Performance of a Bayesian Predictor-Based Procedure of an Additional Trial After a Non-Significant Result

Takemi Yanagimoto and Chizuru Kobayashi

***Abstract*—We attempt to contribute a fundamental issue of the significance test in terms of the Bayesian predictor-based credible region. The present view provides us with a new insight of this difficult problem. In this paper we examine numerically the performance of the proposed procedure. In light of actual setups we discuss the one-sided equivalence test for two means and incidence probabilities under the normal and the binomial errors. The proposed procedure is favorably compared with two existing naive test procedures.**

***Index Terms*— Discrete data; false discovery rate; performance of a statistical method; significance test**

I. Introduction

An explanation of the result of the significance test is often controversial, especially in the field of clinical trials. It is often believed that any additional trial is not formally allowed, if a trial ends without rejecting a null hypothesis. In fact, Cornfield (1966), a well-known Bayesian and biostatistician, stated

" *A trial finishes without showing significant difference.*
However, he still believes the hypothesis false,
and hopes to continue the trial.
If he asks a statistician an additional sample size,
then the answer is 0; he should give up."

His assertion is really right. In fact, the standard theory of the significance test indicates that no conclusion is available from a non-significant result; an additional trial is to be allowed.

On the other hand, the preservation of the nominal level in the significance test is important to avoid unnecessary excess of spurious discovery of useless findings. Various techniques to preserve the nominal level have been proposed, see Pocock (1977), DeMets (1987) and DeMets and Lan (1994), for example. Thus this problem is really tough, and there are many to be done to facilitate this fundamental problem. In practical applications this problem is critical in governmental agencies relating licensing such as approval of new therapeutics.

In order to contribute this subject, Bayesian methods looks promising, since they are flexible enough to meet complex requirements. Bayesian methods are becoming acceptable for frequentists. Though the significance test was a typical frequentist procedure, it is becoming covered by Bayesian methods, see Bolstat (2010) and Yanagimoto and Ogura (2012). In biostatistical fields Bayesian methods are employed in analyzing complex datasets, as seen in Spiegelhalter, Abrams and Myles (2004) and Berry et al. (2010).

This study is an attempt to contribute this problem. A possible procedure was discussed in Yanagimoto and Ohnishi (2009a). It is based on the prediction through the e-mixture (Yanagimoto and Ohnishi 2009b). These works are theoretical, and detailed works are necessary to examine the performance of a proposed procedure under various criteria are essential.

To contribute this fundamental problem, we are attempting to propose a novel procedure in order to make an additional trial possible. This attempt requires large amount of elaboration, but is worth to be attacked, but is worth to be attacked.

Another motivation of the present work is to pursue fundamental issues of Bayesian inference. The present interests of Bayesian researchers focus on developing practical methods, as so did the first author in Yanagimoto and Yanagimoto (1987). Our deeper understanding of fundamentals of Bayesian inference is an urgent subject.

II. Proposed procedure and its competitors

We consider a situation where a previous trial did not succeed in showing a significant result but a researcher hopes to continue the trial. The proposed procedure is based on a simple idea. The previous result is employed to elicit a prior distribution, and then a predictor-based credible region is applied to yield a Bayesian counterpart of a critical region. A specific form of a predictor-based credible region was discussed in Yanagimoto and Ohnishi (2009a). This treatment allows us to use the previous result in an indirect way. Recent developments of Bayesian methods in conjunction with the significance test were reviewed in Bolstad (2007). An earlier work of Bayesian test in relation with the Fisher exact test is seen in Altham (1969).

The proposed procedure is outlined as follows. Let $p(x|\theta_0)$ and $p_1(x|\theta_1)$ be two sampling densities, and prior densities $\pi(\theta_0)$ and $\pi(\theta_1)$. Consider the one sided equivalence test for H_0: $\theta_0 = \theta_1$ against H_0: $\theta_0 < \theta_1$. Our problem is the analysis of an additional trial, when the original trial results in a non-significance result. Let $x = (x_1, ..., x_n)$ and $y = (y_1, ... y_m)$ be samples of size *n* and *m* of an additional trial. In terms of the original trial we can obtain a posterior density, which can be

Manuscript received March 06, 2013; revised April 06, 2013 This work was supported in part by JSPS KAKENHI GRANT (C) No.23500357.

T. Yanagimoto is with *Department of Industrial and Systems Engineering, Chuo University*, Tokyo, Japan. (phone: 81-3-3817-1940; fax: 81-3-3817-1943; e-mail: yanagmt@indsys.chuo-u.ac.jp).

C. Kobayashi is with *Department of Industrial and Systems Engineering, Chuo University*, Tokyo, Japan. e-mail: chizuru_0511@yahoo.co.jp).

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

used as a new prior density, which yields the predictive density based on the new posterior density. A density is induced from the predictive densities. Consequently, the probability of the alternative hypothesis under the density is compared with 1-α where α is the significance level.

To examine the performance of the proposed procedure, we consider the two naive test procedures for the reference: One is to use only the result of the additional trial by discarding the previous result, and the other is to combine simply the results of the previous and the additional trials. We will call them the separate and the combined methods, respectively. Obviously, there are various reservations about these naive methods. It is to be avoided to discard an observation, and all the observations are hoped to be taken into account. In contrast, the combined test neglects the result of previous test completely. This method is directly related with the repeated test, which results in the excess of the actual level. We understand that the proposed test also is subject to criticism. In fact, there is no actual procedure free from the criticism. However, it is necessary to pursue a procedure in an explicit way, as discussed in Introduction. Thus the comparison study with existing methods is our major concern.

First, we consider the simplest case where the underlying error distribution is normal and the mean parameter θ_0 is known. Then simple , explicit forms of the critical regions of the three procedures in study are available. They are given in Table I, which are of familiar forms. The proposed procedure depends on the previous result through the combined sample mean. It does not depend on the sample size of the previous trial, which is to be compared with the total sample size in the combined method. This form of the proposed critical region looks satisfactory

Table 1. The rejection and incredible regions of the three methods

Method	Test statistic	Critical value
Proposed method:	$\sqrt{n}(\bar{z} - \mu_0)$	z_α
Separate method:	$\sqrt{n}(\bar{x} - \mu_0)$	z_α
Combined method:	$\sqrt{n+m}(\bar{z} - \mu_0)$	z_α

$\bar{z}$ denotes $(nx + my)/(n + m)$

Since a usual error distribution is the binomial distribution in the comparative study, we should be careful in computations, and numerical evaluations are necessary.

III. Numerical Computation

As is usual in analyzing a discrete data set, the proposed procedure can require large amounts of computation. Suitable techniques are required to reduce the computational load、since exhaustive combinatorial computation always appears. An example is illustrated in Figure 1, where critical regions are given for the original and the additional trials. We observe that there are monotone boundaries between the critical and the non-critical regions in both cases. This observation can be analytically proven. This finding reduced sharply amounts of computation, since the computation for boundaries are enough to determine the critical regions.

Other detailed techniques are applied to reduce amounts of computation. Some of which was given in a manuscript (Yanagimoto and Ogura, 2002). Consequently, there is no serious burden to apply the proposed procedure. Asymptotic approximation is accurately applied, when sample sizes are large.

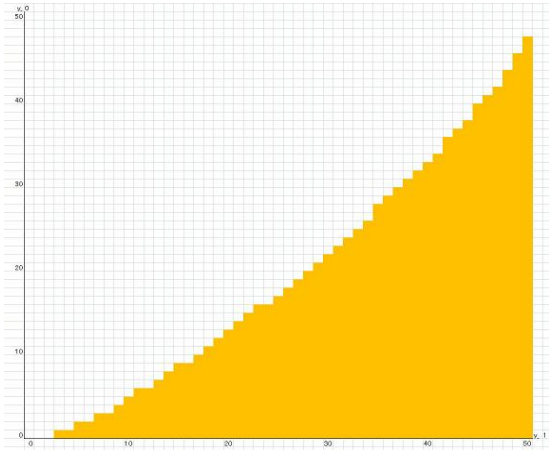

Fig. 1-1.A sample critical region of the equivalence test of two binomial populations: Case of n=m= 60

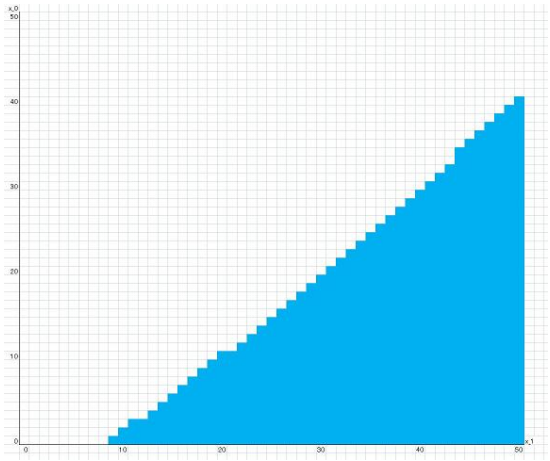

Fig. 1-2.A sample critical region of the proposed test of two binomial populations: Case of n=m= 60

IV. Performance

The performance is examined under various practical settings. In addition to samples sizes, the result of the original trial is important for the present comparison study.

4.1 Power

First, we compare powers of the proposed procedure and that of the separate method. The power is divided into two terms; the original significance level and the actual level. Our interest is in the excess of the actual level caused by the test procedure of the additional trial, and also is in the expected powers gained by the test. It is hoped that the excess is small and the power is large.

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

A numerical result is given in Table II, indicating a very small excess and comparatively large powers under alternative models. Notably, the excess of the proposed method 0.0038, which is much less than that of the separate method 0.0475. In contrast, the powers of these method becomes close, when the true alternative model is largely different from the null model. These observation s are satisfactory, since we do not hope to obtain the significant result, when the null model is true. On the other hand, we do hope to reject the null hypothethis, when the true model is largely different from the null model. This means that the application of the proposed procedure does not result in unexpected increase of the power, when the true incidence probability is small. The primary goal of the proposed procedure is that a highly effective therapy is likely to be accepted through the trials, even if an earlier trial unexpectedly shows an unlucky result.

TABLE II
The excess powers of the proposed and the separate methods under the common sample sizes. The left and the right columns denotes the excess/the power and the ratio of the power to the excess, respectively.

	Proposed		Separate	
H_0				
0.05	0.0038		0.0475	
H_1				
0.1	0.0125	3.289	0.09	1.895
0.2	0.0385	10.132	0.16	3.368
0.3	0.0696	18.316	0.21	4.421
0.4	0.1000	26.316	0.24	5.053
0.5	0.1250	32.895	0.25	5.263
0.6	0.1400	36.842	0.24	5.053
0.7	0.1404	36.947	0.21	4.421
0.8	0.1215	31.974	0.16	3.368
0.9	0.0775	20.395	0.09	1.895

4.2. Effect of the previous trial

Next, we examine the effect of the result of the original trial. When the result showed a large *p*-value, it is expected to be less likely to continue an additional trial. Since the separate method means that an additional trial is conducted independently from the result of the original trial. This view presents a strong reservation about the use of the separate method.

We consider the cases where the previous trial results in (20,20), (20, 22), (20,24), (20,26) and (20,28) under the condition in Fig 1.1. The last case is the closest to the critical region. A numerical result is given in Figure 2, showing a sensitive behavior of the proposed method by the result of the original trial. The case of (20,28) in the solid line shows close powers to the separate method. The case of (20,20) in the bold dashed line at the bottom shows lowepowers throughout.

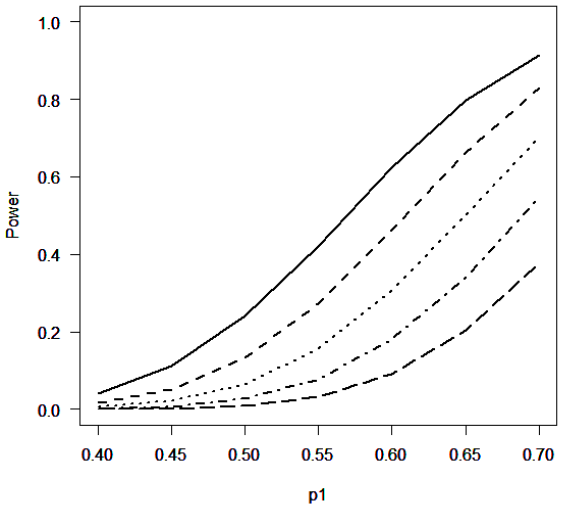

Fig. 2..Operation curves of the proposed test by different non-significant results of the original trial under the condition in Figure 1.1

This fact comes from the poor result of the previous trial. In all the cases, we observe that powers decrease as the previous trial finished poorer.

4.3 False discovery rate.

Another view can be obtained by the false discovery rate. It was proposed by Benjamini and Hochberg (1995), which provides us with a criterion for evaluating practical performance in this case. The criterion is largely different from the traditional theory of the significance test. This is because the key problem concerns the excess of the actual level. The results in the previous two subsections suggest better performance of the proposed method.

To consider practical situations, we introduce the motivation level. When the *p*-value of the result of the previous trial is close to the significance level, a researcher is likely to hope to continue a trial. The trial will stop it, otherwise. Thus we assume that a researcher continue the trial, if and only if the *p*-value of the test of the previous trial is less than the motivation level.

Three numerical results are described in Figure 3 in the next page, showing small values of the false discovery rate in the proposed method. We assumed a small incidence probability of the control group. Then the false discovery rate decreases as the incidence probability of the treated group increases. It may be observed that the false discovery rate becomes large, when the incidence probability of the treated group is large. Note, however, this behavior comes from the fact that the power at the original trial is large. This behavior comes from the fact that the previous trial shows the significant result with a high probability in such a case. We observe that the combined method results in large false discovery rates. The results of the separate method are not serious, but show greater rates than the proposed method throughout.

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

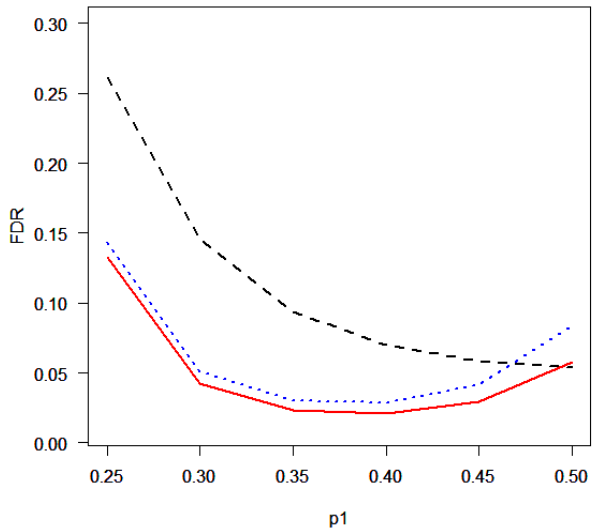

Fig. 3-1.T he false discovery rates of the proposed method (in solid line) , separates test (in dotted line) and the combined method (in dashed line) The x-axis p1 denotes the incidence probability of the alternative: Case of p0=0.2 and motivation level 0.1

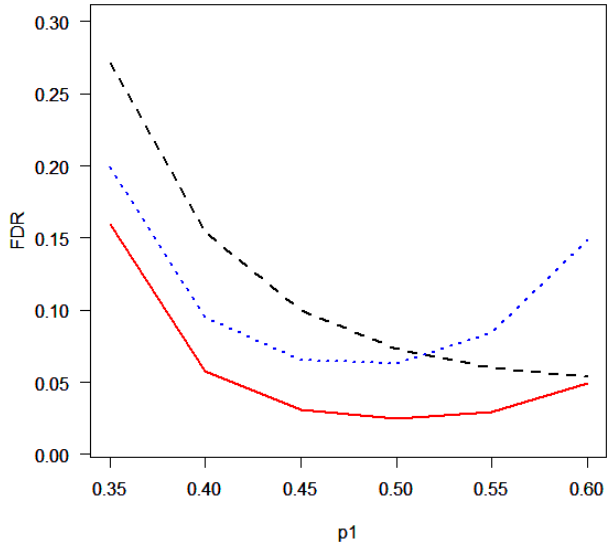

Fig. 3-2.Case of p0=.0.2 and motivation level 0.13

V Conclusion

The proposed procedure of analyzing the result of an additional trial is found to be computationally feasible, and also to perform favorably. Though the present attempt is challenging, and further detailed studies may be necessary to reach a definite recommendation of the proposed procedure. However, this attempt is worth enough to eliminate unnecessary restrictions on the experimental design.

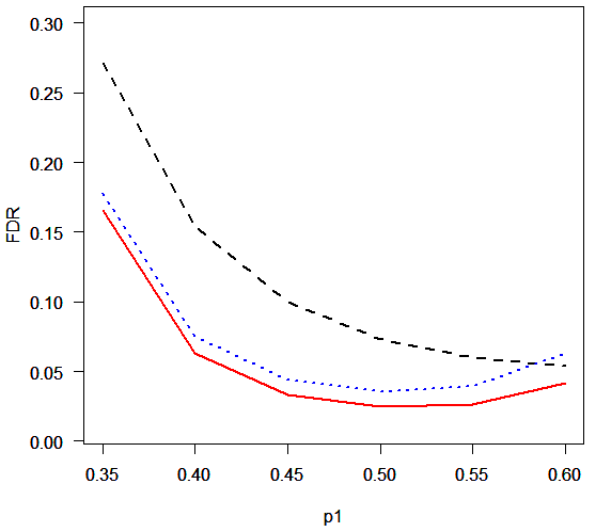

Fig. 3-3.Case of p0=.0.3 and motivation level 0.1

References

[1] P.M.E. Altham. "Exact Bayesian analysis of a 2x2 contingency table, and Fisher's "exact" significance test". Journal of the Royal Statistical Society: Series B, 31, pp 261-269, 1969.

[2] Y. Benjamini and Y. Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing.” Journal of Royal Statistical Society Series B. 57, pp 289-300, 1995.

[3] S.M. Berry et al. "Bayesian Adaptive Methods for Clinical Trials". Chapman & Hall, Boca Raton, 2010.

[4] V. W.Bolstad “Introduction to Bayesian Statistics. “ Wiley: New York , 2007..

[5] J. Cornfield. “Sequential trials, sequential analysis and the likelihood principle.” American Statistician, 29 pp 18-23, 1966.

[6] D.L. DeMets. "Practical aspects of data monitoring: A brief review". Statistics in Medicine, 6, pp753-760 1987.

[7] D.L DeMets and K.K.G. Lan. "Interim analysis: The alpha spending function approach". Statistics in Medicine, 13, pp 1341-1352, 1994.

[8] S.J. Pocock. "Group sequential methods in the design and analysis of clinical trials". Biometrika, 64, 191-199, 1977.

[9] D.J. Spiegelhalter, K.R. Abrams and J.P. Myles. "Bayesian Approaches to Clinical Trials and Health-Care Evaluation". Wiley, Chichester, 2004.

[10] T. Yanagimoto and T. Ogura. “Powerful test of two proportions by assuming a registered prior density” ISM-RM- 1161 (Abstract is available at http://www.ism.ac.jp/editsec /resmemo/resmemo-file /resm1161.htm) , 2012.

[11] T. Yanagimoto and T. Ohnishi. “Predictive credible region for Bayesian diagnosis of a hypothesis with applications.” Japanese Journal of Statistical Association, 39, pp 111-131, 2009a.

[12] T. Yanagimoto and T. Ohnishi. "Bayesian prediction of a density function in terms of e-mixture". Journal of Statistical Planning and Inference, 139, pp 3064-3075, 2009b.

[13] T. Yanagimoto and M. Yanagimoto. "The use of the marginal likelihood for a diagnostic test for the goodness of fit of the simple regression model". Technometrics, 29, pp95-101, 1987.

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)