Letters to the Editor

Am. J. Hum. Genet. 72:496, 2003

Simulation-Based *P* Values: Response to North et al.

To the Editor:

North et al. (2002) discussed the estimation of a *P* value on the basis of computer (i.e., Monte Carlo) simulations. They emphasized that such a *P* value is an estimate of the true *P* value. This is essentially their only point with which we agree. The letter from North et al. is more likely to confuse than enlighten.

Consider an observed test statistic, x, that under the null hypothesis follows some distribution, f. Let X be a random variable following the distribution f. We seek to estimate the *P* value, $p = \Pr(X \geq x)$. Let $y_1, \ldots, y_n$ be independent draws from f, obtained by computer simulation. Let $r = \#\{i: y_i \geq x\}$ (i.e., the number of simulated statistics greater than or equal to the observed statistic). Let $\hat{p} = r/n$ and $\tilde{p} = (r + 1)/(n + 1)$.

North et al. (2002) stated that $\hat{p}$ is "not strictly correct" and that $\tilde{p}$ is "the most accurate estimate of the *P* value." They further called $\tilde{p}$ "the true *P* value."

We strongly disagree with this characterization. First, minor differences in *P*-value estimates on the order of Monte Carlo error should not be treated differently in practice, and so it is immaterial whether one uses $\hat{p}$ or $\tilde{p}$. Second, $\hat{p}$ is a perfectly reasonable estimate of p. Indeed, in many ways $\hat{p}$ is superior to $\tilde{p}$. Given the observed test statistic, x, r follows a binomial (n,p) distribution, and so $\hat{p}$ is unbiased, whereas $\tilde{p}$ is biased. (The bias of $\tilde{p}$ is $(1 - p)/(n + 1)$.) Further, $\hat{p}$ has smaller mean square error (MSE) than $\tilde{p}$, provided that $p < n/(1 + 3n) \approx 1/3$. (The MSE of $\hat{p}$ is $p(1 - p)/n$, whereas that of $\tilde{p}$ is $(1 - p)(np + 1 - p)/(n + 1)^2$.)

These results are contrary to those of North et al. (2002) because they evaluate the performance of $\tilde{p}$ under the joint distribution of both the observed and Monte Carlo data, whereas we prefer to condition on the observed value of the test statistic. Evaluating *P*-value estimates conditionally on the observed data is widely accepted when the estimation is performed via analytic approximations.

Regarding the question of how many simulation replicates to perform, we recommend consideration of the precision of the estimate, $\hat{p}$, using the properties of the binomial distribution, rather than adherence to a rule such as $r \geq 10$. Standard statistical packages, such as R (Ihaka and Gentleman 1996), allow one to calculate a CI for the true *P* value and to perform a statistical test, such as whether the true *P* value is <.01.

KARL W. BROMAN AND BRIAN S. CAFFO
Department of Biostatistics
Johns Hopkins University
Baltimore

References

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comp Graph Stat 5:299–314

North BV, Curtis D, Sham PC (2002) A note on the calculation of empirical *P* values from Monte Carlo procedures. Am J Hum Genet 71:439–441

Address for correspondence and reprints: Dr. Karl W. Broman, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205. E-mail: kbroman@jhsph.edu

0002-9297/2003/7202-0032$15.00

Am. J. Hum. Genet. 72:496–498, 2003

On Estimating *P* Values by Monte Carlo Methods

To the Editor:

North et al. (2002) propose a new formula for the empirical estimation of *P* values by Monte Carlo methods to replace a standard conventional estimator. They claim that their new formula is "correct" and "most accurate" and that the conventional formula is "not strictly correct," repeating this claim many times in their letter. The claim, however, is incorrect, and the conventional formula is the correct one.

The North et al. claim arises when a test statistic (called here "t") takes a certain numerical value (called here "t^*") when calculated from data from some experiment, and it is required to find an unbiased estimate of the *P* value corresponding to t^* by Monte Carlo simulation. This is done by performing n Monte Carlo simulations, all performed under the null hypothesis tested

in the original experiment and with the same sample size and other characteristics as for the original experiment. Suppose, to be concrete, that sufficiently large positive values of the test statistic t are significant. Then, we define "r" as the number of simulations in which the simulation value of t is greater than or equal to the observed value t^*. North et al. claim that an unbiased, and thus preferred, estimate of the P value arising from these simulations is $(r + 1)/(n + 1)$ instead of the conventional estimate r/n. This claim is incorrect.

Strangely, North et al. (2002) themselves show by algebra that the mean value of their estimator $(r + 1)/(n + 1)$ is $(nP + 1)/(n + 1)$, where "P" is the P value to be estimated. Since this is not equal to P, their P value estimator is biased. Further, their calculation also shows that the mean value of the conventional estimator r/n, whose use they do not recommend, is the desired value P. Thus, the conventional estimator is unbiased. Thus, there is an internal inconsistency in their argument, and their algebraic calculations contradict their claim and the argument leading to it. The algebraic calculations are correct. It is important to see why the argument given in North et al. (2002) is incorrect, since the reasoning involved relates to the theory and practice of Monte Carlo simulation procedures that are performed increasingly in genetics, in particular to questions surrounding P values and type 1 errors.

The incorrect argument given by North et al. (2002) is that if the original data were generated under the null hypothesis tested, then, in all, $n + 1$ "experiments" were conducted, of which one is real and n simulation. With r as defined above, in $r + 1$ of these, the value of the statistic t is either equal to the observed value t^* or is greater than this value. It is then claimed that the estimator $(r + 1)/(n + 1)$ is an unbiased estimator of the null hypothesis probability that the test statistic t exceeds t^* when the null hypothesis is true.

The error in this argument is, perhaps, best demonstrated by considering parallel reasoning used in the genetic ascertainment sampling context, exemplified as follows. Suppose that we wish to estimate the proportion of girls in a population, using a sample of families from that population. However, the sampling procedure is such that only families in which the oldest child is a girl are included in the sample. Clearly, using all children in the sample to estimate the proportion of girls in the population is incorrect, and the sample proportion of girls will overestimate the population proportion. The oldest child in each family, automatically included in the category of interest (girls), must be excluded in the estimation process. The analogy with the Monte Carlo case is that the observed value of the test statistic found from the actual data must be excluded in estimating a P value, since it is similarly automatically included in the category of interest (greater than or equal to itself). Any mathematical calculation concerning P values that does take this into account will be incorrect.

It now appears that North et al. (2002) used mistaken terminology, and that the claim that they wished to make does not concern P value estimation, but that use of $(r + 1)/(n + 1)$ "provides the correct type 1 error rate." More precisely, if the type 1 error is chosen to be α, then it is claimed that rejecting the null hypothesis when $(r + 1)/(n + 1) < \alpha$ leads to the desired type 1 error of 5%.

To see this in formal statistical terms, the null hypothesis is rejected, with the notation and assumptions given above, if the value of r is "too low." More specifically, with the chosen type 1 error of α, the null hypothesis is rejected if $r < K$, where K is chosen so that Prob($r < K$, given null hypothesis is true) $= \alpha$.

The one "experimental" and n simulation values of t, leading to a total of $n + 1$ values, can be listed in ascending order. The event that $r < K$ is identical to the event that the experimental value of t lies among the highest $K + 1$ of these $n + 1$ values. The null hypothesis probability of this is $(K + 1)/(n + 1)$. Equating the probability$(K + 1)/(n + 1)$ with α, we get $K = (n + 1)\alpha - 1$ $K = (n + 1)\alpha - 1$. The event $r < K$ is, thus, the same as the event $(r + 1)/(n + 1) < \alpha$, and this is the criterion that North et al. give.

This procedure does not, however, imply, as claimed by North et al. (2002), that $(r + 1)/(n + 1)$ is an unbiased estimate of the P value. It is best to keep the questions of unbiased estimation of the P value and the nature of the testing procedure that leads to a desired type 1 error separate. Pursuing this point, it is not clear in what sense North et al. relate, as they do, a P value estimate to a type 1 error. They claim, for example, that when $r = 0$, so that the standard procedure P value estimate r/n is also 0, it is implied, under the standard procedure, that the type 1 error is also 0. This claim is incorrect. A type 1 error in statistics is set in advance, typically 5% or 1%, and the value so chosen for it is not in any way determined by or estimated from the observed value of any statistic.

Warren J. Ewens

Department of Biology
University of Pennsylvania
Philadelphia

References

North BV, Curtis D, Sham PC (2002) A note on the calculation of empirical P values from Monte Carlo procedures. Am J Hum Genet 71:439–441

Address for correspondence and reprints: Dr. Warren J. Ewens, Department of Biology, University of Pennsylvania, Philadelphia, PA 19104-6018. E-mail: wewens@sas.upenn.edu

0002-9297/2003/7202-0033$15.00

Am. J. Hum. Genet. 72:498–499, 2003

A Note on the Calculation of Empirical *P* Values from Monte Carlo Procedures

To the Editor:

We welcome the opportunity to correct our mistaken terminology in referring to $(r + 1)/(n + 1)$ as an unbiased estimate of P, where a Monte Carlo procedure has been carried out with n simulations, of which r exceed the observed statistic obtained from the real data set. As we ourselves pointed out (North et al. 2002), this estimate is indeed slightly biased. What we intended to write was that using this estimate is valid in the sense that it produces the correct type 1 error rate. According to Cox and Hinkley (1974), the observed P value of a study, denoted as P_{obs}, is defined as $\Pr(T \geq t_{obs}; H_0)$, the probability that the test statistic T is greater than or equal to its actual value t_{obs} from the observed data, if the null hypothesis, H_0, is true. Their interpretation of the P value is that it is "the probability that we would mistakenly declare there to be evidence against H_0, were we to regard the data under analysis as just decisive against H_0." Since $P \leq P_{obs}$ if and only if $T \geq t_{obs}$, it follows that $\Pr(T \geq t_{obs}; H_0) = \Pr(P \leq P_{obs}; H_0) = P_{obs}$. In other words, we should obtain a P value of .05 (or lower) with frequency 0.05, and a P value of .01 (or lower) with frequency 0.01, and so on, if the null hypothesis is true. If a test procedure produces P values of .05 (or lower) with greater frequency than 0.05, when the null hypothesis is true, then the procedure is anticonservative.

Our article (North et al. 2002) was motivated by the recognition that the common practice of using r/n as the P value from a Monte Carlo procedure is, in fact, anticonservative, whereas the use of $(r + 1)/(n + 1)$ provides the correct type 1 error rate. There is nothing novel about the use of $(r + 1)/(n + 1)$—it is published in a standard textbook on Monte Carlo methods (Davison and Hinkley 1997), and we merely sought to give it greater prominence and to investigate its implications. We accept that it is mildly counterintuitive, and so some people may find the reasons for its usage difficult to grasp. Nevertheless, we remain convinced that it is far preferable to use an estimate that is slightly biased but yields the correct type 1 error rate than one that is unbiased but is demonstrably anticonservative.

One way to understand the justification for using $(r + 1)/(n + 1)$ rather than r/n is as follows. When the null hypothesis is true, the actual value of the test statistic and the n replicate values based on simulations constitute $n + 1$ independent realizations of the same random variable. All possible ranks of the actual test statistic among these $n + 1$ values, from rank 1 to rank $n + 1$ in descending order of magnitude, are, therefore, equally probable. The probability of the actual test statistic being exceeded in exactly r of n simulated replicates (i.e., of being ranked $r + 1$) is, therefore, $1/(n + 1)$. Likewise, the probability of the actual test statistic being exceeded in r or fewer of n simulated replicates (i.e., of being ranked $r + 1$ or higher) is $(r + 1)/(n + 1)$.

For those who are not convinced by the above argument, we present a more mathematical derivation. The probability that the actual test statistic is exceeded in exactly r simulations, conditional on any particular value of P, is given by the binomial distribution with parameters n and P. The unconditional probability that the actual test statistic is exceeded in exactly r simulations is obtained by integrating the product of this conditional probability and the density function $f(P)$ of P, over the possible range of P. Therefore,

$$\begin{aligned}\Pr(r; \mathrm{H}_0) &= \int_0^1 \frac{n!}{(n-r)!r!} p^r (1-p)^{n-r} \mathrm{f}(p)\mathrm{d}p \\ &= \frac{n!}{(n-r)!r!} \int_0^1 p^r (1-p)^{n-r}\,\mathrm{d}p \\ &= \frac{n!}{(n-r)!r!} \frac{(n-r)!r!}{(n+1)!} \\ &= \frac{1}{n+1}\end{aligned}$$

for $r = 0,1,\ldots,n$. The second step in the derivation depends on the density function of P being uniform in [0,1] under the null hypothesis, whereas the third step is due to the recognition that the integral is a beta function with parameters $n - r + 1$ and $r + 1$. From the fact that the probability of achieving any particular value of r is $1/(n + 1)$, it follows that the probability of the actual test statistic being exceeded in r or fewer of n simulated replicates (i.e., of being ranked $r + 1$ or higher) is $(r + 1)/(n + 1)$.

For anyone who continues to remain skeptical in spite of these theoretical arguments, it is trivial to carry out simulation procedures that demonstrate that using r/n is anticonservative, whereas using $(r + 1)/(n + 1)$ does indeed yield the correct type 1 error rate. Anybody who