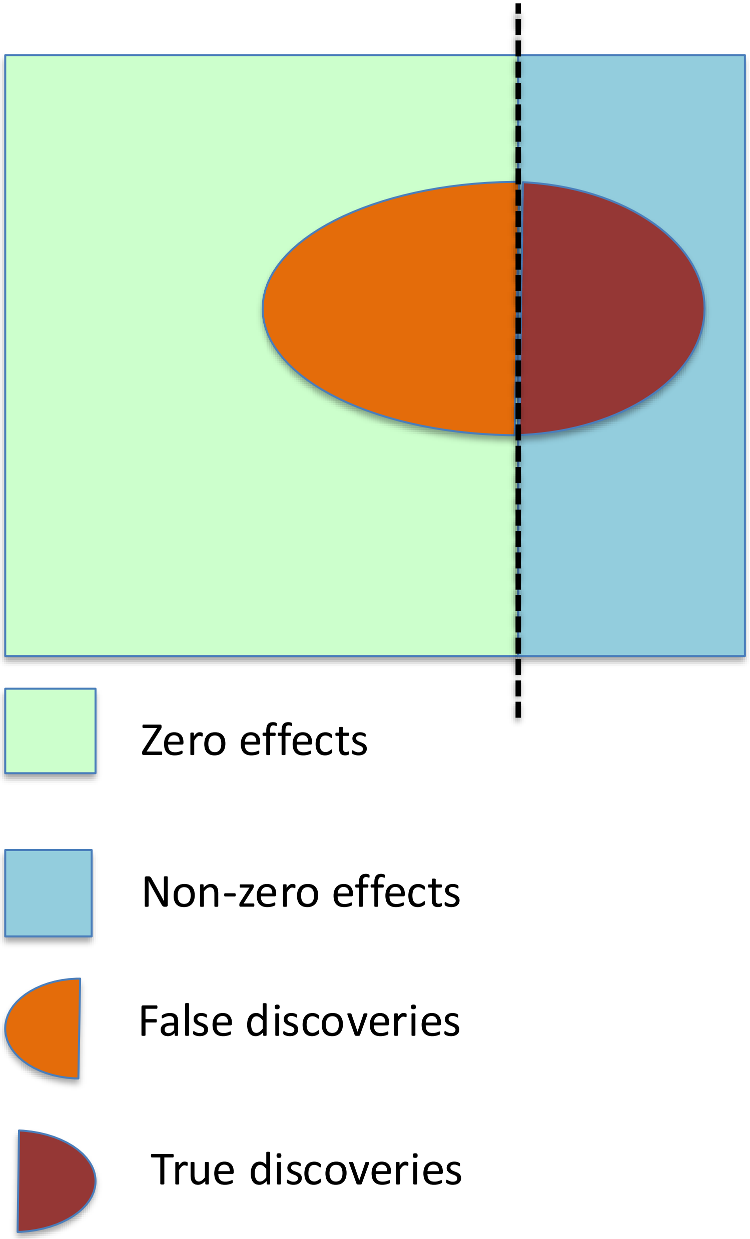
High dimensional statistics

Q-value

Matti Pirinen
University of Helsinki



We have p hypotheses to test, one for each variable/feature.

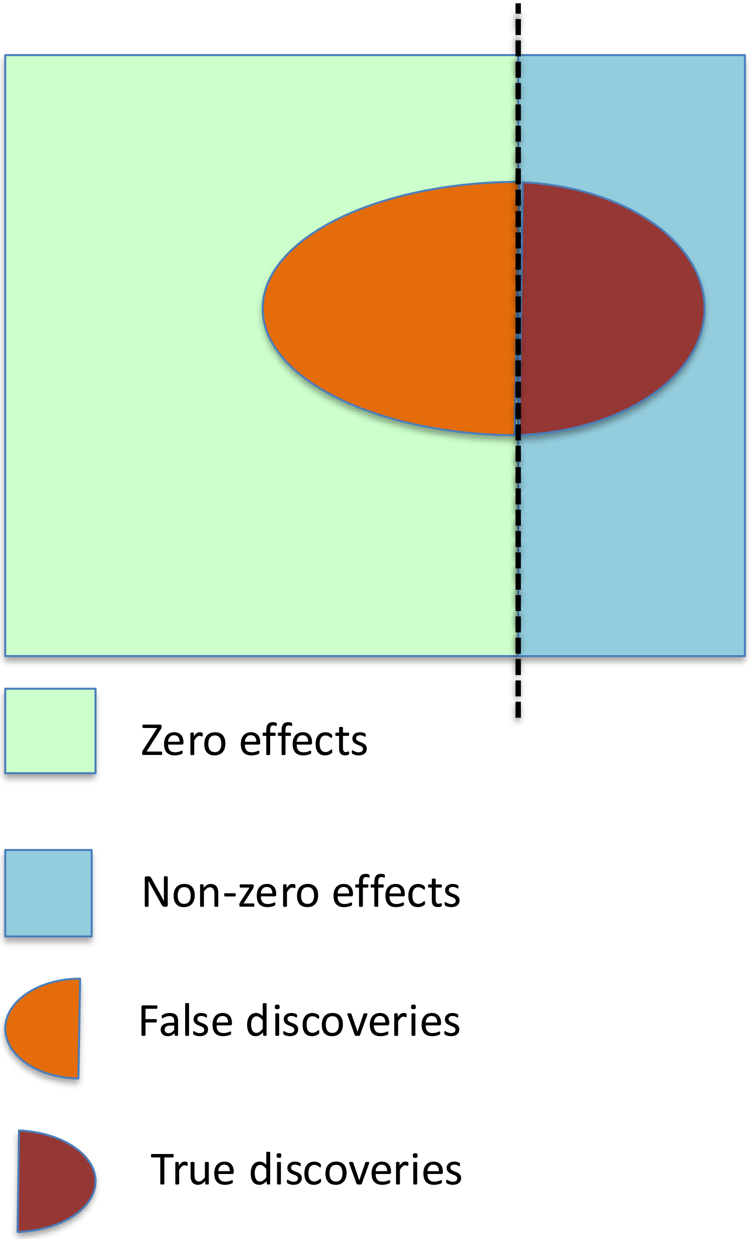
Each null hypothesis states that the effect is zero.

Hypotheses are represented by the square that is divided into two parts corresponding to truly zero and non-zero effects.

A statistical procedure is used to label each variable as a discovery or non-discovery. (“Discovery” = “significant variable” = “rejected null hypothesis”)

Some discoveries are true (non-zero effects) and some are false (zero effects).

Controlling false positive rate at level α



Zero effects

Non-zero effects

False discoveries

True discoveries

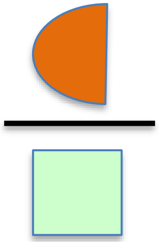
Proportion of false discoveries out of all zero effects ≤ α

Nothing is said about true discoveries.

Can be done by thresholding *P*-value at ≤ α

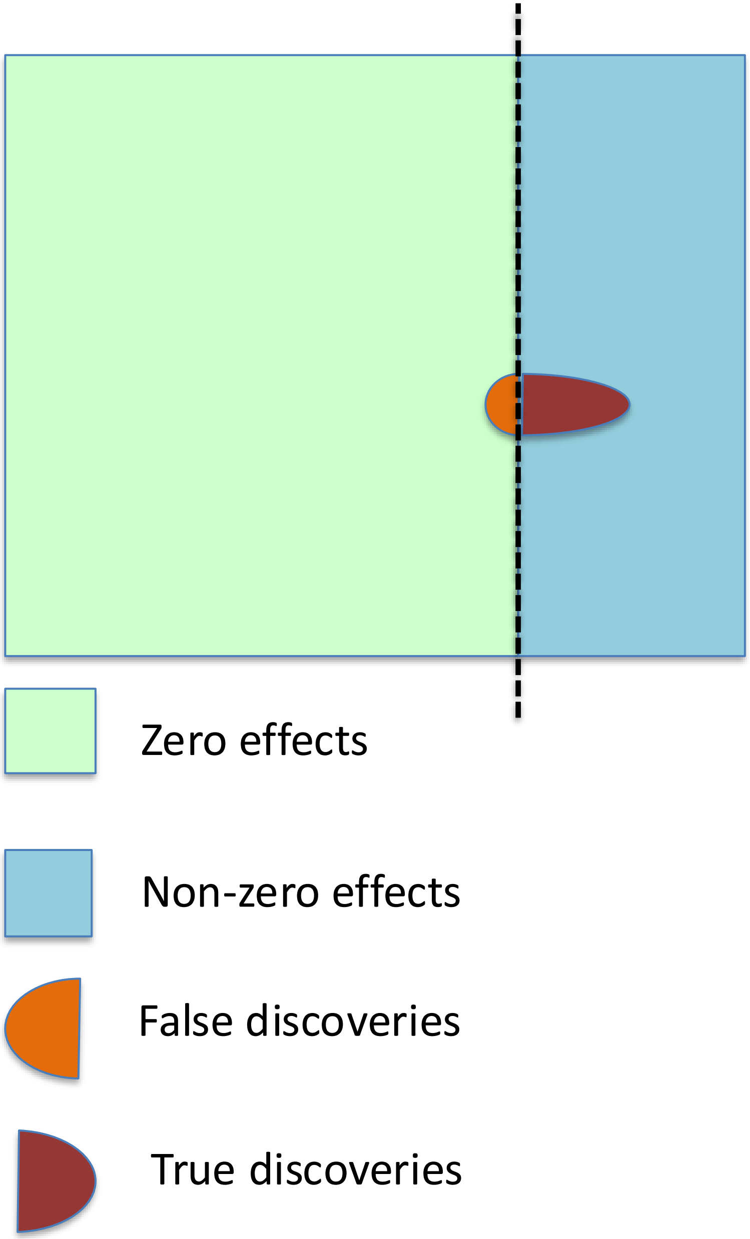
This is the standard way of doing hypothesis testing in statistics.

Empirical estimate from Figure:



≤ α

Controlling family-wise error rate at level α


🟩 Zero effects

🟦 Non-zero effects

◖ False discoveries

◗ True discoveries

Probability of at least one false discovery ≤ α

Nothing is said about true discoveries.

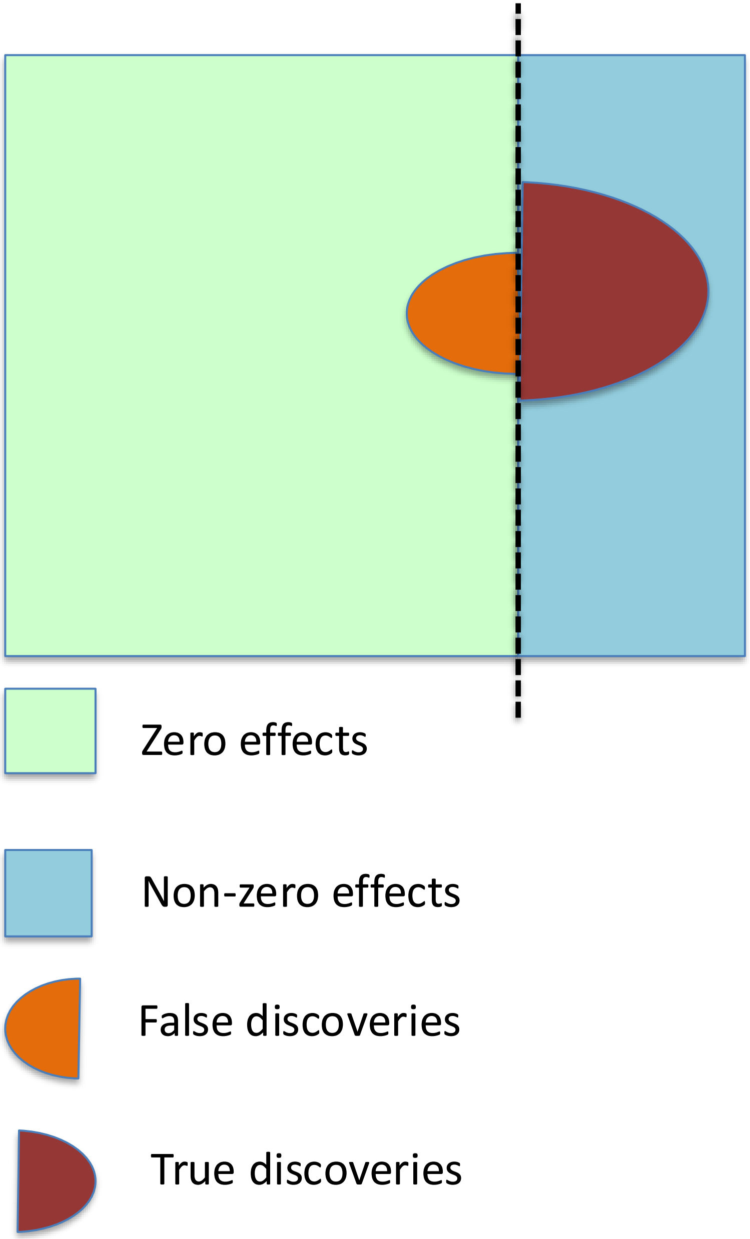
Can be done by
1. thresholding *P*-value at ≤ α/p (Bonferroni)
2. Holm method

Very stringent requirement and discovers usually only at most a few non-zero effects.

Empirical estimate from Figure:

Whether there is any element in ◖

Controlling false discovery rate at level α



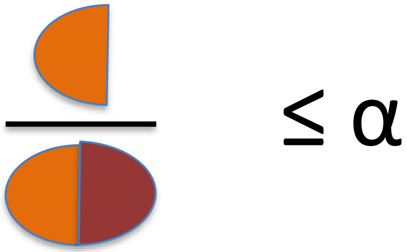
Proportion of false discoveries out of all discoveries ≤ α

Can be done by

1. Benjamini-Hochberg (independence)
2. Benjamini-Yekutieli (any dependence)

This approach correctly discovers many non-zero effects and keeps the proportion of zero effects low among the discoveries

Empirical estimate from Figure:



Definition of false discovery rate

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

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

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

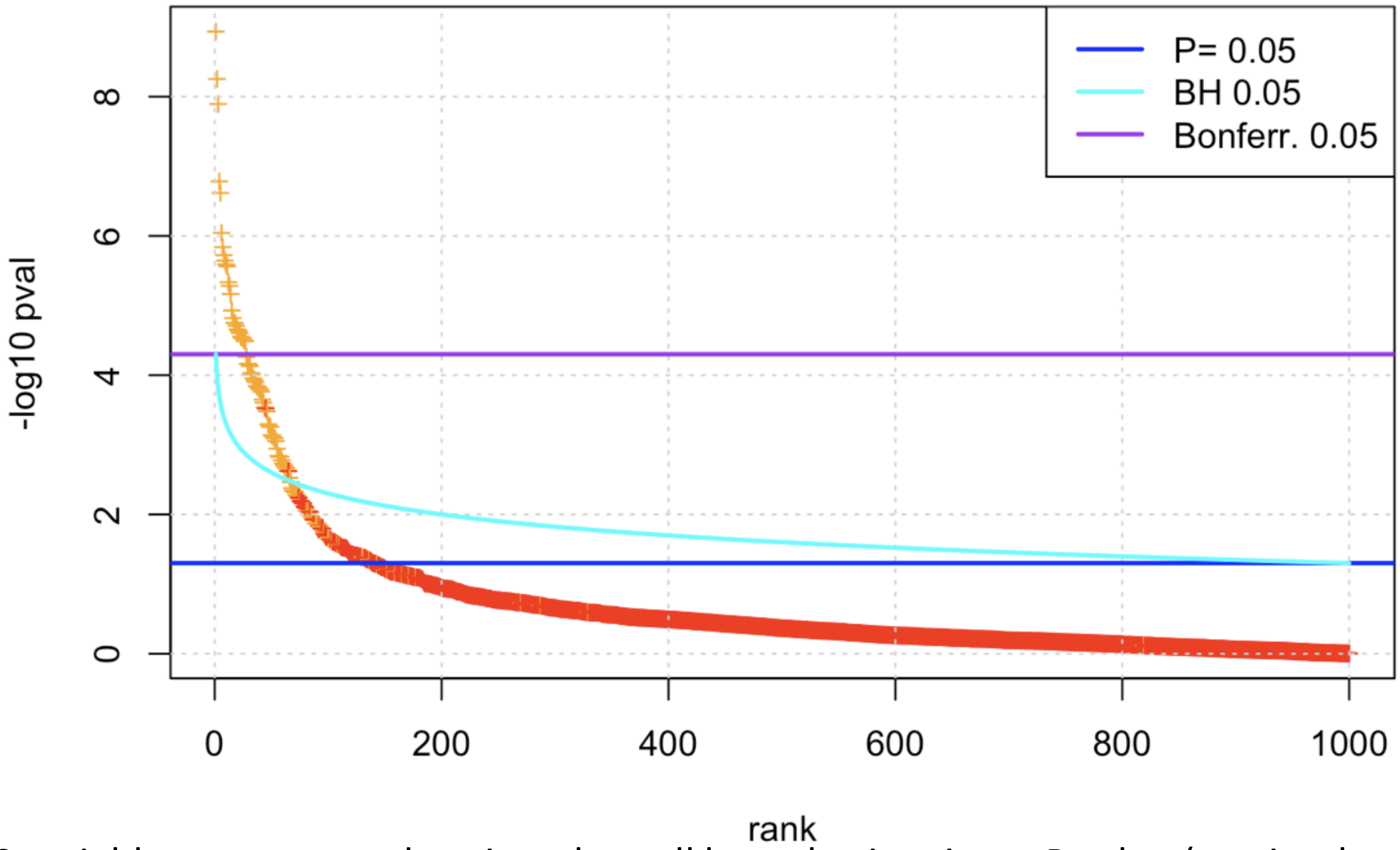
$$\mathrm{FDR} = \mathrm{E}(\mathrm{FDP}).$$

Benjamini-Hochberg (BH) 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 *j*th *P*-value. BH procedure at level α_F (BH(α_F)) is to

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

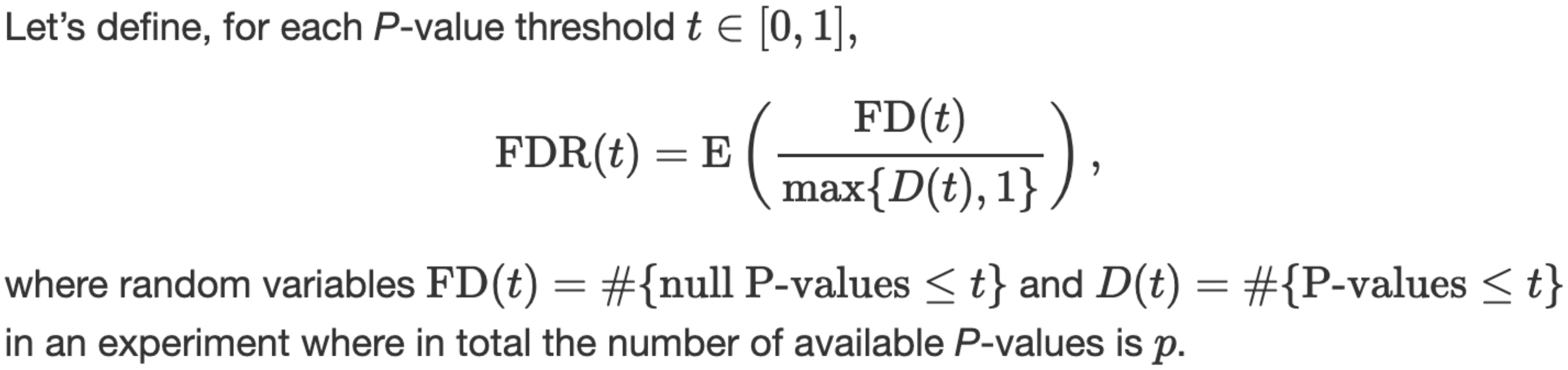
Theorem (BH). For independent tests and for any configuration of false null hypotheses, BH(α_F) controls the FDR at level α_F.



1000 variables are assessed against the null hypothesis using a *P*-value (y-axis, –log10 scale)
Three inference methods are shown by threshold lines/curve.
For significance testing ($P < 0.05$) and FWER ($P < 0.05/1000$) the threshold is fixed for all tests.
For Benjamini-Hochberg, the threshold gets less stringent with the rank of the P-value,
and we look for the highest ranking *P*-value that still is below its threshold (= above its –log10 *P*-value in the Figure) and declare all lower ranking tests as discoveries.
The red points are truly null variables and orange are truly non-null.

Today

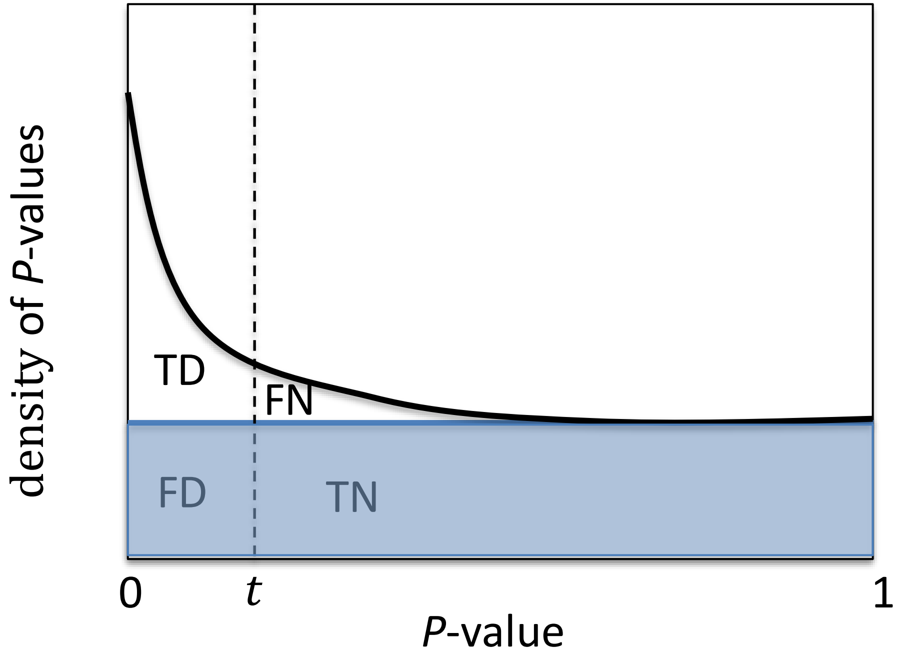
- Refine BH method by estimating p_0 from the observed distribution of *P*-values
  - BH method assumes $p_0 = p$ which leads to a conservative FDR method when $p_0 < p$, i.e. the number of false discoveries is clearly below the target level α_F when p_0 is clearly below p
- Define *Q*-value that is similar for FDR control as *P*-value is for false positive rate control
  - Thresholding variables by $\mathrm{Q} < \alpha_F$ gives FDR < α_F



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

$$\mathrm{FDR}(t) = \mathrm{E}\left(\frac{\mathrm{FD}(t)}{\max\{D(t),1\}}\right),$$

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



TD = true discoveries
FD = false discoveries
FN = false non-discoveries
TN = true non-discoveries

D = TD + FD counts all discoveries

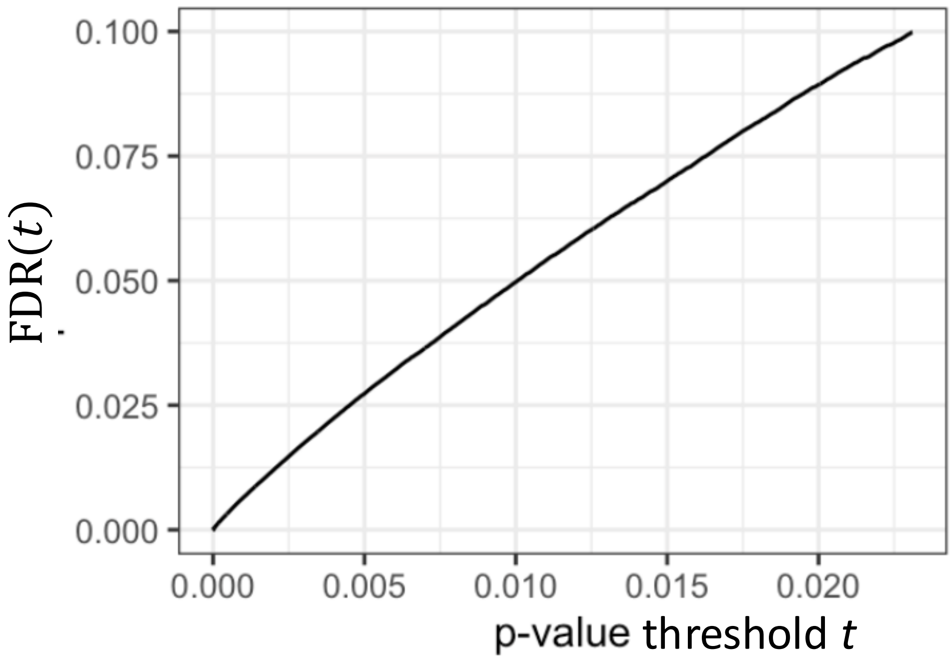
FDR(t) = FD(t)/(TD(t) + FD(t))

Note: Null P-values are uniformly distributed as described by the blue block.

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

$$\mathrm{FDR}(t) = \mathrm{E}\left(\frac{\mathrm{FD}(t)}{\max\{D(t),1\}}\right),$$

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



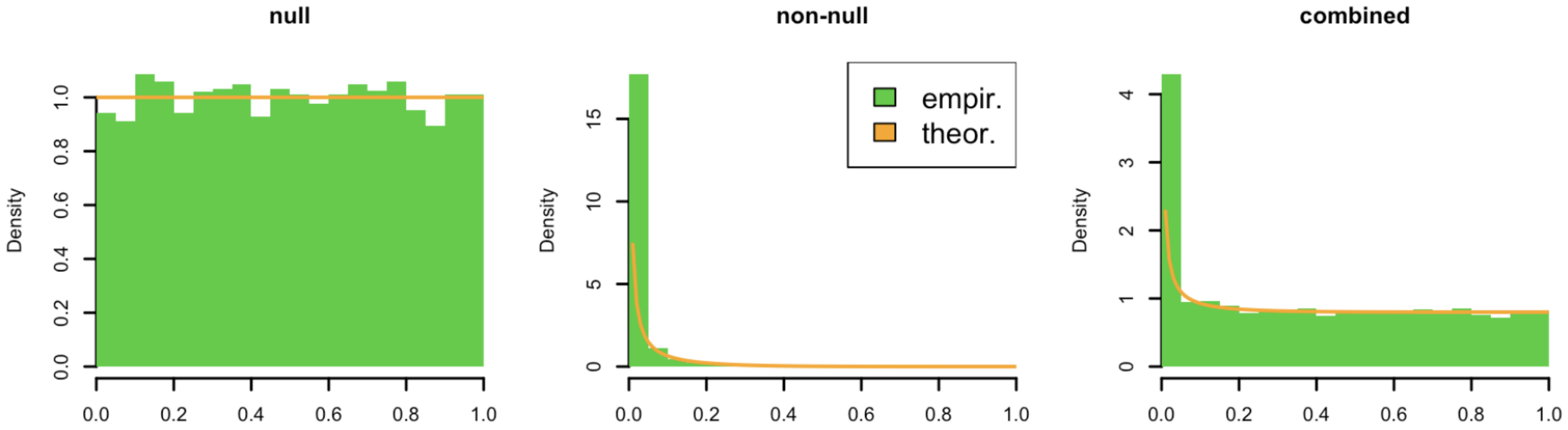
Example:
If we label tests with *P* < 0.01 as discoveries, then we expect that 5% of discoveries are false discoveries, while with *P* < 0.20, we expect about 9% to be false discov.

How can we estimate $\mathrm{FDR}(t)$ from the observed P-value distribution?

Mixture distribution of *P*-values

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 a mixture proportion π_0 for the null distribution. In other words, the cdf Φ and pdf ϕ of the *P*-values are

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



Here non-null *P*-values come from distribution Beta(0.1, 4.9) and $\pi_0 = 0.80$.

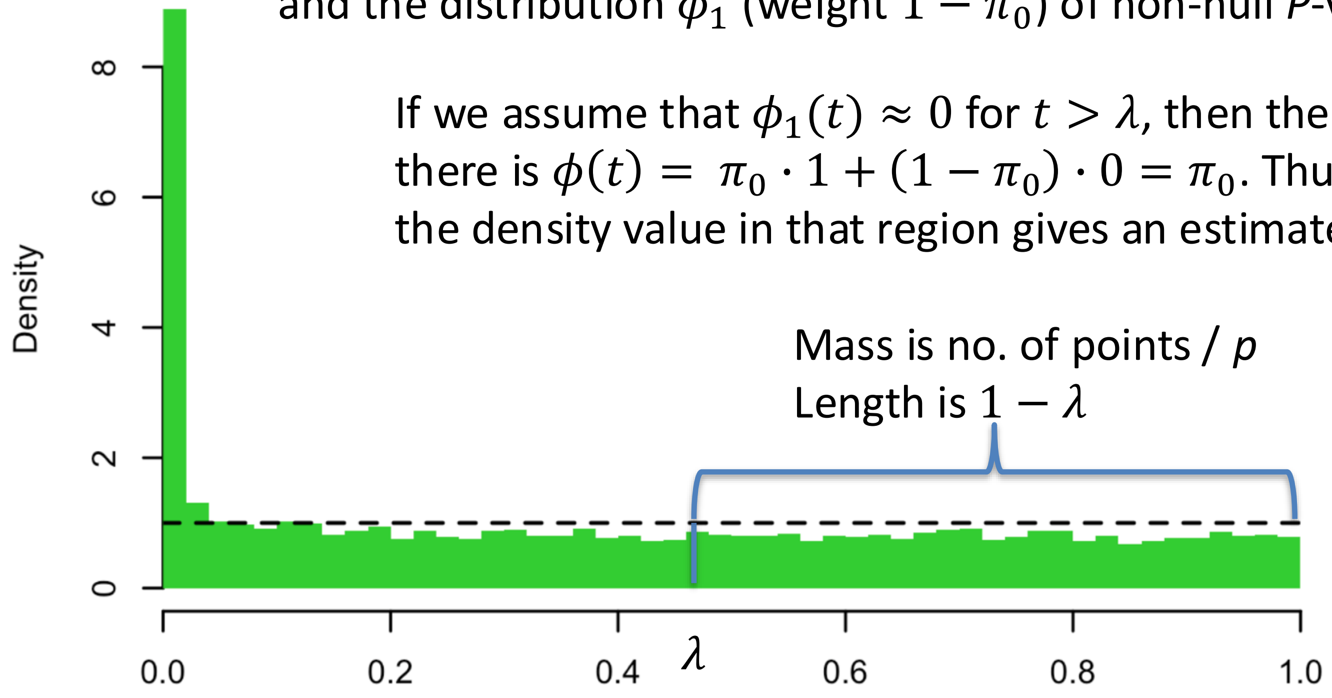
Estimating $\mathrm{FDR}(t)$

- $\mathrm{FDR}(t) \approx \mathrm{E}\left(\frac{\mathrm{FD}(t)}{D(t)}\right) \approx \frac{\mathrm{E}\left(\mathrm{FD}(t)\right)}{\mathrm{E}(D(t))} \approx \frac{p\pi_0 t}{\widehat{D}(t)}$
  - $\mathrm{E}\left(\mathrm{FD}(t)\right) = p_0 t = p\pi_0 t$
  - $\mathrm{E}\left(D(t)\right)$ is replaced by the observation $\widehat{D}(t)$
- We still need to estimate π_0

Estimating π_0

Distribution of *P*-values is a mixture of Uniform(0,1) (weight π_0) and the distribution ϕ_1 (weight $1-\pi_0$) of non-null *P*-values.

If we assume that $\phi_1(t) \approx 0$ for $t > \lambda$, then the density function there is $\phi(t) = \pi_0 \cdot 1 + (1-\pi_0) \cdot 0 = \pi_0$. Thus, estimate of the density value in that region gives an estimate of π_0.

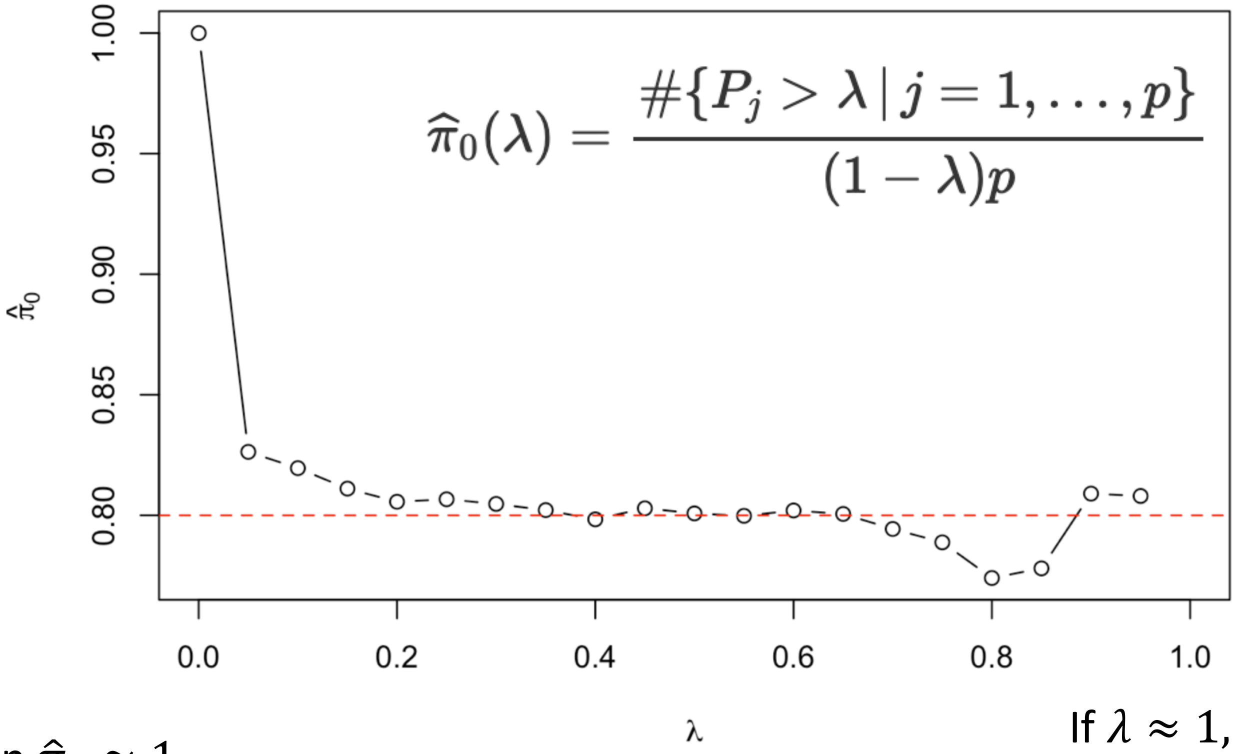


Mass is no. of points / *p*
Length is $1-\lambda$

Estimate of a constant density function in a region is the probability mass in the region divided by the size of the region.

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

Estimating π_0



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

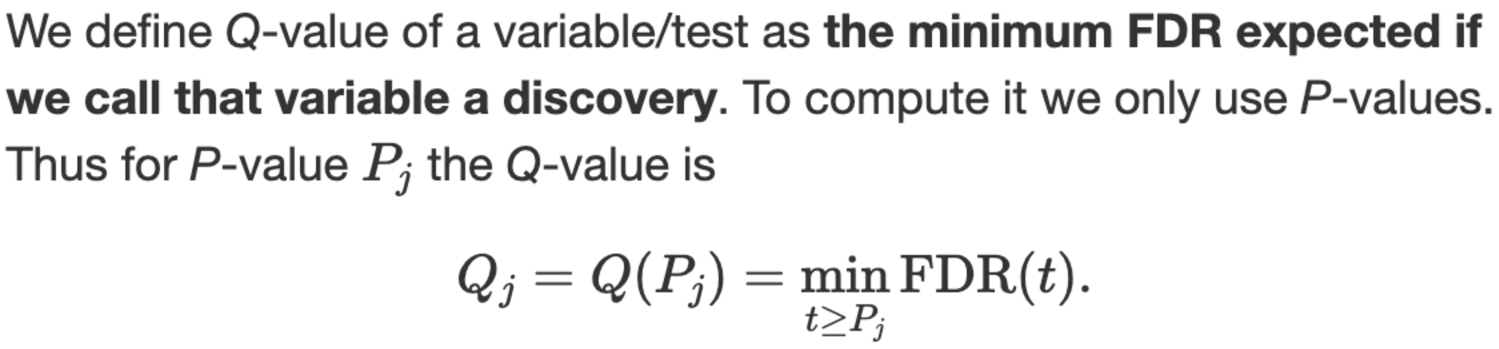
If $\lambda \approx 0$, then $\hat{\pi}_0 \approx 1$
typically an overestimate

$\lambda \approx 0.5$ is usually a good choice.

If $\lambda \approx 1$,
then $\hat{\pi}_0$ is unstable.

With $\hat{\pi}_0$, we can estimate $$\widehat{\mathrm{FDR}}(t) = \frac{\widehat{\mathrm{FD}}(t)}{\widehat{D}(t)} = \frac{p \cdot \widehat{\pi}_0 \cdot t}{\widehat{D}(t)}$$

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

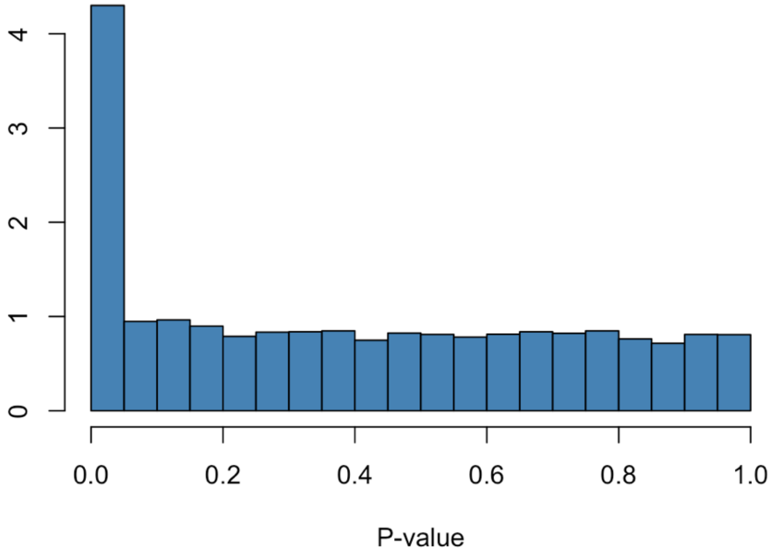
$$Q_j = Q(P_j) = \min_{t \geq P_j} \mathrm{FDR}(t).$$

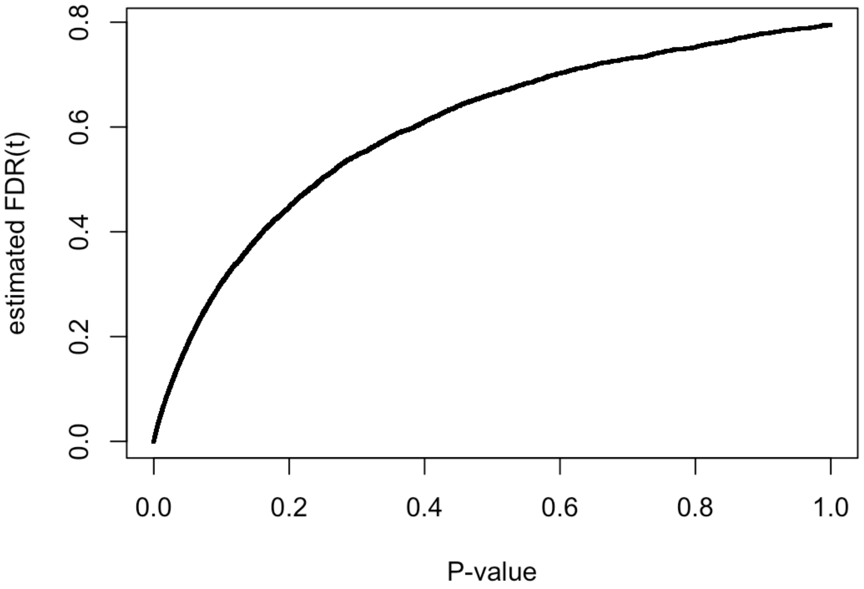
- The *Q*-value for a particular test is the expected proportion of false positives incurred when calling all tests with at most as large *Q*-values as significant/discoveries.

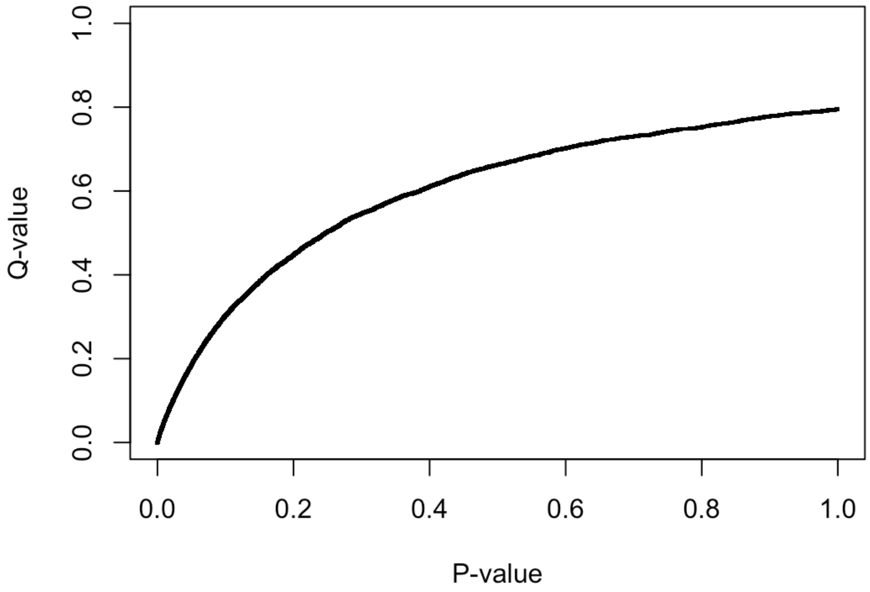
- Therefore, calculating the *Q*-values for each test and thresholding them at the *Q*-value level α produces a set of significant variables among which a proportion of α is expected to be false positives.

- *Q*-value of a variable depends on the *P*-values of other variables.

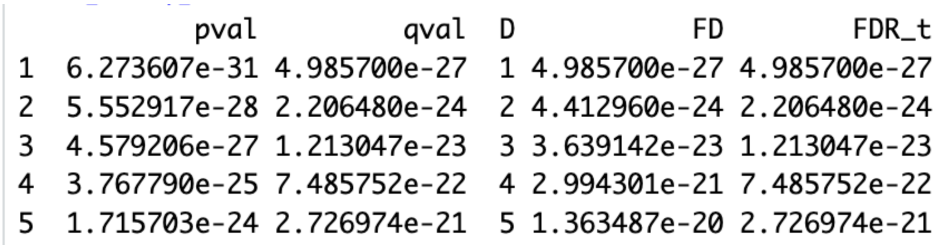
Example 1





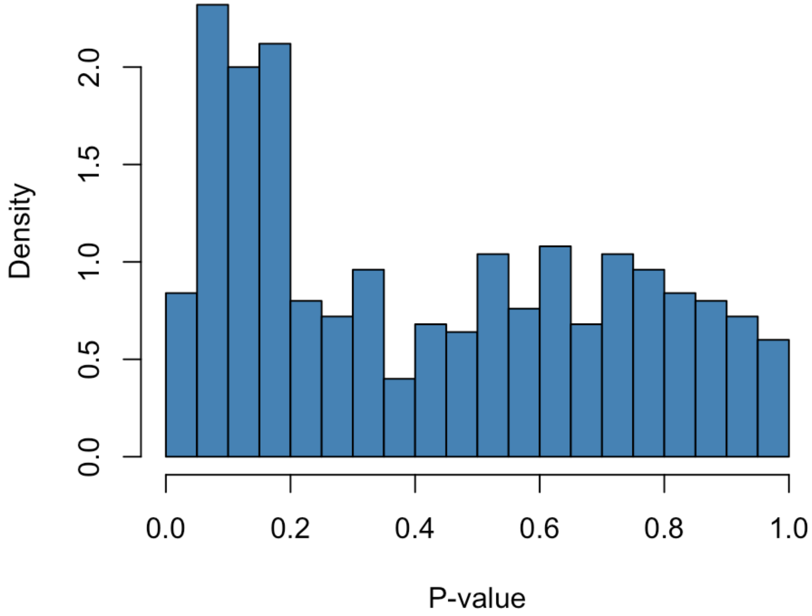


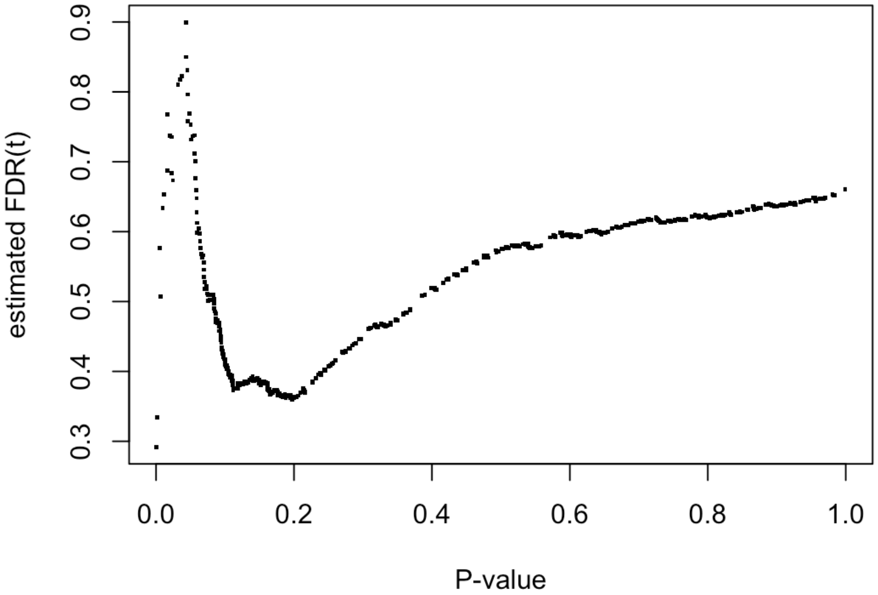
Here, the non-nulls have low *P*-values and both FDR(t) and *Q*-values are increasing as function of *P*-values.

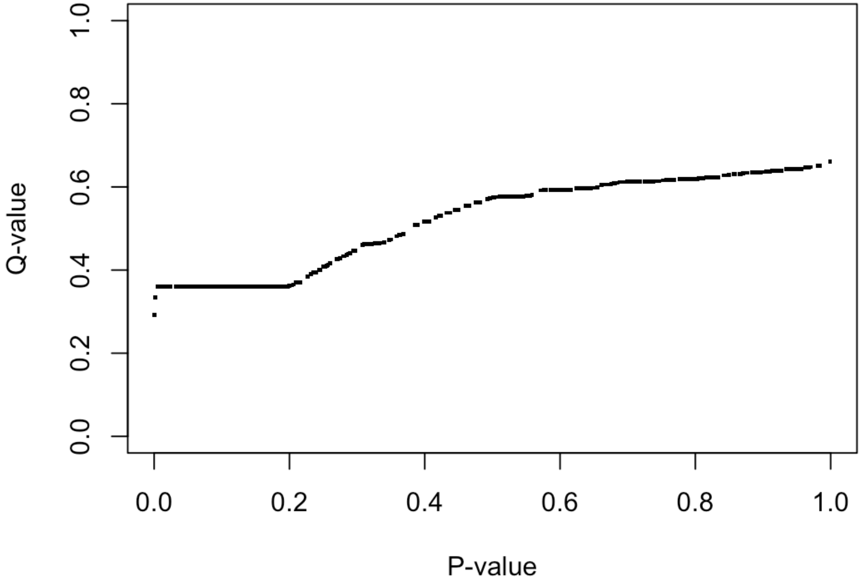


```
          pval         qval D           FD        FDR_t
1 6.273607e-31 4.985700e-27 1 4.985700e-27 4.985700e-27
2 5.552917e-28 2.206480e-24 2 4.412960e-24 2.206480e-24
3 4.579206e-27 1.213047e-23 3 3.639142e-23 1.213047e-23
4 3.767790e-25 7.485752e-22 4 2.994301e-21 7.485752e-22
5 1.715703e-24 2.726974e-21 5 1.363487e-20 2.726974e-21
```

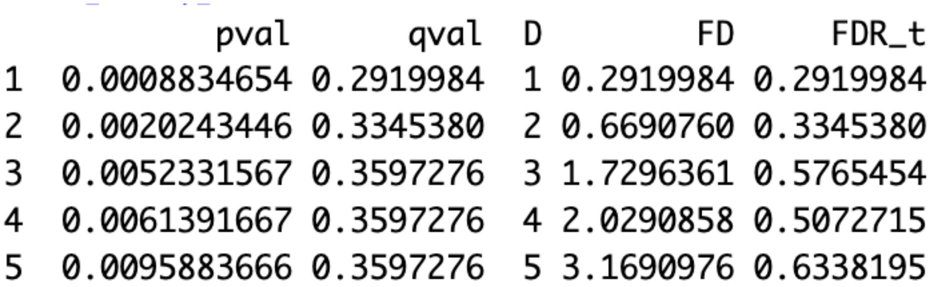
Example 2







These data have non-null P-values in [0.05,0.2].
Consequently FDR(t) goes down after 0.05
And many *Q*-values are equal (0.3597).



```
          pval      qval  D        FD     FDR_t
1 0.0008834654 0.2919984  1 0.2919984 0.2919984
2 0.0020243446 0.3345380  2 0.6690760 0.3345380
3 0.0052331567 0.3597276  3 1.7296361 0.5765454
4 0.0061391667 0.3597276  4 2.0290858 0.5072715
5 0.0095883666 0.3597276  5 3.1690976 0.6338195
```