The empirical 0.005 rule

true true

Madrid, Nov 2017

Motivation

- A real statistical practical situation:
  - consider a test for a null hypothesis H_0
  - p-values are the most (**mis**)*used* tools for questioning H_0;
- After ASA Statement on p-values (Wasserstein and Lazar 2016) debate grows;
- (Benjamin 2017) recommend:

1. questioning the null by using a sound statistical reasonament;
2. if p has to be used, a good *statistical practice* would be:

$$\textit{Reject } H_0 \textit{ for } p < 0.005 \Leftarrow 0.005 \textit{ rule}.$$

Rationality FOR $p < 0.005$ rule (SBB Calibration)

- From (Sellke, Bayarri, and Berger 2001, M. J. Bayarri and Berger (2000)) suppose:

$$\begin{cases} H_0: & p \sim f_0 \equiv U(0,1) \\ H_1: & p \sim f_1 \equiv \text{Beta}(\xi, 1) \textit{ for } 0 < \xi \leq 1 \end{cases}$$

- and $B_\pi(p) = \frac{1 \times I_{[0,1](p)}}{\int_0^1 f(p|\xi)\pi(\xi)d\xi}$.
- Then

$$\underline{B}(p) = \inf_{\forall \pi} B_\pi(p) = \frac{1}{\sup_\xi \xi p^{\xi-1}} = \begin{cases} -ep\log p & p < e^{-1} \\ 1 & \textit{otherwise} \end{cases}.$$

- $\underline{\alpha}(p) = 1/(1 + \underline{B}(p))$ is the LB of the frequentist (*conditional*) Type I error

Rationality for $p < 0.005$ rule

p	$\underline{\alpha}(p)$	$1/\underline{B}(p)$
0.050	0.289	2
0.005	0.067	14

The last is the 0.005 rule.

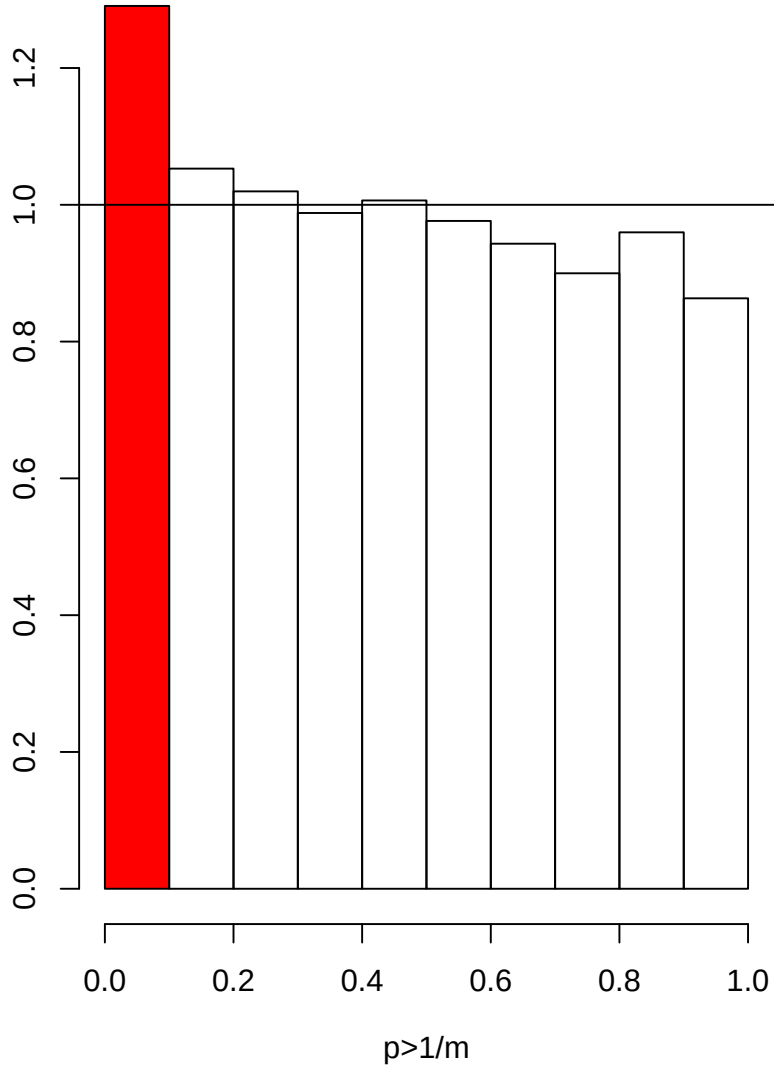
Outline

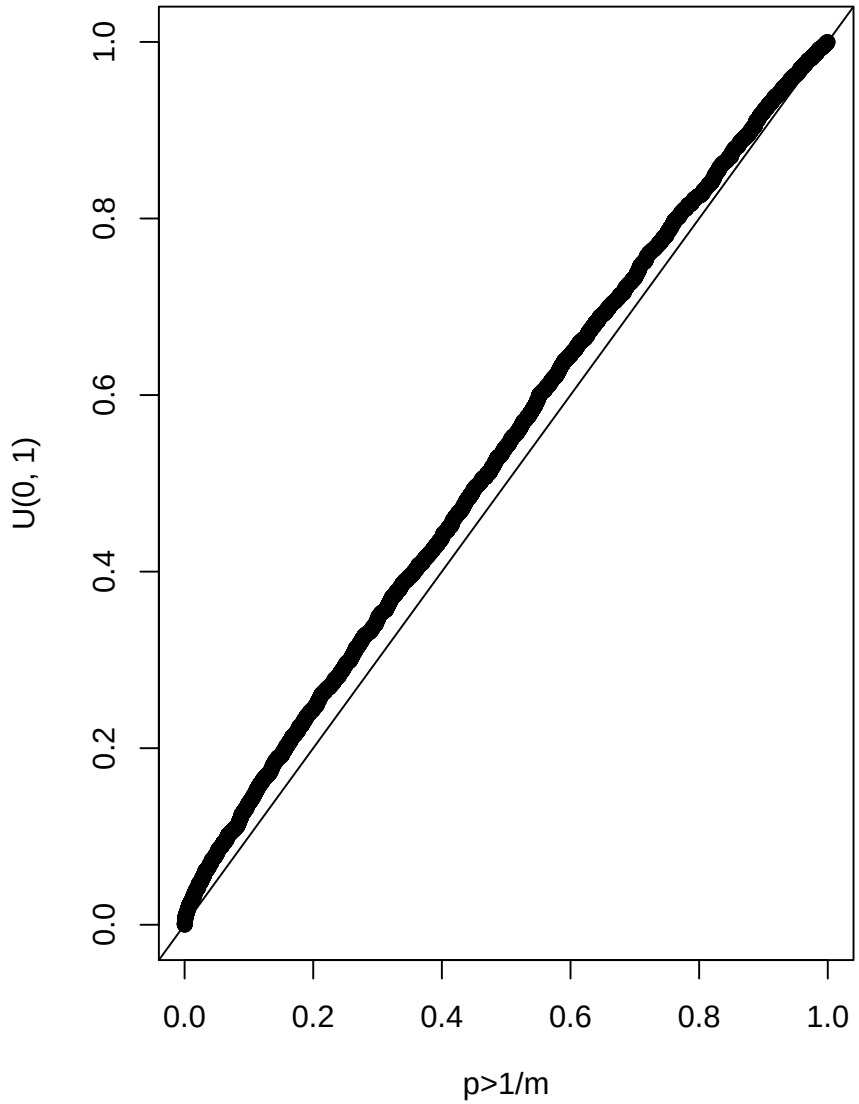
1. Uniformity assumptions behind SBB Calibration
2. The *empirical null* and the *empirical 0.005 rule*
3. Estimation in Multiple Testing
4. Examples in real data
5. Final remarks

Example 1: Prostate cancer data

- A microarray Experiments with:
- $m = 6033$ genes
- 102 patients: $n_x = 50$, $n_y = 52$
- T-Test ($H_0 : \mu_x = \mu_y$) with Welch correction to compare condition X, versus Y

Example 1: Prostate cancer data





These genes are supposed to be not expressed thus under $H_0 : \mu_x = \mu_y$

Example 1: Prostate cancer data

- ... good modelling is a far more important issue than p-value thresholds ...
- We should question: $H_0 : p \sim f_0 \equiv U(0, 1)$ in SBB Calibration ...
- ... and this advocates for an empirical 0.005 rule.

The *empirical null* and the *empirical 0.005 rule*

- We assume

$$\begin{cases} H_0: & p \sim \widehat{f}_0 \equiv \text{Beta}(\widehat{\xi}_0, 1) = \widehat{\xi}_0 p^{\widehat{\xi}_0 - 1},\ \widehat{\xi}_0 = -m_0 / \sum_{i=1}^{m_0} \log p_i \\ H_1: & p \sim f_1 \equiv \text{Beta}(\xi, 1)\ \textit{for}\ 0 < \xi \leq 1 \end{cases}$$

- then

$$\underline{B}^*(p) = \begin{cases} -\widehat{\xi}_0 p^{\widehat{\xi}_0} e \log p & p < e^{-1/\widehat{\xi}_0} \\ 1 & \textit{otherwise} \end{cases}.$$

- and $\underline{\alpha}^*(p) = 1/(1 + \underline{B}^*(p))$ is the *empirical* LB of the frequentist (*conditional*) Type I error
- $m_0 = \#p_i > \widetilde{p}$, i.e. $\widetilde{p} = 1/m$ ... $\#$ of tests under H_0.

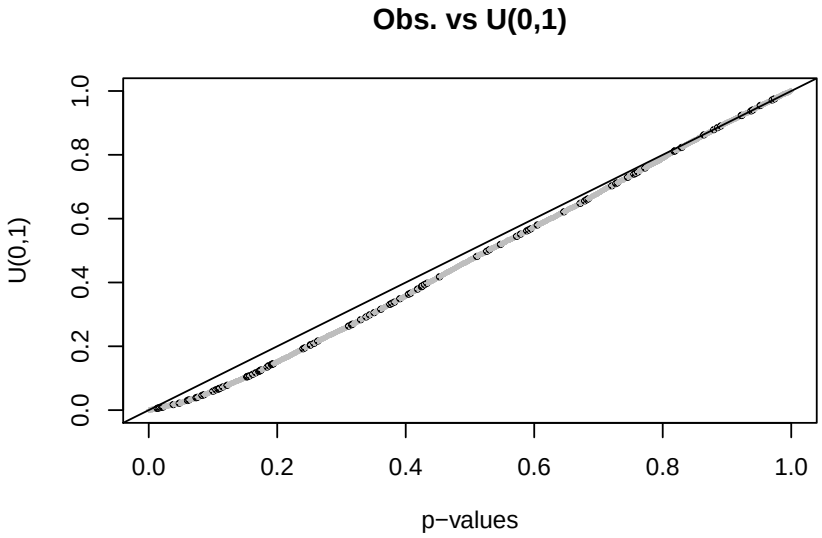
The *empirical null* and the *empirical 0.005 rule*

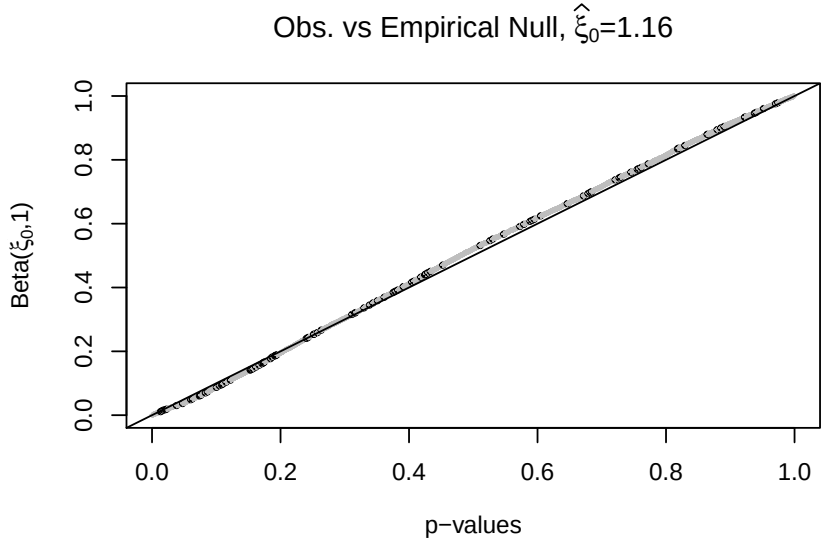
- For $D = \widehat{\xi}_0 - 1$ the following illustrates the empirical calibration:
  Shiny applications not supported in static R Markdown documents

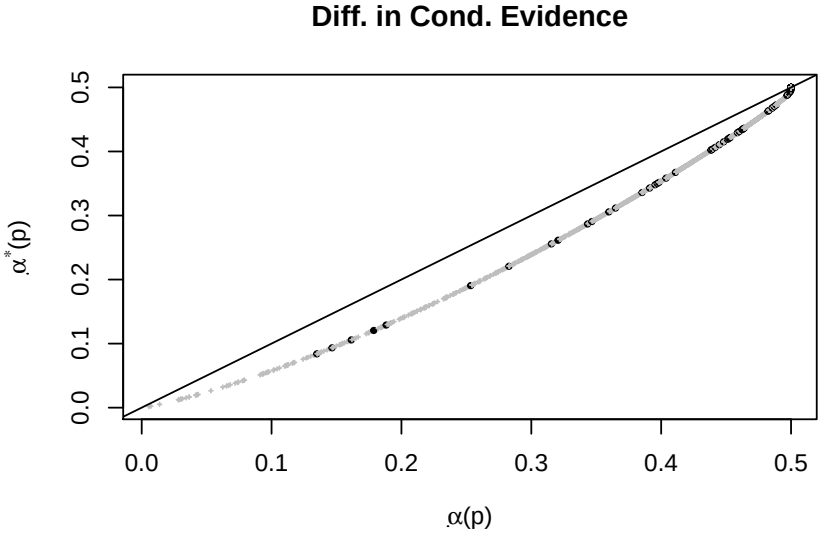
Behrens-Fisher problem in small samples

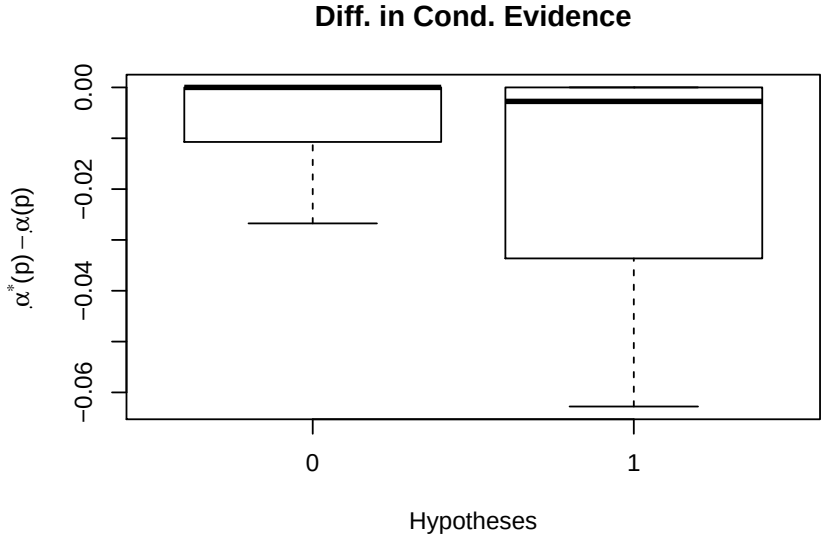
- For:
  - $X_i \sim N(\mu_{X_i}, \sigma^2_{X_i})$ and $Y_i \sim N(\mu_{Y_i}, \sigma^2_{Y_i})$ for $i = 1, 2, \ldots, m$.
- Suppose to test: $\{H_{0i} : \mu_{X_i} = \mu_{Y_i} = \mu_i$ *versus* $H_{1i} : \mu_{X_i} \neq \mu_{Y_i}, \forall\, \sigma^2_{X_i} > 0, \forall\, \sigma^2_{Y_i} > 0\}$.
- p-values from Student t-test with the Welch correction for $\sigma^2_{X_i} \neq \sigma^2_{Y_i}$.
- Simulations:
  - $m = 10000$ tests and sample sizes $n_x = n_y = 5$
  - Under H_{0i} we have $\mu_{X_i} = \mu_{Y_i} = 0$ (99% of tests)
  - Under H_{1i} we have $\mu_{X_i} = 2$ and $\mu_{Y_i} = 0$ (1% of the tests)
  - Homoschedasticity: $\sigma^2_{X_i} = \sigma^2_{Y_i} = 1$

. . . a simulated dataset









- $\underline{\alpha}^*(p)$ increases the chance to separate H_0 from H_1 more than $\underline{\alpha}(p)$;

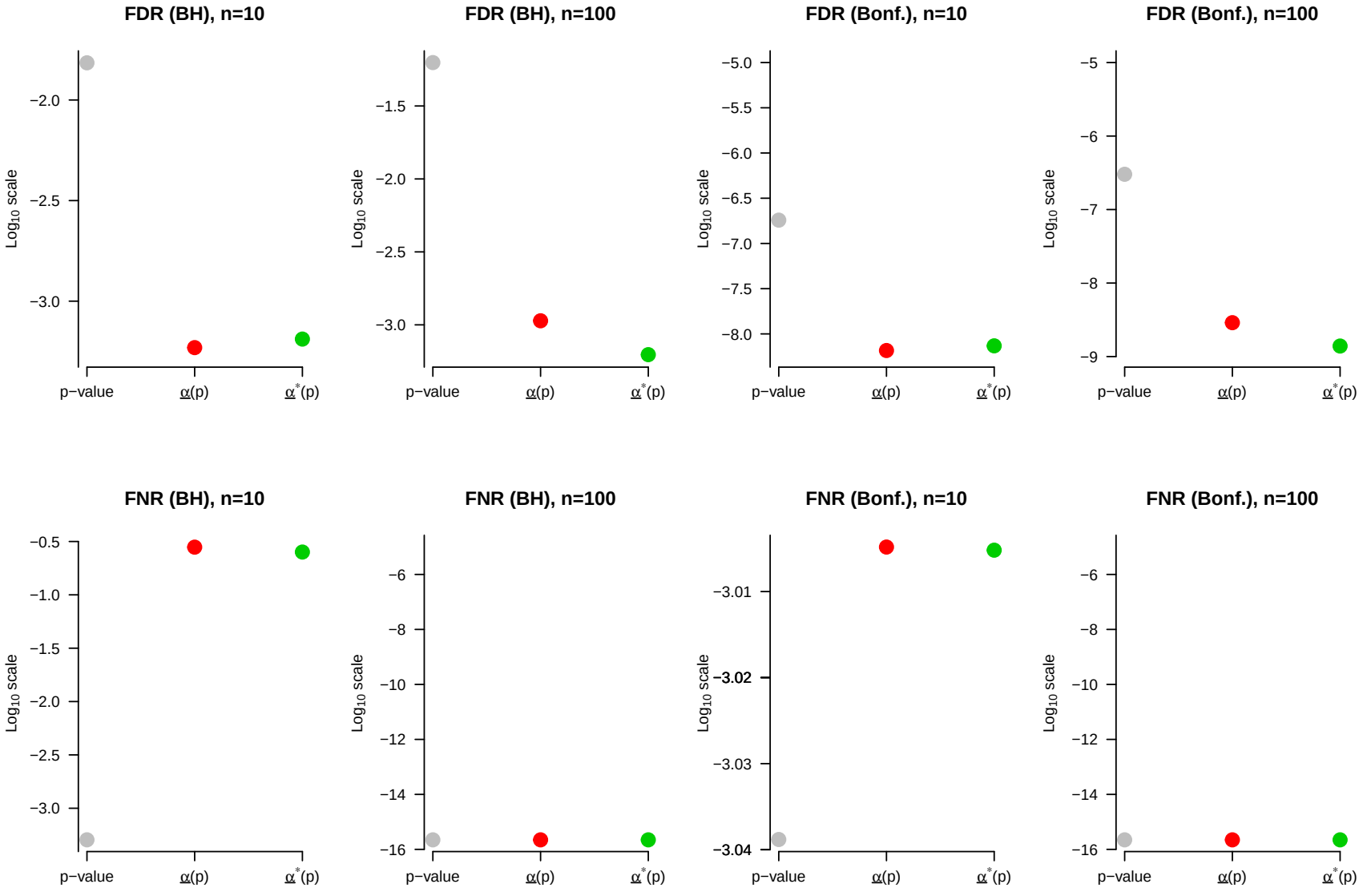
Multiple Hypotheses Testing

- Estimating $\widehat{\xi_0}$ is possible is in Multiple Testing;
- in what follows we consider Multiple Testing procedures for $p_{(1)} < \ldots < p_{(i)} < \ldots < p_{(m)}$:
  1. Benjamini-Hochberg (BH): Reject all $H_i 0$ such that $p_{(i)} < q(i/m)$
  2. Bonferroni: Reject all $H_i 0$ such that $p_{(i)} < q(1/m)$
- based on:
  - $\underline{\alpha}^*(p)$ (our recommendation)
  - $\underline{\alpha}(p)$
  - the observed p-value $p_1, \ldots, p_m$

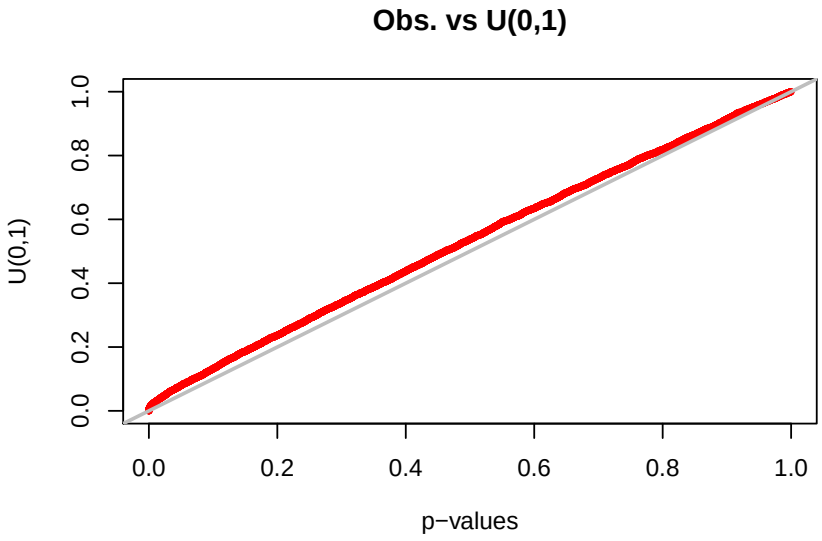
Multiple Hypotheses Testing (some simulations)

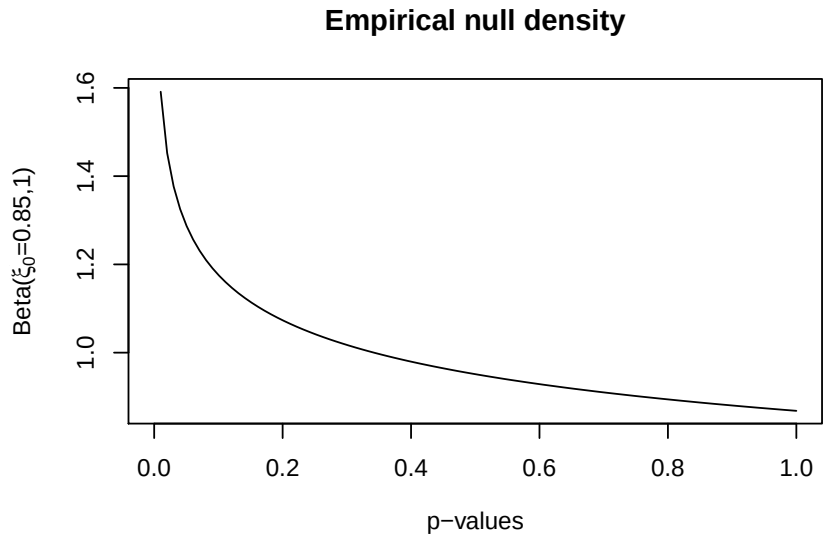
- The following has been simulated 1 million times:
  - $m = 100000$
  - $m_1 = 100$ tests under $H_0 : Y_i, X_i \sim Gamma(1, 1)$ and $H_1 : Y_i \sim Gamma(5, 1)$
  - $n_X = n_Y = 10,100$
  - p-value from T-Student with Welch correction
- This mimics *model mispecification*
- For each simulated data set we apply:
  1. BH procedure which controls the False Discovery Rate (FDR);
  2. Bonferroni which controls the Family Wise Error Rate (FWER $<$ FDR);
- We measure the actual FDR and False Non Rejection Rate (FNR)

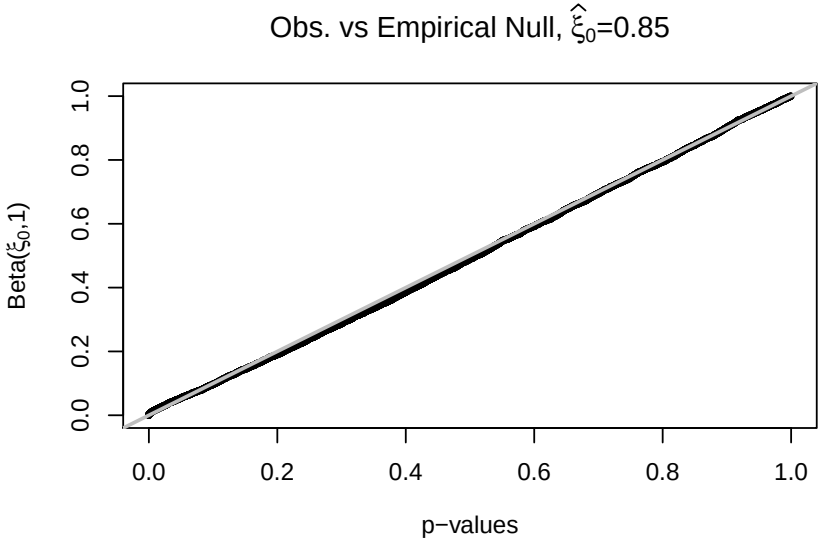
Multiple Hypotheses Testing (some simulations)

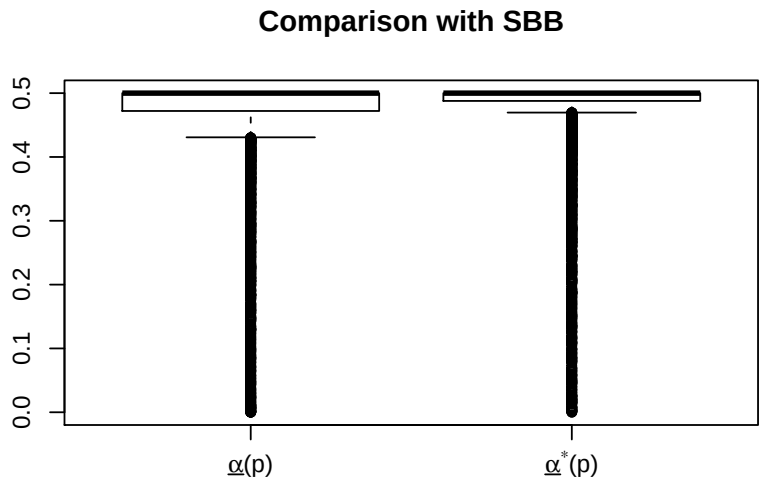


Example 1: Prostate cancer data









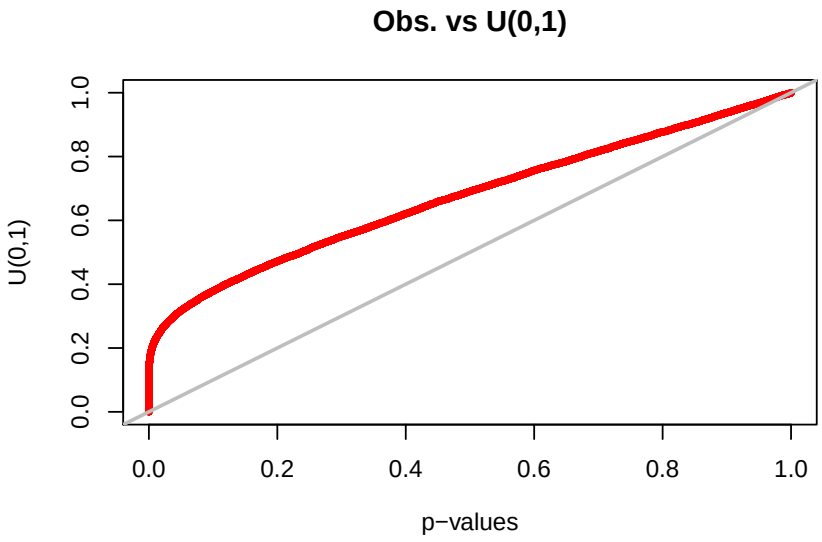
Example 1: Prostate cancer data

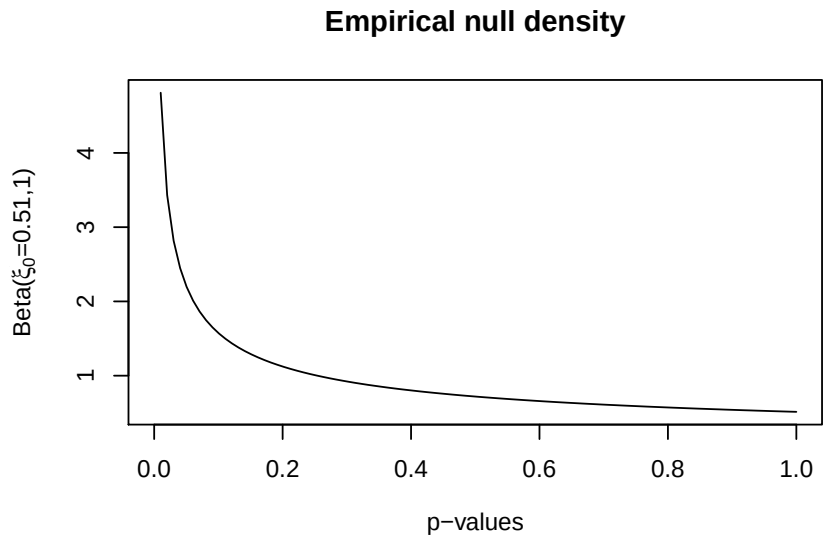
- We apply BH and Bonferroni with $q = 0.05$;
- How many genes should be claimed as differentially expressed ?
  - 1 according $\underline{\alpha}(p)$
  - 0 according $\underline{\alpha}^*(p)$

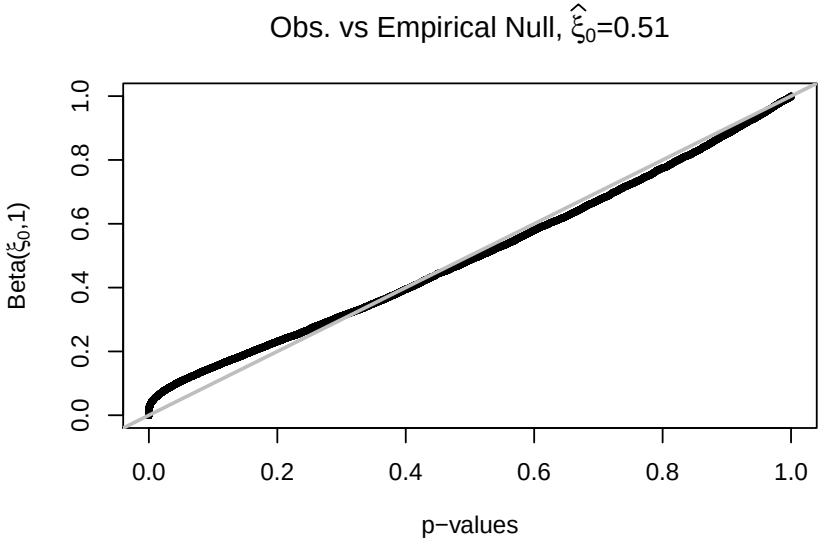
Example 2: *Mycobacterium bovis* infection

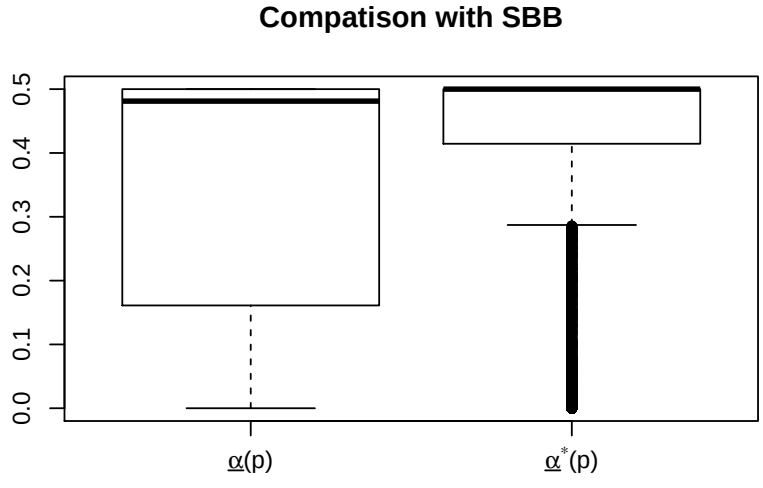
- Raw data consists of 3.6 trillion reads of RNA sequences for comparing bovines infected and non-infected by *Mycobacterium bovis* ;
- After data normalization, there are $m = 11131$ genes with corresponding p-values (Nalpas et al. 2013);

Example 2: *Mycobacterium bovis* infection









Example 2: *Mycobacterium bovis* infection

- 2584 genes have been declared as differentially expressed (Nalpas et al. 2013) with the BH procedure $q = 0.05$.

	Bonferroni (FWER)	BH (FDR)
$\underline{\alpha}(p)$	728	1490
$\underline{\alpha}^*(p)$, $\widetilde{p} = 0$	72	154
$\underline{\alpha}^*(p)$, $\widetilde{p} = 1/m$	339	675
$\underline{\alpha}^*(p)$, $\widetilde{p} = 2/m$	355	703

- are the differences w.r.t. (Nalpas et al. 2013) due to a statistical practice ?

Shiny App for your own analysis

- A very friendly App is available at:
  - `https://stefano-cabras.shinyapps.io/p-value_calibration/`
  - you can upload your p-values and analyse them.
- more details can be found in (Cabras and Castellanos 2017).

Remarks

1. Is this the unique variant of the 0.005 rule ?
   - no: we could postulate many empirical null models, $\widehat{f}_0$;
   - the proposed approach tries to keep formulation as simple as possible:

     $$-ep\log p$$

     versus

     $$-e\widehat{\xi}_0 p^{\widehat{\xi}_0}\log p$$

2. The Bayesian way of thinking may mitigate a bad statistical practice: the use of p-value ($<$**0.005 ?**)

References

Bayarri, M. J., and J. O. Berger. 2000. "P Values for Composite Null Models." *J. Am. Stat. Assoc.* 95 (452): 1127–42.

Benjamin, et al., Berger. 2017. "Redefine Statistical Significance." *PsyArXiv* https://osf.io/preprints/psyarxiv/mky9j/: 1–18. doi:10.17605/OSF.IO/MKY9J.

Cabras, Stefano, and Maria Eugenia Castellanos. 2017. "P-Value Calibration in Multiple Hypotheses Testing." *Statistics in Medicine* 36 (18): 2875–86.

Nalpas, Nicolas C., Stephen D E. Park, David A. Magee, Maria Taraktsoglou, John A. Browne, Kevin M. Conlon, Kévin Rue-Albrecht, et al. 2013. "Whole-Transcriptome, High-Throughput Rna Sequence Analysis of the Bovine Macrophage Response to Mycobacterium Bovis Infection in Vitro." *BMC Genomics* 14: 230. doi:10.1186/1471-2164-14-230.

Sellke, T., M.J. Bayarri, and J. O. Berger. 2001. "Calibration of P-Values for Testing Precise Null Hypotheses." *Am. Stat.* 55 (1): 62–71.