Confidence intervals for the Berger & Boos' procedure in the 2x2 Binomial Trial

Enrico Ripamonti and Piero Quatto

Abstract The Berger & Boos' procedure [1] consists in deriving the attained size of a test by maximizing the null power function over a confidence set for the nuisance parameter. This contrasts with the original Lehmann's procedure [2], which maximizes the null power function over the entire nuisance parameter space. We adopt both these procedures in the Suissa & Shuster's test [3], which is an unconditional test for comparing sample proportions that considers either the Wald's or the score test statistic. The use of unconditional tests for comparing hypotheses on the 2×2 binomial trial is still not widespread in the applications, despite these preserve the significance level and are usually more powerful than conditional exact tests for moderate to small samples [4]. Previously, this was due to the bigger computational demand of this approach with respect to the conditional approach. Today, softwares can easily compute the p-values of both conditional and unconditional tests. We've developed a new R algorithm aimed to calculate exact unconditional p-values. Optimal values for the confidence level of the Berger & Boos' procedure are derived for different degrees of imbalance of the sample sizes.

Key words: Binomial trial, Berger & Boos' procedure, Suissa & Shuster's test

1 Introduction

In this paper we consider the problem of calculating the attained sizes for unconditional tests on the 2×2 binomial trial when the power function depends on nuisance

Enrico Ripamonti
Department of Statistics, University of Milan-Bicocca, Via Bicocca degli Arcimboldi, 8 - 20126 Milano, e-mail: e.ripamonti8@campus.unimib.it

Piero Quatto
Department of Statistics, University of Milan-Bicocca, Via Bicocca degli Arcimboldi, 8 - 20126 Milano e-mail: piero.quatto@unimib.it

parameters. Unconditional methods for testing statistical hypotheses represent an appropriate approach for, at least, three reasons. First, the unconditional approach allows a researcher to handle data in a pertinent way when only the marginal rows are fixed by design. Second, as claimed by [3], the use of the unconditional approach permits a more natural and intuitive interpretation of the results (also for the non statisticians) than the conditional approach. Third, the unconditional tests generally lead to achieve more power than the conditional tests [4].

In the case of the 2×2 binomial trial, the main problem of the unconditional approach is that the power function depends on a nuisance parameter (p, the common success probability under the null hypothesis), which has to be eliminated in order to calculate the attained sizes. Curiously, even if from a historical point of view several approaches to solve the elimination problem have been proposed, normally only the method put forth in [2] is used in the applications:

$$\begin{aligned} & sup_{0 \leq p \leq 1} P(Z(X,Y) \geq Z(x,y)) \\ = & sup_{0 \leq p \leq 1} \sum_{(a,b) \in R(x,y)} Bi(a;m,p) Bi(b;n,p) \end{aligned}$$

where Z is the test statistic, Bi is the probability function of the Binomial random variable, $R(x,y) = \{(a,b) : (a,b) \in \mathscr{X} \text{ and } Z(a,b) \geq Z(x,y)\}$. This approach eliminates the dependence upon the nuisance parameter by maximizing the null power function over the *entire* nuisance parameter space. In this way, *valid* p-values [1] can be calculated.

Nevertheless, it has been shown by [1] that the method proposed in [2] calculates the attained sizes using values of the nuisance parameter which can be very unusual in the light of the observations. It might be the case that the maximum of the null power function on the nuisance parameter space is reached for values of p that are strictly close to 0 or to 1. Consequently, [1] proposed a new approach for the computation of the attained sizes, for which these are obtained maximizing the null power function over a confidence set (calculated at a fixed level $(1-\gamma)$) for the nuisance parameter and summing up the result of this maximization with the value of γ:

$$\sup_{p \in C_\gamma} P(Z(X,Y) \geq Z(x,y); p) + \gamma$$

where C_γ is a $100(1-\gamma)$ per cent confidence interval for p. It can be demonstrated that the p-values calculated with this restricted maximization procedure are *valid* [1]. Moreover, it is shown by several examples [1] that these attained sizes are improved (in the sense of less conservatorism) with respect to those calculated with the original unrestricted maximization procedure.

Several authors have compared the degree of conservatorism and the power achieved by both conditional and unconditional tests calculated with different methods. Nevertheless, as recently stated in [4], no research has been yet conducted on the use of different confidence levels. In [1], it is suggested to fix γ at 0.001 whereas [4] claim that in the applications most authors fix γ at either 0.001 or 0.0001 (with the relevant exception of the popular software StatXact 8, which sets 0.000001 as default value). All these proposals appear as cryptic suggestions, since no investigation has been so far conducted in order to find optimal values of γ.

We propose a new R algorithm aimed to calculate the attained sizes and the power of the original Suissa & Shuster's test for both balanced and imbalanced sample sizes. We've considered both the unpooled Z test, which is directly treated in [3], and the pooled Z test, which has been found to achieve higher levels of power than the unpooled test [5].

2 A Monte Carlo Study

The original Fortran algorithm used in [3] is a complicated two-steps procedure involving: i) an analytical calculation on the derivative of a null power function; ii) a numerical routine aimed to produce a least upper bound on the null power function.

We've implemented an R algorithm in order to directly calculate both the attained sizes and the power of the test using either the unpooled or the pooled Z statistics in the case of both balanced and imbalanced sample sizes. This code has been used for the computation of the attained sizes of the test for the relevant cases of $\alpha = 0.05, 0.025, 0.01$. These sizes have been calculated using both the unrestricted [2]'s procedure and the restricted [1]'s procedure, fixing the confidence level at 0.001, 0.0001, 0.00001. Asymptotic confidence sets for the nuisance parameter have been computed, using Monte Carlo simulations from binomial random variables with different success probability parameters (P=0.10; 0.25; 0.50; 0.75; 0.90).

In [3] it is reported that, using the [2]'s maximization procedure, the unpooled and the pooled Z tests are equivalent in the case of balanced sample sizes and we've replicated this result using the new R algorithm. In the case of imbalanced sample sizes we've found that the pooled Z test is less conservative than the unpooled Z test. Furthermore, generally the pooled Z test achieves more power than the unpooled Z test. This result has also been reported by previous works (e.g. [5]), but we've studied more cases, varying the degree of imbalance of the two sample sizes. Nevertheless, the pooled Z test is not uniformly more powerful than the unpooled Z test, since when the imbalance is slight (e.g. $n_1 = 10, n_2 = 20$) the former proves to be less powerful than the latter.

With respect to the unpooled Z test, using the [1]'s procedure in order to calculate the attained sizes leads to less conservative tests, especially when the probability parameter in the population on which is calculated the confidence set is not extreme ($P = 0.25; 0.50; 0.75$). Moreover, we've found that, in terms of conservatorism, it is not useful to calculate an interval at a larger confidence level (i.e. $\gamma = 0.00001$), but the best performances are obtained when $\gamma = 0.001$ or $\gamma = 0.0001$.

With respect to the pooled Z test, the use of the [1]'s procedure to calculate the attained sizes leads to less conservative tests when the probability parameter in the population is not extreme ($P = 0.25; 0.50; 0.75$). On the contrary, when $P = 0.10$ or $P = 0.90$, the use of the classic [2]'s procedure is more appropriate. As far as the

[1]'s procedures are concerned, it is not relevant (in terms of conservatorism) to fix a larger confidence level (i.e. $\gamma = 0.00001$), since the best performances have been obtained when $\gamma = 0.001$ or $\gamma = 0.0001$.

The use of the [1]'s procedure leads to more powerful tests than the [2]'s procedure in case of high imbalanced designs (e.g. $n_1 = 20, n_2 = 70$). Only very slight differences emerge with respect to the use of different confidence levels ($\gamma = 0.0001 \approx \gamma = 0.00001 > \gamma = 0.001$). Hence, we suggest to fix a not too large level of confidence for the [1]'s procedure (i.e. $\gamma = 0.001$ or $\gamma = 0.0001$), thus obtaining a less conservative test. In cases of low imbalanced designs (e.g. $n_1 = 20, n_2 = 30$; $n_1 = 30, n_2 = 50$) the two maximization procedures lead to equally powerful tests, regardless of the level of confidence.

These pattern of results have been obtained with all the levels of α that we've fixed ($\alpha = 0.05$, $\alpha = 0.025$, $\alpha = 0.01$), for both the pooled and the unpooled test statistics.

3 Conclusion and further directions

We've studied by means of Monte Carlo simulations the problem of deriving the attained sizes and the power of the test developed by [3] using: i) either the unpooled or the pooled Z statistics; ii) either the restricted [1] or the unrestricted [2] maximization procedures; iii) either balanced or unbalanced samples. Globally, our results strongly support the choice of the restricted maximization procedure in cases of high imbalanced sample sizes both in terms of less conservatorism and of higher levels of power.

We are assessing the [1]'s procedures using different methods to construct the confidence set (e.g. Clopper-Pearson, Bayesian). Moreover, we're planning to use other test statistics (e.g. Fisher-Boschloo's test; Lancaster's unconditional test; Liebermeister's unconditional test) and to thoroughly compare both the degree of conservatorism of the p-values and the power achieved by these unconditional tests.

References

1. Berger, R. and Boos, D.: P-values maximized over a confidence set for the nuisance parameter. Journal of the American Statistical Association. **89**(427), 1012–1016 (1994).
2. Lehmann, E.: Testing Statistical Hypotheses. Springer, New York (1959).
3. Suissa, S. and Shuster, J.: Exact unconditional sample sizes for the 2 × 2 binomial trial. Journal of the Royal Statistical Society. Series A (General), **148**(4), 317–327 (1985).
4. Lydersen, S., Fagerland, M. and Laake, P. Recommended tests for association in 2×2 tables. Statistics in Medicine, **28**(7), 1159–1175 (2009).
5. Berger. R. Power comparison of exact unconditional tests for comparing two binomial proportions. Institute of Statistics Mimeo Series (1994).