M A T H E M A T I C A L E C O N O M I C S

No. 10(17) 2014

CONTROLLING THE EFFECT OF MULTIPLE TESTING IN BIG DATA

Sabina Denkowska

Abstract. Big Data poses a new challenge to statistical data analysis. An enormous growth of available data and their multidimensionality challenge the usefulness of classical methods of analysis. One of the most important stages in Big Data analysis is the verification of hypotheses and conclusions. With the growth of the number of hypotheses, each of which is tested at α significance level, the risk of erroneous rejections of true null hypotheses increases. Big Data analysts often deal with sets consisting of thousands, or even hundreds of thousands of inferences. FWER-controlling procedures recommended by Tukey [1953], are effective only for small families of inferences. In cases of numerous families of inferences in Big Data analyses it is better to control FDR, that is the expected value of the fraction of erroneous rejections out of all rejections. The paper presents marginal procedures of multiple testing which allow for controlling FDR as well as their interesting alternative, that is the joint procedure of multiple testing MTP based on resampling [Dudoit, van der Laan 2008]. A wide range of applications, the possibility of choosing the Type I error rate and easily accessible software (MTP procedure is implemented in R *multtest* package) are their obvious advantages. Unfortunately, the results of the analysis of the MTP procedure obtained by Werft and Benner [2009] revealed problems with controlling FDR in the case of numerous sets of hypotheses and small samples. The paper presents a simulation experiment conducted to investigate potential restrictions of MTP procedure in case of large numbers of inferences and large sample sizes, which is typical of Big Data analyses. The experiment revealed that, regardless of the sample size, problems with controlling FDR occur when multiple testing procedures based on minima of unadjusted p-values ($minP$) are applied. Moreover, the experiment indicated the serious instability of the results of the MTP procedure (dependent on the number of bootstrap samplings) if multiple testing procedures based on minima of unadjusted p-values ($minP$) are used. The experiment described in the paper and the results obtained by Werft, Benner [2009] and Denkowska [2013] indicate the need for further research on MTP procedure.

Keywords: multiple testing, FDR, Big Data.

JEL Classification: C12, C14, C15.

DOI: 10.15611/me.2014.10.01.

Sabina Denkowska
Department of Statistics, Cracow University of Economics, Rakowicka Street 27, 31-510 Kraków, Poland.
E-mail: sabina.denkowska@uek.krakow.pl

1. Introduction

Big Data sets are becoming increasingly available as a result of the dynamic development of techniques of automatic collecting and archiving data from industrial systems, telecommunication networks, social networks and – a recent phenomenon – the IoT (Internet of Things). Until recently most data which underwent electronic processing were keyed into computer systems manually by operators. At present, data are generated and aggregated by microchips and software in a more automatic way (e.g. RFID cards, gateways, cameras, sensors, etc.). Moreover, the advancement of IT technologies allows for collecting new categories of data, which several years ago were unavailable for processing or even non-existing (e.g. the number of "likes" on a social network in terms of geographical distribution).

Such enormous incrementing of data calls for investigating information hidden in them, and their scope and multidimensionality require new ways of their processing. The information techniques (databases) and statistical methods used so far should be adapted to the new reality governed by Big Data. Thus, Big Data poses new challenges for statistical data analysis.

The task of Big Data analysts is to discover significant dependencies by skilfully using various analytical methods, drawing on experts' knowledge, and expanding their source data by additional external information. Information obtained from these gigantic datasets increases the chances of taking more effective decisions in many areas of the economy and stimulates the advancement of science.

Big Data analysts often have to deal with sets containing thousands or even hundreds of thousands of inferences. Obviously, the greater the number of hypotheses to be tested, each at the significance level α, the greater the risk of rejecting the true null hypotheses. In case of 14 independent true null hypotheses, each of which is tested at the $\alpha = 0.05$ significance level, it is more likely to make at least one Type I error than to state failure to reject all 14 null hypotheses (which is the correct statement). In case of 100 independent true null hypotheses, the probability of making at least one Type I error equals 0.994! In practice, analysts rarely deal with independent tests, which makes controlling the effect of multiple testing even more challenging.

The most common Type I error rate for the family[1] of inferences which enables to control the effect of multiple testing is FWER (*Family-Wise Error Rate*). It is defined in the following way:

$$\text{FWER} = P(V > 0), \tag{1}$$

where V denotes the number of true null hypotheses rejected while testing m null hypotheses. Controlling FWER refers to the traditional approach to testing statistical hypothesis. The procedures controlling FWER at a given level α ensure fulfilling the condition that the probability of rejecting at least one true null hypothesis will not exceed α. In his monograph *The Problem of Multiple Comparisons* Tukey [1953] compared various Type I error rates for sets of inferences and claimed that "controlling FWER should be a standard"[2] in multiple testing. However, when Tukey recommended controlling FWER the word 'multiple' carried a different meaning for statisticians than it does today. In the past, families of inferences consisted of only several null hypotheses and corresponding alternative hypotheses, while now sets of inferences can contain thousands of inferences. Unfortunately, Tukey's [1953] recommendations have lost their validity for numerous sets of inferences, because FWER-controlling procedures lack power if a great number of inferences is taken into account. In cases of very numerous families of inferences, individual testing is conducted at such low significance levels that in practice many important dependencies may remain undetected.

The FDR (*False Discovery Rate*) proposed in 1995 by Hochberg and Benjamini, offers a completely different approach to controlling Type I errors in multiple testing. When using FDR, an analyst allows for a certain number of erroneous rejections among all the rejections, but gains an improvement of power, which seems the golden mean between the lack of control of the effect of multiple testing and the conservative nature of FWER in analysing very numerous families of inferences.

The paper presents the marginal procedures of multiple testing which allow for controlling FDR as well as their interesting alternative, that is the joint procedure of multiple testing MTP based on resampling [Dudoit, van der Laan 2008]. A wide range of applications, the possibility of choosing

[1] This term was introduced by Hochberg and Tamhane [1987, p. 5], who proposed treating "any collection of inferences for which it is meaningful to take into account some combined measure of errors" as a family.

[2] See: [Hochberg, Tamhane 1987].

the Type I error rate and easily accessible software (the MTP procedure is implemented in R *multtest* package) are their obvious advantages. Unfortunately, the results of the analysis of MTP procedure obtained by Werft and Benner [2009], revealed problems with controlling FDR in cases of numerous sets of hypotheses and small samples. The paper presents a simulation experiment conducted to investigate the potential restrictions of the MTP procedure in cases of large numbers of inferences and large sample size, which is typical of Big Data analyses. The experiment revealed that, regardless of the sample size, problems with controlling FDR occur when multiple testing procedures based on minima of unadjusted *p*-values ($minP$) are applied. Moreover, the experiment indicated the serious instability of the results of the MTP procedure (dependent on the number of bootstrap samplings) if multiple testing procedures based on minima of unadjusted *p*-values ($minP$) are used.

2. FDR (False Discovery Rate)

Benjamini and Hochberg [1995], suggested controlling not the number of erroneous rejections, but the expected value of the proportion of Type I errors among the rejected hypotheses. Their FDR (*False Discovery Rate*) is defined as follows:

$$\mathrm{FDR} = \mathrm{E}\left(\frac{V}{max\{R,1\}} \right), \tag{2}$$

where V denotes the number of Type I errors and R – the number of rejected null hypotheses.

Thus, FDR procedures have much greater power than FWER-controlling procedures. The difference between FDR and FWER is illustrated by the following example. Let us consider a family consisting of 1000 inferences and compare the following situations:

I. rejecting 2 hypotheses one of which is true,

II. rejecting 100 null hypotheses one of which is true,

III. rejecting 500 null hypotheses five of which are true.

From the perspective of FWER, all three situations are equally disadvantageous, because at least one true null hypothesis is rejected, but when FDR is considered, only situation I is unwelcome because it results in 50% of erroneous rejections, in situations II and III it is only 1%.

When an analyst chooses controlling FDR, he/she accepts a tiny fraction of erroneous rejections out of all the rejections, but in return obtains considerable improvement of power in comparison to FWER-controlling procedures.

In order to present marginal FDR-controlling procedures, let us adopt the following assumptions and symbols. We will consider a family m of minimal null hypotheses $H_{0,1}, H_{0,2}, \ldots, H_{0,m}$ with corresponding raw *p*-values $p_1, p_2, \ldots, p_m$. Let us order *p*-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ and let $H_{(0,1)}, H_{(0,2)}, \ldots, H_{(0,m)}$ denote corresponding null hypotheses.

2.1. FDR-controlling marginal procedures

Together with FDR, Benjamini and Hochberg [1995], proposed a procedure which enables to control FDR at an *a priori* chosen level q ($q = \alpha$). This means that when we use this procedure, we allow for $q100\%$ erroneous rejections of null hypotheses out of all rejections.

The algorithm of the Benjamini-Hochberg procedure (BH) takes the following form:

Stage 1. We appoint $k = max\left\{i : p_{(i)} \leq iq / m\right\}$.

Stage 2. If such k exists, we reject k hypotheses $H_{(0,1)}, H_{(0,2)}, \ldots, H_{(0,k)}$, otherwise, we do not reject any hypotheses.

The testing process can be considerably simplified by direct comparison of the assumed q with adjusted *p*-values obtained for the Benjamini-Hochberg (BH) procedure from the following formulas:

$$\tilde{p}_{(m)} = p_{(m)},$$

$$\tilde{p}_{(m-j)} = min\left(\tilde{p}_{(m-j+1)}; \frac{m}{m-j} p_{(m-j)} \right) \text{ for } \quad j = 1, \ldots, m-1 \tag{3}$$

Benjamini and Hochberg [1995], demonstrated that if test statistics are independent, their procedure controls FDR at level $q\frac{m_0}{m} \leq q$, where m_0 is an unknown number of the true null hypotheses. This means that if half of the null hypotheses are true and $q = 0.05$, the BH procedure *de facto* controls FDR at the 0.025 significance level.

The improvement of power can be obtained by applying the following two-stage modification of the BH procedure:

Stage 1. Estimate $\hat{m}_0$.

Stage 2. If $\hat{m}_0 = 0$, reject all hypotheses, otherwise apply the BH procedure at $q\frac{m}{\hat{m}_0}$.

The most common modifications of the Benjamini-Hochberg procedure are its adaptive version ABH and the two-stage procedure TSBH. Both modifications of the BH procedure are based on the initially estimated number of true null hypotheses, which in the ABH procedure is estimated directly on the basis of raw *p*-values p_i (see: [Benjamini, Hochberg 2000]), while in the TSBH procedure it is estimated on the basis of the results obtained from the initial application of the BH procedure [Benjamini, Krieger, Yekutieli 2006].

Independent test statistics rarely appear in practical studies. Bejamini and Yekutieli[3] showed that the BH procedure ensures FDR control for test statistics with more general dependence structures, such as positive regression dependence. The condition ensuring controlling FDR is the condition of positive regression dependency (PRDS)[4] on the subset of test statistics corresponding to true null hypotheses, which solves many practical problems[5]. Benjamini and Yekutieli [2001] and Yekutieli [2008a; 2008b] quoted examples of studies in which the BH procedure controls FDR, even though test statistics are not independent and are not positive regression dependent; one such example is pairwise comparisons for means, in which simulation studies indicated the conservative nature of controlling FDR by the BH procedure [Yekutieli 2008b].

Benjamini and Yekutieli [2001], proposed a conservative modification of the BH procedure, which controls FDR for test statistics with arbitrary joint distribution, regardless of the type of dependency between them. Adjusted *p*-values are obtained from the following formulas:

$$\tilde{p}_{(m)} = \min\left(1; p_{(m)} \sum_{i=1}^{m} \tfrac{1}{i}\right),$$

[3] Here and later see: [Bejamini, Yekutieli 2001].

[4] Property PRDS on I_0 (*Positive Regression Dependency on each one from a Subset* I_0) means that for any increasing set D, and for each for each $i \in I_0$, $P\left((X_1,\ldots,X_n) \in D \middle| X_i = x\right)$ is nondecreasing in x. (Set D is called increasing if $x \in D$ and $y \geq x$, implying that $y \in D$ as well.)

[5] See: [Bejamini, Yekutieli 2001].

$$\tilde{p}_{(m-j)} = \min\left(\tilde{p}_{(m-j+1)}, p_{(m-j)} \frac{m}{m-j} \sum_{i=1}^{m} \frac{1}{i} \right) \text{ for } j = 1, ..., m-1. \qquad (4)$$

In case of a great number of inferences m, calculations can be simplified by assuming[6]:

$$\sum_{i=1}^{m} \frac{1}{i} \cong 0.5772156649 + \ln m. \qquad (5)$$

R *multtest* package offers the function *mt.rawp2adjp,* which allows for obtaining adjusted *p*-values for the Benjamini-Hochberg (BH) procedure, adaptive version of the BH procedure (ABH), two-stage version of the BH procedure (TSBH) and the Benjamini-Yekutieli procedure (BY).

2.2. FDR-controlling resampling-based joint multiple testing procedures

An unquestionable advantage of resampling-based joint multiple testing procedures is the fact that they can be used in the case of the lack of normality and regardless of the type of dependencies between test statistics. Additionally, since they account for dependencies between test statistics, they have more power than versatile marginal procedures.

Westfall, Young [1993], proposed joint FWER-controlling procedures based on maxima of test statistics (*maxT*) or minima of unadjusted *p*-values (*minP*). A serious flaw of these procedures is the condition of the *subset pivotality*, which means that for any subset of null hypotheses $I\{1,\ldots,m\}$ the joint distribution of test statistics corresponding to these hypotheses must be identical under the restrictions $\bigcap_{i\in I} H_{0,i}$ and the complete null H_0^C. Westfall and Young [1993], procedures are based on data generating null distribution, which satisfies the complete null hypothesis that all null hypotheses are true. However, data generating null distribution may result in a joint distribution of the test statistics that has a different dependence structure than their true distribution (if the condition of the subset pivotality is not met). For example, the subset pivotality fails for tests regarding correlation coefficients and for tests regarding regression coefficients.

Dudoit and van der Laan [2008], proposed joint procedures of multiple testing based on the null distribution for the test statistics. Thanks to this

[6] We use the fact that Euler's constant is defined as the limit of the sequence $\lim_{n\to\infty}\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}-\ln n\right)$ and its numerical value approximately equals 0.5772156649.

approach, Type I error control does not rely on a restrictive assumption of the subset pivotality, and these procedures can be applied to pairwise comparisons of mean, to test the significance of regression coefficients in the regression model, to test the significance of correlation coefficients, and in many other studies. These procedures are implemented in R *multtest* package and called MTP. The multiple testing procedure MTP is defined by the choice of test statistics *test statistics* (these statistics are determined by the choice of the test, e.g. *t.twosamp.equalvar, t.cor, f*), *the method of estimation of the test statistics null distribution*[7] (e.g. bootstrap with centering and scaling *boot.cs*, quantile-transformed bootstrap *boot.qt*), Type I error rate (e.g. FWER, FDR) and *the joint procedure of multiple testing*[8] based on maxima of test statistics or minima of unadjusted *p*-values in a single step version (*SSmaxT, SSminP*) or a step-down version (*SDmaxT, SDminP*) which are used to control the chosen Type I error rate. FDR control is obtained by the augmentation of FWER-controlling procedures in which suitable null hypotheses are added to a set of hypotheses already rejected by the initial FWER-controlling procedure MTP [Dudoit, van der Laan 2008; Werft, Benner 2009].

3. Simulation experiment

Simulation studies presented by Werft and Benner [2009] and Denkowska [2013], revealed that the MTP procedure does not always guarantee control of selected Type I error rates. Werft, Benner [2009], reported problems with controlling FDR in genetic studies in cases of small samples and a large number of tests, while the simulation study conducted by Denkowska [2013] indicated problems with controlling FWER.

In Big Data analyses, families of inferences can be numerous, reaching even thousands of inferences. In order to further investigate problems with controlling FDR in Big Data analyses, a simulation experiment was conducted in which a family consisting of 1000 inferences was considered. In the experiment, $m = 1000$ samples of size n were independently generated from normal distribution $N(0,1)$ and the following hypotheses were tested:

$$H_{0,i} : \mu_i = 0 \quad vs. \quad H_{A,i} : \mu_i \neq 0 \quad (i = 1,\ldots,m). \tag{6}$$

[7] See: [Dudoit, van der Laan 2008].
[8] See: [Dudoit, van der Laan 2008; Westfall, Young 1993; Denkowska 2013].

The MTP function implemented in R *multtest* package was used in the experiment. The parameters of the MTP function were, among others, Student's t-test for the expected value (*t.one.samp)*, and the verified value was set at 0.

It was assumed that FDR = 0.05. Taking into consideration the fact that when all null hypotheses are true, the following equation is satisfied:

$$\mathrm{FDR} = \mathrm{E}\left(\frac{V}{max\{R,1\}}\right) = \mathrm{P}(V > 0) = \mathrm{FWER}, \tag{7}$$

in the experiment, which was repeated 500 times[9], the probability of recognizing that all null hypotheses are true was estimated depending on:

- the sample size (n = 30, 100, 500),
- the method of estimation of the test statistics null distribution (*boot.cs*, *boot.qt*),
- the joint procedure of multiple testing (*SSmaxT*, *SDmaxT*, *SSminP*, *SDminP*),
- the number of bootstrap samplings (B = 1000, 5000).

The results of the simulation tests are presented in Table 1 and Table 2. Table 1 contains the results obtained with the default number of bootstrap samplings in MTP (B =1000). It was revealed that the probability of recognizing that all hypotheses are true, estimated with the use of joint procedure of multiple testing based on minima of unadjusted p-values ($minP$) does not exceed 0,428. Moreover, the increase in the sample size did not result in the improvement of the evaluations.

Table 1. The results of the simulation study for the default number of bootstrap size sample (B = 1000)

n	*SSmaxT* boot.cs	*SDmaxT* boot.cs	*SSminP* boot.cs	*SDminP* boot.cs	*SSmaxT* boot.qt	*SDmaxT* boot.qt	*SSminP* boot.qt	*SDminP* boot.qt
30	0.988	0.986	0.400	0.428	0.97	0.968	0.338	0.354
100	0.972	0.966	0.346	0.358	0.974	0.974	0.366	0.326
500	0.990	0.992	0.366	0.322	0.992	0.992	0.370	0.362

Source: own calculations.

[9] In case of numerous families of inferences, simulation studies using the multiple testing procedures based on resampling are very time-consuming and that is why 500 repetitions are considered enough and frequently used in simulation studies (e.g. by [Dudoit, Gilbert, van der Laan 2008]).

The experiment was repeated for 5000 bootstrap samplings. The experiment turned out to be very time-consuming, and that is why it was limited to small sample sizes ($n = 30$), taking into consideration the fact that with B = 1000 no considerable improvement of results was observed when sample sizes increased. When bootstrap sample sizes were increased 5 times, the results (Table 2) of *minP* procedures improved considerably, reaching 0.88 probability of recognizing that all hypotheses are true in case of "null transformation" based on scaling and centering (*boot.cs*). In quantile transformation (*boot.qt*) the improvement was also noted, although the results cannot be considered satisfying.

Table 2. The results of the simulation study for B = 5000 samplings

n	*SSmaxT* boot.cs	*SDmaxT* boot.cs	*SSminP* boot.cs	*SDminP* boot.cs	*SSmaxT* boot.qt	*SDmaxT* boot.qt	*SSminP* boot.qt	*SDminP* boot.qt
30	0.982	0.984	0.88	0.878	0.972	0.972	0.802	0.788

Source: own calculations.

The experiment revealed a serious instability of the results of MTP procedure dependent on the number of bootstrap samplings in using multiple testing procedure based on *minP*. For a family consisting of 1000 inferences, with a default setting of bootstrap samplings (B = 1000), the results were unsatisfactory (Table 1). Increasing the number of bootstrap samplings considerably improved the results (Table 2), however, users often use default settings, unaware of the negative consequences of such decision. In the experiment when the number of samplings was increased 5 times, the results were still not satisfactory (Table 2), thus we should consider the number of samplings which will guarantee controlling FDR with the use of joint procedure *minP*. This issue is addressed by e.g. Werft and Benner [2009], who reported a problem with controlling FDR in genetic studies with a large number of hypotheses and a small sample size. In the experiment described in this paper, increasing the sample sizes did not result in the improvement of the probability of recognizing that all the null hypotheses are true (Table 1). The experiment also revealed that a joint procedure based on maxima of test statistics (*maxT*) controls FDR, but the estimated probabilities indicate the conservative nature of this control.

In a parallel simulation study on marginal multiple testing procedures, both marginal procedure BH and its two-stage modifications ABH and TSBH, obtained probability 0.95, regardless of the sample size. Only the

Benjamini-Yekutieli procedure yielded probability 0.992, which confirmed the conservative nature of the BY procedure in comparison to the BH procedure.

4. Conclusion

Uncontrolled multiple testing results in detecting dependencies which, in fact, do not exist. Controlling FWER recommended by Tukey (1953) is not a sensible solution in Big Data, because in cases of numerous families of inferences FWER-controlling procedures display a drastic loss in power. In such families of inferences controlling FDR seems the best option, that is controlling the expected value of the proportion of Type I errors among the rejected hypotheses at an a priori chosen level q ($q = \alpha$). FDR-controlling procedures allow a low percentage of erroneous rejections out of all rejections (q100%), but are not as conservative as FWER-controlling procedures. For independent test statistics or ones with positive regression dependence, a simple Benjamini-Hochberg procedure or one of its two-stage variants are recommended. In more complicated studies, the joint procedure MTP based on resampling [Dudoit, van der Laan 2008] is worth considering. A wide range of applications, the possibility of choosing the Type I error rate and easily accessible software implemented in R *multtest* package are their obvious advantages. Unfortunately, the simulation experiment described in the paper revealed that in cases of numerous families of inferences, the MPT procedure does not control FDR if the multiple testing procedure based on minima unadjusted *p*-values *minP* and a default number of bootstrap samples in MTP procedure are used. Increasing the number of bootstrap samplings considerably improves the results (although they are still not satisfactory), however, such instability of results is a cause for concern and indicates the need for further research on the MPT procedure.

In 2001 Benjamini [2001] wrote: "Even though FDR departs from classical multiple comparisons I believe it is one of the cornerstones in the bridge that 'multiple comparisons' can offer between traditional statistical thinking and modern problems". Nowadays FDR is widely accepted and recommended both by proponents of classical frequentist statistics and proponents of the Bayesian approach (see: e.g [Efron 2010; Dudoit, Gilbert, van der Laan 2008]), as it offers a rational solution to the problem of controlling multiple testing in large-scale research when Big Data are used. Regardless of the approach preferred, all statisticians share the same objective, that is adapting statistical tools to the challenges of the 21st century.

References

Benjamini Y. (2001). *False Discovery Rate in Large Multiplicity Problem*. www.math.tau.ac.il/~ybenja/Temple.ppt (6.12.2014).

Benjamini Y., Hochberg Y. (1995). *Controlling the false discovery rate: A practical and powerful approach to multiple testing*. Journal of the Royal Statistical Society. Ser. B. 57 (1). Pp. 289-300.

Benjamini Y., Hochberg Y. (2000). *On the adaptive control of the false discovery rate in multiple testing with independent statistics*. J. Behav. Educ. Statist. Vol. 25. Pp. 60-83.

Benjamini Y., Krieger A.M., Yekutieli D. (2006). *Adaptive linear step-up procedures that control the false discovery rate*. Biometrika. Vol. 93. Pp. 491-507.

Benjamini Y., Yekutieli D. (2001). *The Control of the False Discovery rate in multiple testing under dependency*. Annals of Statistics 29. Pp. 1165-1188.

Denkowska S. (2013). *Non classical procedures of multiple testing (Nieklasyczne procedury testowań wielokrotnych)*. Przegląd Statystyczny. Z. 4. Pp. 461-476.

Dudoit S., Gilbert H.N., van der Laan M. (2008). *Resampling-based empirical Bayes multiple testing procedures for controlling generalized tail probability and expected value error rates: focus on the false discovery rate and simulation study*. www.ncbi.nlm.nih.gov/pubmed/18932138.

Dudoit S., van der Laan M. (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer Series in Statistics.

Efron B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.

Hochberg Y., Tamhane A.C. (1987). *Multiple Comparison Procedures*. John Wiley & Sons. New York.

Tukey J.W. (1953). *The problem of multiple comparisons*. In: H.I. Braun. (1994). *The Collected Works of John W. Tukey*. Vol. VIII: *Multiple Comparisons: 1948-1983*. Chapman & Hall. New York. Pp. 1-300.

Westfall P. H., Young S.S. (1993). *Resampling Based Multiple Testing*. Wiley. New York.

Werft W., Benner A. (2009). www.iscb2009.info/RSystem/Soubory/Prez%20Monday/S10.4%20Werft.pdf

Yekutieli D. (2008a). *Comments on: Control of the false discovery rate under dependence using the bootstrap and subsampling*. Test 17 (3). Pp. 458-460.

Yekutieli D. (2008b). *False discovery rate control for non-positively regression dependent test statistics*. Journal of Statistical Planning and Inference 138 (2). Pp. 405-415.