Using pwrFDR in design and analysis of a multiple testing experiment (Version 3.2.4)

Grant Izmirlian

2025-02-13

1 Introduction

The package `pwrFDR` allows computation of sample size required for average power or for TPX Power under various sequential multiple testing procedures such as the Benjamini-Hochberg False Discovery Rate (BH-FDR) procedure. Before we begin, we first load some libraries and then provide a brief review of multiple testing and sequential procedures.

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

In addition to the `pwrFDR` library, we load `ggplot2`, an advanced plotting package many readers will be familiar with, and `TableMonster`, and easy to use frontend to `xtable` for generating publication quality tables in LaTeX.

In short, statistical hypothesis testing whereby a single p-value is compared to a threshhold value, α, to determine statistical significance assures that the resulting conclusion has false positive rate less than α. This guarantee applies in the context of a single statistical hypothesis test only. Adjustment for multiple tests of hypotheses provides an algorithm whereby the comparison thresholds are adjusted so that some aggregate false positive rate is guaranteed. We will review several multiple testing procedures and discuss the aggregate false positive rate or target of *protected inference* which each one controls. Consider a multiple testing experiment with m simultaneous tests of hypotheses. The most widely used multiple testing procedure is Bonferroni's [3] procedure which guarantees control of the family-wise error rate (FWER). This is the probability that of one or more of the hypothesis tests results in a false positive. It is applied by referring all p-values to the common threshold α/m. This quickly becomes overly conservative as the number of tests, m, becomes larger. The Benjamini-Hochberg [2] procedure guarantees control of the false discovery rate (FDR). This is the expected proportion of hypothesis tests declared significant which are false positives. It is applied by sorting p-values in increasing order, comparing the i^{th} largest to $\alpha i/m$ and then declaring all tests statistically significant which correspond to p-values not larger than the largest exceeded by its threshold. Thus the Bonferroni procedure and the BH-FDR procedure control different targets of protected inference. 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 such as the FWE rate, and lower costs (thresholding in –omics studies) allowing for a less conservative target of control, such as the FDR.

We now discuss sequential multiple testing procedures in general. The application of a sequential procedure in a multiple testing experiment 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 sequence of multiple testing penalties, $\psi_m(j)$. All procedures begin with marking rows for which the sorted p-value is less than its corresponding criterion value.

Sequential procedures are defined by two distinguishing features which in turn, provide a recipe for their application. First is the chosen sequence of multiple testing penalties, and second, 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. Table 1 below shows p-values for 10 simultaneous tests of hypotheses, and a sequence of threshold criterion values, $\alpha i/m$, with $\alpha = 0.05$. This sequnce of threshold criterion values should be familiar as it is the one used in the BH-FDR procedure. Also shown in the table is an indicator of whether or not each p-value is less than or equal to its corresponding threshold value. A step up procedure based upon the given criterion sequence will call statistical tests corresponding to the smallest 4 p-values significant. Notice that this includes the third smallest p-value which was not smaller than its threshold value. A step-down procedure based upon the given criterion sequence would call only the first two tests 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. As mentioned above in our

xi	X	P	crit	Marked
1	3.931	0.000	0.005	1
1	2.744	0.007	0.010	1
1	2.414	0.018	0.015	0
1	2.385	0.019	0.020	1
1	2.232	0.028	0.025	0
0	-1.904	0.060	0.030	0
0	1.124	0.264	0.035	0
0	-1.007	0.317	0.040	0
0	0.933	0.353	0.045	0
0	0.901	0.370	0.050	0

Table 1: Sorted p-values, their test statistics, population indicators, and BH-FDR threshold in 10 simultaneous tests

example, table 1 above, $R = 4$ under the step-up procedure and $R = 2$ under the step-down 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, $V + T = R$. We see that $V = 0$ under both the step-up and step-down procedures, while $T = 3$ under the step-up procedure and $T = 2$ under the step-down procedure. 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). We see that FDP $= 0$ under both the step-up and step-down procedures, while TPP $= 3/5$ under the step-up procedure, and TPP $= 2/5$ under the step-down procedure. We note in passing that the BH-FDR procedure is the step-up procedure based upon the given criterion sequence.

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: $\mathbb{P}\{\text{TPP} > x\} = \mathbb{P}\{T/M > x\}$.

As previously noted, the BH-FDR procedure is a step-up procedure with multiple testing penalty sequence $\psi_m(j) = j/m$. 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. Our methodology assumes a common mixture distribution for the p-value CDF. This means that each test statistic is distributed according to the alternative hypothesis with probability, r_1, and distributed according to the null hypothesis with the complementary probability. 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, 5]). The last argument, which was not specified, `FDP.control.method`, takes its default value, `"BHFDR"`, as we here desire.

You can use this vignette file to follow along or if you prefer, open the companion script file (all supporting text removed) at `/usr/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:

```
> ss.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. The calculation is done using the infinite tests consistent limit approximation. This consistent limit exists for procedures controlling the FDR and for procedures controlling the FDX, but not for procedures controlling the family-wise error rate (FWER).

As is the case with the base R library supplied power functions like `power.t.test`, the routine will solve for any missing parameter, except in this case, α must be specified. This means that the routine will calculate the average power or TPX power (see below) under the specified procedure at the given alpha at the specified effect size and prior probility. It will also find the required sample size, effect size, or prior probability required for specified average power or TPX power for given values of the other parameters. For example the following 4 lines of code return essentially the same result, but calculate, in order listed, the average power, the sample size required for average power, the effect size required for average power, and the prior probability required for average power, respectively, for given values of the other parameters.

```
> find.avp <- pwrFDR(effect.size=0.79, alpha=0.15, n.sample=ss.fdr.r05$n.sample, r.1=0.05)
> find.ss <- pwrFDR(effect.size=0.79, alpha=0.15, r.1=0.05, average.power=0.80)
> find.es <- pwrFDR(alpha=0.15, n.sample=ss.fdr.r05$n.sample, r.1=0.05, average.power=0.80)
> find.r1 <- pwrFDR(effect.size=0.79, alpha=0.15, n.sample=ss.fdr.r05$n.sample, average.power=0.80)
```

The point is that of the four parameters, desired power (average or TPX), effect size, sample size and prior probability, the user must specify α together with three of these and the missing one will be calculated. See the help documentation for more information.

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, $r.1 = 0.10$.

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

The following line generates a publication ready table.

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

The first six lines are user specified parameters or default values, and the seventh through tenth lines are calculated by the function. As for the last two lines, sample size and power, as is usually the case, one is specified and the other is calculated. The first line indicates that calculations were done according to the theoretical method, which in the case of average power under FDR control is the infinite tests consistent limit approximation. Lines 2 and 4 are here default values. The default method of FDP control is `"BHFDR"` as mentioned above and the value `Inf` for `N.tests` signifies the infinite tests consistent limit is being used, and in this case specification of `N.tests` is not required. Not to belabor an obvious point, but this means that all quantities derived are independent of the number of simultaneous tests. We shall discuss when and how this assumption breaks down below. Lines 3, 5, and 6, being α, r_1 and `effect.size`, respectively, are user specified as discussed above. Recall that result a and b differ only by the specified value for prior

Parameter	result a	result b
method	Theoretical	Theoretical
FDP.control	BHFDR	BHFDR
alpha	0.15	0.15
N.tests	Inf	Inf
r.1	0.05	0.1
effect.size	0.79	0.79
gamma	0.0466	0.0925
sigma.rtm.Rom	0.281	0.386
sigma.rtm.VoR	1.74	1.2
sigma.rtm.ToM	2.11	1.51
n.sample	42	37
average.power	0.8	0.8

Table 2: $m = \infty, r_1 = 0.05, 0.10$

probability, r_1, being 0.05 and 0.10, respectively. The first of the derived values, γ, on line 7, is the infinite tests consistent limit of R/m. This limit is the rejection rate, or expected proportion of all tests which are declared significant. The next three lines are the asymptotic standard deviations of the rejection proportion, R/m, the false discovery proportion, V/R, and the true positive proportion, T/M. We will see below why it is useful to know these. Lines 11 and 12 are the sample size and average power, respectively. In this case we specified average power and the function calculated required sample size given the other parameter values. Comparing results "a" and "b" it is clear that doubling the prior probability, r_1, results in roughly a doubling of the rejection rate, γ, and roughly three quarters the sample size required for 80% average power.

The package provides a simulation method as a check on the variety of theoretical methods used. In this case we must specify the number of tests. The simulation method will not find sample size required for specifed power as it is impractical given the use of a back-solver resulting in more than 20 calls to the function. Thus we must instead request a computation of power (average power in this case) given specified sample size. The simulation routine generates replicate data-sets, each containing m full data records, where each of these consist of a population indicator (bernouli, probability r_1), a test statistic distributed under the alternative or null corresponding to the value of the population indicator, and a corresponding p-value. For each simulation replicate dataset, 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 need not be recorded as it can be found via subtraction: $V = R - T$. These per simulation replicate statistics are in the `reps` component of the `detail` attribute, which can be obtained for given a `pwrFDR` object, `result`, via the expression `detail(result)[["reps"]]`. In the following code block we call `pwrFDR` via the `"simulation"` method at the parameter settings used in `ss.fdr.r10` above when the number of simultaneous tests is 10,000, 1,000, and 100, respectively in the three following lines of code.

```
> avgpwr.fdr.sim.r10.m1e5 <- pwrFDR(effect.size=0.79, alpha=0.15, r.1=0.10,
+                                   n.sample=ss.fdr.r10$n.sample, N.tests=10000,
+                                   meth="sim")
> avgpwr.fdr.sim.r10.m1e3 <- update(avgpwr.fdr.sim.r10.m1e5, N.tests=1000)
> avgpwr.fdr.sim.r10.m100 <- update(avgpwr.fdr.sim.r10.m1e5, N.tests=100)
```

Parameter	result a	result b	result c
method	Simulation	Simulation	Simulation
FDP.control	BHFDR	BHFDR	BHFDR
alpha	0.15	0.15	0.15
N.tests	10000	1000	100
r.1	0.1	0.1	0.1
effect.size	0.79	0.79	0.79
emp.FDR	0.135	0.136	0.138
emp.FDX	0.108	0.338	0.435
gamma	0.0935	0.0939	0.0959
se.Rom	0.00383	0.012	0.0379
se.VoR	0.0121	0.0365	0.122
se.ToM	0.0145	0.046	0.157
n.sample	37	37	37
average.power	0.808	0.807	0.801

Table 3: Results of simulation calls with varying ‘m’.

Comparing the simulation results in each of the three columns in table 3 with the theoretical approximation at the same design parameters in the second column of table 2, the only differences in derived results beyond that expected from simulation error are the designation that the `"Simulation"` method was used, what appear to be standard errors of the rejection proportion, false discovery proportion and true positive proportion as opposed to asymptotic standard deviations, and appearance of two new derived quantities, `emp.FDR` and `emp.FDX`. First, when the number of tests, `N.tests`, is specified, the function returns estimated standard errors instead of asymptotic standard deviations, these being the latter divided by the square root of the number of tests. Secondly, the two new derived quantities are the empirical FDR and FDX derived as simulation estimates. The latter estimates $\mathbb{P}\{\text{FDP} > \alpha\}$, the probability that the false discovery proportion exceeds α. Notice that the empirical FDR’s corresponding to the differing numbers of simultaneous tests are identical to within simulation error, but the standard error of the false discovery proportion, as well as those of the other two ratios increase in proportion to the ratio of the square root of number of tests. While the location, i.e. the mean of the FDP distribution remains more or less constant as the number of tests decreases from 10,000 to 100, the width of the distribution grows. The point is that the BH-FDR procedure guarantees that the mean of the FDP will be less than α, but not what the width of the distribution will be. BH-FDR control means that if a multiple testing experiment is repeated then the FDP's corresponding to each experiment will have average value less than α. Per experiment values of FDP's may vary wildly and in fact, have high probability of being unacceptably large. Here we see that while for 10,000 tests, the probabiliy that the FDP exceeds α is roughly 11%, it becomes quite large as the number of tests decreases, being 34%, 44% when the number of tests is 1,000 and 100, respectively. This raises the question as to whether BH-FDR control is appropriate for a moderate to small number of simultaneous tests.

2 Caveats Arising from FDR control and Use of Average Power

This point regarding the appropriateness of BH-FDR control for a moderate to small number of simultaneous tests is made clearer by having a look at the distribution of the FDP as the number of tests decreases. We will re-run the above simulations for 6 multiple testing experiments at the same design parameter settings when the number of tests is 10,000, 2,000, 1,000, 500, 250, and 100, respectively.

```
> ss <- pwrFDR(effect.size = 0.79, average.power=0.80, r.1 = 0.10, alpha = 0.15)
> avgp <- update(ss, average.power=NULL, n.sample=ss$n.sample)
```

	10000	2000	1000	500	250	100
$P(\text{FDP} \geq 0.20)$	0	0.006	0.047	0.107	0.207	0.334
$P(\text{TPP} < 0.70)$	0	0.003	0.021	0.058	0.144	0.2

Table 4: Simulation estimates of indicated probabilities of FDX and TPX for indicated values of m when $\alpha = 0.15, r_1 = 0.20$, effect size 0.79 with average power 80%

A sample of 37 is required for 79.99% average power. Figure 1 below shows violin plots of the FDP distribution for varying values of m from 10,000 down to 100 when the FDR is 15% and the other parameters are as indicated above. It is clear that the spread of the FDP distribution goes from very narrow to very disperse. In the more dispersed cases for 250 and 100 tests, it is clear that controlling the mean of the FDP distribution offers little assurance as to the value of the FDP. Table 4 shows the probability that the FDP exceeds 20% for each of the indicated values of m. At the most extreme level of dispersion when $m = 100$, the probability that the FDP exceeds 20% is roughly 20%. It is easy to be lulled into a sense that FDR control at 15% means that the FDP will be less than 15% but here is probability 133% of this value that it exceeds a value 133% of the target. At the larger numbers of tests, the simulation estimate of exceedence probability is zero, suggesting that FDR control is a good indication that the FDP is controlled when the number of tests is larger.

Similar caveats arise from the use of the average power to define the power for a multiple testing experiment. The average power is the expected value of the TPP, which can be thought of as the average TPP over many identical multiple testing experiments. That a given sample size guarantees average power says very little about what the TPP will be for any one given multiple testing experiment. Figure 2 below shows violin plots of the TPP distribution for varying values of m from 10,000 down to 100 when the average power is 80% and the other parameters are as indicated above. Once again, it is clear that the spread of the TPP distribution goes from very narrow to very disperse. In the more dispersed cases for 250 and 100 tests, it is clear that controlling the mean of the TPP distribution offers little assurance as to the value of the TPP. Table 4 shows the probability that the TPP is less than 70% for each of the indicated values of m. At the most extreme level of dispersion when $m = 100$, the probability that the TPP is less than 80% is roughly 15%. It is easy to be lulled into a sense that a sample size required for average power 80% means that the TPP will 80% but here is probability 15% that the TPP is less than 70%. At the larger numbers of tests, the simulation estimate of the probability that the TPP is less than 70% is zero, suggesting that the use of average power to size a multiple testing experiment will result in an equally high TPP when the number of tests is large.

3 FDX control and the TPX Power

When the FDP distribution is too dispersed as we saw above in the case of only several hundred tests, a more reliable method of controlling the value of the FDP is to control the probability that the FDP exceeds a given threshold, $\mathbb{P}\{\text{FDP} > \delta\} \leq \alpha$, known as FDX control. A procedure due to Lehmann, Romano and Shaikh, [4, 6], controls the FDX. It is a step-down procedure with multiple testing penalty sequence,

$$\psi_m(j;\delta) = \frac{1 + \lfloor \delta j \rfloor}{m + 1 + \lfloor \delta j \rfloor - j} \tag{1}$$

Lets compute the sample size required for 80% average power under the Lehmann-Romano-Shaikh procedure when $\alpha = 0.15$. This call is exactly the same as the very first sample size we computed above except that we specify `FDP.control.method="Romano"` to override its default value, “FDR” .

```
> ss.Rom <- pwrFDR(effect.size = 0.79, average.power=0.80, r.1 = 0.20, alpha = 0.15,
+                  FDP.control.method="Romano")
```

The following table was generated

Parameter	result a	result b	result c
method	Simulation	Simulation	Simulation
FDP.control	Romano	Romano	Romano
alpha	0.15	0.15	0.15
delta	0.15	0.15	0.15
N.tests	500	250	100
r.1	0.2	0.2	0.2
effect.size	0.79	0.79	0.79
emp.FDR	0.0221	0.0224	0.0244
emp.FDX	0	0	0.001
gamma	0.165	0.166	0.171
se.Rom	0.0195	0.0265	0.0453
se.VoR	0.0164	0.0231	0.0364
se.ToM	0.0461	0.0637	0.1
n.sample	46	46	46
average.power	0.806	0.806	0.821

Table 5:

```
> ss.BHFDX.500 <- pwrFDR(effect.size = 0.79, average.power=0.80, r.1 = 0.20, alpha = 0.15,
+                         FDP.control.method="BHFDX", N.tests=500)
> ss.BHFDX.250 <- update(ss.BHFDX.500, N.tests=250)
> ss.BHFDX.100 <- update(ss.BHFDX.500, N.tests=100)
> avgp.BHFDX.500.sim <- update(ss.BHFDX.500, n.sample=ss.BHFDX.500$n.sample, average.power=NULL,
+                               method="sim")
> avgp.BHFDX.250.sim <- update(ss.BHFDX.250, n.sample=ss.BHFDX.250$n.sample, average.power=NULL,
+                               method="sim")
> avgp.BHFDX.100.sim <- update(ss.BHFDX.100, n.sample=ss.BHFDX.100$n.sample, average.power=NULL,
+                               method="sim")
```

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.

Parameter	result a	result b	result c
method	Simulation	Simulation	Simulation
FDP.control	BHFDX	BHFDX	BHFDX
alpha	0.15	0.15	0.15
delta	0.15	0.15	0.15
N.tests	500	250	100
r.1	0.2	0.2	0.2
effect.size	0.79	0.79	0.79
alpha.star	0.142	0.127	0.1
emp.FDR	0.112	0.104	0.0781
emp.FDX	0.135	0.158	0.153
gamma	0.182	0.179	0.177
se.Rom	0.0228	0.0311	0.047
se.VoR	0.0357	0.0467	0.0679
se.ToM	0.0478	0.0642	0.107
n.sample	32	33	35
average.power	0.808	0.806	0.81

Table 6:

[2] Benjamini, Y and Hochberg, Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J Roy Stat Soc B*, 57(1):289–300, 1995.

[3] Bonferroni, CE. Teoria statistica delle classi e calcolo delle probabilità. Technical report, R Ist Sup Sci Econ Com Fir, 1936.

[4] EL Lehmann and JP Romano. Generalizations of the familywise error rate. *Ann Stat*, 33(3):1138–1154, 2005.

[5] 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.

[6] JP Romano and AM Shaikh. Stepup procedures for control of generalizations of the familywise error rate. *Ann Stat*, 34(4):1850–1873, 2006.

List of Figures

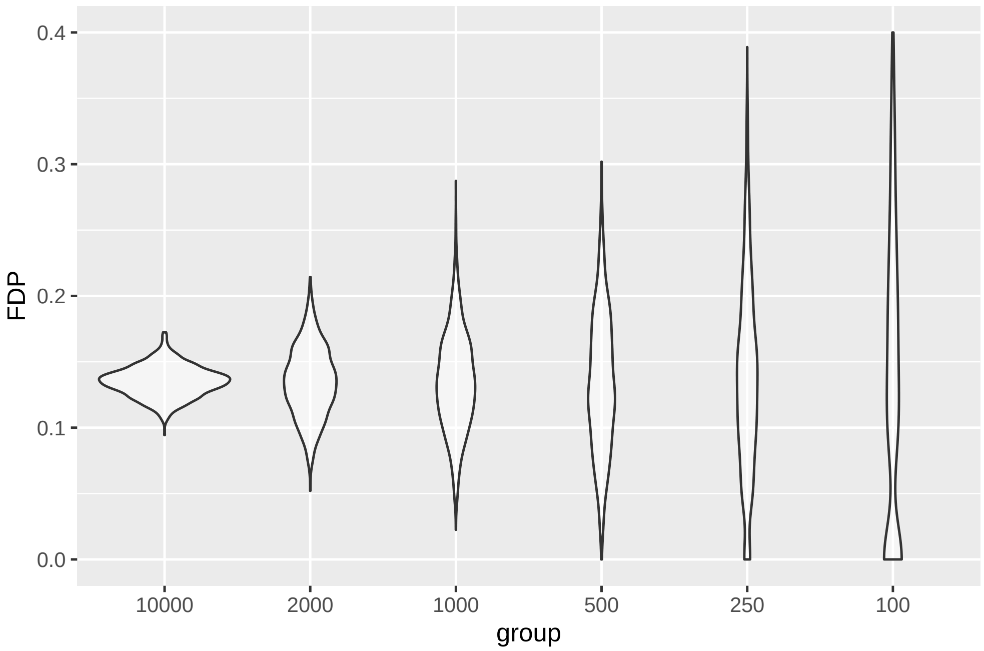

Figure 1: Violin plots of FDP distribution for numbers of simultaneous tests varying from 10,000 down to 100, effect.size=0.79, n.sample=47, r.1=0.20, alpha=0.15

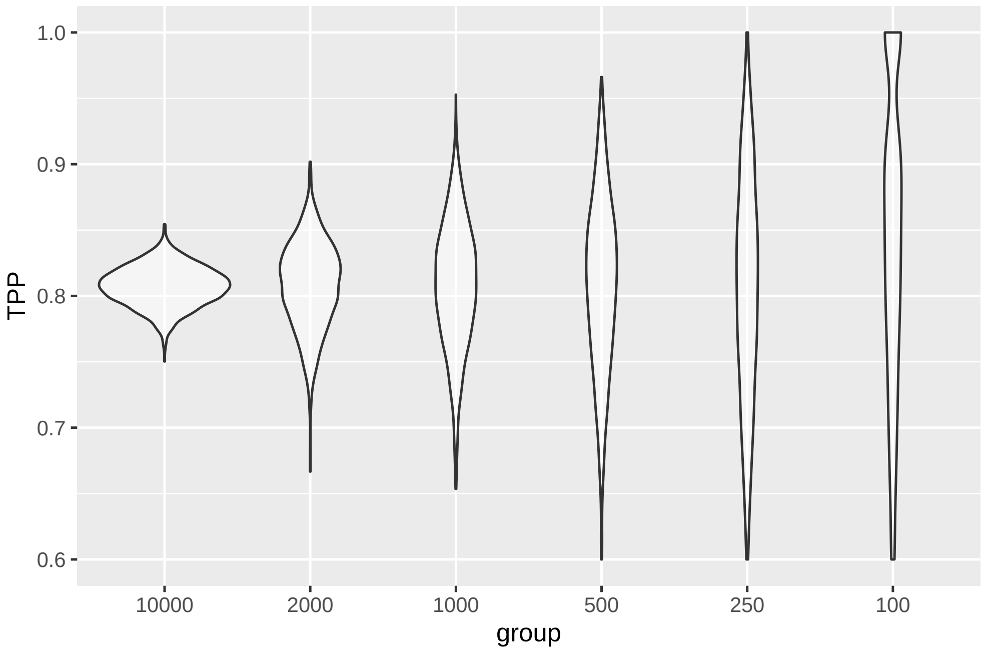

Figure 2: Violin plots of TPP distribution for numbers of simultaneous tests varying from 10,000 down to 100, effect.size=0.79, n.sample=47, r.1=0.20, alpha=0.15