Monte Carlo test under general conditions: Power and number of simulations

I.R.Silva*, R.M.Assunção

Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Abstract

The statistical tests application demands the probability distribution of the test statistic U under the null hypothesis. When that distribution can not be obtained analytically, it is necessary to establish alternative methods to calculate the p-value. If U can be simulated under the null hypothesis, Monte Carlo simulation is one of the ways to estimate the p-value. Under some assumptions about the probability distribution and the power function of U, the literature has obtained thin upper bounds for the power difference between exact test and Monte Carlo test. The motivation of this paper is to dispense any assumptions and to demonstrate that the Monte Carlo test and exact test have same magnitude in power, for any test statistic, even for moderated m. This demonstration is possible by exploring the trade-off between the power and the significance level.

Keywords:
MC Test, Exact Test, Power, p-value, Significance Level

1. Introduction

A current obstacle in the construction of a test statistic U is the fact that in several cases is impossible to obtain analytically its probability distribution function. Regularly, even its asymptotic distribution can not be found. An example of that situation is the Scan statistic, Kulldorff (2001), which is applied to detect spatial clusters. Even when it is possible to deduce the asymptotic distribution F_d of U the problem is not properly solved. An undesirable consequence of using F_d to obtain the p-value is that the real type one error and the power loss compared to the exact test[1] are not controlled for finite samples.

If it is possible to generate samples from the test statistic under the null hypothesis H_0, the maximum-likelihood estimator for the p-value is given by the ratio between the number of simulated statistics u_i's greater than or equal to the observed value u_0, $i = 1, 2, ..., m-1$. The null hypothesis is rejected if the estimated p-value is smaller than or equal to the desired significance level α_{mc}. That approach is known as Monte Carlo test (MC test), proposed by Dwass (1957), introduced by Barnard (1963), and extended by Hope (1968) and Birnbaum (1974). A first property which must be mentioned is that estimating the p-value in that way and rejecting H_0 if the estimative is smaller than or equal to α_{mc} induces a probability of type one error smaller than or equal to α_{mc}. An additional positive property of MC test is that the number m of simulations can be random, Besag and Clifford (1991), without power loss, Silva et al. (2009). Another intensifier aspect of MC test application, showed by Dufour (2005), is the fact that MC test is viable even when, under H_0, the test statistic involves nuisance parameters.

*Corresponding author
Email addresses: irs@ufmg.br (I.R.Silva), assuncao@est.ufmg.br (R.M.Assunção)

[1] Exact test here is based on the p-value obtained directly from the real distribution of U under the null hypothesis

Methods directed to data spatial analysis are strongly cited in the literature to justify the theoretical development of MC test. Applications of MC test on spatial analysis are seen in Ripley (1992), Kulldorff (2001), Assunção and Maia (2007) or Peng et al. (2005). Applications are also sorted in a variety of areas, as we can see in Booth and Butler (1999), Caffo and Booth (2003) or Wongravee et al. (2009).

An important researching aspect involving MC tests is its power compared to that from the exact test. Hope (1968) has considered MC tests on situations when the likelihood ratio $f_1(u)/f_0(u)$ is a monotone increasing function of a fixed test statistic value u, with $f_1(u)$ and $f_0(u)$ the densities of the test statistic under the alternative and null hypothesys, respectively . Using $\alpha_{mc} = j/m$ and $j \in N$, he has showed that the uniformly most powerfull test (UMP) based on U exists and, consequently, the power of MC test converges, with m, to the power of the UMP. With a less restrictive condition of concavity for the exact test power function, Jockel (1986) has proved that the MC test power converges uniformly to the one of the exact test, and establishes upper bounds to the ratio between them for finite m. Marriott (1979) has used a normal distribution to approximate the power of MC test, what made possible to deduce that its convergence to the exact test power is fast. Fay and Follmann (2002) treated the probability of the sequential Monte Carlo test to conduce to different decisions about rejecting or not H_0 faced to the decisions by using the exact test, called resampling risk (RR). They proposed a sophisticated sequential implementation of MC tests and bound the RR by considering a certain class of distributions to the random variable p-value. Sequential implementation of MC tests are important for saving time when test statistic is computationally intense. Other efficient designs to implement MC tests sequentially has been proposed, as well the study of the expected number of simulations under the alternative hypothesis and the control of the resampling risk, and concerning these aspects, we can cite Fay and Follmann (2002), Fay et al. (2007), Gandy (2009) and Kim (2010). However, our focus here is to treat MC tests with fixed m, which is the simpler design when the test statistic is computationally light. We believe that extensions of our results to the sequential implementation is a natural path for future explorations.

We shall show here an approach for establishing upper bounds to the power difference between the exact test and MC test alternative to those described previouslly. Our proposal produces an upper bound sufficiently slim, such that, in practical terms, the possibility of losing power by using MC test should be unconsidered. This result was possible by considering α_{mc} slightly greater than the significance level α of the exact test. Assumptions involving the distribution of the test statistic, the exact power function or over the likelihood in U, are not required here.

Choosing a significance level greater than the initialy planned is not something enjoyable, neither it is a new idea use this artifice for obtaining a better power. But, if we consider that the required differences between α_{mc} and α, sufficient to guarantee slim upper bounds, are negligible even with short m, and such differences are controlled by the user, we conclude that this proposal offers an appeal to treat satisfactorily an inconvenient arbitration demanded in MC tests, that is the choice of m combined to the necessity of having a power equal to that from the respective exact test.

When the control of type one and type two errors is imperative, a possible interpretation of the results here is that MC test is as useful as the own exact test. Therefore, under that aspect, the asymptotic approximation of the distribution of U could be replaced by Monte Carlo procedure, once the test based in asymptotic distributions has no analitical control of the probability of the type one error and upper bounds for the power loss comparatively to the exact test for each sample size must be studied, analitically or empirically, for each specific application.

It is easy to see that, for having a probability of type one error equal to α_{mc} in MC test, it is necessary to choose m as a multiple of $1/\alpha_{mc}$, for rational α_{mc}. However, non-compliance with the rule $m = j/\alpha_{mc}$, j integer, results in power reduction of MC test as m increases. Let $\pi(m, \alpha_{mc}, F_P)$ be the MC test power for fixed significance level α_{mc} and F_P be the probability distribution of the p-value. The existence of F_P depends solely on the existence of the distributions of U under H_0 and the alternative hypothesis H_1, what

can be well understood in Sackrowitz and Samuel-Cahn (1999). It is intuitive that, as larger is m, larger is the power. But that is not true at the generic case. Silva et al. (2009) offers an example about how MC test power can decrease with m, even when $m = j/\alpha_{mc}$. We shall show here that $\pi(m, \alpha_{mc}, F_P)$ has potential for increasing with m only if $m = j/\alpha_{mc}$. Hope (1968) showed that, when the likelihood ratio is a monotone increasing function of u, for $m = j/\alpha_{mc}$, $\pi(m, \alpha_{mc}, F_P)$ is a monotone increasing function of m. Concerning this topic, our focus is to show what happens with the MC power if m is not a multiple of $1/\alpha$ for any likelihood ratio shape, and with a very simple reasoning, it can be showed that $\pi(m, \alpha_{mc}, F_P)$ is always non-increasing for m on the range $[\lfloor j/\alpha_{mc} \rfloor, \lfloor (j+1)/\alpha_{mc} \rfloor)$.

The next section develops a proposal to obtain upper bounds for the power difference between the exact test and MC test. In practice, those upper bounds allow the construction of MC tests that present same power that the exact tests. Section 3 shows that m must be a multiple of $1/\alpha_{mc}$ and Section 4 finishes this paper with a discussion.

2. Upper Bound for the Power Difference Between Exact Test and Monte Carlo Test

The calculation of the p-value depends on how the alternative hypothesis is defined. Without loss of generality, this paper works only the cases where the alternative hypothesis is formulated so that large values of U lead to the rejection of H_0.

As adopted by Silva et al. (2009), for a given test statistic U, with observed value u_0, let consider the event $[U_i \geq u_0]$ as a success, where $U_1, U_2, ..., U_{m-1}$ are independent copies from U under the null hypothesis. Let F be the probability distribution function of U_i under the null hypothesis. Then, the probability $\mathbb{P}(U_i \geq u_0) = 1 - F(u_0)$ is the p-value.

Let Y be the number of successes for a fixed m. The random variable Y has a binomial distribution with $m-1$ essays and success probability equal to the observed p-value. We shall use P for denoting the p-value as a random variable. Thus, given $P = p$, the MC procedure leads to the rejection of H_0 with probability

$$\mathbb{P}(Y \leq m\alpha_{mc} - 1 \mid P = p) = \pi(m, \alpha_{mc}, p) = \sum_{x=0}^{\lfloor m\alpha_{mc} \rfloor - 1} \binom{m-1}{x} p^x (1-p)^{m-1-x}. \tag{1}$$

The power of MC test is obtained by integrating, in continuous case, or by summing, in discrete case, (1) with respect to the distribution F_P. Without loss of generality, we shall work in this paper only with the continuous case. The discret case can be treated by applying the randomized test. Therefore, the power of MC test is:

$$\pi(m, \alpha_{mc}, F_P) = \int_0^1 \sum_{x=0}^{\lfloor m\alpha_{mc} \rfloor - 1} \binom{m-1}{x} p^x (1-p)^{m-1-x} F_P(dp). \tag{2}$$

Let α be the significance level of the exact test. The probability of rejecting H_0 by using the exact test is 1, if $p \leq \alpha$, and it is 0, otherwise. The power of the exact test can be expressed as follows:

$$\pi(\alpha, F_P) = \int_0^\alpha F_P(dp). \tag{3}$$

It is easy to prove the convergence of the MC test power to the exact power. Take $\alpha_{mc} = \alpha$. When $m \to \infty$, $Y/m \overset{ae}{\to} p$, then, for $p < \alpha$, $\mathbb{P}(Y \leq m\alpha - 1 \mid P = p) \overset{ae}{\to} 1$, and if $p \geq \alpha$, $\mathbb{P}(Y \leq m\alpha - 1 \mid P = p) \overset{ae}{\to} 0$.[2] Thus, according to the dominated convergence theorem, $\lim_{m\to\infty} \pi(m, \alpha, F_P) = \pi(\alpha, F_P)$. However, under the pratical point of view, it is more important to understand the relation between the exact and MC powers

[2]The abbreviation ae means almost everywhere convergence

for finite m.

Bounding the power loss of MC test, compared to the exact test, demands a cost to pay, and in the literature the usual payment is to restrict the p-value behaviour. That is a very expensive price, because, in practice, and under the main context where MC tests are required, assumptions about the unknown p-value density behaviour are rarely feasible to verify. We propose here a more practical exchange currency. We suggest to change the restrictions over the p-value density by using significance levels for the MC test slightly larger than the exact test ones.

From expressions (1) and (2), the trade-off between the significance level and the power can be manipulated. With that, we propose to sacrifice the significance level α_{mc} in an irrelevant way comparatively to α, $\alpha_{mc} = \alpha + \delta$, $\delta > 0$ and, at the same time, producing satisfatory upper bounds for the power loss of MC procedure.

Formally, the comparation in power between two or more tests presupposes identity between the associated significance levels. But, we perform here the comparation in power by using different significance levels α_{mc} and α. Nevertheless, the abuse of terminology is justified, because sufficiently thin upper bounds for the power loss of MC test are obtained by using despicable magnitudes for δ under the practical aspect.

The power difference between the exact test and the MC test is:

$$E[D(P)] = \int_0^1 D(p) F_P(dp) \tag{4}$$

with $D(p)$ given by

$$D(p) = 1_{(0,\alpha]}(p) - \sum_{x=0}^{\lfloor m\alpha_{mc} \rfloor - 1} \left(\begin{array}{c} m-1 \\ x \end{array} \right) p^x (1-p)^{m-x-1} \tag{5}$$

where $1_{(0,\alpha]}(p)$ is the step function of $p \in (0, \alpha]$. As $\pi(m, \alpha_{mc}, p)$ decreases monotonously with p, the maximum value for $D(p)$ is:

$$b(m, \alpha, \alpha_{mc}) = \max_p \{D(p)\} = 1 - \sum_{x=0}^{\lfloor m\alpha_{mc} \rfloor - 1} \left(\begin{array}{c} m-1 \\ x \end{array} \right) \alpha^x (1-\alpha)^{m-x-1} \tag{6}$$

that is the maximum distance between the step function $1_{(0,\alpha]}(p)$ and the power function of the MC test $\pi(m, \alpha_{mc}, p)$.

Table 1 offers upper bounds $b(m, \alpha, \alpha_{mc})$ for some values of m, α and α_{mc} by applying directly Equation (6), and it highlights that just an unexpressive increase of α_{mc} is sufficient to lead MC test to a power pratically equal to that from exact test. It is important to note that these upper bounds are thin even with small m values, for example, observe $\alpha_{mc} = 0.015$ and $\alpha = 0.01$ or $\alpha_{mc} = 0.06$ and $\alpha = 0.05$, with $m = 3000$, where the upper bounds are smaller than 0.006 and 0.008, respectively.

The expression (6) can be used to find a value for m that, except by a known constant $\epsilon > 0$ chosen by the user, it induces the same power for MC test and the exact test, pointing out that the probability of type one error $\alpha_{mc} = \alpha + \delta(\epsilon)$, with desirable $\delta(\epsilon)$. With that, is possible to calibrate the upper bounds for the

m	$\alpha = 0.01$ $\alpha_{mc} = 0.013$	$\alpha_{mc} = 0.015$	$\alpha = 0.05$ $\alpha_{mc} = 0.055$	$\alpha_{mc} = 0.060$
1000	0.2065387	0.0818942	0.2505276	0.0855520
2000	0.1106199	0.0211332	0.1633004	0.0247740
3000	0.0636350	0.0059943	0.1124737	0.0078300
4000	0.0378388	0.0017746	0.0795444	0.0025768
5000	0.0229464	0.0005386	0.0571672	0.0008684
6000	0.0140999	0.0001661	0.0415376	0.0002973
7000	0.0087467	0.0000518	0.0304238	0.0001029
8000	0.0054652	0.0000163	0.0224209	0.0000359
9000	0.0034342	0.0000052	0.0166038	0.0000126
10000	0.0021679	0.0000016	0.0123450	0.0000045

Table 1: Upper bounds for the power differences between the exact test, with significance levels α= 0.01, 0.05, and MC test, with significance levels α_{mc}.

power loss according to the convenience of the user, who has the prior expertise to define subjectively what is a despicable value for $\delta(\epsilon)$. That fact confirms the assertion from Jockel (1984), which is, MC test has potential to compete against the tests based on asymptotic theory, mainly for small sample sizes.

As m increases, smaller is the $\delta(\epsilon)$ sufficient for ensuring an acceptable value for $b(m, \alpha, \alpha_{mc}) = \epsilon$. To see that, consider a random variable $X \sim Bin(n, p)$. Okamoto (1958) has established that, for $p < 1/2$, $\mathbb{P}(\sqrt{X/n} - \sqrt{p} \geq c) < \exp(-2nc)$, where c is a constant. Manipulating the last inequality, we have the following:

$$\mathbb{P}(Y \leq \alpha_{mc}m - 1 \mid P = p) \geq 1 - \exp\left\{-2(m-1)\left[\left(\frac{\alpha_{mc}m}{m-1}\right)^{1/2} - \sqrt{p}\right]^2\right\}. \tag{7}$$

Fixing $b(m, \alpha, \alpha_{mc}) = \epsilon$, from 6 and 7,

$$\epsilon \leq \exp\left\{-2\left[\sqrt{\alpha_{mc}m} - \sqrt{\alpha(m-1)}\right]^2\right\}. \tag{8}$$

The right hand side of the inequality (8) is decreasing with m for $m > (1 - \alpha/\alpha_{mc})^{-1}$, then, as m soars, smaller is $\delta(\epsilon)$ to keep an arbitrary ϵ. However, we must emphasize the importance of increasing m by multiples of $1/\alpha_{mc}$, which is the subject treated in the next section.

3. m as a Multiple of $1/\alpha_{mc}$

Hope (1968), by considering a monotone increasing function for the likelihood ratio in U and $m = j/\alpha_{mc}$, with $j \in N$, has studied the power behavior of MC test when m soars. Under such conditions, he has proved both, the existence of the uniformily more powerful test (UMP) and the convergence of MC test power to the UMP one. We offer here an additional argument for the using of m in the form j/α_{mc} that is the fact of $\pi(m, \alpha_{mc}, F_P)$ decreasing with m for $\lfloor (j-1)/\alpha_{mc} \rfloor < m < \lfloor j/\alpha_{mc} \rfloor$, $j > 1$.

We reject H_0 if, among the $(m-1)$ simulated U_is, the number of values greater than or egual to u_0 is not greater than $\alpha_{mc}m - 1$. Obviously, that requires $m \geq 1/\alpha_{mc}$, because, otherwise, H_0 is never rejected.

Consider two MC tests which differ from the number of simulations, m and $m + k$, $k > 0$. For the first MC test, which is based on m, we reject H_0 only if the number of simulated values U_is exceeding u_0 is at most $\lfloor \alpha_{mc}m \rfloor - 1$, whereas, for the second test, with $(m + k)$ simulations, H_0 is rejected if such number is

at most $\lfloor \alpha_{mc}(m+k) \rfloor - 1$. According to (1), for any observed $P = p$, the power of the second test is greater than that from the first test if

$$\sum_{y=0}^{\lfloor \alpha_{mc}(m+k) \rfloor - 1} \left(\begin{array}{c} m+k-1 \\ y \end{array} \right) p^y (1-p)^{m-y-1} > \sum_{y=0}^{\lfloor \alpha_{mc} m \rfloor - 1} \left(\begin{array}{c} m-1 \\ y \end{array} \right) p^y (1-p)^{m-y-1} \tag{9}$$

Observe that this inequality is equivalent to

$$\mathbb{P}(X \leq \lfloor \alpha_{mc}(m+k) \rfloor - 1) > \mathbb{P}(Y \leq \lfloor \alpha_{mc} m \rfloor - 1)$$

where $X \sim \text{Bin}(m+k-1, p)$ and $Y \sim \text{Bin}(m-1, p)$. If $0 < k < 1/\alpha$, then

$$\mathbb{P}(X \leq \lfloor \alpha_{mc}(m+k) \rfloor - 1) = \mathbb{P}(X \leq \lfloor \alpha_{mc} m \rfloor - 1)$$

and the inequality (9) becomes

$$\mathbb{P}(X \leq \lfloor \alpha_{mc} m \rfloor - 1) > \mathbb{P}(Y \leq \lfloor \alpha_{mc} m \rfloor - 1)\,,$$

which is not valid. Thus, if $0 < k < 1/\alpha$, $\pi(m+k, \alpha_{mc}, F_P) \leq \pi(m, \alpha_{mc}, F_P)$. For example, for $\alpha_{mc} = 0.01$, a MC test with $m_1 = 1050$ is less powerful than a test with $m_2 = 1000$.

4. Discussion

A considerable parcel of the researchers, who are devoted to the development of statistical tests, are restricted to study the limit distribution of the test statistic to guide the criterium decision about acceptance/rejection the null hypothesis. Such works should consider Monte Carlo test application with the same entusiasm that is currently devoted to asymptotic study. What support this assertion is that Monte Carlo approach preserves power comparatively to the exact test and controls the real probability of the type one error, what is not always true in the asymptotic treatment, particularly for moderated sample sizes.

References

Assunção, R., Maia, A., 2007. A note on testing separability in spatial-temporal marked point processes. Biometrics 63 (1), 290–294.

Barnard, G., 1963. Discussion of professor bartlett´s paper. J. R. Statist. Soc. 25B (294).

Besag, J., Clifford, P., 1991. Sequential monte carlo p-value. Biometrika 78, 301–304.

Birnbaum, Z., 1974. Computers and unconventional test-statistics. in: F. proschan and r.j. serfling. Reliability and Biometry, 441–458.

Booth, J., Butler, R., 1999. An importance sampling algorithm for exact conditional tests in log-linear models. Biometrika 86, 321–332.

Caffo, B., Booth, J., 2003. Monte carlo conditional inference for log-linear and logistic models: a survey of current methodology. Statistical Methods in Medical Research 12, 109–123.

Dufour, J., 2005. Monte carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. Jornal of Econometrics 133 (2), 443–477.

Dwass, M., 1957. Modified randomization tests for nonparametric hypotheses. Annals of Mathematical Statistics 28, 181–187.

Fay, M., Follmann, D., 2002. Designing monte carlo implementations of permutation or bootstrap hypothesis tests. The American Statistician 56 (1), 63–70.

Fay, M., Kim, H.-J., Hachey, M., 2007. On using truncated sequential probability ratio test boundaries for monte carlo implementation of hypothesis tests. Journal of Computational and Graphical Statistics 16, 946–967.

Gandy, A., 2009. Sequential implementation of monte carlo tests with uniformly bounded resampling risk. Journal of the American Statistical Association 104 (488), 1504–1511.

Hope, A., 1968. A simplified monte carlo significance test procedure. Journal of the Royal Statistical Society 30B, 582–598.

Jockel, K., 1984. Application of monte-carlo tests - some considerations. Biometrics 40 (1), 263–263.

Jockel, K., 1986. Finite sample properties and asymptotic efficiency of monte carlo tests. The Annals of Statistics 14, 336–347.

Kim, H.-J., 2010. Bounding the resampling risk for sequential monte carlo implementation of hypothesis tests. Journal of Statistical Planning and Inference (140), 1834–1843.

Kulldorff, M., 2001. Prospective time periodic geographical disease surveillance using a scan statistic. Journal of Royal Statistical Society 164A, 61–72.

Marriott, F., 1979. Bernard's monte carlo test: How many simulations? Applied Statististics 28, 75–77.

Okamoto, M., 1958. Some inequalities relating to the partial sum of binomial probabilities. Annals of the Institute of Statistical Mathematics 10, 29–35.

Peng, R., Schoenberg, F., Woods, J., 2005. A space-time conditional intensity model for evaluating a wildfire hazard index. Journal of the American Statistical Association 100 (469), 26–35.

Ripley, B., 1992. Applications of monte-carlo methods in spatial and image-analysis. Lecture Notes in Economics and Mathematical Systems 376, 47–53.

Sackrowitz, H., Samuel-Cahn, E., 1999. P values as random variables - expeted p values. The American Statistician 53, 326–331.

Silva, I., Assunção, R., Costa, M., 2009. Power of the sequential monte carlo test. Sequential Analysis 28 (2), 163–174.

Wongravee, K., Lloyd, G., Hall, J., Holmboe, M., Schaefer, M., Reed, R., Trevejo, J., Brereton, R., 2009. Monte-carlo methods for determining optimal number of significant variables. application to mouse urinary profiles. Metabolomics 5 (4), 387–406.

Acknowledgements

We are grateful to Martin Kulldorff for very useful comments and suggestions on an earlier draft of this paper. This research was partially funded by the National Cancer Institute, grant number RO1CA095979, Martin Kulldorff PI. The second author was partially supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). This research was partially carried out while the first author was at the Department of Ambulatory Care and Prevention, Harvard Medical School, whose support is gratefully acknowledged.