Letter to the Editor

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 takes the trouble to do this cannot fail to discover this for himself. For example, here is a simple C program that demonstrates the phenomenon:

```
#include <stdio.h>
#include <stdlib.h>
float p1,p2,m1,m2,r,alpha=0.01;
int x,j,n=500;
long i,nsim=1000000;
int main(int argc, char *argv[])
{
for (i=0;i<nsim;++i)
{
x=rand();
for (r=0,j=0;j<n;++j)
   if (rand()>=x) ++r;
if (r/n<=alpha) ++m1;
if ((r+1)/(n+1)<=alpha) ++m2;
}
printf("Using r/n, Type 1 error =
   %f\n",m1/nsim);
printf("Using (r+1)/(n+1), Type 1 error =
   %f\n",m2/nsim);
}
```

As the theory predicts, when the number of simulations is 500, using r/n and $(r + 1)/(n + 1)$ provide an empirical P value of .01 (or lower) with frequency 0.012 and 0.010, respectively. One can readily use a range of different values to see that the argument holds in all situations.

Although lack of bias is desirable, it is not so crucial a property as that of providing the correct type 1 error interpretation. The estimator r/n is unbiased but anticonservative, and its usage can lead, for example, to the absurd assertion that when $r = 0$, then the type 1 error rate is 0, implying that the results are impossible under the null hypothesis and, therefore, must be rejected. Because r/n and $(r + 1)/(n + 1)$ are both linear functions of r, they are perfectly correlated with each other. Using $(r + 1)/(n + 1)$ introduces only a small bias, being $(1 - p)/(n + 1)$, which diminishes with increasing n. Proponents of using r/n might argue that it should be regarded merely as an estimate of the true P value, and not as an empirical P value. In our view, this is unnecessarily cumbersome, since $(r + 1)/(n + 1)$ *can* be interpreted directly as an empirical P value, which will have the correct type 1 error rate.

B. V. North,[1] D. Curtis,[1] and P. C. Sham[2]
[1]*Academic Department of Psychiatry, St. Bartholomew's and Royal London School of Medicine and Dentistry, and* [2]*Department of Psychological Medicine, Institute of Psychiatry, London*

References

Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, London

Davison AC, Hinkley DV (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge

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. D. Curtis, Department of Adult Psychiatry, Royal London Hospital, London E1 1BB, United Kingdom. E-mail: dcurtis@hgmp.mrc.ac.uk

0002-9297/2003/7202-0034$15.00