HDS 3. Q-value

Matti Pirinen, University of Helsinki

21.10.2021

So far we have used P-values for inference based on false positive rate (this is the definition of P-value itself), on family-wise error rate (FWER; Bonferroni and Holm) and on false discovery rate (FDR; Benjamini-Hochberg and Benjamini-Yekutieli). See slides.

Here we will consider a quantity called Q-value that can be attached to each test and that gives an empirical estimate of FDR among all tests whose Q-values are smaller or equal to the given Q-value. We will follow Storey and Tibshirani (2003, PNAS 100: 9440-9445), whose work led to `qvalue` R-package. Before defining Q-value we will first refine BH method by empirically estimating p_0, the number of null tests. Remember that the proof of BH theorem showed that BH(α_F) method actually controls FDR at level $p_0/p \cdot \alpha_F$. When p_0 is considerably smaller than p, we can improve the accuracy of the FDR method by estimating p_0.

Let's define, for each P-value threshold $t \in [0,1]$,

$$\text{FDR}(t) = \text{E}\left(\frac{\text{FD}(t)}{D(t)}\right),$$

where random variables $\text{FD}(t) = \#\{\text{null P-values} \leq t\}$ and $D(t) = \#\{\text{P-values} \leq t\}$ in an experiment where in total p P-values are available.

To refine our understanding of FDR methods, let's think that random variables $\text{FD}(t)$ and $D(t)$ result from p draws of P-values from a mixture distribution between Uniform(0,1) (for null P-values) and an alternative distribution with cdf Φ_1 and pdf ϕ_1 (for non-null P-values), with mixture proportion π_0 for the null. In other words, cdf Φ and pdf ϕ of the P-values are

$$\Phi(t) = \pi_0 \cdot t + (1-\pi_0)\Phi_1(t), \qquad t \in [0,1], \tag{1}$$
$$\phi(t) = \pi_0 \cdot 1 + (1-\pi_0)\phi_1(t), \qquad t \in [0,1]. \tag{2}$$

We can interpret sampling from such a mixture distribution as a 2-step process. Namely, we first choose between the null distribution (with probability π_0) and the alternative distribution (with probability $1-\pi_0$), and second, conditional on the chosen distribution, we sample a P-value from the chosen distribution.

Example 3.1. Suppose we do $p = 10000$ tests of which $m = 2000$ are non-null, i.e., we can model this as a mixture distribution with $\pi_0 = \frac{p-m}{p} = 0.80$. Null P-values come from Uniform(0,1) and non-null P-values come from distribution Beta(0.1,4.9). (Beta(0.1,4.9) gives values in [0,1] that have mean $0.1/(0.1+4.9) = 0.02$ and a skew to right; there is no particular reason why in real data non-null P-values had exactly this distribution, but this is used for a demonstration purpose here; see figures below.) Let's plot theoretical (orange) and empirical (green) density functions for (1) null tests, (2) non-null tests and (3) for all tests combined (that is, the mixture distribution of the null and alternative distributions).

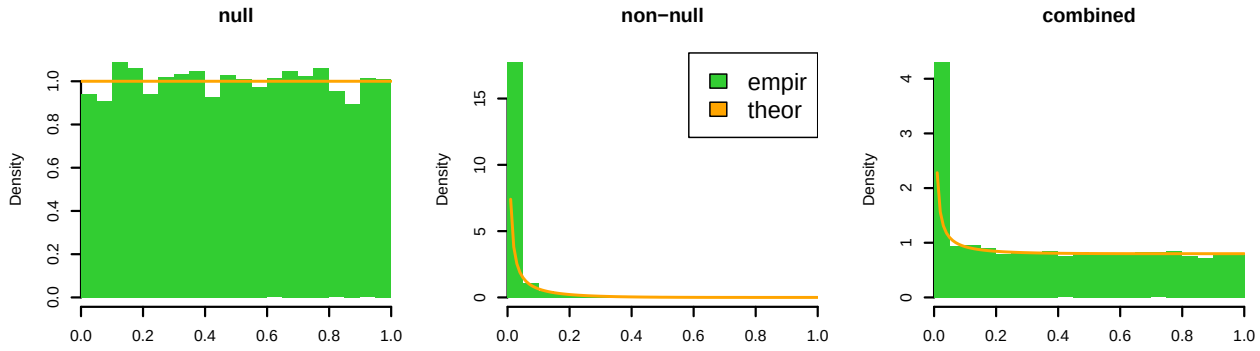
```
p = 10000
m = 2000
beta.1 = 0.1 # weight for unit interval's end point 1
beta.0 = 4.9 # weight for unit interval's end point 0
null.pval = runif(p-m, 0, 1)
```

```
alt.pval = rbeta(m, beta.1, beta.0) #non-null = alternative distribution
pval = c(alt.pval, null.pval) #all P-values together
eff = c(rep(T, m), rep(F, p - m)) #indicator for non-null effects

par(mfrow = c(1,3)) #Empirical histogram and theoretical curve for (1), (2), (3)
hist(null.pval, breaks = 20, prob = T, col = "limegreen", main = "null", xlab = "",
     xlim = c(0,1), sub="", lwd = 1.5, lty = 2, xaxs="i", border = NA)
curve(dunif(x, 0, 1), 0, 1, add = T, col = "orange", lwd = 2)

hist(alt.pval, breaks = 20, prob = T, col = "limegreen", main = "non-null", xlab = "",
     xlim = c(0,1), sub="", lwd = 1.5, lty = 2, xaxs="i", border = NA)
curve(dbeta(x, shape1 = beta.1, shape2 = beta.0), 0, 1, add = T, col = "orange", lwd = 2)
legend("topright", fill = c("limegreen","orange"),
       legend = c("empir","theor"), cex = 1.5)

hist(pval, breaks = 20, prob = T, col = "limegreen", main = "combined", xlab = "",
     xlim = c(0,1), sub="", lwd = 1.5, lty = 2, xaxs="i", border = NA)
curve((p-m)/p*1 + m/p*dbeta(x, shape1 = beta.1, shape2 = beta.0), 0, 1, add = T,
      col = "orange", lwd = 2)
```



Density plots show how the null P-values follow the uniform distribution and non-null P-values are enriched for small values near 0.

```
summary(null.pval)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000108 0.2521960 0.4999947 0.5002426 0.7467176 0.9998005
```

```
summary(alt.pval)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000000 0.0000001 0.0001608 0.0196240 0.0081581 0.7202532
```

When these two distributions are combined in the rightmost panel, the enrichment of small P-values is still present but has a smaller weight of 20% compared to the weight of 80% given to the null distribution.

Since $\Phi(t)$ is the probability that a particular P-value from this mixture distribution $\leq t$, the random

variables $D(t)$ and $\mathrm{FD}(t)$ are distributed as

$$\begin{aligned} D(t) &\sim \mathrm{Bin}(p, \Phi(t)) && (3)\\ \mathrm{FD}(t)\,|\,D(t) &\sim \mathrm{Bin}(D(t), \theta_t), \text{ where} && (4)\\ \theta_t &= \Pr(\mathrm{NULL}\,|\,P \le t) = \frac{\Pr(\mathrm{NULL})\Pr(P \le t\,|\,\mathrm{NULL})}{\Pr(P \le t)} && (5)\\ &= \frac{\pi_0 t}{\pi_0 t + (1-\pi_0)\Phi_1(t)}. && (6) \end{aligned}$$

Using the law of total expectation: $\mathrm{E}_Y(Y) = \mathrm{E}_X(\mathrm{E}_Y(Y\,|\,X))$, we have that

$$\mathrm{FDR}(t) = \mathrm{E}\left(\frac{\mathrm{FD}(t)}{D(t)}\right) = \mathrm{E}\left(\mathrm{E}\left(\frac{\mathrm{FD}(t)}{D(t)}\,\middle|\,D(t)\right)\right) = \mathrm{E}\left(\frac{1}{D(t)}\mathrm{E}(\mathrm{FD}(t)\,|\,D(t))\right) = \mathrm{E}\left(\frac{1}{D(t)}\theta_t\,D(t)\right) = \theta_t.$$

On the other hand, $\mathrm{E}(\mathrm{FD}(t)) = \mathrm{E}(\mathrm{E}(\mathrm{FD}(t)\,|\,D(t))) = \mathrm{E}(D(t)\theta_t) = \theta_t \cdot \mathrm{E}(D(t))$ Thus

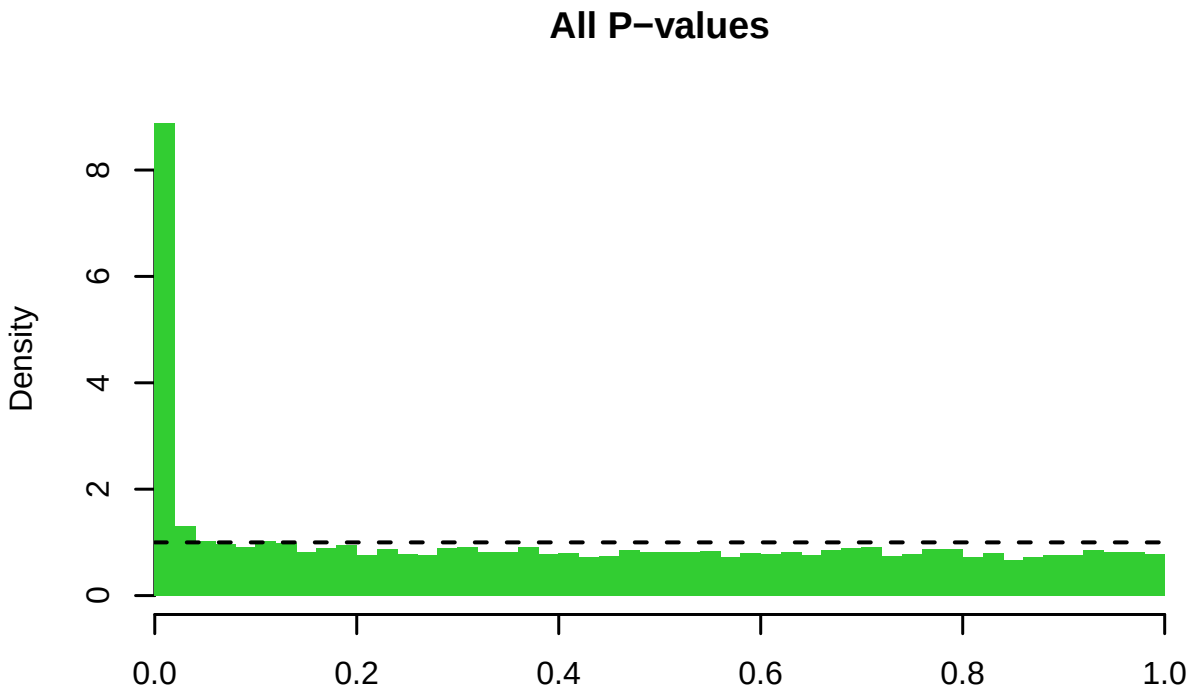
$$\frac{\mathrm{E}(\mathrm{FD}(t))}{\mathrm{E}(D(t))} = \frac{\theta_t \cdot \mathrm{E}(D(t))}{\mathrm{E}(D(t))} = \theta_t = \mathrm{FDR}(t).$$

So we can estimate $\mathrm{FDR}(t)$ by the ratio of expectations of $\mathrm{FD}(t)$ and $D(t)$. And what could we use as estimates for these expectations?

For each P-value threshold t, denote the number of all discoveries at threshold t by $\widehat{D}(t) = \#\{\text{P-values} \le t\}$. We use this to estimate $\mathrm{E}(D(t)) \approx \widehat{D}(t)$.

To estimate $\mathrm{E}(FD(t))$, we remember that the null P-values are uniformly distributed and hence $\mathrm{E}(FD(t)) = p_0 \cdot t = \pi_0 \cdot p \cdot t$. In estimating π_0 we again rely on the fact that the null P-values are uniformly distributed and that most P-values near 1 are expected to be from the null distribution. Let's remind us how our P-value distribution looked like.

```
hist(pval, breaks = 40, prob = T, col = "limegreen", main = "All P-values", xlab = "",
     xlim = c(0,1), sub = "", lwd = 1.5, lty = 2, xaxs = "i", border = NA)
abline(h = 1, lty = 2, lwd = 2)
```



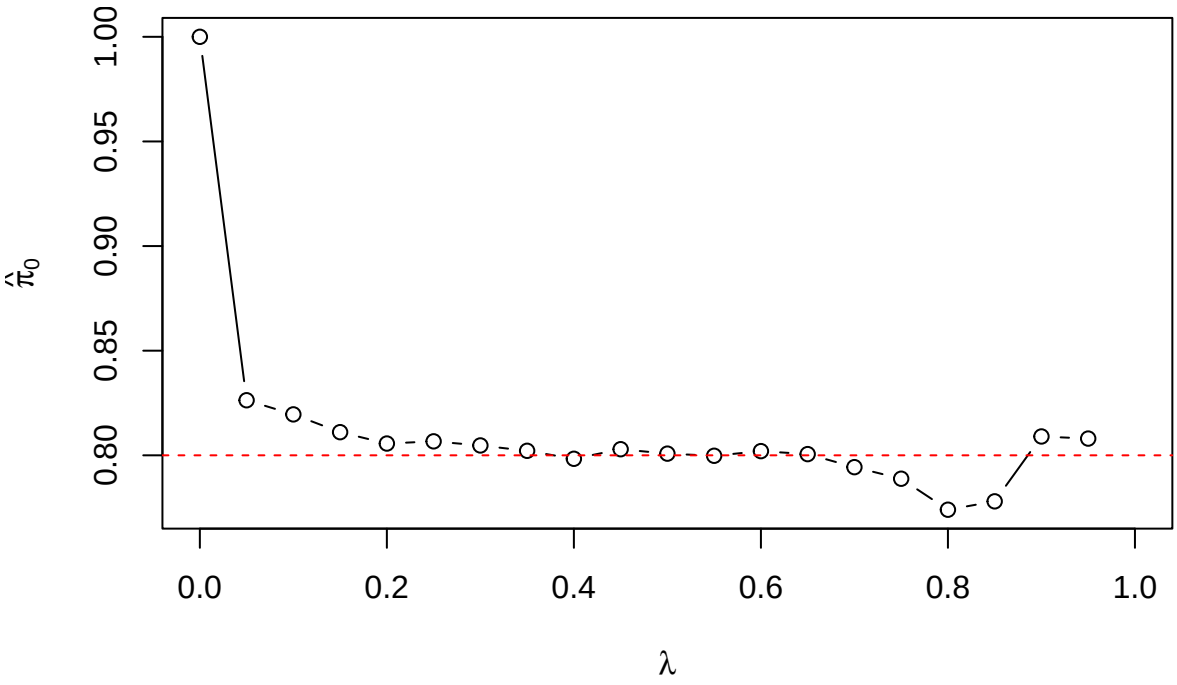
We can see that the density of P-values > 0.2 looks fairly flat. If we assume that the density of non-null P-values is ≈ 0 in this region, then the density function of P-values in this region is approximately $\pi_0 \cdot 1 + (1 - \pi_0) \cdot 0 = \pi_0$. Thus the value of the density function in this flat part gives an estimate of the overall proportion of null P-values π_0. The constant density function over any interval $(\lambda, 1]$ is the probability mass in the interval divided by the length of the interval $(1 - \lambda)$. With empirical data, this probability mass is
$\#\{P_j > \lambda \,|\, j = 1, \ldots, p\}/p$. We can apply such an estimate for any value λ to yield an estimator

$$\widehat{\pi}_0(\lambda) = \frac{\#\{P_j > \lambda \,|\, j = 1, \ldots, p\}}{(1 - \lambda)p}.$$

Parameter λ should be large enough that there are not many true effects with P-values $> \lambda$ but simultaneously λ should be small enough that there are enough null P-values $> \lambda$ so that the estimate has a good precision. An inclusion of a few non-null P-values $> \lambda$ makes this estimate conservative, that is, an overestimate of π_0. That keeps us on the safe side and leads to an overestimate $\mathrm{FDR}(t)$.

If we take $\lambda = 0$, then $\widehat{\pi}_0(\lambda) = 1$ which is usually much too conservative when a considerable proportion of tests are expected to be truly non-null. On the other hand, as we set λ closer to 1, the variance of $\widehat{\pi}_0(\lambda)$ increases due to decreasing number of P-values exceeding λ, which makes the estimate more unreliable. Therefore, as a simple, practical choice for λ we could use 0.5, but optimal threshold would depend on the data set. Let's see see how the choice of λ affects the estimate $\widehat{\pi}_0(\lambda)$ in our ongoing example data.

```
lambda = seq(0, 1, by = 0.05)
pi.0 = sapply(lambda, function(x) {sum(pval > x)}) / p / (1 - lambda)
plot(lambda, pi.0, t = "b", xlab = expression(lambda), ylab = expression(hat(pi)[0]))
abline(h = 1 - m/p, col = "red", lty = 2) #this is the true value
```



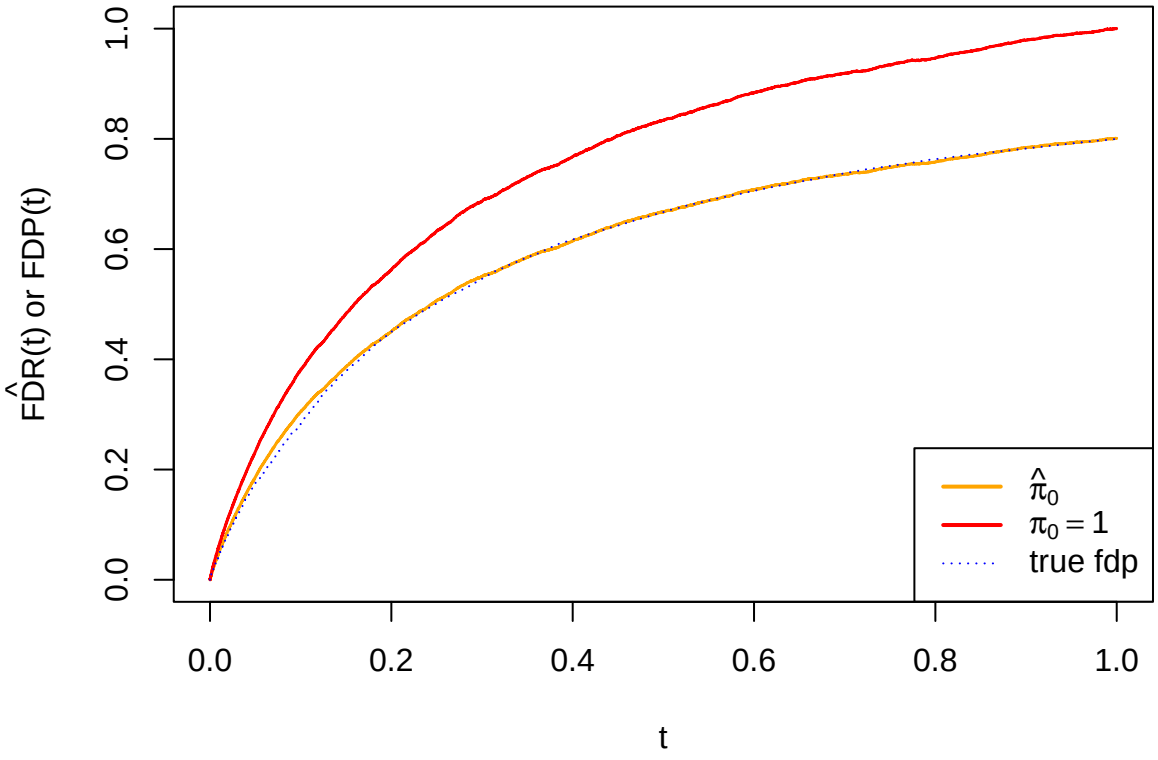
Indeed, we see that with values near $\lambda = 0.5$, the estimate $\widehat{\pi}_0(\lambda)$ is a good one whereas its quality deteriorates towards either of the endpoints of the interval.

With this estimator for $\widehat{\pi}_0$, we have the estimate

$$\widehat{\mathrm{FDR}}(t) = \frac{\widehat{\mathrm{FD}}(t)}{\widehat{D}(t)} = \frac{p \cdot \widehat{\Pr}(\text{NULL P-value}) \cdot \Pr(P \leq t \,|\, \text{NULL P-value})}{\widehat{D}(t)} = \frac{p \cdot \widehat{\pi}_0 \cdot t}{\widehat{D}(t)}.$$

Let's plot our estimate $\widehat{\text{FDR}}(t)$ in orange as function of t by using the value $\widehat{\pi}_0(0.5)$ estimated above. Let's also plot a curve that corresponds to conservative assumption $\pi_0 = 1$ in red and let's add the true empirical FDP(t) curve in blue, which we are able to do since we know which P-values were simulated from the null.

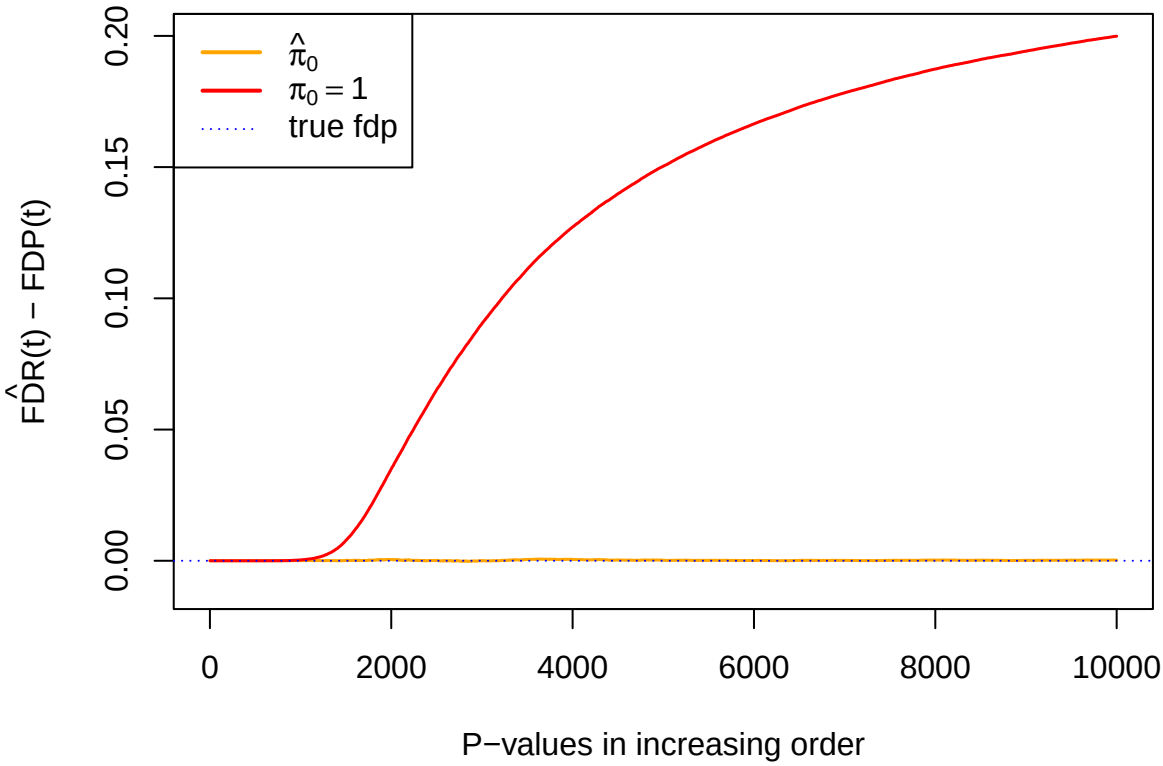
```
pval.sorted = sort(pval)
pi.0 = pi.0[which(lambda == 0.5)]
par(mar = c(5,6,2,1)) # widen left margin for text
plot(pval.sorted, p * pi.0 * pval.sorted / (1:p), xlab = "t",
     ylab = expression(paste(hat(FDR),"(t) or FDP(t)")), t = "l", col = "orange",
     xlim = c(0,1), ylim = c(0,1), lwd = 1.5)
#Add line for conservative assumption pi.0 = 1:
lines(pval.sorted, p * 1 * pval.sorted / (1:p), xlab = "t",
      col = "red", lwd = 1.5)
#Add empirical FDP line
fdp = cumsum(!eff[order(pval)]) / (1:p) #true false discovery proportion in data
lines(pval.sorted, fdp, xlab = "t", lty = 3, col = "blue", lwd = 1)
legend("bottomright", leg = expression(hat(pi)[0], pi[0] == 1 ,paste("true fdp")),
       col = c("orange", "red", "blue"), lty = c(1, 1, 3), lwd = c(2, 2, 1))
```



Note that the derivation of BH approach (in lecture notes HDS2) made the conservative assumption $\pi_0 = 1$ (or equivalently $p_0 = p$) which corresponds to the red curve in the figure above. Now we have improved that estimate by empirically estimating $\widehat{\pi}_0 < 1$ (orange curve). In this particular simulated data set, we know the actual false discovery proportion for any threshold t (blue curve). We see that, as expected, red line is clearly conservative: $\widehat{\text{FDR}}(t) > \text{FDP}(t)$. Orange line instead becomes an accurate estimate for FDP(t). In practice, empirically estimated $\widehat{\pi}_0$ will sometimes lead to higher and sometimes lower values of $\widehat{\text{FDR}}(t)$ compared to the true FDP(t), whereas assumption $\pi_0 = 1$ is expected to be conservative in general.

Let's also check whether $\widehat{\text{FDR}}(t)$ estimated using empirical $\widehat{\pi}_0$ averages to FDP(t) over many data sets.

```
R = 1000 #replications of data generation
lambda = 0.5
res.fdr.lambda = matrix(NA, nrow = R, ncol = p ) #lambda = 0.5
res.fdr.1 = matrix(NA, nrow = R, ncol = p ) #lambda = 1
for(rr in 1:R){
  pvs = c(rbeta(m, beta.1, beta.0), runif(p-m)) #m non-nulls and p-m nulls
  pvs.sorted = sort(pvs)
  pi.0 = sum(pvs > lambda)/p/(1-lambda)
  fdr.lambda = p * pi.0 * pvs.sorted / (1:p)
  fdr.1 = p * 1 * pvs.sorted / (1:p)
  fdp = cumsum(!eff[order(pvs)]) / (1:p)
  res.fdr.lambda[rr,] = fdr.lambda - fdp
  res.fdr.1[rr,] = fdr.1 - fdp
}
par(mar = c(5,6,2,1)) # widen left margin for text
plot(1:p, colSums(res.fdr.lambda) / R, col = "orange", t="l", lwd = 1.5,
     ylim = c(-0.01,0.2), ylab = expression(paste(hat(FDR),"(t) - FDP(t)")),
     xlab = "P-values in increasing order")
lines(1:p, colSums(res.fdr.1)/R, col = "red", t = "l", lwd = 1.5)
abline(h = 0, lty = 3, col = "blue", lwd = 1)
legend("topleft", leg = expression(hat(pi)[0], pi[0] == 1 ,paste("true fdp")),
       col = c("orange","red","blue"), lty = c(1, 1, 3), lwd = c(2, 2, 1))
```



We see that on average $\widehat{\text{FDR}}(t)$ estimated using $\hat{\pi}_0$ is very close to true FDP(t) at any point t, (but in any one instance it can be either smaller or larger). As expected, Figure also shows that use of $\pi_0(\lambda = 0) = 1$ results in overestimates for FDP throughout the threshold values t. Why does the red curve grow end at value 0.2?

Questions.

1. What was the piece of information from observed P-value distribution that together with the number of tests p can be used to estimate π_0, the proportion of null P-values?

2. Which assumption about π_0 was done in earlier derivation of BH method?

3. In which case could an empirical estimate of π_0 be misleading?

Definition of Q-value We define Q-value of a variable/test as **the minimum FDR expected if we call that variable a discovery**. To compute it we only use P-values. Thus for P-value P_j the Q-value is

$$Q_j = Q(P_j) = \min_{t \geq P_j} \mathrm{FDR}(t).$$

Note that the minimum is needed because we will call j a discovery, if we will call any other variable with at least as large P-value a discovery. This idea is exactly the same as the step-up procedure property of the BH method. In practice, we use estimate for $\widehat{\mathrm{FDR}}(t)$ to compute Q-values.

The Q-value for a particular hypothesis test is the expected proportion of false positives incurred when calling all tests with at most as large Q-values as significant/discoveries. Therefore, calculating the Q-values for each test and thresholding them at Q-value level α_F produces a set of significant variables so that a proportion of α_F is expected to be false positives. Typically, P-value is described as the probability of a null test statistic being at least as extreme from the null hypothesis as the observed one. Q-value of a particular test can be described as the expected proportion of false positives among all tests with at least as extreme Q-value as the observed one.

Suppose that variables with Q-values $\leq 5\%$ are called significant. This results in a FDR of 5% among the significant variables. A P-value threshold of 5% yields a false positive rate of 5% among all null variables in the data set. In light of the definition of the false positive rate, a P-value cutoff says little about the content of the variables actually called significant. The Q-values provide a more meaningful measure among the variables called significant. Because significant variables will likely undergo some subsequent verification, a Q-value threshold can be phrased in practical terms as the proportion of significant variables that turn out to be false leads.

`qvalue` package The definition of Q-value is independent of any algorithm to compute Q-values, and therefore using 'Q-values' doesn't strictly speaking imply which method has been used to compute them. However, often Q-value is associated with a common approach to compute Q-values from P-values using the `qvalue` package by John Storey. `qvalue()` uses an empirical estimate $\widehat{\pi}_0$ in defining Q-values through $\widehat{\mathrm{FDR}}(t)$. The only difference to our derivation above, where we computed $\widehat{\pi}_0$ at $\lambda = 0.5$, is that by default `qvalue()` first computes $\widehat{\pi}_0(\lambda)$ at a grid of λ values < 0.95 and then fits a spline from which it estimate $\widehat{\pi}_0$ as the spline-predicted value $\widehat{\pi}_0(\lambda = 1)$. (One can override this default smoothing by passing to `qvalue` function a fixed value of λ through parameter `lambda`.)

Let's apply `qvalue` and see what its basic output functions produce.

```
#do this the first time to install qvalue package from Bioconductor:
#if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")
#BiocManager::install("qvalue")
library(qvalue)
q = qvalue(pval) #this makes Q-values out of P-values
str(q)
```

```
## List of 8
##  $ call      : language qvalue(p = pval)
##  $ pi0       : num 0.795
##  $ qvalues   : num [1:10000] 9.10e-08 2.24e-01 4.52e-01 1.52e-14 2.24e-04 ...
```

```
##  $ pvalues   : num [1:10000] 3.92e-09 6.47e-02 2.03e-01 1.15e-16 2.33e-05 ...
##  $ lfdr      : num [1:10000] 7.17e-07 7.56e-01 9.39e-01 7.17e-07 2.15e-03 ...
##  $ pi0.lambda: num [1:19] 0.826 0.82 0.811 0.806 0.807 ...
##  $ lambda    : num [1:19] 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 ...
##  $ pi0.smooth: num [1:19] 0.819 0.816 0.814 0.811 0.808 ...
##  - attr(*, "class")= chr "qvalue"
```

We have fields for the estimated value of π_0 (`pi0`), for P-values (`pvalues`) that were given as input and estimated Q-values (`qvalues`). Additionally there are values of $\widehat{\pi}_0(\lambda)$ at the grid given in `lambda`, and their smoothed estimates from the spline fitted to these points. Local false discovery rate (`lfdr`) will be discussed in next lecture.

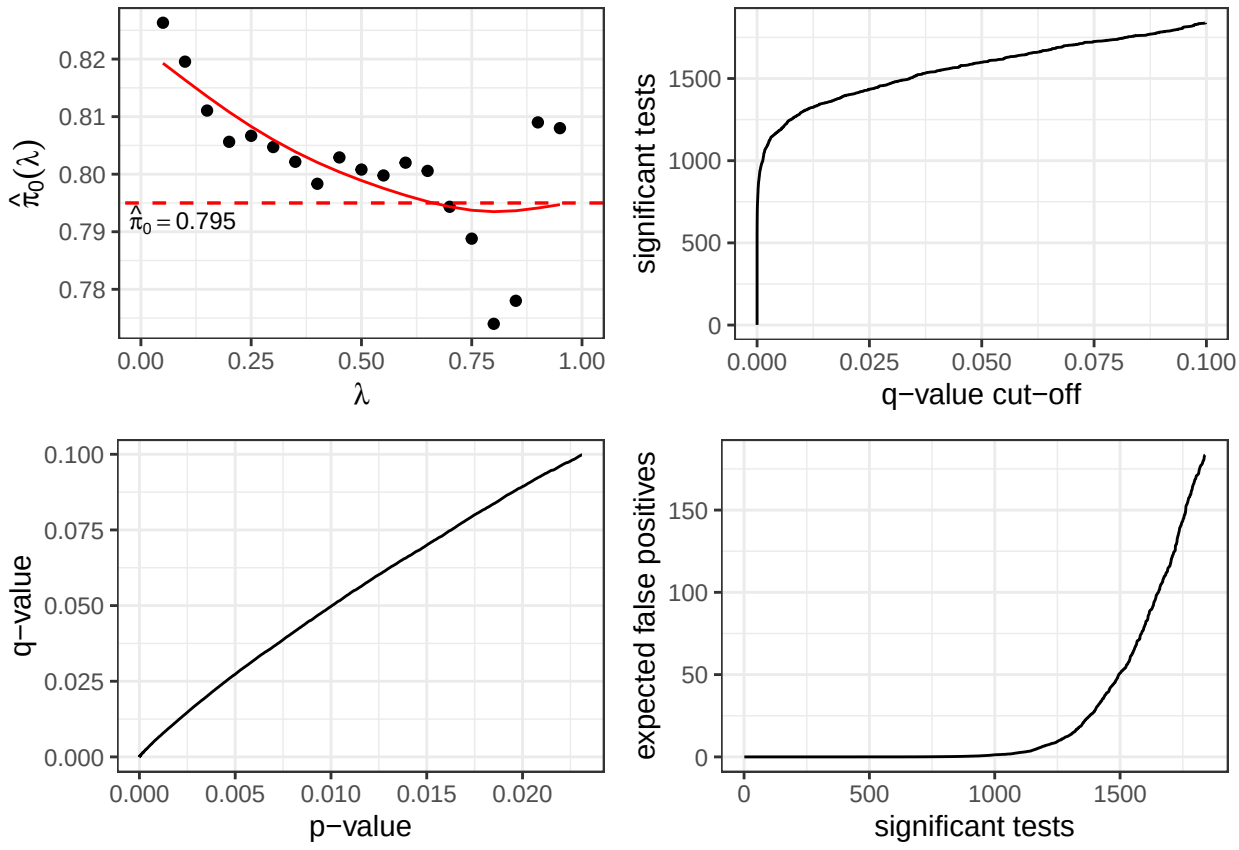
These Q-values were computed using π_0 value

```
signif(q$pi0, 3)
```

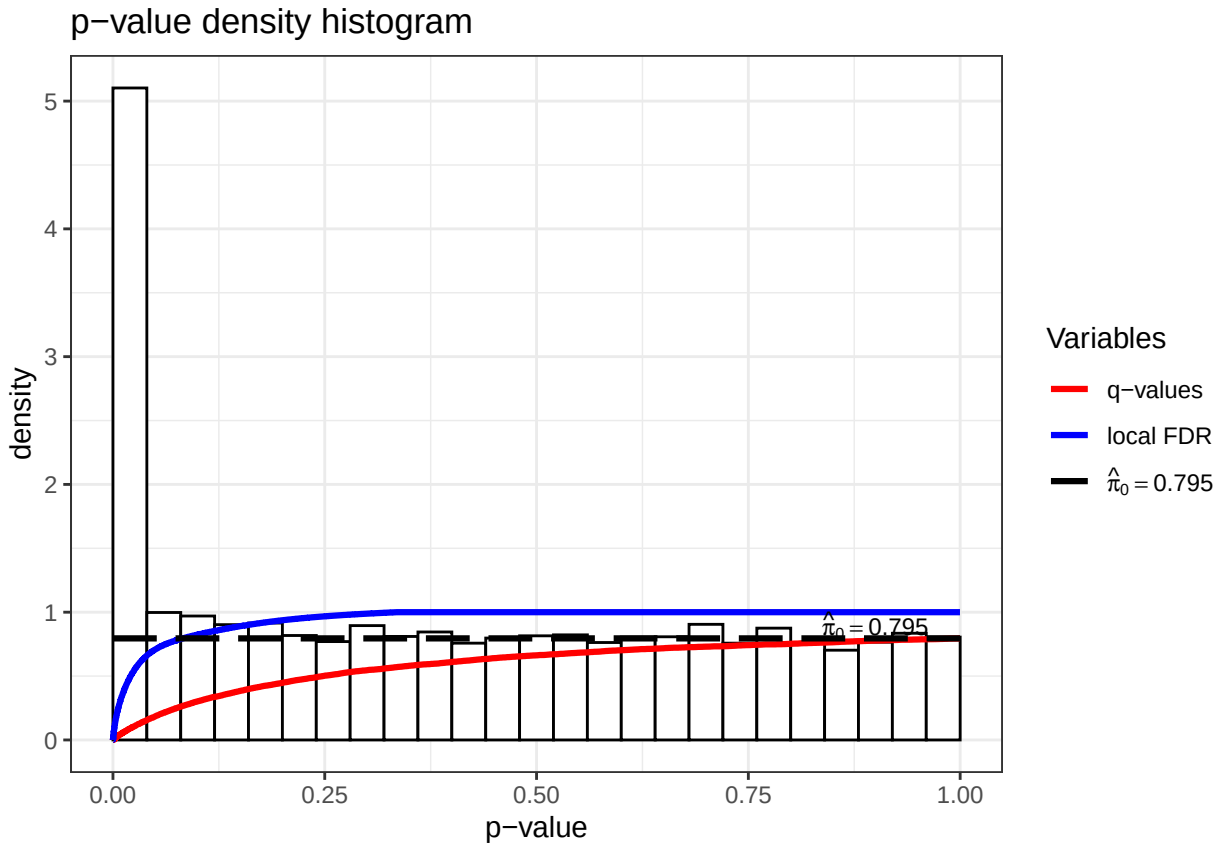
```
## [1] 0.795
```

Basic plots of `qvalue`

```
plot(q)
```




```
hist(q)
```




```
summary(q)
```

```
##
## Call:
## qvalue(p = pval)
##
## pi0: 0.7947102
##
## Cumulative number of significant calls:
##
##           <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1    <1
## p-value      960   1224  1598   1859  2150 2624 10000
## q-value      747    979  1296   1434  1598 1839 10000
## local FDR    589    747   986   1114  1188 1307  4671
```

This table tells, for example, that there are 2150 P-values below 0.05, but it is estimated that if we want to make a set of discoveries that had at most 5% false positives among them, we should only choose 1596 of the smallest P-values. From P-value vs Q-value plot above, we see that Q-value 0.05 would correspond P-value of about 0.01 in these data.

In the previous lecture, we used Benjamini-Hochberg method to define thresholds for a given FDR level. Let's see how `qvalue`compares to BH method in determining the discoveries at a given FDR level. Here we will use $\alpha_F = 0.1$.

```
alpha = 0.1
BH.pval = p.adjust(pval, method = "BH")
```

```
bh = c(sum(BH.pval < alpha), sum(BH.pval[(m+1):p] < alpha))
qval = c(sum(q$qvalues < alpha), sum(q$qvalues[(m+1):p] < alpha))
data.frame(D = c(bh[1],qval[1]),
           TD = c(bh[1],qval[1])-c(bh[2],qval[2]),
           FD = c(bh[2],qval[2]),
           FDP = c(bh[2],qval[2]) / c(bh[1],qval[1]),
           row.names=c("BH","qvalue"))
```

```
##           D   TD  FD        FDP
## BH     1736 1611 125 0.07200461
## qvalue 1839 1661 178 0.09679173
```

Here `qvalue` gave 50 more true positives and 53 more false discoveries than BH, and was the more accurate method to approximate the target false discovery proportion of 10% (9.6% vs. 7.2%). We will study more systematically BH and `qvalue` in home exercises.

The contents of the course so far is summarized by a 7-page document of multiple testing and FDR by John Storey: False Discovery Rates, 2010. The last bit in there about "Bayesian Derivation" is the topic of our next lecture.

1. What does it mean when a test statistic has Q-value of 0.05 and how is it different from having a P-value of 0.05?
2. Which one is typically smaller, Q-value or P-value? (Or neither?)

Extra: Using covariate information with FDR Recently, the standard FDR methods (BH, Storey's Q-values) have been extended to include additional covariate information. For example, when studying differences between brain measurements under two conditions, the physical coordinates of each measured spot in brain may associate with the proportion of true positive differences between the conditions. In such a case, tests could be binned depending on the coordinates, and a dependency between $\widehat{\pi}_0$ and physical coordinates could be estimated. Examples of such methods are Independet Hypothesis Weighting (IHW) and `swfdr`.