ON ESTIMATING THE PROPORTION OF FALSE HYPOTHESES IN MULTIPLE TESTING PROCEDURE

Bobosharif Shokirov[1]

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

Robust 2010
31.01 – 5.02.2010, Králíky

Content

Introduction
FWER and it Modifications

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

Introduction

FWER and it 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)

Introduction

FWER and it 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;

Introduction

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

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

Introduction

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

Introduction

- 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 λ for the estimate of the proportion of false hypotheses, with the property

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

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

Introduction

- 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 λ for the estimate of the proportion of false hypotheses, with the property

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

  where $1 - \alpha$ is a given confidence level.
- The message: proportion of false (null) hypotheses is at least $\hat{\lambda}$.

Motivation

Motivation Example

Let us have n points (rv's) from interval $[0, 1]$. Take a random variable $x \in [0, 1]$ and test the hypothesis:

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

Then the share of points from the null hypothesis greater than x would approximately be equal to $1 - x$:
(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: (number of points $< x)/(n - k) \approx x$. Then the total number of points which are less then x approximately equals to $x(n - k) + k$ and the total number of points which are greater then 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

Replace Uniform by df $F(x)$ and $G(x)$

Having replaced $\mathbb{U}[0,1]$ by df $F(x)$ and $G(x)$ such that

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

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

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

$$p^*(x) = 1 - \frac{T_Z(x)}{n(1-F(x))}, \tag{2}$$

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 $\{Z_j > x\}$.

Main Results

Estimator the ratio k/n

Lemma

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

$$\mathbb{E}[p^*(x)] = p\left[1 - \frac{1-G(x)}{1-F(x)}\right], \tag{3}$$

where $p = k/n$.

Main Results
Properties of the Estimator $p^*(x)$

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

$$\frac{F'(x)}{1-F(x)} \leq \frac{G'(x)}{1-G(x)}, \tag{4}$$

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

Main Results

Properties of the Estimator $p^*(x)$

Corollary

Moreover, since

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

then $0 \le \mathbb{E}[p^*(x)] \le p \; \forall x \in [0, 1]$.

Main Results

Standard Deviation of $p^*(x)$

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

Main Results

Standard Deviation of $p^*(x)$

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

Theorem

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

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

Main Results

Standard Deviation $p^*(x)$

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

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
- Benjamini, Y., Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing, *J. R. Statist. Soc.*, **Vol. 57**, 1995.
- Benjamini, Y., Hochberg, Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics, *Journal of Educational and Behavioral Statistics*, **Vol. 25**,2000.

References

more references

- Farcomeni, A. Multiple testing procedures under dependence, with applications. Ph.D. thesis, Univ Roma “La Sapienza”, 2004.
- Hochberg, Y. and Tamhane, A. Multiple Comparison Procedures. New York, Wiley, 1987.
- Holm, S. A simple sequentially rejective multiple procedure, *Scand. J. Statist.*, **Vol, 6**, 1979.
- Klebanov, L. B. Yakovlev, A. Diverse correlation structures in gene expression data and their utility in improving statistical inference, *Statistics and Probability Letters*, **Vol. 31**, 2000.
- Klebanov, L. B. Yakovlev, A. A nitty-gritty Aspects of correlation and network inference from gene expression data, *Biology Direct*, available at: http://www.biology-direct.com/content/3/1/35.

References

- Lehmann, E. L., Romano, J. P. Generalization of he Familywise Error Rate, *Annals of Statistics*, **Vol. 34**, 2006.
- Meinhausen, N., Rice, J. P. Estimating the proportion of false null hypotheses among alarge number of independently tested hypotheses, *Annals of Statistics*, **Vol. 33**, 2006.
- Meinhausen, N., Bu¨hlmann, P. Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures, *Biometrika*, **Vol. 92**, 2005.
- Qiu, X., Brooks, A. I., Klebanov, L. B. and Yakovlev, A., The effect of normalization on the correlation structure of microarray data *BMC Bioinformatics*, **Vol. 6**, 2005.

References

- Storey, J. D., A direct approach to false discovery Rate, *Journal of Royal Statistical Society,*, **Vol. 64**, 2002.
- Wu, W. B., Nonlinear system theory: Another look at dependence, *Proc. Natl. Acad. Sci. USA*, **Vol. 102**, 2005.
- Wu, W. B., On false discovery control under dependence, *The Annals of Statistics*, **Vol. 36**, 2008.
- Westfall, P. H., Young, S. S. Resampling-based multiple testing: Examples and Methods for p-value Adjustment, Wiley, New York, 1993.