Advanced Regression: 2c Multiple testing

Verena Zuber

Epidemiology and Biostatistics, Imperial College London

24th January 2019

Multiple testing: What is the problem?

Bonferroni correction

False discovery rate (FDR)

- The concept of FDR
- Benjamini and Hochberg Fdr
- Tail area based Fdr: The q-value
- Local fdr

FDR in practice

Multiple tests: What is the problem?

1. Assume we want to perform one statistical test. What can go wrong?

		True Null hypothesis (H_0) is True	True Null hypothesis (H_0) is False
Decision	Accept	True Negative H_0	Type II error (False negative)
	Reject H_0	Type I error (False positive)	True Positive H_1

- The type I error rate or significance level is the probability of rejecting H_0 given that it is true (False positive).
- By fixing a significance threshold like e.g. $\alpha = 0.05$ we ensure that the probability to have a false positive finding is small.

Multiple testing: What is the problem?

2. Assume we want to perform two tests.
   - What is the probability that we do not make *any* false positive in any of the two tests?

$$(1-\alpha)*(1-\alpha) = 0.95*0.95 = 0.9025$$

3. Assume we want to perform $N = 15$ tests.
   - What is the probability that we do not make *any* false positive in any of the 15 tests?

$$(1-\alpha)^{15} = 0.95^{15} = 0.4632912$$

We do not control the type 1 error when performing multiple tests.

Bonferroni correction

- Bonferroni correction is a simple tool to adjust for multiple testing.
- Assume we consider N tests.
- There are two equivalent ways of performing Bonferroni correction:
  1. Adjust your p-values: Multiply your p-values by N

$$p_i^{\text{Bonferroni}} = p_i N$$

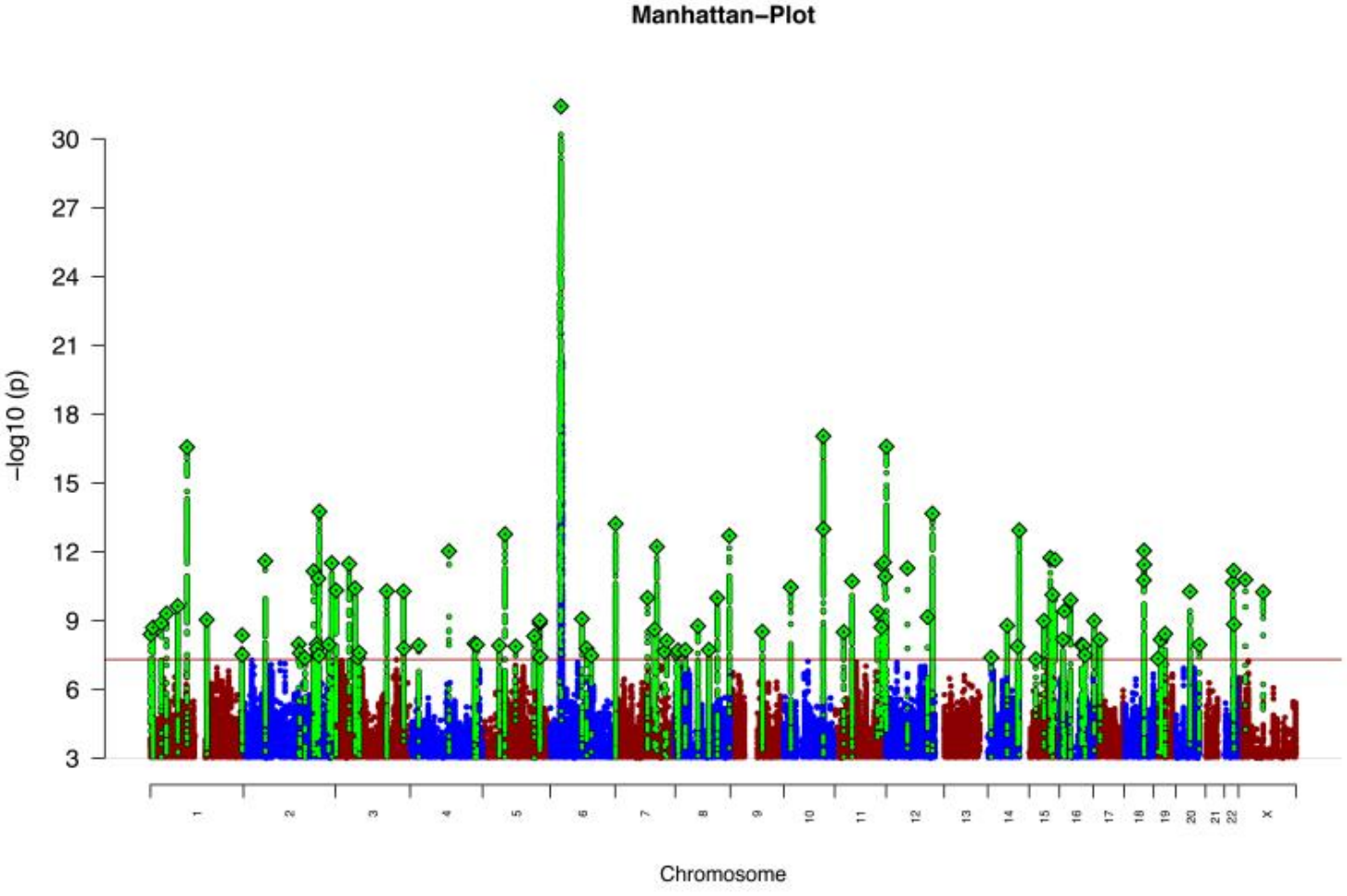
  2. Adjust your significance threshold: Divide your α by N

$$\alpha^{\text{Bonferroni}} = \alpha / N$$

- R-command: `p_vals` is a vector of p-values of length N: `p.adjust(p_vals, method = 'bonferroni')`
- Example: Genome-wide significance threshold $p < 5 \times 10^{-8}$ is a Bonferroni correction that accounts for $1m$ independent SNPs.

Genome-wide significance

Manhattan plot of association with schizophrenia



- Publicly available summary data from the PGC consortium

Bonferroni correction

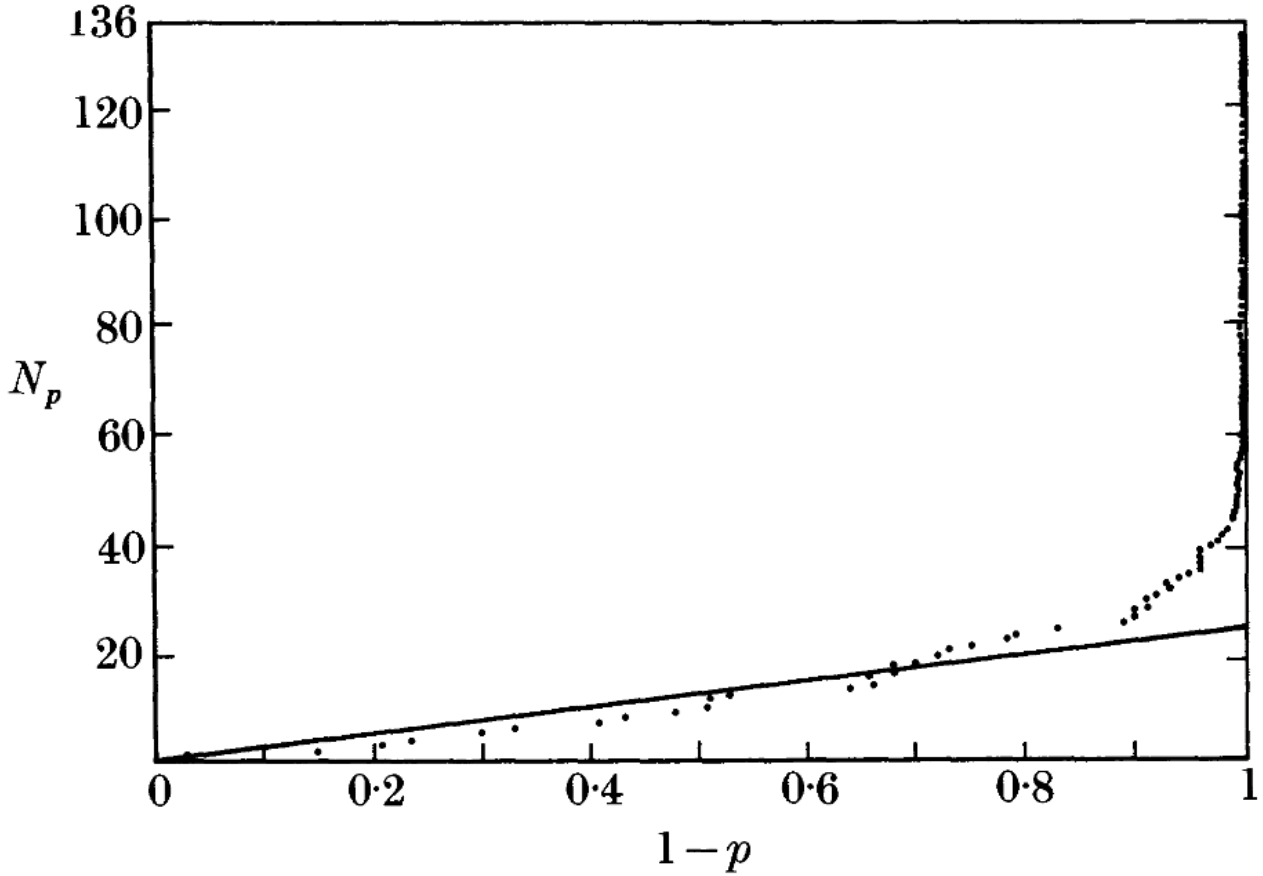
Advantages:

- Very easy to perform.
- Non-parametric

Disadvantages:

- Very conservative $\rightarrow$ Many false negatives!
- Number of tests is a penalty: The more tests that are performed, the stricter the threshold.

Plot of p-values



- Under the Null $p-$values follow a straight line.

Schweder and Spjotvoll 1982

The concept of false discovery rate (FDR)

What we are actually interested in are the discoveries, ie the variables we declare to be Non-Null.

- We are not really interested in the Null variables.
- Within the discoveries (declared as Non-Null) we distinguish between
  - True discoveries
  - False discoveries

Aims:

- Focus on the discoveries only.
- Control the false discovery rate.

A quick note on notation

- FDR: General concept of false discovery rate
- Fdr: Tail-area based FDR (including Benjamini and Hochberg Fdr and q-value)
- fdr: Local FDR

What does Null and Non-Null mean?

- Null: Noise, H_0, acceptions, or not interesting variables
- Non-Null: Signal, H_1, rejections, alternative, or interesting variables

True and false discoveries

		True, Actual Null	Non-Null	
Decision	Null	$N_0 - a$	$N_1 - b$	$N - R$
	Non-Null	a	b	R
		N_0	N_1	N

- The total sample size is N, the number of all tests performed
- a is the number of false discoveries
- b is the number of true discoveries
- a/R is the proportion of false discoveries among all discoveries

Benjamini-Hochberg step-up procedure

1. The first step in FDR calculation is to sort the $p-$values of N tests

$$p_{[1]} \leq p_{[2]} \leq \ldots \leq p_{[i]} \leq \ldots \leq p_{[N-1]} \leq p_{[N]}$$

   where $p_{[1]}$ is the smallest and $p_{[N]}$ is the largest $p-$value.
2. Fix $q \in (0, 1)$ as the level at which to control the FDR. Often $q = 0.05$ is used as convention.
3. Select $i_{\max}$ as the largest i for which the following holds

$$p_{[i]} \leq \frac{i}{N} q$$

4. All $i \leq i_{\max}$ are considered as discoveries and all $i > i_{\max}$ are considered as Null.

Benjamini and Hochberg 1995

Another look at Benjamini-Hochberg

- This is equivalent to adjusting the $p-$values as follows

$$p_i^{\text{BH}} = p_i \frac{N}{\text{order}(i)}$$

where order (i) is the rank of the ith variable, which equals 1 for the smallest p-value and m for the largest.

$$p_i \leq p_i^{\text{BH}} = p_i \frac{N}{\text{order}(i)} \leq p_i^{\text{Bonferroni}} = p_i N$$

Strimmer 2008

Multiple testing using `p.adjust()`

- `method = 'bonferroni'`
- `method = 'BH'`

```
> pvec
 [1] 0.000100000 0.001000000 0.007550354 0.322375702 0.325354898 0.345921639
 [7] 0.577974827 0.619889967 0.721198847 0.809660355
> p.adjust(pvec,method="BH")
 [1] 0.00100000 0.00500000 0.02516785 0.57653607 0.57653607 0.57653607
 [7] 0.77486246 0.77486246 0.80133205 0.80966035
> p.adjust(pvec,method="bonferroni")
 [1] 0.00100000 0.01000000 0.07550354 1.00000000 1.00000000 1.00000000
 [7] 1.00000000 1.00000000 1.00000000 1.00000000
```

Multiple testing using `p.adjust()`

	unadjusted	BH	Bonferroni
1	0.0001	0.0010	0.0010
2	0.0010	0.0050	0.0100
3	0.0076	0.0252	0.0755
4	0.3224	0.5765	1.0000
5	0.3254	0.5765	1.0000
6	0.3459	0.5765	1.0000
7	0.5780	0.7749	1.0000
8	0.6199	0.7749	1.0000
9	0.7212	0.8013	1.0000
10	0.8097	0.8097	1.0000

Tail area based Fdr

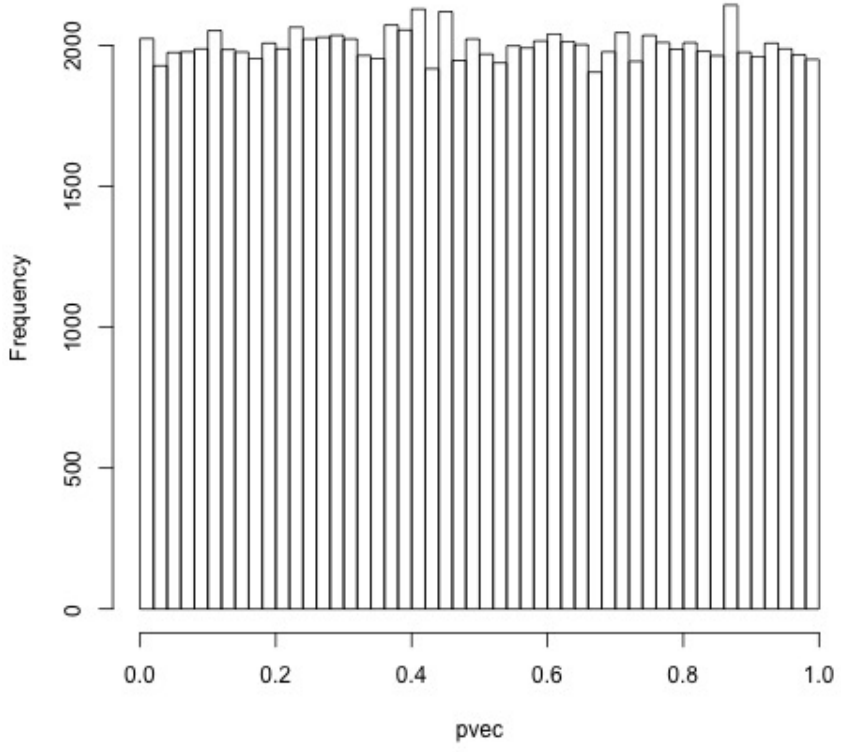
We can consider that our vector of p-values is generated from a mixture distribution. We model the cumulative density function as

$$F(p) = \underbrace{\pi_0 F_0(p)}_{\text{Null component}} + \underbrace{\pi_1 F_A(p)}_{\text{Signal component}}$$

where

- π_0: Null proportion
- $\pi_1 = 1 - \pi_0$: Non-Null proportion
- F_0 : Cumulative density function under the Null
- F_A : Cumulative density function under the Alternative

Null-distribution for $p-$values



- Under the Null, $p-$values follow a uniform distribution between 0 and 1.
- Cumulative density function of the uniform distribution simplifies to $F(p_i) = p_i$

Tail area based Fdr

- This allows us to define the tail-area based Fdr as

$$Fdr(p_i) = \mathbb{P}(\text{“}uninteresting\text{”} \mid P \leq p_i) = \frac{\pi_0 F_0(p_i)}{F(p_i)} = \frac{\pi_0 p_i}{F(p_i)}.$$

- Read: Probability that variable i is uninteresting given that the p-value is as small or smaller than the observed p_i.
- Estimate of the cumulative density function is

$$\hat{F}(p_i) = \frac{\text{order}(i)}{N}.$$

- Thus the estimated Fdr simplifies to

$$\hat{Fdr}(p_i) = \frac{\pi_0 p_i}{\hat{F}(p_i)} = \pi_0 \times p_i \frac{N}{\text{order}(i)} = \pi_0 \times p_i^{\text{BH}}.$$

Fdr and Benjamini-Hochberg

$$\hat{Fdr}(p_i) = \pi_0 \times p_i^{\text{BH}}$$

- π_0 and π_1 can be estimated from the data.
- Interpretation: Proportion of true Non-Null variables
- The Benjamini-Hochberg procedure is equivalent to Fdr, where $\pi_0 = 1$.
- Interpretation: Benjamini-Hochberg is more conservative than Fdr estimates, where $\pi_0 \leq 1$ is estimated from the data.

Storey's q-value

- One of the most popular implementations of the Fdr is the $q-$value in the bioconductor R-package `qvalue`.
- In order to fit the Fdr use:
  `qobj = qvalue(p=pvec)`
- Values
  - Proportion of true Null variables `qobj$pi0`
  - Fdr values `qobj$qvalues`

Storey and Tibshirani, 2003

Local fdr

Again we consider that our vector of p-values is generated from a mixture distribution. The local fdr considers the density function as

$$f(p) = \underbrace{\pi_0 f_0(p)}_{\text{Null component}} + \underbrace{\pi_1 f_A(p)}_{\text{Signal component}}$$

where

- π_0: Null proportion
- $\pi_1 = 1 - \pi_0$: Non-Null proportion
- f_0 : Density function under the Null
- f_A : Density function under the Alternative

Local fdr vs tail-are based Fdr

1. Local fdr

$$fdr(p_i) = \mathbb{P}(\text{“}uninteresting\text{”} \mid P = p_i) = \frac{\pi_0 f_0(p_i)}{f(p_i)}$$

Interpretation:

- Probability of the null model conditional on the observed test statistic $p_i \rightarrow$ Empirical Bayesian posterior probability for a variable to be Null given the observed data

2. Tail-are based Fdr

$$Fdr(p_i) = \mathbb{P}(\text{`}uninteresting\text{'} \mid P \leq p_i) = \frac{\pi_0 F_0(p_i)}{F(p_i)} = \frac{\pi_0 p_i}{F(p_i)}$$

Interpretation:

- Controls the number of false discoveries
- Provides an adjusted $p-$value

Variables are treated as observations

- In order to fit a mixture model we benefit from more tests.
- The more tests we have the more precise we can fit the components of the mixture model.

Variables are treated as observations.

Course becomes a blessing.

- Implicit assumption: The signal is sparse, ie there are much more Null than Non-Null variables.
- This ensures the identifiability of the mixture model.

Adding more variables

- Assume we have tested $N = 100,000$ SNPs on a customised SNP-chip.
- Our technician found a new imputation method to double the coverage of SNPs to $N = 200,000$.

How would this impact different multiple testing approaches?

Adding more variables

- Assume we have tested $N = 100,000$ SNPs on a customised SNP-chip.
- Our technician found a new imputation method to double the coverage of SNPs to $N = 200,000$.

How would this impact different multiple testing approaches?

1. Bonferroni: We need to lower our significance threshold from $\alpha/100,000$ to $\alpha/200,000$.
   $\rightarrow$ The number of tests is a penalty.
2. FDR: We would need to refit our FDR, but since the FDR is only concerned with the 'discoveries' and the rate of false discoveries within the discoveries, there is no penalty in adding more variables.

Genetic associations with schizophrenia

- Psychiatric Genetics Consortium has published genome-wide summary data on genetic association with schizophrenia.
- This study is based on a meta-analysis including 36,989 cases and 113,075 controls.
- The summary-level data is public and free to download from `https://www.med.unc.edu/pgc/results-and-downloads`

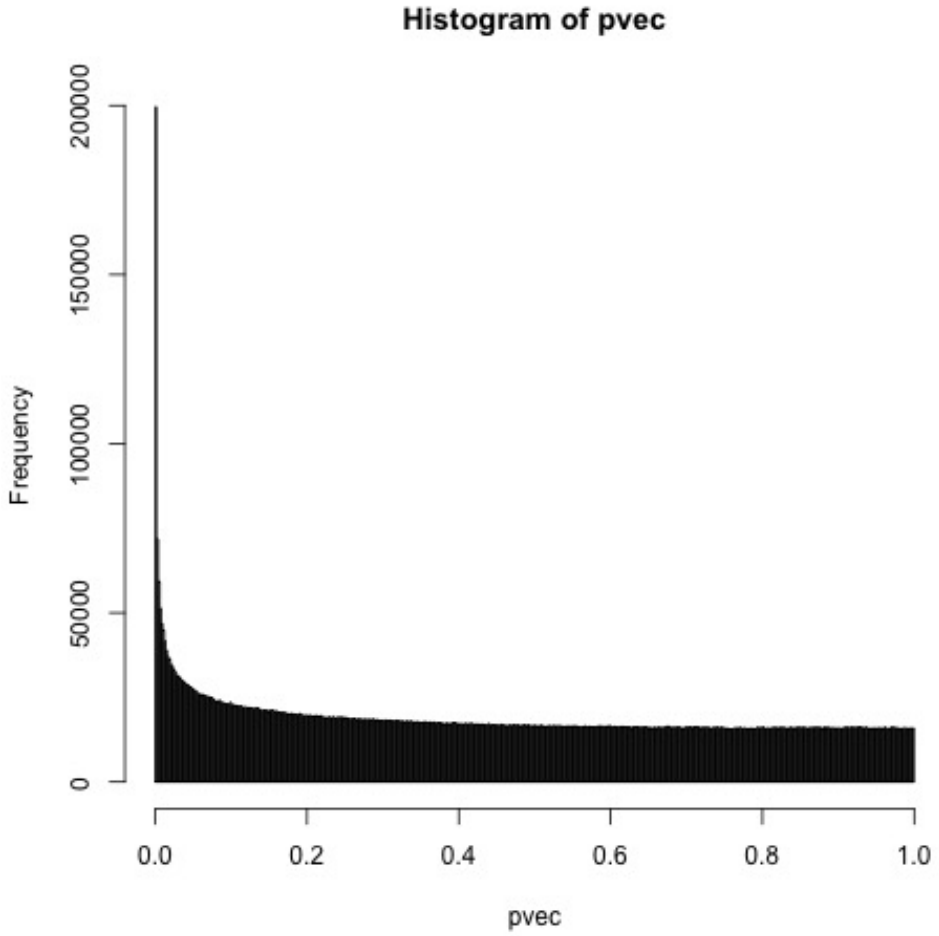
Genetic associations with schizophrenia

- After downloading the datafile and un-zipping, we can read the file 'ckqny.scz2snpres' into R.
- The summary data file includes regression coefficients, standard errors, and $p-$values on $N = 9,444,230$ SNPs.

```
> head(data)
  hg19chrc       snpid a1 a2     bp  info      or     se      p ngt
1     chr1   rs4951859  C  G 729679 0.631 0.97853 0.0173 0.2083   0
2     chr1 rs142557973  T  C 731718 0.665 1.01949 0.0198 0.3298   0
3     chr1 rs141242758  T  C 734349 0.666 1.02071 0.0200 0.3055   0
4     chr1  rs79010578  A  T 736289 0.649 0.98748 0.0193 0.5132   0
5     chr1 rs143225517  T  C 751756 0.853 0.99681 0.0164 0.8431   0
6     chr1   rs3094315  A  G 752566 0.881 0.99601 0.0149 0.7870  36
```

Distribution of p-values

- There are $12,897$ that have a $p-$value smaller than 5×10^{-8}.
- That is 0.00137% of all SNPs.



Adjusting for multiple testing: `p.adjust()`

- Bonferroni correction
  - `p_bonferroni=p.adjust(p=pvec, method='bonferroni')`
  - There are $9,323$ SNPs significant at a level of 0.05.
  - That is 0.00099% of all SNPs.
- Benjamini Hochberg
  - `p_bh=p.adjust(p=pvec, method='BH')`
  - There are $124,955$ SNPs significant at a level of 0.05.
  - That is 0.01323% of all SNPs.

Adjusting for multiple testing: `qvalue()`

- The `qvalue` package is on Bioconductor.
- We first fit a qvalue object `qobj = qvalue(p=pvec)`.
  - $q-$values: `qvalues = qobj$qvalues`
  - Proportion of Null SNPs: `pi0 = qobj$pi0`
- Results:
  - There are $141,181$ SNPs significant at a $q-$level of 0.05.
  - That is 0.01494892% of all SNPs.
  - Proportion of Null SNPs is 0.8394974, thus the proportion of Non-Null SNPs is 0.1605026.

Adjusting for multiple testing: local fdr

- An estimator of the local fdr is in the `fdrtool` package.
- We first fit a fdrtool object `lfdr_out = fdrtool(x=pvec, statistic='pvalue')`.
  - local fdr: `lfdr_out$lfdr`
  - Proportion of Null SNPs: `lfdr_out$param`

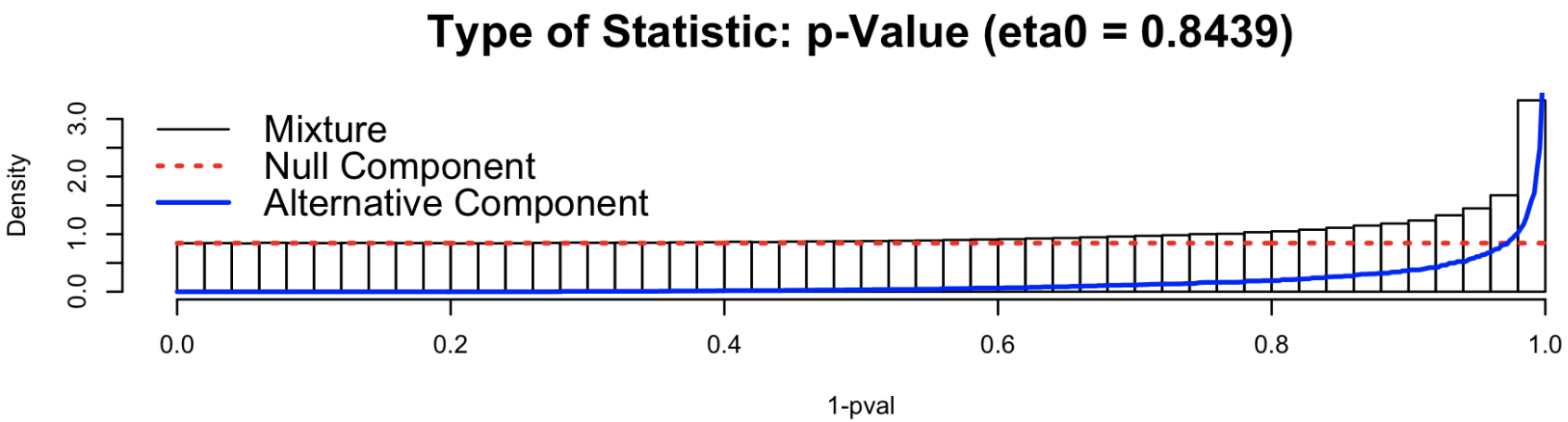
```
> lfdr_out$param
        cutoff  N.cens      eta0      eta0.SE
[1,] 0.7401648 2070932 0.8439201 0.0005181626
```

- Results:
  - There are 212,322 SNPs significant at a local fdr level of 0.2.
  - That is 0.02248166% of all SNPs.
  - Proportion of Null SNPs is 0.8439201, thus the proportion of Non-Null SNPs is 0.1560799.

Adjusting for multiple testing: local fdr

Plot of model fit

- f_A: alternative density in blue
- f_0: Null density (Uniform distribution) in red



Comparison of multiple testing methods

discoveries	GWAS-significant	Bonferroni	BH	q-value Fdr	local fdr
#	12,897	9,323	124,955	141,181	212,322
%	0.00137	0.00099	0.01323	0.01495	0.02248

MENU nature genetics

Letter | Published: 17 July 2017

Association analyses based on false discovery rate implicate new loci for coronary artery disease

Christopher P Nelson, Anuj Goel […] Panos Deloukas

Nature Genetics **49**, 1385–1391 (2017) | Download Citation

- GWAS in coronary artery disease had identified 66 loci at 'genome-wide significance' ($P < 5 \times 10-8$).
- There is a much larger number of putative loci at a Fdr of 5%.
- Increasing the sample size of the GWAS by using UK Biobank yielded 13 new loci at genome-wide significance.
- Of these 13 new loci, 12 had been discovered in Fdr analysis on the smaller sample size.
- 'Thus providing strong support that the remaining loci identified by FDR represent genuine signals.'

Take away: Multiple testing

- When performing multiple tests it is essential to correct for multiple testing.
- Bonferroni adjustment is simple, but very conservative and has a high false negative rate.
- FDR approaches provide more powerful approaches for multiple testing.
- Benjamini-Hochberg were the first to provide a new paradigm how to look at data ('variables become observations').
- The $q-$value is more powerful than the BH adjustment as it allows to estimate the proportion of Null variables.
- The local fdr is an Empirical Bayesian posterior probability for a variable to be Null.

Outlook: Next week

Practical:

- Perform variable ranking on high-dimensional data-sets
- Compute Bonferroni multiple testing correction
- Compute and contrast various FDR approaches (Benjamini-Hochberg, q-value and local fdr)

Lectures:

- Principles of high-dimensional data analysis
- Penalised regression models (ridge, lasso, elastic net)
- Prediction accuracy and cross-validation

Reading list

Practical topic on the epigenetic clock:

- The clock watcher: Biomathematician Steve Horvath has discovered a strikingly accurate way to measure human ageing through epigenetic signatures.
  `https://www.nature.com/news/biomarkers-and-ageing-the-clock-watcher-1.15014`

More on the FDR:

- False Discovery Rate Control chapter 4 in Computer Age Statistical Inference by Brad Efron and Trevor Hastie:
  `https://web.stanford.edu/~hastie/CASI_files/PDF/casi.pdf`