Contemporary Clinical Trials 107 (2021) 106491

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

A comprehensive open-source library for exact required sample size in binary clinical trials

Chris J. Lloyd [a], Enrico Ripamonti [b,c,*]

[a] Melbourne Business School, The University of Melbourne, Melbourne, Australia
[b] Karolinska Institutet – A Medical University, Stockholm, Sweden
[c] Milan Center for Neuroscience NeuroMi, Milan, Italy

ARTICLE INFO

Keywords:
RCT
Sample size
Exact computation
R library

ABSTRACT

We describe how we are creating a new and comprehensive R library solving the problem of exact sample size determination of RCTs. A crucial prerequisite for the trial protocol is a priori sample sizes that bound the test size below a target (often 5%) and the test power above a target (often 80%). Approximate formulas are available for binary trials but the target test size and power are often violated by standard methods for even quite large sample sizes. Moreover, adjusting standard tests to take account of their size bias can reduce power substantially. This has been well known for several decades. Exact and quasi-exact tests are now available and can be computed in a few seconds for a single data set. However, calculating the exact power and size of such tests requires computing them for all possible outcomes. Searching for minimum samples sizes that achieve a given target requires doing this for a wide range of sample sizes. This becomes computationally infeasible very quickly; to compute required sample sizes for a target size of 5% and power of 80% would, on a standard computer, take several months. Computation time increases as the size and clinically relevant difference decreases. After having presented the main operative challenges to creating this library, mainly due to the need of summarizing a very large amount of information, we put forward our innovative solutions to deal with this complex problem from a statistical viewpoint. The described library will be released in open source.

1. Introduction

Consider a two-arm randomized clinical trial (RCT), where n participants are randomized into two groups of size n_1 (experimental treatment) and n_0 (control treatment). The unknown probabilities of target binary response are p_1 and p_0 respectively. The researcher chooses a minimal clinically relevant therapeutic difference $\delta = p_1 - p_0$ as well as the a priori target test size α^* and power β^*. We aim to determine minimal sample sizes n_0, n_1 that meet these targets.

Define Y_1 and Y_0 as the number of participants with a positive response in the two groups. Without loss of generality, any test can be defined in terms of a so-called p-value statistic $P(Y_0, Y_1)$ and the rule "reject the null hypothesis if $P(Y_0, Y_1) < \alpha$". Standard theory [1–3] based on a central limit approximation to the distribution of the Wald statistic leads to approximate minimum sample size formulas such as

$$|\delta| = Z_{1-\alpha^*}\sqrt{p(1-p)(1/n_0+1/n_1)} + Z_{\beta^*}\sqrt{(p_0+\delta)(1-p_0-\delta)/n_1 + p_0(1-p_0)/n_0}. \quad (1)$$

where $p = (n_0p_0 + n_1p_1)/n$, Z_q is the upper standard normal q-quantile. For given p_0 and clinical difference δ one then finds a range of solutions for n_0 and n_1.

The problem is that the approximate test that this theory is based on can grossly violate the target error rates α^* and so the target power is achieved, if at all, at the expense of inflated size. One can potentially correct this by computing the exact upper critical value of the test statistic, or equivalently by computing the exact worst-case significance of the test result. However, as explained below this can lead to loss of power. What clinicians need is a powerful test that always satisfies the type 1 error restriction and then required sample sizes to achieve the power target.

The main aim of this article is to explain the computational impossibility of doing this in real time and to describe the process of creating a new and comprehensive R library that gives the exact power and size of a quasi-exact test for a large range of sample sizes and effect sizes. This can be used to extract optimal sample sizes and/or to create a smaller library of sample sizes that achieve target error rates and provided

* Corresponding author at: Milan Center for Neuroscience NeuroMi, University of Milan-Bicocca, Piazza Ateneo Nuovo 1, 20126 Milano, Italy.
E-mail address: enrico.ripamonti@unimib.it (E. Ripamonti).

https://doi.org/10.1016/j.cct.2021.106491
Received 1 January 2021; Received in revised form 17 June 2021; Accepted 18 June 2021
Available online 22 June 2021

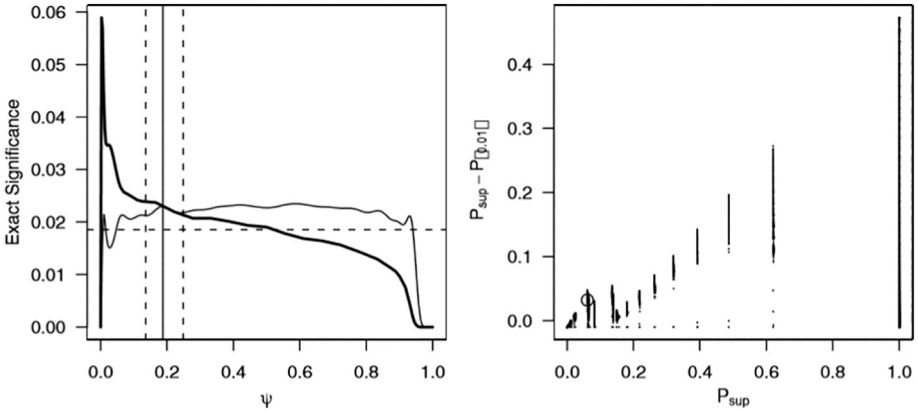

Fig. 1. From Lloyd [6]. *Left.* Observed profile $\Pr(T \geq 2.0847; \psi)$ (thick curve), where T is the pooled T-statistic (score) and ψ is the common probability of success. Vertical dashed lines are the 99% confidence interval for ψ. The horizontal dashed line is the asymptotic p-value. *Right.* (for details see explanation in the text).

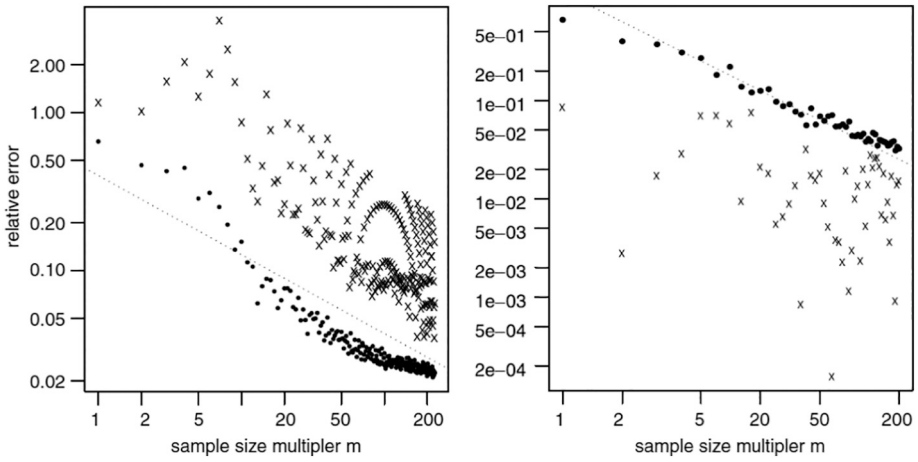

Fig. 2. From Lloyd [8]. Accuracy (average relative error and maximum relative error) for the standard likelihood ratio test (left) and for the E test (right). (for details see explanation in the text).

directly to researchers. Both will be open source.

2. Why exact?

Required sample sizes clearly depend on the test employed. A less powerful test procedure will require larger sample sizes. We also require excellent size control, since a test that violates the type 1 error bound will achieve the target power under false pretences. And as Eq. (1) above shows, the target sample sizes are quite sensitive to the size of the test.

Our library is based on what we call E p-values [4,5], explained in detail in the next section. Tests based on E p-values have size very close to nominal, for all possible values of the unknown response probability p_0. In contrast, asymptotic tests have poorer size control than is typically appreciated. Maximization accounts for the non-pivotality of the test statistic but at the expense of power compared to the E p-value, which is very close to pivotal and almost exact, and so does not require any maximization in practice.

While violations often occur for post-hoc unlikely values of the nuisance parameter this makes no difference in the frequentist framework. The left plot in Fig. 1 (from Lloyd [6]) describes the standard t-test for sample sizes $n_1 = 47$ and $n_0 = 283$ and the particular outcome $y_1 = 14$ and $y_0 = 48$. There is nothing special about these numbers except that the outcome of the experiment is in the "interesting range": the realised value of T is 2.0874 and the approximate p-value is $P_T = 0.0185$. The left plot shows $\Pr(T \geq 2.0874)$, or equivalently $\Pr(P_T \leq 0.0185)$ as a function of the common probability ψ of response, computed exactly using the underlying joint binomial distribution.

The true probability is as large as 0.0611. The estimated value of ψ is the vertical line and gives the estimated p-value of 0.0239. This is what we call the E p-value. The E p-value itself is a statistic and we can calculate the probability of it being equal or smaller than its observed value. This profile is displayed as the flatter curve. It is uniformly close to the quoted value of 0.0239 for all values of ψ. So should we quote 0.0239 or 0.0611?

The right-hand plot compares the p-value based on taking the maximum of the T profile (which is called the M p-value) with the E p-value, across all possible data sets. For almost every data set, the E p-value is smaller than the M p-value. Since both control size, the E p-value is superior. It will have greater power for the same size. The uniformity of this power dominance has been studied in Lloyd [7].

It is worth remarking that the size violations of approximate tests persist for quite large sample sizes whereas size violations of E p-values are practically negligible for all sample sizes. A striking graphical display of the size accuracy of T compared to the E p-value across a range of sample sizes may be found in Fig. 2 of Lloyd [8] reproduced below.

The horizontal axis is sample size (a small 2×2 table is being multiplied up by a factor m). The vertical scale is the relative size inaccuracy. The x symbols are worst violation. The ordinary points are average absolute size error. The left plot is for the standard likelihood ratio test. A value of 1 means the actual size is 100% larger than nominal. Even for a multiplier of 200 (where total sample size is around 1000) you can have 20% maximum size bias. The right plot is for the E p-

value. The maximum of the vertical axis is 0.05 meaning a 5% violation of nominal size. For instance, we claim it is 5% but it is actually 5.25%. With a sample size multiplier of 20 the maximum size violation is around 1%.

In summary, size violation is still an important problem for approximate tests for moderate sample sizes and this has consequences for required sample size calculations. Allowing for the erratic profile of the standard p-value leads to power loss. We use the E p-value because it has excellent size and power properties. And it is required samples sizes for this choice of test that are to be computed.

3. A new library for sample size computation

We denote the probability of any event under the joint binomial model for Y_0, Y_1, by $Pr_{n_0,\ n_1}(\ |\pi)$, where the "event" is indicated as "blank", $\pi = (p_0, p_1)$. Let $\widehat{\pi}_0$ denote the restricted ML estimate under the null hypothesis. For this article, the null hypothesis is $p_0 = p_1$ and $\widehat{\pi}_0 = (\widehat{p}, \widehat{p})$, where $\widehat{p}$ is the total responses over total sample size. For testing non-zero differences, the restricted ML estimate involves solving a cubic equation [9]. The so-called E p-value is then given by

$$\widehat{P}(y_0, y_1) = Pr_{n_0,n_1}\left\{P(Y_0, Y_1) \le P(y_0, y_1)\,|\widehat{\pi}_0\right\} \tag{2}$$

The quasi-exact E p-value is very close to pivotal, which means that it achieves the stated size almost exactly. Computing $\widehat{P}(y_0, y_1)$ for a single data set (y_0, y_1) requires computing all possible values of the approximate p-value $P(y_0, y_1)$ which involves $N = (n_0 + 1)(n_1 + 1)$ evaluations. In our previous example where the E p-value was 0.0239 and was read off the "profile" plot, computing the plot itself require computing all $N = 11{,}472$ possible values of the T statistic. If $P(y_0, y_1)$ is monotonic in y_0, y_1 however we do not have to compute all possible values to determine the tail set in (2). We can instead make use of bisection [6] to reduce the expected evaluations of the T-statistic to $N^* = n_{\min} ln(n_{\max}) = 265$. Even without this computational trick, computing a *single realisation* of $\widehat{P}(y_0, y_1)$ is apparently instant for sample sizes of the order of a hundred and takes a few seconds for samples sizes of the order of five hundred.

To compute the exact error rates of the quasi-exact test based on the rule "reject the null hypothesis if $\widehat{P}(Y_0; Y_1) < \alpha^*$" requires computing *all possible values* of $\widehat{P}(Y_0; Y_1)$ which is a computation of order $O(N^* \times N)$. If the experimental arms n_0, n_1 are both of order $O(n)$, then this is a computation of order $O(n^3\ log\ n)$ at best. This cannot be done in a few seconds. For $n = 50$ it takes about 8 s while for $n = 500$ it takes over four hours. This order of computation does not assume equal allocations; none of these do, though N depends on n_0 and n_1 and is largest when they are equal.

Clinicians need a tool that provides the minimum sample sizes n_0, n_1 for their planned trial, from provided values δ, α^*, β^* as well as a likely value p of the control probability p_0 of response in the control group. The exact power of the test is

$$\beta = Pr_{n_0,n_1}\left\{\widehat{P}(Y_0, Y_1) \le \alpha^*\,|\pi = (p, p+\delta)\right\} \tag{3}$$

depending on n_0, n_1, α^*, δ and p. Of course, when $\delta = 0$ this is the size, which we denote α. Once all possible values of $\widehat{P}(Y_0, Y_1)$ have been calculated, which is a $O(n^3\ log\ n)$ computation, calculation of the exact errors rates are only $O(N)$ and can be done quickly for a grid of values of p, δ and α^*. But each different sample size pair (n_0, n_1) requires another $O(n^3\ log\ n)$ computations.

Depending on the application, the samples sizes need not be equal so the problem is to determine minimum values of the control sample size $N_0(n_1)$ as a function of the treatment sample size such that

$$N_0(n_1) = min\{n_0 : \beta(n_0, n_1) \le \beta^*, \alpha(n_0, n_1) \le \alpha^*\} \tag{4}$$

It is known from extensive numerical work [4,10] that, for the E p-value, the actual size α is extremely close to the target error rate α^*, this being illustrated in the earlier example and the source of the label "quasi-exact". So, the second inequality restriction can be practically ignored. The set of solutions depends on provided values p, δ, α^*, and β^*.

3.1. Computational challenge

Because it is not feasible to compute $\beta(n_0, n_1, p, \delta, \alpha^*)$ for practical values of n_0, n_1 in real time, it is not feasible to allow users to specify p, δ, α^*, β^* and then search for a set of solutions $N_0(n_1)$ as defined in (4). The only practical approach is to store values of β (and less critically α) for a practical but restricted grid of values for p, δ, α^* for all values of n_0, n_1 within a practical range.

The *first challenge* is to get everything relevant into our data base, so as to support any reasonable query from the user about planned sample size. While in the context of RCTs we can be quite restrictive in terms of choice of α and β values, there being commonly established standard for target size and power, it is important to leave flexibility over sample size combinations n_0, n_1, as well as allowing for background knowledge of the control response probability p_0. The *second challenge* is that the final library of results should not be so large that (a) it can never be stored in available packages and (b) extraction of results in real time for the user from the package becomes prohibitively slow. The least important *third challenge* is that creating this final library of results should be executable, by the authors of this article, in a practical time scale for a research project.

3.2. Practical solution

As our approximate p-value $P(y_0, y_1)$ on which the E p-value is based we adopt the standard normal approximation to the signed likelihood ratio statistic, which is straightforward to compute though perhaps less well-known than the standard T statistic. However it has some guaranteed and important monotonicity properties [11]. This choice has almost no impact on the resulting quasi-exact p-value $\widehat{P}(Y_0, Y_1)$ [4]. Indeed, it is an attractive property of E p-values that they hardly depend at all on the initial choice of the user's favourite test statistic.

We start by setting a restricted range for both sample sizes, currently 25 to 500. This can be extended in the future though the computational effort to do so will be extreme. For each pair (n_0, n_1), we first compute all possible values of $\widehat{P}(Y_0, Y_1)$. For smaller sample sizes, this is fast. But computation time grows quickly with sample size as we have seen. For both sample sizes equal to 500, it takes over 4 h on the fastest desktop machine that we have available.

We compute exact size α and power β for the following grid of parameter values: δ = 0,0.05,0.1,0.2, target type 1 error α^* =0.005,0.01,0.02,0.025,0.05,0.10, and control group probabilities in increments of 0.01 from 0 up to $1 - \delta$. The power results can be added into a table for each of the 2214 parameter combinations of p, α^*, δ just described.

We do not store the E p-values however, as their total number across all sample sizes from 25 to 500 will end up being 60,591,902 (details available on request). There are very few computing environments that could interrogate a data base of this size, certainly not R-Studio.

At the end of this process, we have 2214 exact powers for each pair of sample sizes (n_0, n_1) within the designated range. Constructing this library, even breaking the project into 20 parallel sets of computations on different machines, requires *several months* of computation.

Once this library of power data has been accumulated by the researchers, it can be used to compute minimum sample sizes $N_0(n_1)$ by solving (4) subject to $n_0, n_1 \le 500$. If there is no solution then we return the value $N_1(n_0) = 500$ indicating no solution and a larger sample size than 500 required. The solution vectors $N_0(n_1)$ depend on p, α^*, δ but also depends on the target type 2 error β^* which we will limit to four possible values. There are thus $2214 \times 4 = 8856$ solution vectors that

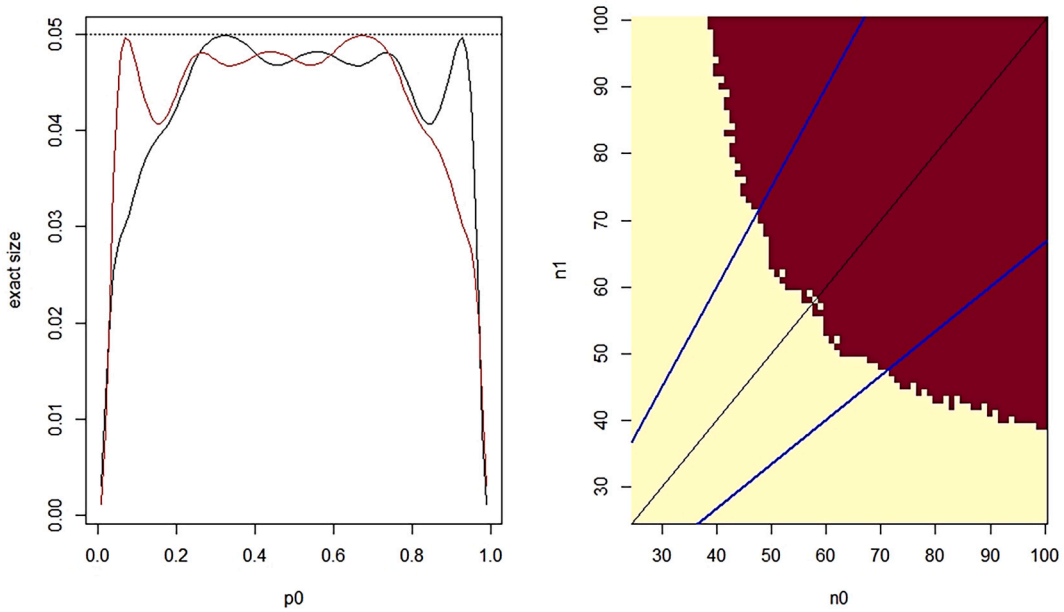

Fig. 3. Illustrative plots. *Left*. Exact size α for $n_0 = 50$, $n_1 = 35$ (black) and $n_0 = 35$, $n_1 = 50$ (red). Worst case type 1 error is almost exactly equal to nominal $\alpha^* = 0.05$. *Right*. Image plot of exact power minimised with respect to p for $\alpha^* = 0.1$, $\beta^* = 0.2$ and $\delta = 0.2$. Balance boundaries are in blue, axis of symmetry in black. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

need to be stored. This is a very manageable amount of data both to download and to be interrogated by the clinician. More likely than a single value p, clinicians will likely give a range R of values for p. In this case, one need only extract the solution vectors $N_0(n_1)$ for discrete values of $p \in R$ and report the largest.

4. Some clarifying illustrative plots

It is simple to show that the exact power β satisfies

$$\beta(n_0, n_1, p, \delta) = \beta(n_1, n_0, 1-p, \delta)$$

so it is sufficient to calculate powers for $n_0 < n_1$ and fill in the results for $n_0 > n_1$ using the above symmetry. A confirmatory plot is in the left panel of Fig. 3 which shows the exact size of the E p-value based test with $\alpha^* = 0.05$ and sample sizes $n_0 = 50$, $n_1 = 35$ in black. The worst-case size is extremely close to the target 5% and this is typical of the E p-value based test. The exact size of the test with sample sizes $n_0 = 35$, $n_1 = 50$ is displayed as a solid red line.

To give an idea of the process involved in extracting optimal sample sizes, suppose a user sets $\alpha = 0.1$, $\beta = 0.2$, $\delta = 0.2$. For illustration, we imagine that both sample sizes are limited to the range 25 to 100. We can do this because we know that there will be many solutions within this range for this favourable choice of α, β and δ. This will not be the case for smaller values and a useful database will need to allow for much larger sample sizes.

With sample sizes limited to 100, there are $76^2 = 5{,}776$ distinct values of n_0, n_1, each with 81 powers corresponding to 0:0.8(0.01). For each pair of samples sizes, we calculated the power, minimised over the 81 values of p. This would be appropriate if the user had no prior information on p. The right image plot shows those values where the minimised power exceeds the 0.8 threshold. As anticipated, it is symmetric about the axis $n_0 = n_1$. Because of the peculiarities of integer arithmetic, the boundary of this region is irregular and is also considerable shifted from the approximate solutions mentioned in the introduction.

Practitioners will want a range of "smallest" solutions, i.e. for fixed values of n_0 the smallest value of n_1 that achieves the target power. There may also be a target range for the balance ratio $\gamma = n_1/n_0$. A typical "balance" restriction is highlighted by the blue lines within which the ratio is between 2/3 and 3/2. Because of the earlier mentioned symmetry in (n_0, n_1) a search would only be required between the blue line and the line of symmetry.

Searching the table can be done very quickly. It is the creation of a table covering larger sample sizes that, in principle, will currently take several months of computing. Allowing sample sizes up to 500 will increase the number of samples size combinations from 5776 to 226,576. However, this will not make the search algorithm impractical. It is the computing burden for filling out the entries of the table itself that will be increased by a factor $5^3\ln(500)/\ln(100) = 168$. Increasing from 500 to 1000 increases total computation to create the expanded table by a factor of 8.89.

5. Conclusion

Exact calculations of sample sizes are performed in other software, for instance PASS. However, when sample sizes become moderate, they resort to normal approximations because of the computational burden. We have pointed out that these approximations are unsuitable for determining sample sizes that are guaranteed to satisfy size and power targets. StatXact also has an algorithm for exact sample sizes, but based on the pooled t-test. For each pair of sample sizes, this is only an O(N) computation. Our library is based on the quasi-exact E p-value because of its distinctly superior power properties as explained in this article. This makes the computational burden $O(N^2)$. Another package PASS explicitly resorts to normal approximation based methods as soon as sample sizes become moderately large, and are in any case based on the pooled t-test.

The final package available online will load a structured library of powers for a range of values of α, δ, p and n_0, n_1 as described in Section 2.2. The user will supply target values α^*, β^*, a clinically relevant difference δ, a range of values of p and an optional balance restriction on n_0/n_1. Importantly, the package will also include a function for calculating the E p-value after experimental results are obtained, as well as the exact size and power of the test if target sample sizes are not met. This should ensure that the researcher uses the test that the sample size calculations are based on.

There may also be a possibility to select an alternative approximate

p-value (default: LR), for instance the more standard Wald or T statistic based p-value; however, a large amount of research shows that this choice has almost no effect. There will also be optional filtering for an expected dropout rate. A search algorithm will then find smallest sample sizes $N_0(n_1)$ as displayed in (4).

In conclusion, this work will provide clinical trialists with an R library that can immediately furnish the required sample size, allowing them to set different parameters values at the same time facilitating their use of a more powerful test with guaranteed properties.

Declaration of competing interests

There are no competing interests related to this submission

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cct.2021.106491.

References

[1] J.M. Lachin, Introduction to sample size determination and power analysis for clinical trials, Control. Clin. Trials 2 (2) (1981) 93–113.
[2] K.F. Hirji, Exact Analysis of Discrete Data, CRC Press, 2005.
[3] A. Agresti, Categorical Data Analysis, John Wiley & Sons, 2013.
[4] C.J. Lloyd, E. Ripamonti, An exhaustive numerical assessment of alternative unconditional tests of a binary treatment effect, J. Stat. Comput. Simul. 88 (11) (2018) 2150–2169.
[5] E. Ripamonti, C. Lloyd, P. Quatto, Contemporary frequentist views of the 2x2 binomial trial, Stat. Sci. 32 (4) (2017) 600–615.
[6] C.J. Lloyd, Exact p-values for discrete models obtained by estimation and maximization, Aust. N. Z. J. Stat. 50 (4) (2008) 329–345.
[7] C.J. Lloyd, A new exact and more powerful unconditional test of no treatment effect from binary matched pairs, Biometrics. 64 (3) (2008) 716–723.
[8] C.J. Lloyd, A numerical investigation of the accuracy of parametric bootstrap for discrete data, Comput. Stat. Data Anal. 61 (2013) 1–6.
[9] C.P. Farrington, G. Manning, Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk, Stat. Med. 9 (12) (1990) 1447–1454.
[10] E. Ripamonti, C. Lloyd, Tests for noninferiority trials with binomial endpoints: a guide to modern and quasi-exact methods for biomedical researchers, Pharm. Stat. 18 (3) (2019) 377–387.
[11] J. Röhmel, U. Mansmann, Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority, Biom. J. 41 (2) (1999) 149–170.