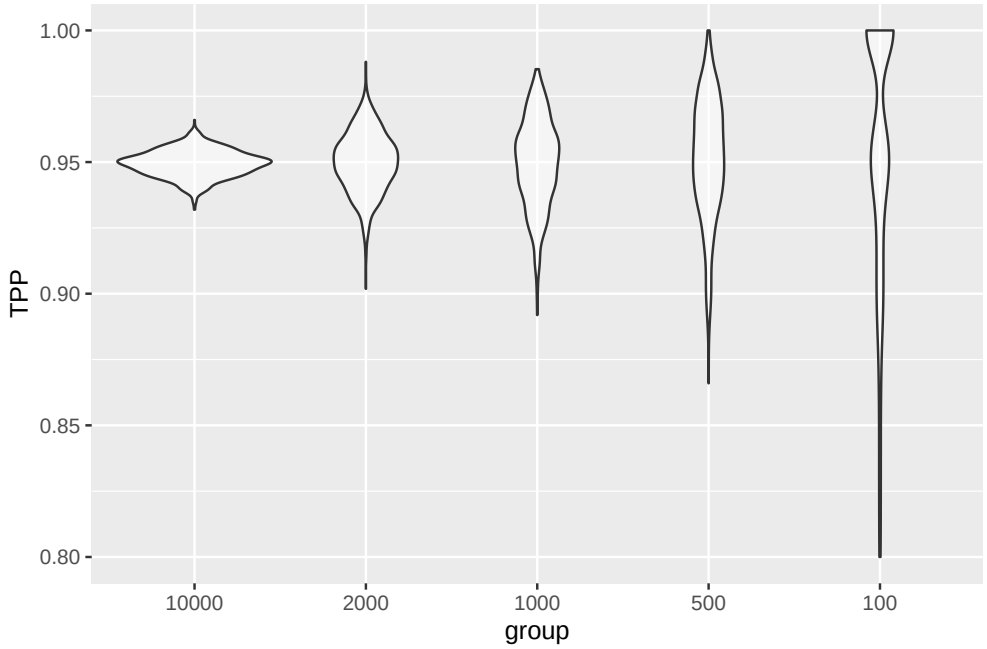

Figure 2: Violin plots of TPP distribution for numbers of simultaneous tests varying from 10,000 down to 100, effect.size=0.79, n.sample=47, r.1=0.20, alpha=0.15