ON ESTIMATING THE PROPORTION OF FALSE HYPOTHESES IN MULTIPLE TESTING PROCEDURE

Bobosharif Shokirov

1Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics
Charles University in Prague

Robust 2010
31.01 – 5.02.2010, Králíky
Introduction

Motivation

Main Results
Family-wise error rate (FWER) - the probability of committing one or more false rejection (Hochnberg and Tahmane, 1987);
Family-wise error rate (FWER)-the probability of committing one or more false rejection (Hochberg and Tamhane, 1987);
False discovery rate (FDR): expected value of false discovery proportion (FDP), (Benjamini and Hochberg, 1992)
**Introduction**

FWER and its Modifications

- Family-wise error rate (FWER)-the probability of committing one or more false rejection (Hochnberg and Tahmane, 1987);

- False discovery rate (FDR): expected value of false discovery proportion (FDP), (Bejamini and Hochnberg, 1992)
  - FDP is the number of false rejections, divided by the total number of rejections;
Family-wise error rate (FWER)-the probability of committing one or more false rejection (Hochnberg and Tahmane, 1987);

False discovery rate (FDR): expected value of false discovery proportion (FDP), (Bejamini and Hochnberg, 1992)
  - FDP is the number of false rejections, divided by the total number of rejections;

Positive false discovery rate (pFDR): conditional expected value of FDP on the event that positive findings have occurred (Storey, 2002).
Introduction
FWER and its Modifications

- Family-wise error rate (FWER)-the probability of committing one or more false rejections (Hochnberg and Tamhane, 1987);
- False discovery rate (FDR): expected value of false discovery proportion (FDP), (Bejamini and Hochnberg, 1992)
  - FDP is the number of false rejections, divided by the total number of rejections;
- Positive false discovery rate (pFDR): conditional expected value of FDP on the event that positive findings have occurred (Storey, 2002).
Family-wise error rate (FWER)-the probability of committing one or more false rejection (Hochnberg and Tahmane, 1987);

False discovery rate (FDR): expected value of false discovery proportion (FDP), (Bejamini and Hochnberg, 1992)
- FDP is the number of false rejections, divided by the total number of rejections;

Positive false discovery rate (pFDR): conditional expected value of FDP on the event that positive findings have occurred (Storey, 2002).
Apart from these concepts for a large number of independently tested hypotheses, based on the empirical distribution function of the \( p \)-values of the tests, Meinhausen (2005) constructed the lower bound \( \lambda \) for the estimate of the proportion of false hypotheses, with the property

\[
\mathbb{P}(\hat{\lambda} \leq \lambda) \geq 1 - \alpha,
\]

where \( 1 - \alpha \) is a given confidence level.
Apart from these concepts for a large number of independently tested hypotheses, based on the empirical distribution function of the $p$-values of the tests, Meinhausen (2005) constructed the lower bound $\lambda$ for the estimate of the proportion of false hypotheses, with the property

$$\mathbb{P}(\hat{\lambda} \leq \lambda) \geq 1 - \alpha,$$

(1)

where $1 - \alpha$ is a given confidence level.

The message: proportion of false (null) hypotheses is at least $\hat{\lambda}$. 
Let us have \( n \) points (rv’s) from interval \([0, 1]\). Take a random variable \( x \in [0, 1] \) and test the hypothesis:

\[
H_0 : x \sim \mathcal{U}[0, 1] \quad \text{against} \quad H_A : x \sim \mathcal{U}[0, 1 - \delta].
\]

Then the share of points from the null hypothesis greater than \( x \) would approximately be equal to \( 1 - x \):

\[
\frac{\text{number of points } > x}{n - k} \approx 1 - x
\]

and the share of points from the null hypothesis which are less than \( x \) would approximately be equal to \( x \):

\[
\frac{\text{number of points } < x}{n - k} \approx x
\]

Then the total number of points which are less than \( x \) approximately equals to \( x(n - k) + k \) and the total number of points which are greater than \( x \) is approximately equal to \((1 - x)(n - k)\). Thus, we have the distribution of the random variable \( x \) on the whole interval \([0, 1]\), under both the null and the alternative hypotheses.
Main Results

Having replaced $[0, 1]$ by df $F(x)$ and $G(x)$ such that

(A1) $G(x) > F(x), \forall x \in [0, 1],$

(A2) $\text{supp } G(x) \subset [0, 1 - \delta], \text{ for some } \delta > 0,$

we obtain the following estimator of the ratio $k/n$ in testing $n$ hypotheses:

$$p^*(x) = 1 - \frac{T_Z(x)}{n(1 - F(x))},$$

(2)

where $T_Z(x) = \mathbb{E}[T_{nZ}(x)]$ and $T_{nZ}(x) = \sum_{j=1}^{n} I\{Z_j > x\};$

$I\{Z_j > x\}$ indicator of the event $\{Z_j > x\}.$
Lemma

If condition (A1) holds, then the expected value of $p^*(x)$ is defined as following:

$$\mathbb{E}[p^*(x)] = p \left[ 1 - \frac{1 - G(x)}{1 - F(x)} \right],$$

where $p = \frac{k}{n}$. 

(3)
Corollary

For \( x \in (1 - \delta, 1] \) \( p^*(x) \) is an unbiased estimator of \( p \).
Main Results
Properties of the Estimator $p^*(x)$

Corollary

For $x \in (1 - \delta, 1]$ $p^*(x)$ is an unbiased estimator of $p$.

Corollary

If in addition to condition (A1) the following condition holds

$$\frac{F'(x)}{1 - F(x)} \leq \frac{G'(x)}{1 - G(x)},$$

(4)

then the expected value of $p^*(x)$ is a monotonic nondecreasing on the interval $[0, 1]$ function.
Corollary

Moreover, since

$$0 \leq 1 - \frac{1 - G(x)}{1 - F(x)} \leq 1,$$

then

$$0 \leq \mathbb{E}[p^*(x)] \leq p \ \forall x \in [0, 1].$$
Main Results

Standard Deviation of $p^*(x)$

$$\sigma_{p^*(x)}^2 = \mathbb{E}[p^*(x)]^2 - [\mathbb{E} p^*(x)]^2.$$ (5)
Main Results

Standard Deviation of $p^*(x)$

$$\sigma^2_{p^*(x)} = \mathbb{E}[p^*(x)]^2 - \mathbb{E}[p^*(x)]^2.$$  \hfill (5)

**Theorem**

*If the random vectors $\mathbb{X}$ and $\mathbb{Y}$ are independent, then standard deviation of the estimator $p^*(x)$ has the form*

$$\sigma^2_{p^*(x)} = \frac{(1 - p)F(x)}{n(1 - F(x))} + \frac{pG(x)(1 - G(x))}{n(1 - F(x))^2}.$$  \hfill (6)
Theorem

Let conditions (A1) and (A2) satisfied. Then the standard deviation \( \sigma^2_{p^*(x)} \), defined in Theorem 1 is a monotonic nondecreasing function of \( x \) for all \( x \in [0, 1] \).
Dekuju za Pozornost
References

Carvajal-Rodriguez, A., Una-Alvarez, J., Rolan-Alvarez, E., A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests, *BMC Bioinformatics*. Available from http://www.biomedcentral.com/content/pdf/1471


References


