The DiscreteFDR Package for Multiple Testing with Discrete Data

useR! 2024

Florian Junge
Sebastian Döhler
Darmstadt University of Applied Sciences
Department of Mathematics and Natural Sciences
Darmstadt Institute of Statistics and Operations Research

July 10, 2024

Agenda

1 Introduction

The Multiple Testing Problem

Big data analysis often involves a huge number m of statistical tests, e.g.

- Gene expression analysis ($m \approx 10^5$)
- Next-generation sequencing ($m \approx 10^6$)

m null hypotheses $H_0^1, \ldots, H_0^m \Rightarrow p$-values $p_1, \ldots, p_m$

- For a given significance level $\alpha \in (0, 1)$ it is expected to *falsely* reject $\alpha \cdot m$ hypotheses.
- $P(\text{at least one false discovery}) = 1 - (1 - \alpha)^m \xrightarrow{m \to \infty} 1$

$\Rightarrow$ **The more tests, the higher the risk of making false discoveries!**

$\Rightarrow$ Need for procedures that keep the number of false discoveries low **with good power**.

False Discovery Rate (FDR)

Definition:

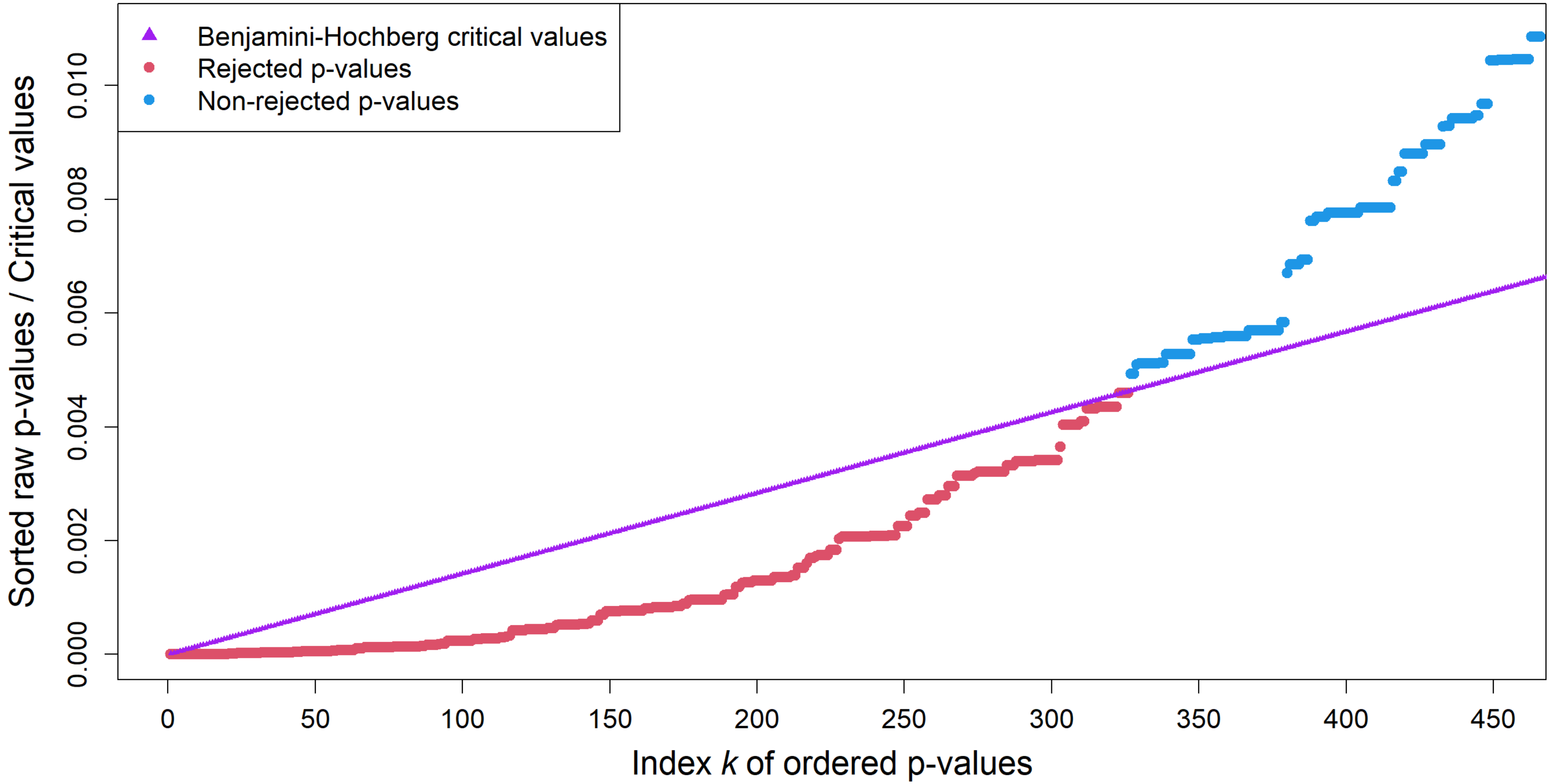
$$FDR := E\left(\frac{\text{number of falsely rejected hypotheses}}{\text{total number of rejected hypotheses}}\right)$$

Benjamini-Hochberg (BH) Procedure

Gold standard:

- Simple
- Powerful
- Guarantees $FDR \leq \alpha$

Idea: Compare ordered p-values $pv_{(k)}$ with critical values $\tau_k^{BH} = \frac{k}{m}\alpha$



Example: 450 smallest p-values out of 3525, BH rejections = 326

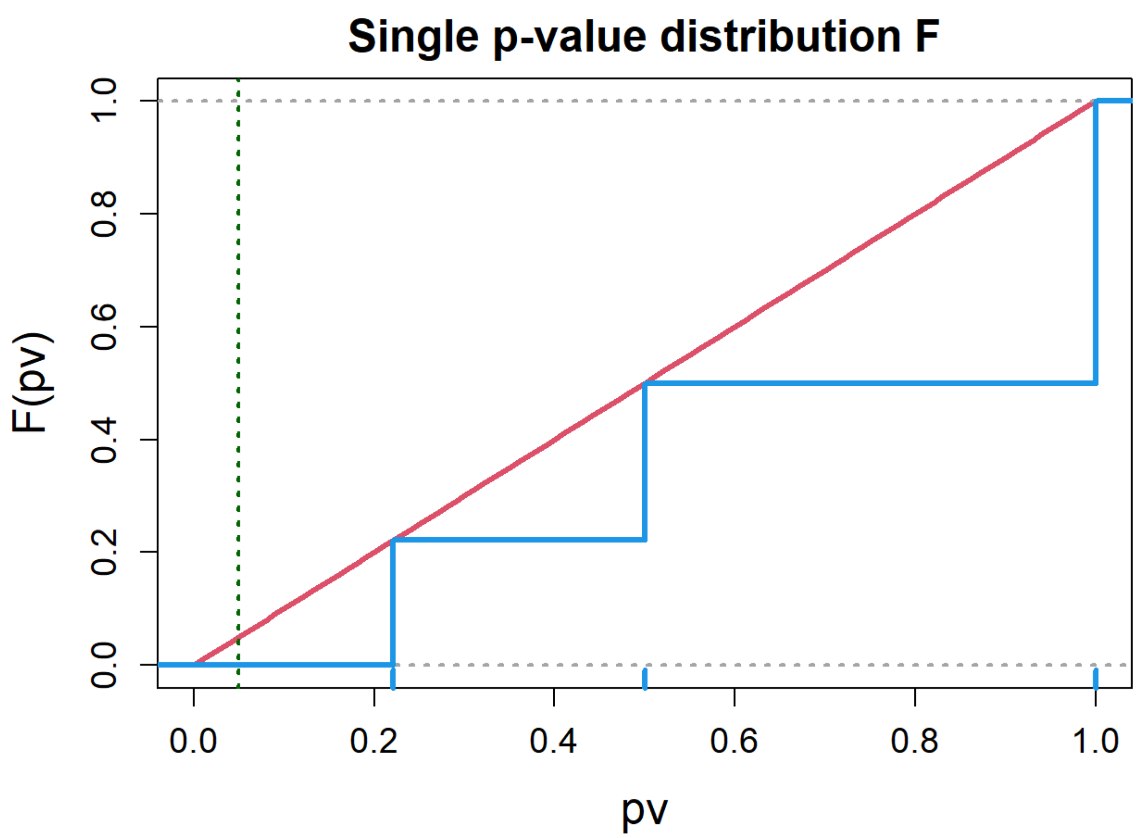
Challenge with Discrete Tests

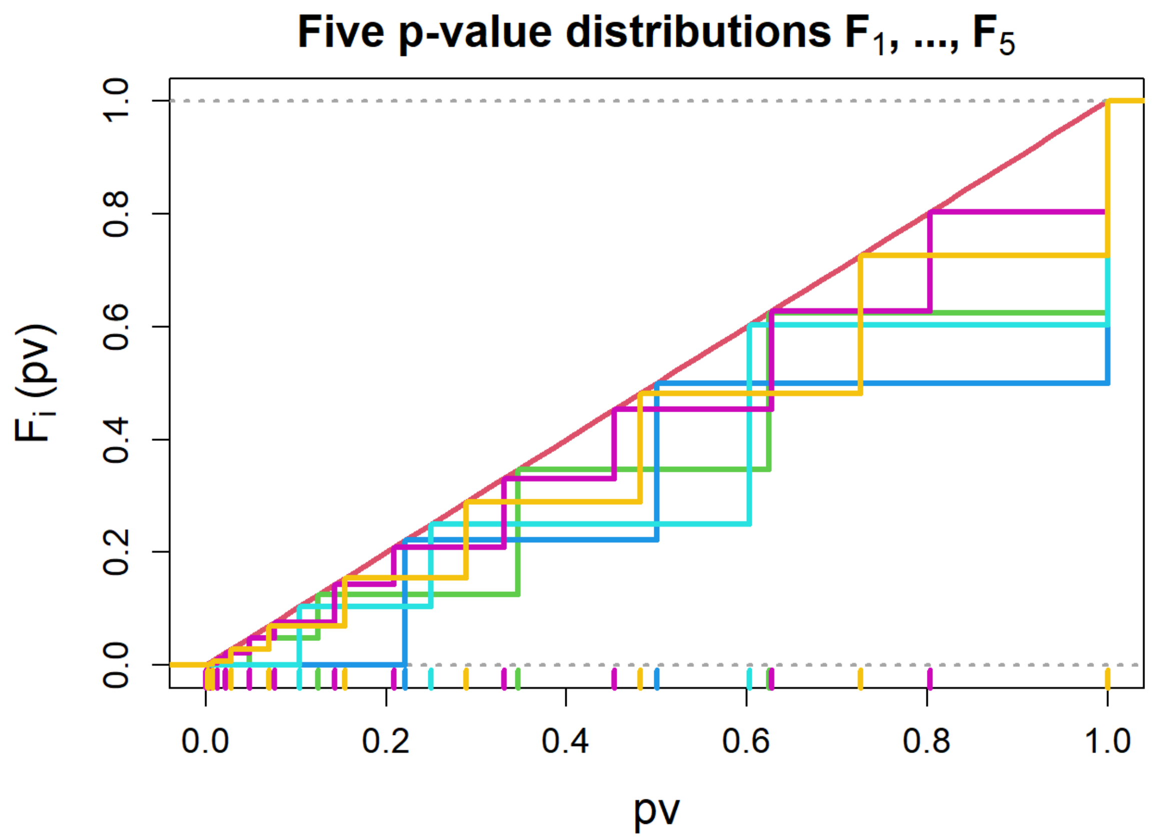
Continuous tests: under H_0, p-values are distributed with $P(PV \leq x) = F(x) = x$

$\Rightarrow$ observable p-value *support*: continuous **interval**

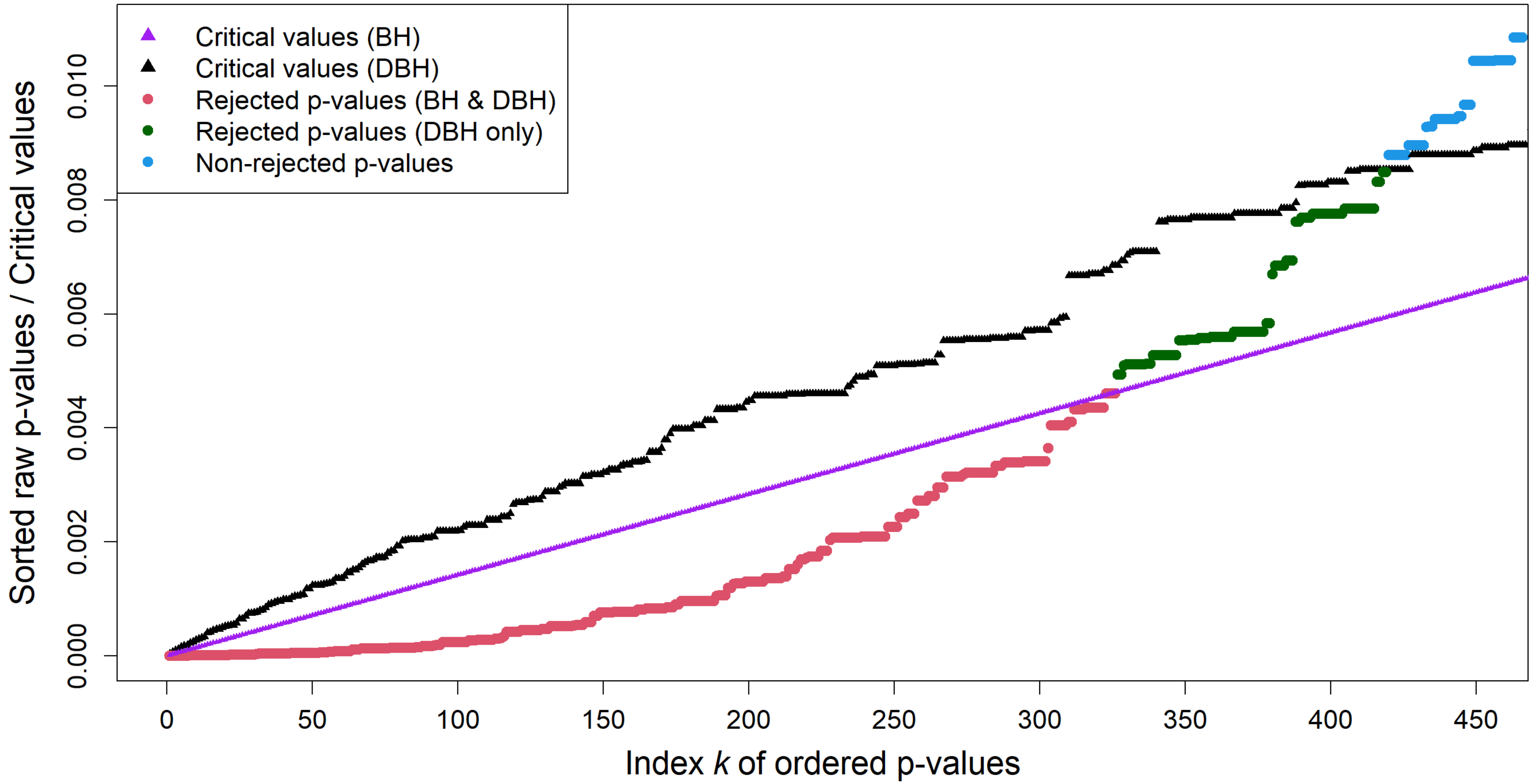
Discrete tests: under H_0, p-values are discretely distributed with $F(x) \leq x$

$\Rightarrow$ observable p-value *support*: discrete **set**





Idea of Discrete Benjamini-Hochberg



Here: BH rejections = 326, DBH rejections = 419

Discrete Benjamini-Hochberg Procedure

From Döhler, Durand & Roquain (2018)

- Central idea: transform p-values for discreteness via $\xi(pv) = \frac{1}{m}\sum_{i=1}^{m}\frac{F_i(pv)}{1-F_i(\tau_m^{DBH})}$
  ($F_1, \ldots, F_m$: p-value distribution functions under their respective nulls)

- Two ways of making decisions:
  a. Use standard BH critical values: $\xi\left(pv_{(k)}\right) \leq \tau_k^{BH}$
  b. Use new discrete critical values: $pv_{(k)} \leq \tau_k^{DBH} = \max\left\{pv \in \mathcal{A} \;\middle|\; \xi(pv) \leq \tau_k^{BH}\right\}$

2 The `DiscreteFDR` R package

Package Family

`DiscreteFDR` is part of a family of packages

- `DiscreteFDR`:
  - originally created and maintained by Guillermo Durand
  - major updates in early July 2024
- `DiscreteTests`:
  - provides vectorised(!) functions for computing exact p-values and their supports of several discrete tests
  - results are output in a `DiscreteTestResults` R6 class object
- `DiscreteDatasets`:
  - provides some benchmark datasets from literature that can be used with `DiscreteTests` and `DiscreteFDR`

Main Function

Generic function `discrete.BH()` with two methods:

```
# "traditional" method
discrete.BH.default()
# new, preferred method
discrete.BH.DiscreteTestResults()
```

most important parameters:

- `test.results`:
  - numeric vector that includes the p-values **or**
  - object of R6 class `DiscreteTestResults` from package `DiscreteTests`
- `pCDFlist`: list of (numeric) support sets; `.default` method only
- `alpha`: FDR level (default: `0.05`)
- `ret.crit.consts`: compute and return critical constants (default: `FALSE`)

Application Example

Preparations

```r
# data
library(DiscreteDatasets)
data("listerdata_four_columns")
print(listerdata_four_columns, max = 12)
```

```
            Col0_Counts.ThisCyto Met13_Counts.ThisCyto
AT1G01070.1                    9                    24
AT1G01070.2                    9                    24
AT1G01150.1                    3                     3
            Col0_Counts.AllOtherCytos Met13_Counts.AllOtherCytos
AT1G01070.1                     34235                      39318
AT1G01070.2                     34235                      39318
AT1G01150.1                     34241                      39339
 [ reached 'max' / getOption("max.print") -- omitted 3522 rows ]
```

```r
# Fisher's exact test (fisher.test.pv returns R6 class object)
library(DiscreteTests)
lister_test <- fisher.test.pv(listerdata_four_columns)
```

Perform Discrete Benjamini-Hochberg

```
library(DiscreteFDR)
# "modern"
DBH_lister <- discrete.BH(lister_test, ret.crit.consts = TRUE)
# "traditional"
DBH_lister <- discrete.BH(lister_test$get_pvalues(), lister_test$get_pvalue_support
                          ret.crit.consts = TRUE)

str(DBH_lister, max.level = 2)
```

```
List of 5
 $ Rejected       : num [1:419] 4.60e-03 4.22e-04 8.79e-06 1.74e-03 5.37e-04 ...
 $ Indices        : int [1:419] 13 17 31 54 61 69 94 101 118 131 ...
 $ Num.rejected   : int 419
 $ Critical.values: num [1:3525] 3.30e-05 6.28e-05 9.29e-05 1.14e-04 1.51e-04 ..
 $ Data           :List of 5
  ..$ Method      : chr "Discrete Benjamini-Hochberg procedure (step-up)"
  ..$ raw.pvalues: num [1:3525] 0.0348 0.0348 1 0.0088 0.2818 ...
  ..$ pCDFlist    :List of 3525
  ..$ FDR.level   : num 0.05
  ..$ Data.name   : chr "lister_test$get_pvalues() and lister_test$get_pvalue_"..
 - attr(*, "class")= chr [1:2] "DiscreteFDR" "list"
```

print and summary methods

```r
# print method
print(DBH_lister)
```

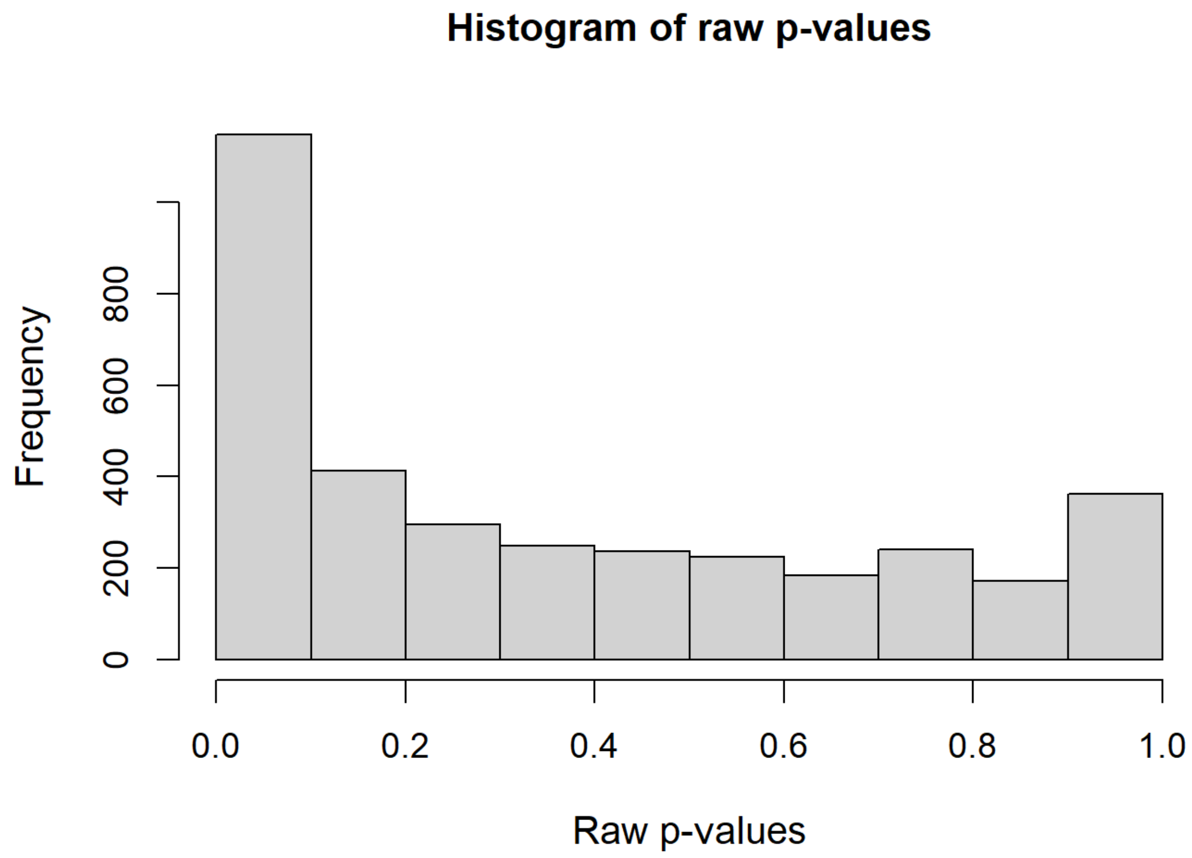
```
	Discrete Benjamini-Hochberg procedure (step-up)

Data:  lister_test$get_pvalues() and lister_test$get_pvalue_supports()
Number of tests = 3525
Number of rejections = 419 at global FDR level 0.05
(Original BH rejections = 326)
Largest rejected p value:  0.008487787
```

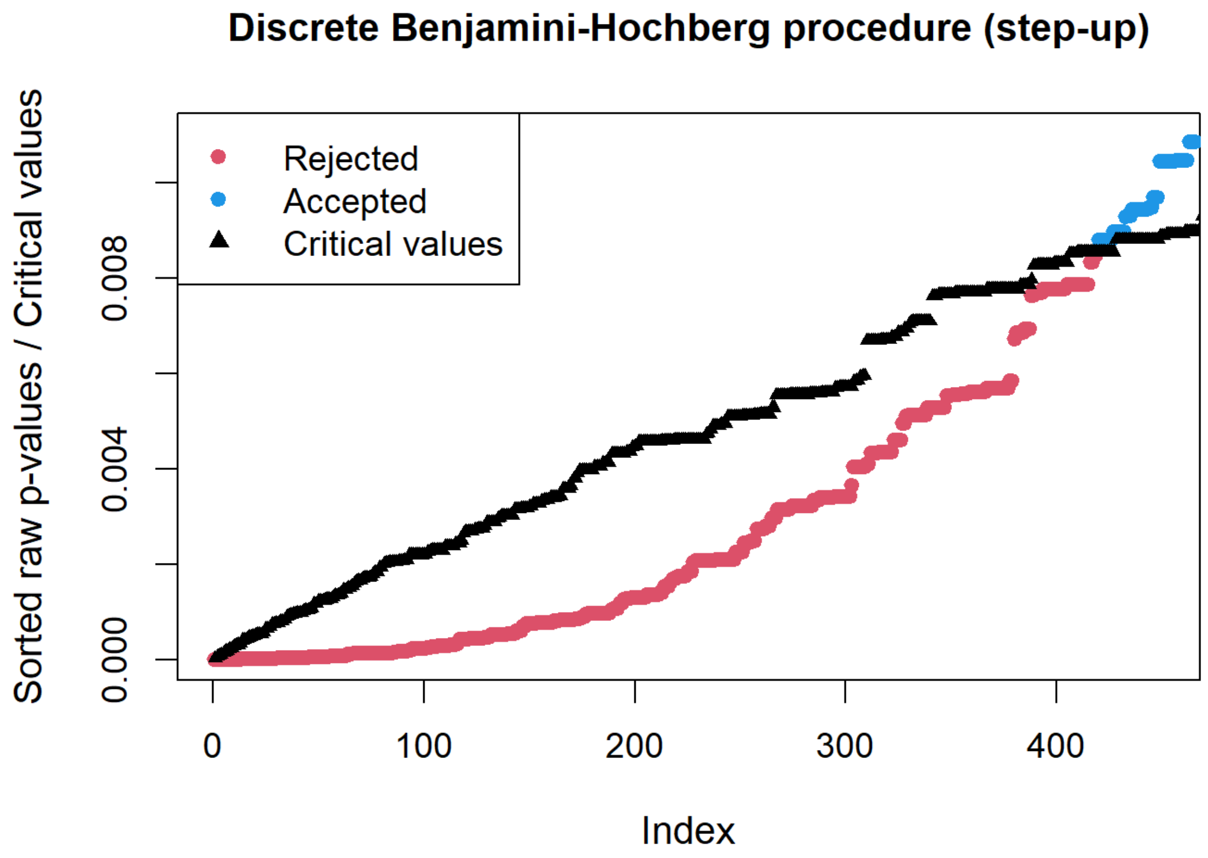
```r
# summary method
summary(DBH_lister)$Table |> head(5)
```

```
  Index      P.value Critical.value Rejected
1   228 1.465774e-07   3.299848e-05     TRUE
2  1705 3.029058e-07   6.277768e-05     TRUE
3  3524 5.742313e-07   9.286340e-05     TRUE
4   942 1.140346e-06   1.141108e-04     TRUE
5  1864 1.179705e-06   1.509283e-04     TRUE
```

hist and plot methods

```
# hist method
hist(DBH_lister)
# plot method
plot(DBH_lister, xlim = c(1, 450), ylim = c(0, 0.011), type.crit = 'p',
     pch = c(20, 20, 17), lwd = c(3, 3, 1), cex = c(1, 1, 0.7), legend = "topleft")
```





Pipes

```r
# pipe
listerdata_four_columns |>
  fisher.test.pv() |>
  discrete.BH() |>
  summary() |>
  with(Table) |>
  head()
```

```
  Index      P.value Rejected
1   228 1.465774e-07     TRUE
2  1705 3.029058e-07     TRUE
3  3524 5.742313e-07     TRUE
4   942 1.140346e-06     TRUE
5  1864 1.179705e-06     TRUE
6  2287 1.179705e-06     TRUE
```

3 Outlook

- `FDX`: apply changes of `DiscreteFDR` package
  - performance improvements
  - enable pipes
- `DiscreteTests`: more discrete tests
- `DiscreteDatasets`: more datasets

Literature

- Döhler, S., Durand, G. and Roquain, E. (2018). New FDR bounds for discrete and heterogeneous tests. *Electronic Journal of Statistics*, *12* (1), pp. 1867-1900. doi: 10.1214/18-EJS1441
- Durand, G., Junge, F., Döhler, S. and Roquain, E. (2019). DiscreteFDR: An R package for controlling the false discovery rate for discrete test statistics. *arXiv*. arXiv: 1904.02054
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*. *57* (1), pp. 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Thank you! Any Questions?