Advanced Regression: Multiple testing

Verena Zuber

Epidemiology and Biostatistics, Imperial College London

25th April 2019

Multiple testing: Motivations

False discovery rate

- The concept of FDR
- Formal definition of the FDR
- Benjamini-Hochberg procedure
- Local and tail-area based FDR

Motivation 1: Controlling type 1 error for multiple tests

- When performing one single test, we fix the type 1 error rate or significance level to for example $\alpha = 0.05$.
- The type 1 error rate is the probability of rejecting H_0 given that it is true (False positive).
- Thus we control the probability for a false positive finding when performing one single test.

Performing more than one test

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?

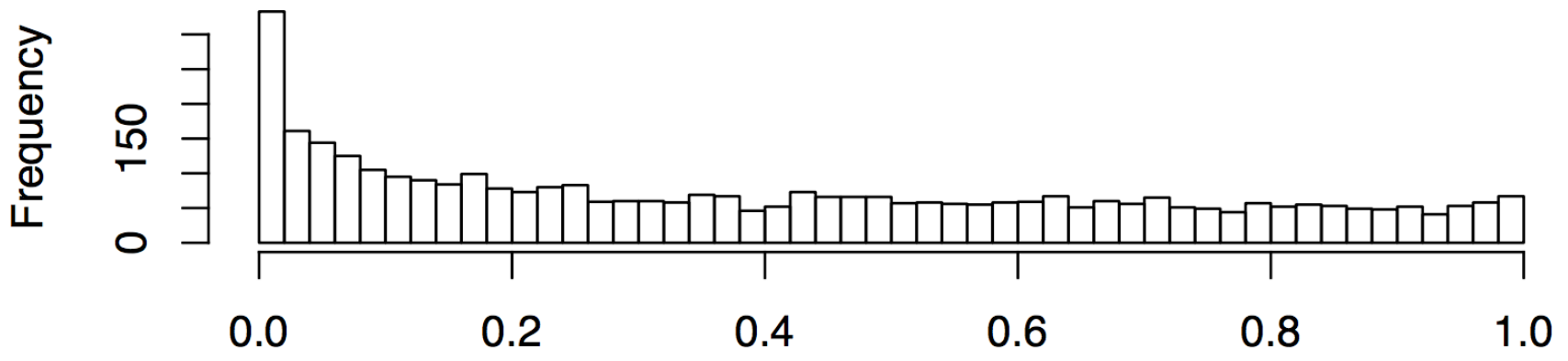
$$(1 - \alpha) * (1 - \alpha) = 0.95 * 0.95 = 0.9025$$

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

Motivation 2: Define the cut-off of a ranked list

- Assume we have list of ranked variables.
- The ranking has been performed in a univariate fashion, ie each feature is evaluated individually (marginally) with respect to its association with the outcome for example in a massively univariate linear model or in a t-test.
- Examples:
  - Differentially expressed transcripts between cases and controls
  - Genetic variants associated with BMI or blood pressure
  - Methylation sites correlated with age (Practicals)

```
hist(pvec, breaks = 50, main="")
```



Epigenetic clock: Which methylation sites to report?

```
##                          pvec
##   [1,] "chr13_43122804" "2.22638293312029e-10"
##   [2,] "chr7_118434508" "2.83341285695823e-08"
##   [3,] "chr8_125759184" "8.08770585470011e-08"
##   [4,] "chr7_62419850"  "1.54558113496926e-07"
##   [5,] "chr7_99238006"  "5.02013794094029e-07"
##   [6,] "chr14_78777456" "1.20194579058332e-06"
##   [7,] "chr12_82783446" "1.2180090770785e-05"
##   [8,] "chr15_60172240" "1.71172214464869e-05"
##   [9,] "chr2_72217019"  "2.100193513036e-05"
##  [10,] "chr6_121015046" "4.01317786827999e-05"
```

How to define the cut-off in a ranked list?

- Multiple testing can be used to define a cut-off in a ranked list.

Example: Epigenetic clock

- Bonferroni-correction detected 7 methylation sites.
- Interpretation: The top 7 methylation sites contain with probability 95% no false positive.
- Benjamini Hochberg-correction detected 35 methylation sites.
- Interpretation: Within the top 35 methylation sites there will be in expectation 5% false positives, in this case less than 2 methylation sites will be false positives.

Motivation 3: pi0 the proportion of Null variables

The cumulative density function of the $p-$values is modeled as a mixture distribution

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

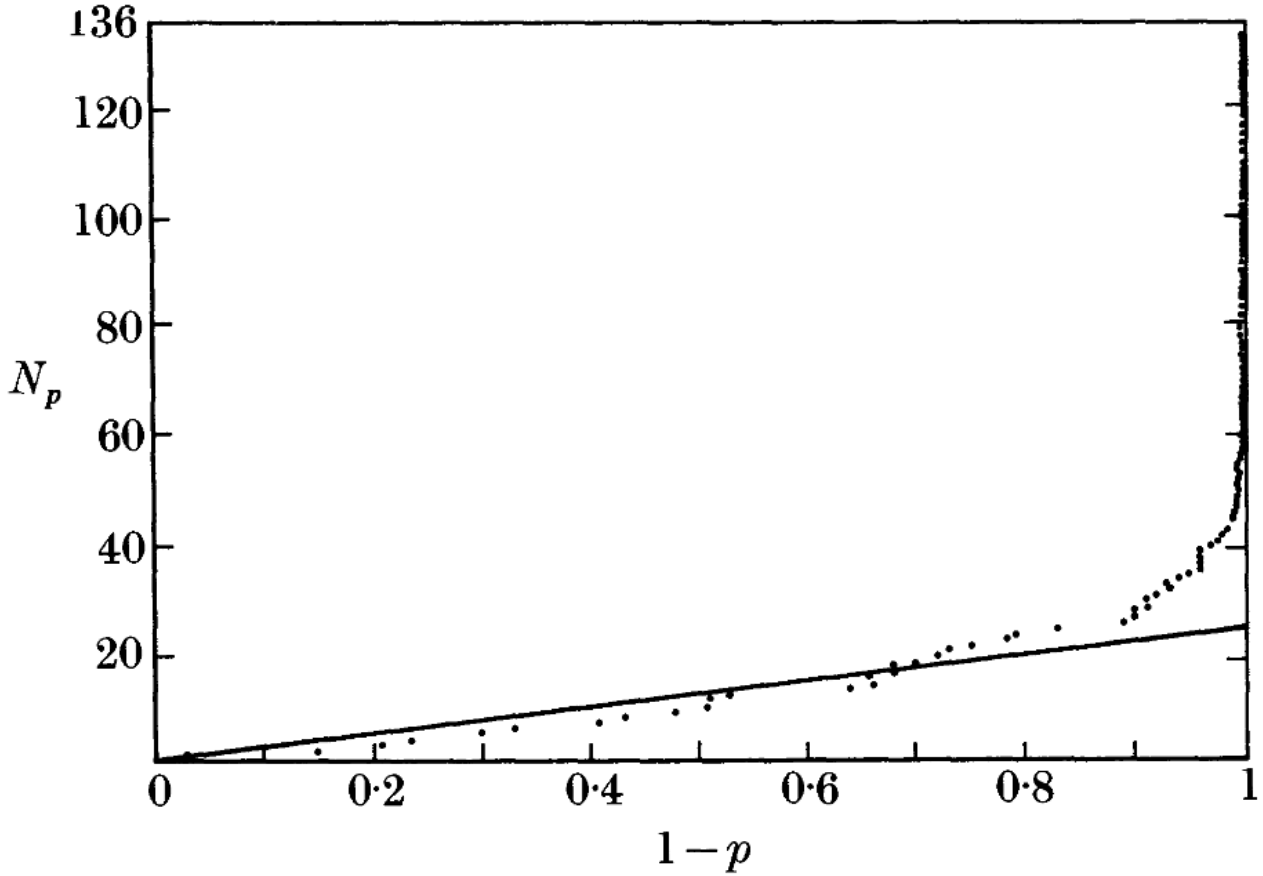
where

- 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

The Null proportion π_0 can be estimated from the data.

- π_0 tells us how much noise there is in our data or equivalently $1 - \pi_0$ tells us about the signal.

Plot of p-values



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

Schweder and Spjotvoll 1982

The concept of false discovery rate (FDR)

- We are not really interested in the Null variables (Bottom of the ranked list).
- Within the discoveries (Top of the list, the features we report or declare 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: 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

Formal Definition: False discovery rate

- a/R is the proportion of false discoveries among all discoveries.
- When working on real data, we do not know what the "true" discoveries are and we need to estimate the false discovery rate (FDR).
- Formally the FDR is defined as the **expected** proportion of false discoveries among all discoveries.

$$FDR = E(a/R)$$

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 $i \leq i_{\text{max}}$ are considered as discoveries and all $i > i_{\text{max}}$ 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

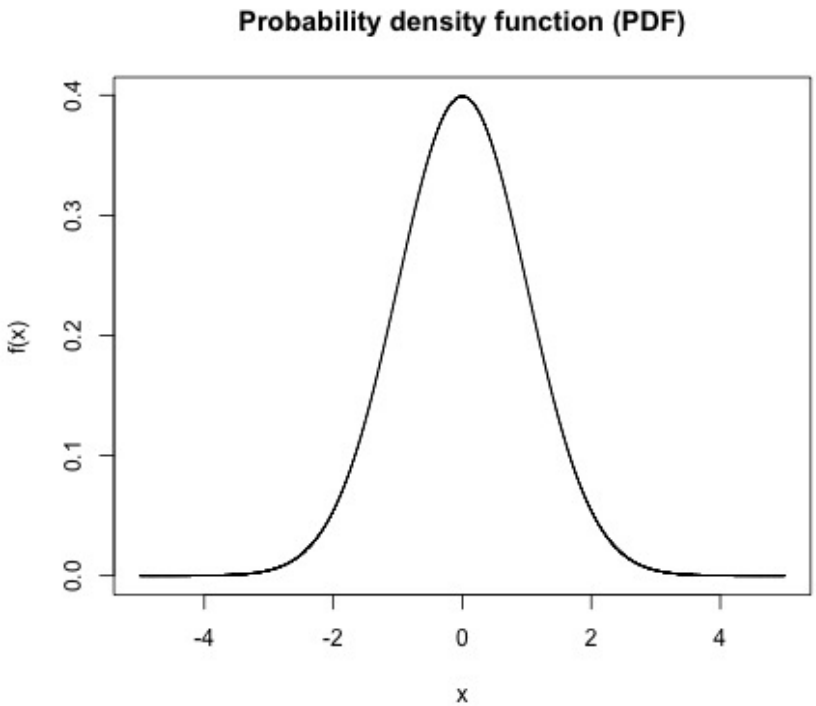
Probability density and cumulative distribution function

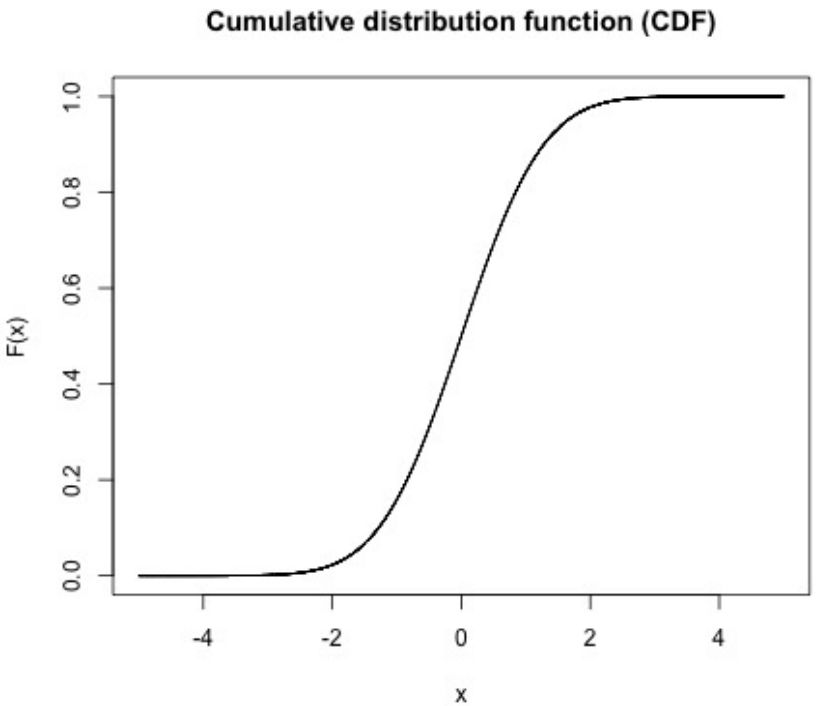
1. Probability density function (PDF) $f_X(x)$
   - Only very specific functions can be pdf's.
2. Cumulative distribution function (CDF)

$$\begin{aligned} F_X(x) &= P(X < x) \\ &= \int_{-\infty}^{x} f_X(t)dt \end{aligned}$$

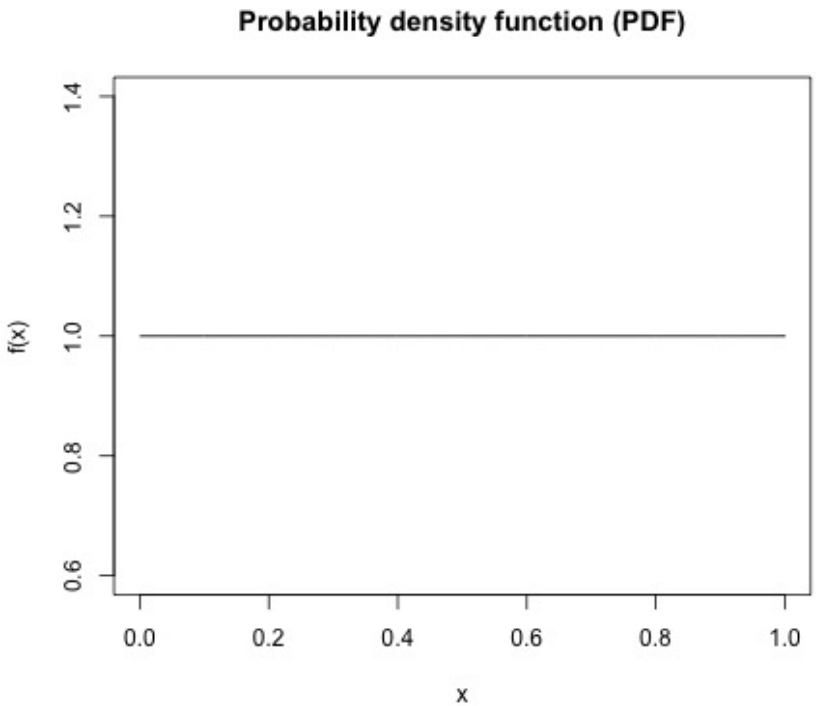
   - Interpretation: What is the probability to observe a value of X equal or smaller than x under the given distribution.

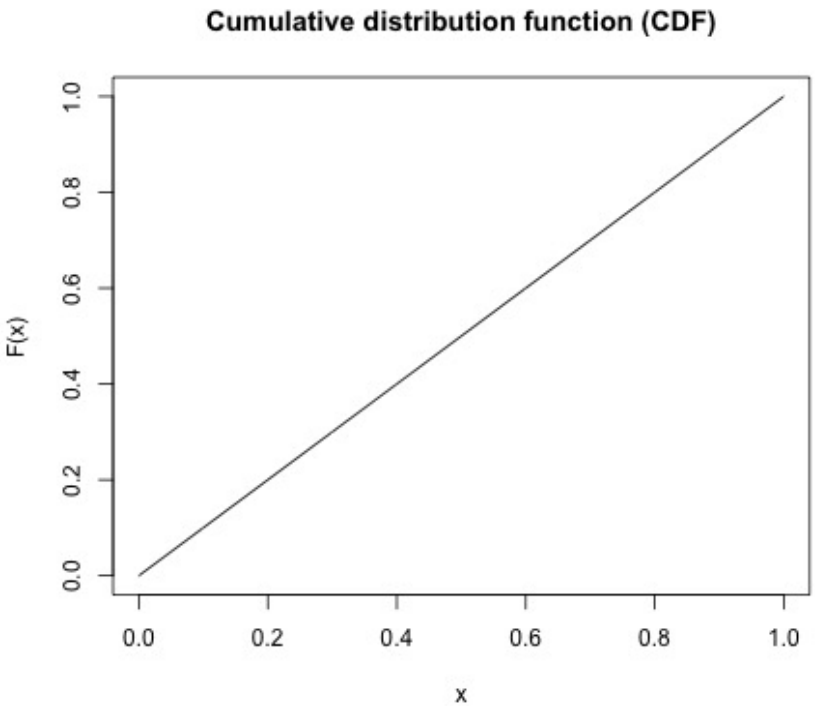
Normal distribution





Uniform distribution





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