Local False Discovery Rates

The Choice of a Null Hypothesis

Overview

- Introduction
- Basics
- Tools
- Local fdr
- Choice of $f_0(z)$
- Example
- Summary

Introduction: Multiple Testing

„Simultaneous Inference“ meant considering up to ten hypothesis tests at the same time.

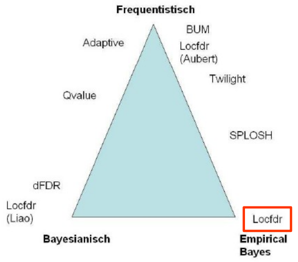
Because of rapid progress in technology, the number of testing problems enlarged up to 5.000 and more.

The more often you test, the more your type I error increases.

Here: Large-scale testing should identify a small percentage of interesting cases that deserve further investigation.

Introduction: local fdr

Different frequentist and Bayesian approaches have been invented. The local fdr is an empirical Bayes method.



Hypothesis and Statistics

- Collection of null hypothesis

  $H_1,\ldots,H_N$ $N > 100$

- Corresponding test statistics

  $Y_1,\ldots,Y_N$ with p values $P_1,\ldots,P_N$

  transformed into z-values

  $z_i = \Phi^{-1}(P_i)$

Theoretical null hypothesis

Because the p values are distributed in the following way:

$P_1,\ldots,P_N \sim$ Unif [0,1] ,

the z-values are distributed

$z_i \mid H_i \sim N(0,1)$ → **theoretical null**

if H_i is exactly true.

Else you have to generate an empirical null hypothesis.

Empirical null hypothesis

The empirical null hypothesis can be generated in different ways.

Just because of large scale testing the number of observations permits an empirical estimation of the null distribution.

→ **empirical null**:

$z_i \mid H_i \sim N(\mu,\sigma^2)$

Probabilities and densities

First we need the probabilities for each class:

p_0 if z_i´s correspond to class "Uninteresting"

$p_1 = 1-p_0$ if to class "Interesting"

and their densities:

$f_0(z)$ density if Uninteresting

$f_1(z)$ density if Interesting

→ **mixture density**

$f(z) = p_0f_0(z)+p_1f_1(z)$

Definition: local fdr

The local false discovery rate is defined as:

$$\boxed{fdr(z) = f_0^+(z) / f(z)}$$

which is the Bayesian a posteriori, using

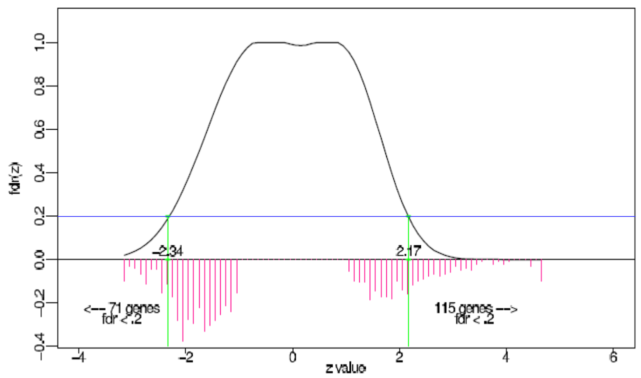
$$f_0^+(z) = p_0 f_0(z).$$

As decision rule you usually choose a threshold

$$fdr(z) < 0.2$$

corresponding to $\alpha \le 0.05$ for univariate cases.

Decision rule



Local FDR

Benjamini and Hochberg developed a different FDR-theory which relies on tail-areas rather than densities.

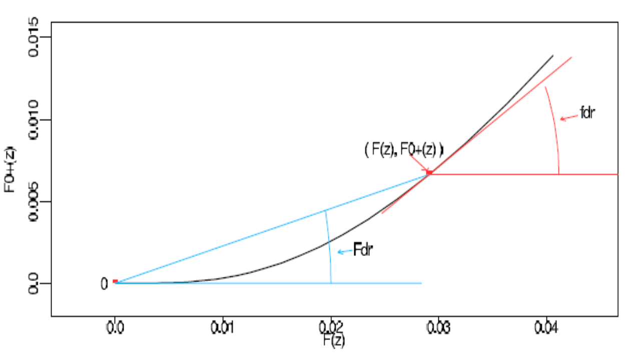
$F_0(z)$,$F_1(z)$, $F(z)$ are the corresponding cdf´s.

→ $FDR(z) = P(\text{ null } | Z \le z) = F_0(z) / F(z) =$

$= E_f(\ fdr(z) \mid Z \le z)$

→ fdr is an advantage in interpreting results for individual cases

Geometrical relationship from FDR to fdr



Estimating the unknown densities

The estimation of f(z) is conducted nonparametric, usually by Poisson-regression.

Assume a parametric null density $f_0(z)$:

Obtain the center and the half-width of the central peak from f(z), defined as

$\delta_0 = \arg\max \{f(z)\}$

$$\sigma_0 = \left[-\frac{d^2}{dz^2} \log f(z) \right]_{\delta_0}^{-\frac{1}{2}}$$

Estimating the probabilities

By definition p_1 is defined as:

$$p_1 = \int_{-\infty}^{\infty} f_1^+(z)\,dz = \int_{-\infty}^{\infty} (1 - fdr(z))\, f(z)\,dz$$

$$\rightarrow \quad p_0 = \left[\int_{-\infty}^{\infty} f_0^+(z)\,dz \right]^{-1}$$

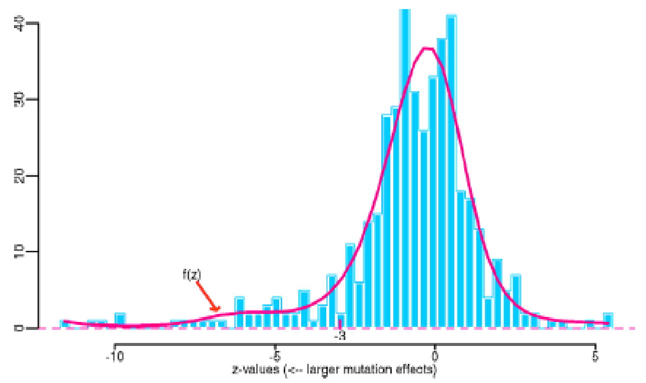
Example: HIV-study

- 1391 patients
- 6 protease inhibitors
- 74 sites on the viral genom
- $\underline{\mathbf{x}} = (x_1,\dots,x_6)$ vector of predictors
- $\underline{\mathbf{v}} = (v_1,\dots,v_{74})$ vector of responses

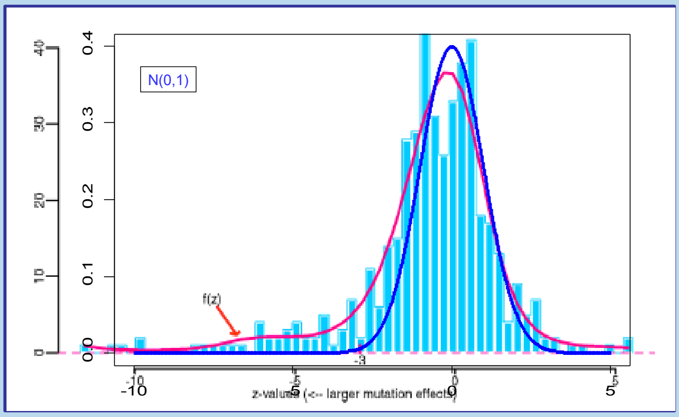
→ 6*74 = 444 z-values

with usual approximation $z_i = y_i/se_i$

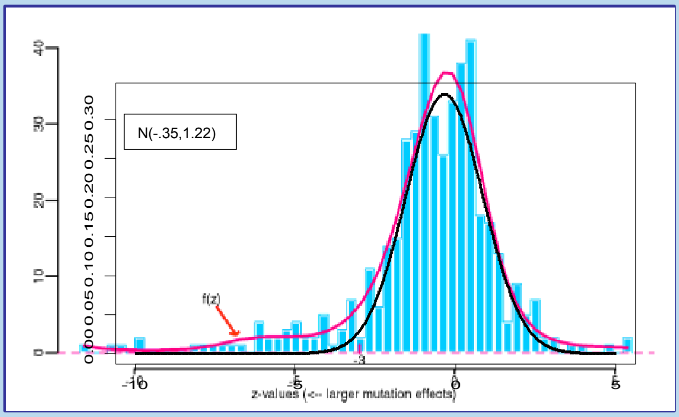
Example: Estimate f(z)



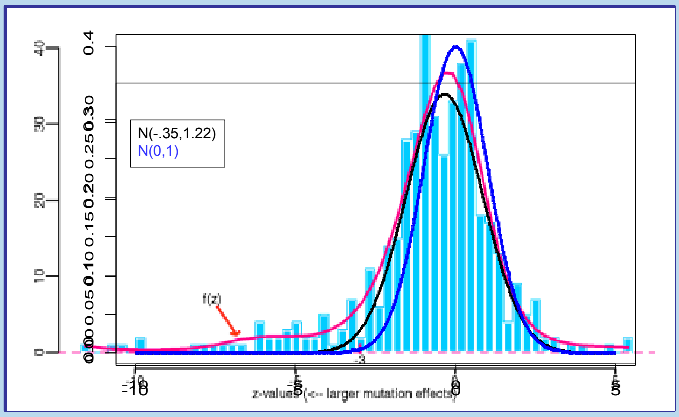
Example: Theoretical null



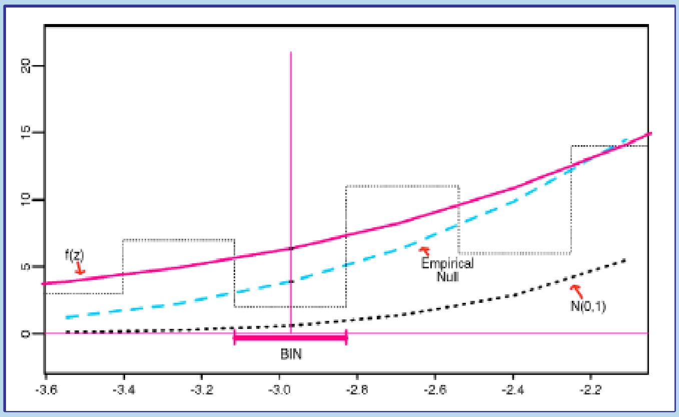
Example: Assign $f_0(z)$ empirically



Example: theoretical vs. empirical



Example: Close-up view for calculation



Summary: Assets

- Large-scale testing intends to identify a small percentage of "Interesting cases"
- Large-scale testing permits the empirical estimation of a null hypothesis
- A minimum of frequentist or Bayesian modeling assumptions are required
- Local fdr calculations provide size and power estimations
- fdr depends only on the marginal distribution of the z-values; independence is not required
- Easy implemetation with familiar software (R)

Summary: Drawbacks

- Microarray observations are usually not independent
- Smoothness of f(z) is an important assumption
- No convention for fdr thresholds yet; increasing it can deliver unacceptably high proportions of false discoveries
- $H_i|z_i$ can differ from $H_i|\underline{\mathbf{z}}$ (only "one at a time" inference)
- Misspecification of the null hypothesis undermines all forms of simoultaneous inference. Using an empirical null avoids this problem but costs estimation efficiency
- Standard deviations for the empirical null are too big for comfort as N exceeds 500