ISSN (print) 2312–4547, ISSN (on-line) 2415–7325.
ВІСНИК ДНУ. Серія "Моделювання". 2017. Вип. 9, № 8. С. 142–151
DOI 10.15421/141604

УДК 519.233:004.738.5

ON THE FREQUENTIST APPROACH TO MULTIVARIATE LANDING PAGE TESTING

I. S. Bondarenko, S. V. Kravchenko

Department of Statistics and Probability Theory, Dnipro National University, Gagarin Avenue, 72, 49010, Dnipro, Ukraine. E-mail: iana.s.bondarenko@gmail.com; serega.kravchenko96@gmail.com

Communicated by Prof. P. I. Kogut

The research deals with the mathematical model of multivariate testing of the landing page of the website. The confidence intervals for the conversion rate difference of the landing page variations are investigated.

Keywords: multivariate testing, confidence interval, hypothesis, family-wise error rate, false discovery rate, conversion rate, landing page.

1. Introduction

The activity of each company is aimed at making profit. At a time when much of the target audience gets information about products and services on-line and making purchases on-line, it is important to have an effective website for every company. In order to increase profit of the company the website is subject to an ongoing and constant optimization. The object of optimization is the conversion rate - the percentage of visitors who completed a targeted action that is desirable for the website owner.

One way to optimize the conversion rate is to conduct the multivariate testing. On the landing page of the site several elements are chosen and their modifications are created. Then all possible combinations of modifications of these elements are formed and thus variations of a page are created. The flow of visitors is distributed randomly and evenly between the landing page variations. Each visitor is invited to view one of these variations and the behavior of visitors is monitored in order to know is the target action performed by the visitor or not.

All possible combinations of landing page elements are simultaneously tested with multivariate testing. It helps to evaluate the impact of each element and their interaction on the conversion rate. The landing page variation (optimal combination of elements), which won the greatest conversion rate is chosen according to the test results. It should be noted, that multivariate testing is expensive and takes time.

Services for multivariate testing Optimizely [1], Visual Website Optimizer [2], Google Analytics Content Experiments [3] and other platforms are using both classical and Bayesian statistical approaches.

The idea of Bayesian approach consists in the fact, that the model parameters are random variables with a prior probability distribution. The Bayes' theorem is used to find a posterior probability distribution of parameters that is using in Bayesian point estimates calculation and in building confidence intervals for unknown parameters. In some cases it is available the information about a priori distribution with the accuracy to unknown hiperparameters that can be evaluated simultaneously with the assessment of unknown parameters. This method won the title of Empirical Bayes method.

By using of the classical frequentist approach the model parameters are considered as unknown constants, point estimates are calculated according to the maximum likelihood estimation. Confidence intervals limits for the unknown parameters are calculated solely on the sample and therefore they do not depend on unknown parameters, and the intervals themselves contain unknown parameters with a set probability.

2. Goal setting

Knowledge systematization about mathematical model of multivariate landing page testing is implemented. Confidence intervals for the conversion rate difference of the landing page variations with correction for multiple comparisons using Bonferroni corrections, the Šidák procedure and the Benjamini–Hochberg procedure are built. The use of confidence intervals allows to present testing visually and promotes the simple interpretation of results.

3. Terminology

Here and after the conventional terminology of Internet marketing will be used.

Landing page is the page where the visitor becomes involved with advertising, search engines, mailing.

Call-to-Action is the element of the landing page (button, text link, image or other item), which leads the visitor to the conversion action. Appearance of the Call-to-Action-item should be allocated among the page's content, because it converts visitor by the user.

Conversion action is the action of the visitor that is significant for owner of the site (free content download, registration, newsletter subscription, purchasing of goods, booking services, etc.).

Unique visitor is the visitor with unique features (IP-address, browser, credentials, etc.), who went to the landing page for some period of time (day, week, month).

User is a visitor who interacts with the landing page of the site.

Original page is basic version of the landing page.

Variation is an alternative version of the landing page.

Conversion is the conversion action that is executed by the visitor on the site.

Conversion rate is the number of conversions divided by the total number of visitors.

Landing page optimization is the process of improving of the landing page elements to maximize the conversion rate.

A/A testing is a method of comparing of the same landing page to verify the accuracy of the test instrument.

A/B testing is a method of optimizing of the original landing page. The original page and its variation must be identical except for one element, whose influence on the perception of visitors is checked during the test.

4. Multiple hypotheses testing under the multivariate testing

Problem formulation. Let independent Bernoulli trials with a probability of success $p_1, p_2, \ldots, p_m$ in each group in one test are conducted in m groups. The probabilities of success $p_1, p_2, \ldots, p_m$ are unknown. Let group 1 be baseline group. Null hypotheses $H_0^{ij} : p_i = p_j$, $i = 1$, $j = 2, \ldots, m$ (group i and group j have similar probability of success) are put forward relatively the unknown parameters. The alternative hypotheses are two-tailed: $H_1^{ij} : p_i \neq p_j$, $i = 1$, $j = 2, \ldots, m$. It is necessary according to the realization of samples from Bernoulli distributions with parameters $p_1, p_2, \ldots, p_m$ make the conclusion – is the hypotheses H_0^{ij} to be rejected or not.

The frequency of success $\hat{p}_i = \mu_i/n_i$ is unbiased, consistent estimator for parameter p_i in group i and the frequency of success $\hat{p}_j = \mu_j/n_j$ is unbiased, consistent estimator for parameter p_j in group j.

Estimator

$$\hat{p} = \frac{\hat{p}_i n_i + \hat{p}_j n_j}{n_i + n_j}$$

serves as combined rate of success in the two groups.

Let us use $z-$test for the equality of two proportions [4]. If the hypothesis $H_0^{ij} : p_i = p_j$ is to reject under

$$Z = \frac{|\hat{p}_i - \hat{p}_j|}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_i} + \dfrac{1}{n_j}\right)}} \geqslant z_{1-\alpha/2}, \tag{4.1}$$

and is not to reject otherwise, then with the probability α the hypothesis H_0^{ij} will be rejected if it is fair (two-tailed alternative $H_1^{ij} : p_i \neq p_j$).

Or, that is the same, the hypothesis $H_0^{ij} : p_i = p_j$ is rejected, if

$$\hat{p}_i - \hat{p}_j \notin \left(-z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}; z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}\right) \tag{4.2}$$

and is not rejected in a reverse situation.

Or, that is the same, according to the distribution of Z statistic the p-value is defined. The $p-value$ is the probability, under the assumption of hypothesis $H_0^{ij} : p_i = p_j$, of obtaining a result equal to or more extreme than what was actually observed

$$p_{ij} = P\{Z \geqslant z | H_0^{ij}\} = 2(1 - N_{0;1}(z)). \tag{4.3}$$

The p-value is compared with significance level α. If $p_{ij} \leqslant \alpha$, so the null hypothesis $H_0^{ij} : p_i = p_j$ is rejected in favour of $H_1^{ij} : p_i \neq p_j$.

Let us calculate the amount of the unique visitors, that is needed for the testing. If simple hypothesis $H_0^{ij} : p_i = p_j$ is unfair, then alternative hypothesis $H_1^{ij} : p_i - p_j = \theta$ is fair and statistic

$$Z_1 = \frac{\hat{p}_i - \hat{p}_j - \theta}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} = \frac{\hat{p}_i - \hat{p}_j - \theta}{SE} \tag{4.4}$$

has a standard normal distribution.

The power of statistical test is equal

$$1 - \beta = P\left\{Z_1 \geqslant z_{1-\alpha/2} - \frac{\theta}{SE}\right\} + P\left\{Z_1 \leqslant -z_{1-\alpha/2} - \frac{\theta}{SE}\right\},$$

$$1 - \beta \approx 1 - N_{0;1}\left(z_{1-\alpha/2} - \frac{\theta}{SE}\right),$$

$$z_{1-\alpha/2} - z_\beta \approx \frac{\theta}{SE}.$$

Let us suppose sample sizes of group i and group j are equal ($n_i = n_j = n$), then

$$z_{1-\alpha/2} - z_\beta \approx \frac{\theta}{\sqrt{\frac{2\hat{p}(1-\hat{p})}{n}}}.$$

Hence the minimal sample size of group i (or group j), which ensures the probability Type I error α, the power of a test $1 - \beta$, expected improvement difference θ and baseline conversion rate $\hat{p}$ under the simple hypothesis checking $H_0^{ij} : p_i = p_j$ against simple alternative $H_1^{ij} : p_i - p_j = \theta$ is equal

$$n \approx \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 \hat{p}(1-\hat{p})}{\theta^2}. \tag{4.5}$$

The unique amount of visitors of the site needed for the testing is equal mn.

While checking of one null hypothesis $H_0^{ij} : p_i = p_j$, the probability of the Type I error is limited with α. Then during the similar checking of $m - 1$ independent

null hypotheses $H_0^{ij} : p_i = p_j$, $i = 1$, $j = 2, \ldots, m$, the probability of the Type I error is limited with value $1 - (1 - \alpha)^{m-1}$, which becomes too big almost when it is small enough m. For the elimination of this effect - the effect of multiple comparisons – it is made a wide range of statistical procedures, which are different according to their power and usage conditions in different situations [5–7].

Family-wise error rate ($FWER$) is the probability of making one or more Type I errors when performing multiple hypotheses tests. This value we want to control, so we need a statistical procedure, that will allow the probability of making one or more Type I errors not more then α. For the completing of the task we use the Bonferroni correction and step-down procedure of multivariate checking of hypotheses – the Šidák method [5, 6].

In cases hundreds or thousands hypothesis are checking, it can be make the certain amount of Type I errors to increase the power of the procedure and reject more unfair hypothesis, so to make less Type II errors. In such cases it is advisable to use the False Discovery Rate (FDR) – the expected proportion of Type I errors among the rejected hypotheses. For any procedure of multivariate hypotheses testing $FDR \leqslant FWER$. By means of that, if to control the FDR, we will get more powerful procedure, so it allows to reject more hypotheses. Method, controlling FDR, is increasing – the Benjamini-Hochberg procedure. One of the terms of this procedure usage lies in the independence of the hypotheses that are checking [7].

5. Implementation of multivariate testing

All possible combinations of elements of the landing page are simultaneously tested under the multivariate testing and the influence of each element and their interinfluence on the conversion rate is estimated. Visitors are offered for the viewing pages $A_1, A_2, \ldots, A_m$. To maintain the purity of the experiment the visitors should be identified during testing and during the next visits to the site they should be offered to view the same page, they saw previously. Null hypotheses of equal conversion rates on the pages A_1 and A_j ($j = 2, \ldots, m$) are put forward. The flow of visitors is modeled. Each visitor with the probability $1/m$ can be involved to one of pages. Once a visitor found himself in one of m groups his behavior is modeled. The visitor's behavior is clearly defined by two events: the success – the visitor executed conversion action – and the failure – the visitor has not fulfilled the conversion action. If it is offered to the visitor the page A_j, success will be with probability p_j.

In figure 1 the result of multivariate testing realization under the parameters $p_1 = 0,2$, $p_2 = 0,225$, $p_3 = 0,186$, $p_4 = 0,209$, $\theta = 0,015$, $\beta = 0,2$ is depicted. The Bonferroni correction is used for control of $FWER$ ($\alpha_i = 0,05/3, i = 1, 2, 3$). The existing of the conversion rate difference outside the limits of the confidence interval when the number of visitors is reached 14890 in each group means the rejection of the hypothesis H_0^{ij}. Confidence limits without correction are depicted with green color. Limits with Bonferroni correction are depicted with red color.

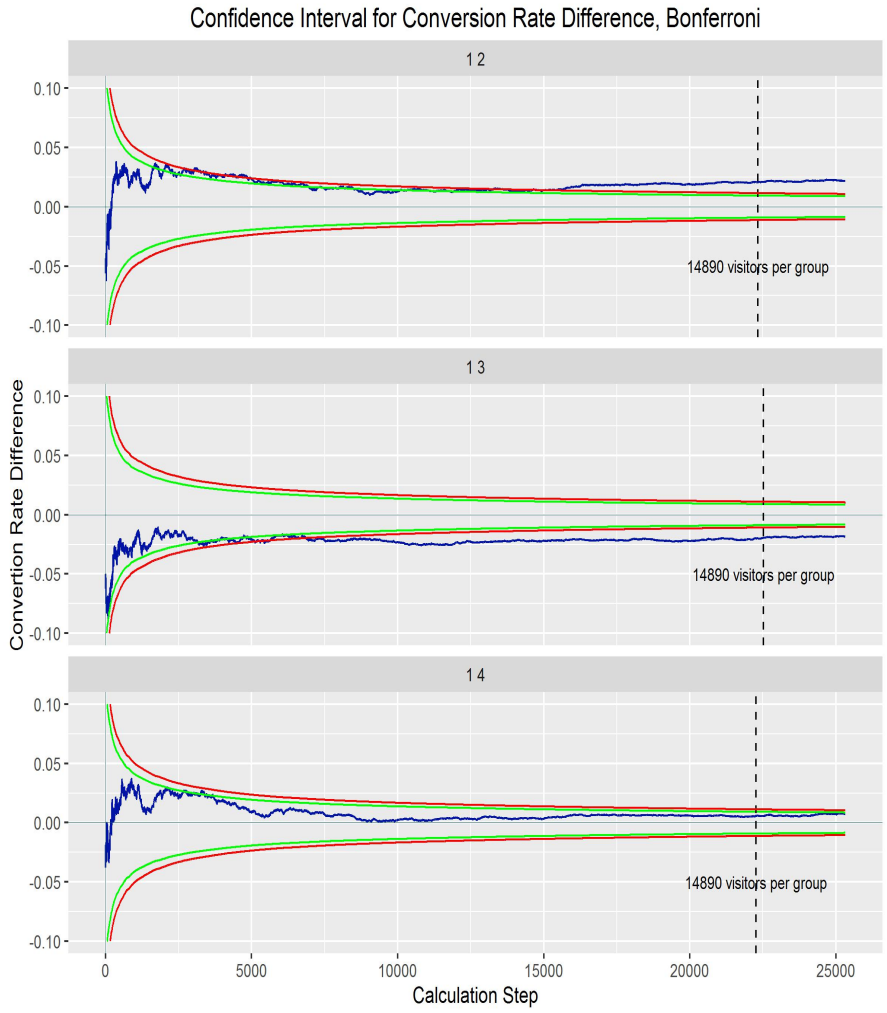

Fig. 1. Confidence intervals with Bonferroni correction for conversion rate difference on the pages A_1 and A_j $(j = 2, 3, 4)$

The Bonferroni correction is simple in realization, it is universal method – it is not dependent on the character of hypotheses and their interconnections. But this method has one sufficient disadvantage: the power of this method is decreased while the amount of statistical hypotheses that are checking is increasing.

In the figure 2 the result of multivariate testing realization under the parameters $p_1 = 0,2$, $p_2 = 0,225$, $p_3 = 0,186$, $p_4 = 0,209$, $\theta = 0,015$,$\beta = 0,2$ is depicted. We use the Šidák method for control of $FWER$ $(\alpha_i = 1-(1-0,05)^{1/3}, i = 1,2,3)$. The existing of the conversion rate difference outside the limits of the confidence interval when the number of visitors is reached 14835 in each group means the rejection of the hypothesis H_0^{ij}. Confidence limits without correction are depicted with green color. Confidence limits with Šidák correction are depicted with red color.

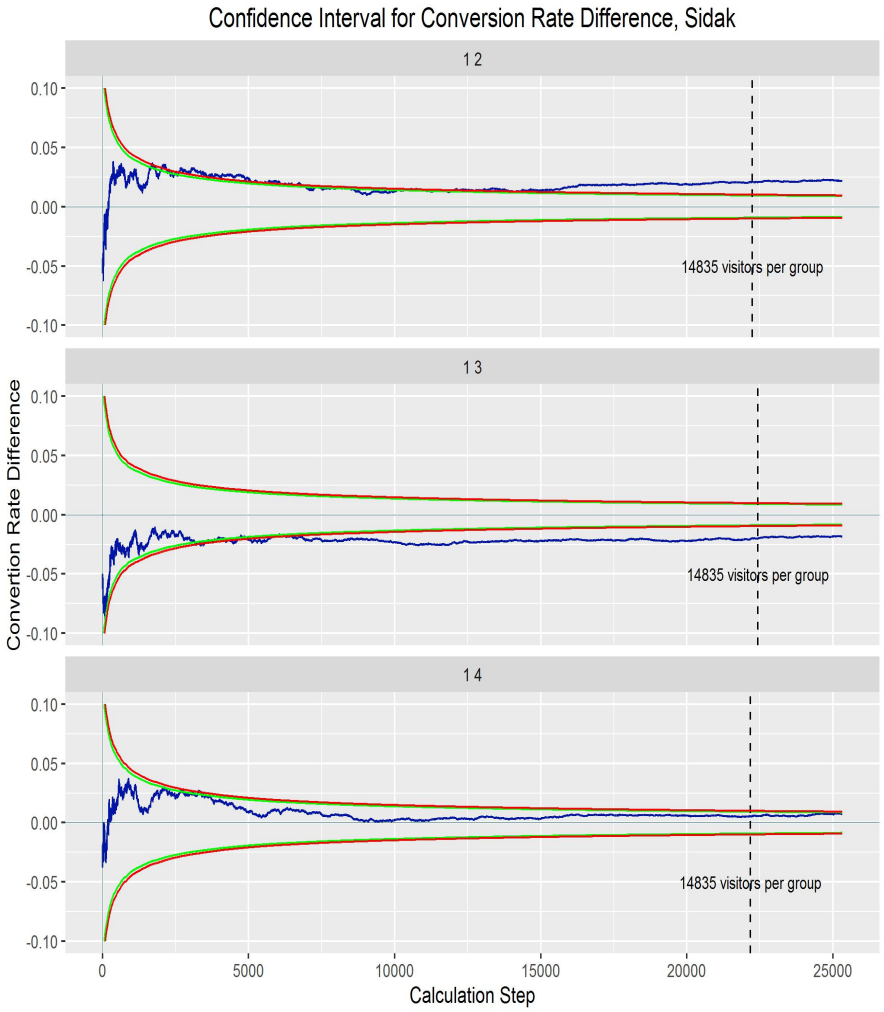

Fig. 2. Confidence intervals with Šidák correction for conversion rate difference on pages A_1 and A_j $(j = 2, 3, 4)$

The Šidák correction is derived by assuming that the statistical hypotheses are independent. Šidák method has a higher power than the Bonferroni method, but when testing a large number of hypotheses its power may be insufficient in terms of not rejection of hypotheses that are potentially interesting for detailed study and whatever had to be rejected.

In figure 3 the result of multivariate testing realization under the parameters $p_1 = 0,2$, $p_2 = 0,225$, $p_3 = 0,186$, $p_4 = 0,209$, $\theta = 0,015$, $\beta = 0,2$ is depicted. For the control of FDR we use Benjamini–Hochberg procedure ($\alpha_1 = 0,05/3, \alpha_2 = 2 \cdot 0,05/3, \alpha_3 = 0,05$). The existing of the conversion rate difference outside the limits of the confidence interval when the number of visitors is reached 12933 in each group means the rejection of the hypothesis H_0^{ij}. Confidence limits without correction are depicted with green color and limits with Benjamini–Hochberg are

depicted with red color. Due to statistical procedure confidence limits are not monotone.

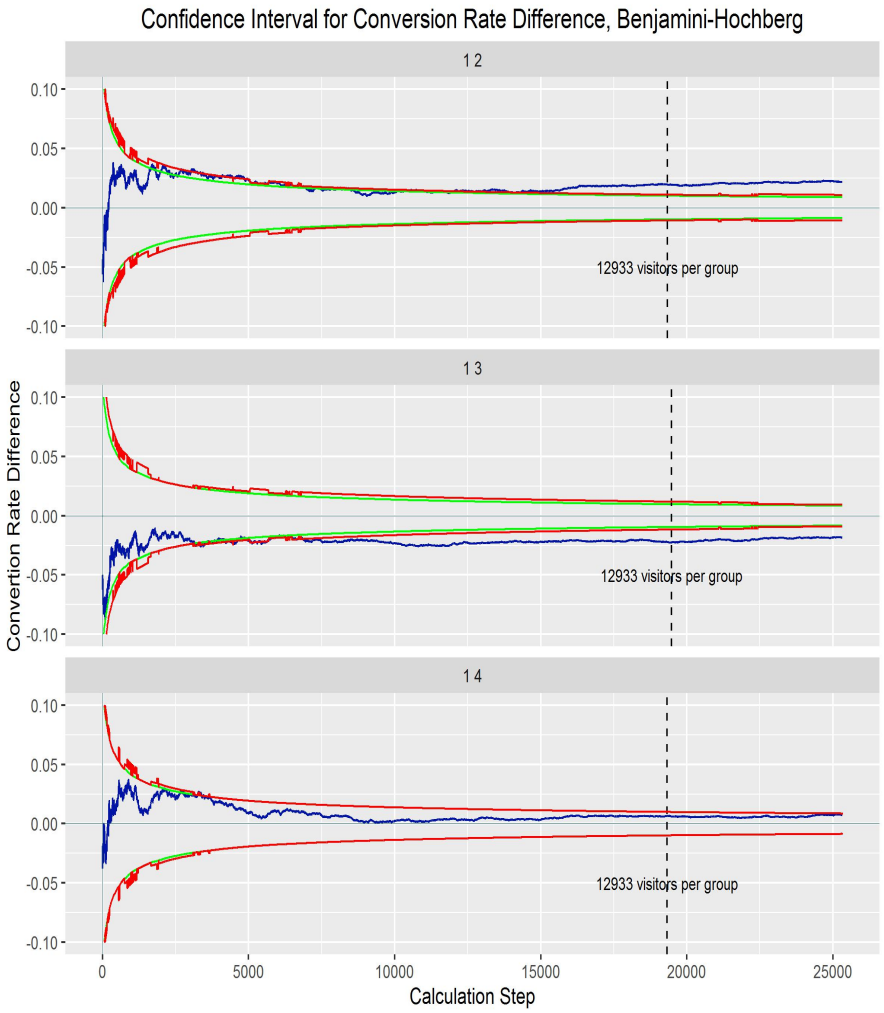

Fig. 3. Confidence intervals with Benjamini–Hochberg correction for conversion rate difference on pages A_1 and A_j $(j = 2, 3, 4)$

The results of multivariate testing under the parameters $p_1 = 0,2$, $p_2 = 0,225$, $p_3 = 0,186$, $p_4 = 0,209$, $\theta = 0,015$, $\beta = 0,2$ are presented in the table 1.

The winner is the landing page variation based on the improvement metric — the relative difference (in percentages) in performance of landing page variation concerning the baseline one [1].

The winner is the landing page variation with the conversion rate 22,53% and improvement 9,26%. Multiple comparisons with the usage of Benjamini–Hochberg correction lead to the least errors under the null hypotheses H_0^{ij} rejection, namely 0,0003; 0,0001 for hypotheses H_0^{12}, H_0^{13} correspondingly.

Table 1. The results of multivariate testing

Hypothesis	Conversion Rate	Improvement	Multiple Comparison Adjustment Method		
			Bonferroni	Šidák	Benjamini Hochberg
			Statistical Significance		
H_0^{12}	$\hat{p}_1 = 0,2044$ $\hat{p}_2 = 0,2261$	10,63 %	0,9999	-	-
H_0^{13}	$\hat{p}_1 = 0,2044$ $\hat{p}_3 = 0,1848$	-9,56 %	0,9999	-	-
H_0^{14}	$\hat{p}_1 = 0,2044$ $\hat{p}_4 = 0,2107$	3,10 %	0,4814	-	-
H_0^{12}	$\hat{p}_1 = 0,2049$ $\hat{p}_2 = 0,2261$	10,33 %	-	0,9999	-
H_0^{13}	$\hat{p}_1 = 0,2049$ $\hat{p}_3 = 0,1848$	-9,79 %	-	0,9999	-
H_0^{14}	$\hat{p}_1 = 0,2049$ $\hat{p}_4 = 0,2106$	2,77 %	-	0,4699	-
H_0^{12}	$\hat{p}_1 = 0,2063$ $\hat{p}_2 = 0,2253$	9,26 %	-	-	0,9997
H_0^{13}	$\hat{p}_1 = 0,2063$ $\hat{p}_3 = 0,1837$	-10,93 %	-	-	0,9999
H_0^{14}	$\hat{p}_1 = 0,2063$ $\hat{p}_4 = 0,2123$	2,91 %	-	-	0,7692

6. Conclusions

Multivariate testing is used to optimize landing pages in order to find the most effective combination of elements that has the highest conversion rate, as well as evaluating the effectiveness of each element of the landing page. Multivariate testing allows to maximize the conversion of any resource with high traffic.

We studied and implemented mathematical model of multivariate landing page testing, built confidence intervals for the conversion rate difference of the landing page variations with correction for multiple comparisons using Bonferroni corrections, the Šidák procedure and the Benjamini–Hochberg procedure.

Software implementation of multivariate testing is developed in the programming language R.

References

1. *L. Pekelis, D. Walsh, R. Johari*, *The New Stats Engine*, Whitepaper, Optimizely, 2015.
2. *C. Stucchio*, *Bayesian A/B Testing at VWO*, Whitepaper, Visual Website Optimizer, 2015.
3. *S. L. Scott*, *Multi-armed bandit experiments in the online service economy*, Applied Stochastic Models in Business and Industry, **31**(1)(2015), 37–45.
4. *G. K. Kanji*,*100 statistical tests*, SAGE Publications, London, 2006.
5. *F. Bretz, T. Hothorn, P. Westfall*, *Multiple Comparisons Using R*, CRC Press, 2016.
6. *Z. K.* Šidák, *Rectangular Confidence Regions for the Means of Multivariate Normal Distributions*, Journal of the American Statistical Association, **62**(318)(1967), 626—633.
7. *Y. Benjamini, Y. Hochberg*, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society, **57**(1)(1995), 289-–300.

Надійшла до редколегії 28.02.2017