A Simplified Monte Carlo Significance Test Procedure

By ADERY C. A. HOPE

Petroleum Recovery Research Institute, Alberta, Canada

[Received January 1968. Revised May 1968]

SUMMARY

The use of Monte Carlo test procedures for significance testing, with smaller reference sets than are now generally used, is advocated. It is shown that, for given $\alpha = 1/n$, n a positive integer, the power of the Monte Carlo test procedure is a monotone increasing function of the size of the reference set, the limit of which is the power of the corresponding uniformly most powerful test. The power functions and efficiency of the Monte Carlo test to the uniformly most powerful test are discussed in detail for the case where the test criterion is $N(\gamma, 1)$. The cases when the test criterion is Student's t-statistic and when the test statistic is exponentially distributed are considered also.

1. INTRODUCTION

MONTE CARLO significance test procedures consist of the comparison of the observed data with random samples generated in accordance with the hypothesis being tested. A test criterion is chosen to facilitate this comparison. The outcome of the test is determined by the rank of the test criterion of the observed data relative to the test criteria of the random samples forming the reference set.

The test criterion may be the sample distribution or a statistic derived from it which is appropriate for testing the hypothesis under consideration. It is not necessary to define the test criterion precisely before generating the reference set but a vague definition of relevant test criteria should be given. (For example, to test the hypothesis that the process is Poisson against the alternative that the model is one of clustering by the simplified Monte Carlo test procedure for the data in Section 2.2, relevant test criteria are those that measure clustering.) It may be difficult, initially, to ascertain which test criterion is most suitable; but, having obtained the reference set, several test criteria based on different, but appropriate, aspects of the distribution could be considered. There would be little doubt as to the conclusion if the results of the Monte Carlo tests were the same for each of these test criteria.

It is preferable to use a known test of good efficiency instead of a Monte Carlo test procedure assuming that the alternative statistical hypothesis can be completely specified. However, it is not always possible to use such a test because the necessary conditions for applying the test may not be satisfied, or the underlying distribution may be unknown or it may be difficult to decide on an appropriate test criterion. Also, it is possible that only a physical model can be obtained which cannot be expressed in mathematical terms. Further, it may be that only vague alternative hypotheses exist and hence, only a vague definition of the test criteria can be given. In such cases the use of Monte Carlo procedures of significance testing makes it possible tɔ obtain some information about a situation that would not be attainable or easily attainable otherwise.

The use of randomization tests for testing statistical hypotheses, first introduced by Fisher (1935), has become more prominent with the advent of high-speed electronic computers. However, many of the recently introduced procedures appear to be unnecessarily complicated involving the simulation of an excessive number of random samples. Monte Carlo procedures similar to those presently in use, except that the size of the reference set is smaller, will give tests that are easier to conduct. The size of the reference set depends on the significance level chosen for testing and the procedure adopted for judging significance. Such a simplified procedure for tests of significance using Monte Carlo methods was suggested by Barnard (Bartlett, 1963, Discussion).

2. A Simplified Monte Carlo Test Procedure

2.1. *The Procedure*

The size of the reference set which consists of random samples generated in accordance with the null hypothesis is smaller for the simplified Monte Carlo test procedure than for the Monte Carlo test procedures now generally used. Otherwise the procedures are similar.

For the simplified Monte Carlo test procedure, the size of the set of simulated samples—the reference set—is determined by the level of significance at which testing is to be carried out, α, and the procedure adopted for judging significance. Let $t_0(\mathbf{x})$ denote the test criterion based on the observed data. The procedure for judging significance in the case of the one-sided test consists of rejecting the hypothesis being tested if $t_0(\mathbf{x})$ is the $N\alpha$th most extreme or more extreme test criterion relative to the corresponding quantities based on the random samples of the reference set. Then the reference set consists of $(N-1)$ samples generated in accordance with the null hypothesis, where α is a positive rational in $(0,1)$ and $N\alpha$ and N are positive integers.

If the test is two-sided the procedure for judging significance is adjusted accordingly and hence, so is the size of the reference set. Discussion will be limited here to the case of the one-sided test. The arguments for the two-sided test are similar.

For the simplified Monte Carlo procedure the most extreme test criteria will either be those with the largest ranks or those with the smallest ranks, depending on the test criterion chosen and the form of the alternative hypothesis. The null hypothesis is rejected at the α significance level if the test criterion of the observed data, $t_0(\mathbf{x})$, is greater (less) than the corresponding values of $(N-N\alpha)$ or more of the $(N-1)$ simulated samples forming the reference set. The Type I error is exactly α. For fixed α, increasing $N\alpha$ and, therefore, increasing N, would, intuitively, increase the power of the Monte Carlo test procedure—this will be discussed in further detail below.

2.2. *Illustration of the Procedure*

Data quoted in a paper by Bartlett (1963, page 280), consisting of times to the nearest tenth of a second at which cars pass an observation point on a two-lane rural Swedish road, will be used to illustrate the simplified Monte Carlo test procedure. These data will be referred to as Bartlett's data. Without loss of generality, the data, comprising 129 observations over an interval of 2023·5 seconds, are transformed so that the interval between the first and last arrivals is unity. All future references to Bartlett's data will be to this transformed data unless otherwise specified.

Assuming that the model for traffic is Poisson, the distribution of car arrival times is uniformly distributed over the interval $(0, W_n)$, given the total number of cars, n, and the arrival time of the last car, W_n ($W_n = 1$ for Bartlett's data). However, on inspection, it appears that a clustering model would be more appropriate for the data being considered than the Poisson model. Hence, when employing a Monte Carlo test procedure only test criteria that measure clustering need be considered.

The reference set

To test the hypothesis that the process is Poisson using the simplified Monte Carlo test procedure, samples, forming the reference set, are generated such that each sample consists of 129 observations where (i) the gap time between successive arrivals, the inter-arrival time, is greater than zero, (ii) the arrival times of the first and last cars are 0 and 1, respectively, and (iii) the other 127 arrival times are chosen at random from a distribution uniformly distributed over the unit interval. Working at the 0·05 significance level and judging the test to be significant if the test criterion of the observed data is more extreme than the corresponding values of all members of the reference set, the reference set consists of 19 random samples generated in accordance with the hypothesis of no bunching.

Test criteria and significance testing

Several test criteria can be considered for ranking Bartlett's data relative to the members of the reference set. By the appropriate choice of test criteria different aspects of the underlying distribution can be tested. In this example, where testing is carried out at the 0·05 significance level, the null hypothesis of no bunching is rejected if the test criterion for Bartlett's data is the most extreme of the 20 values corresponding to Bartlett's data and the 19 random samples of the reference set. Most of the test criteria considered here would be expected to be greater, numerically, for Bartlett's data than for the members of the reference set if the process was not Poisson.

To carry out a Monte Carlo test of statistical significance it is only necessary to rank the sample of observed data relative to the samples forming the reference set which may possibly be accomplished by direct comparison of the distributions (that is, "by eye"). In this example, direct comparison indicates that there is more bunching, or clustering, in Bartlett's data than in the 19 random samples generated in accordance with the hypothesis of no bunching.

The χ^2 statistic was calculated for each sample assuming that the process was Poisson and hence, that the number of car arrivals per sub-interval of width 0·05 (omitting "time" 1) should be 6·4. Also, the χ^2 statistic based on the distribution of the number of observations per interval of width 0·05 (under the null hypothesis this distribution is Poisson with parameter $\lambda = 6{\cdot}4$) was obtained for each sample. In both cases the value of the test criterion, the χ^2 statistic, based on the observed data was greater than all the corresponding values based on the random samples of the reference set, leading to the rejection of the hypothesis of no bunching. Instead of these test criteria it would be equivalent to consider the test criterion to be, in each case, the probability of getting a χ^2 value equal to or greater than the one obtained assuming the hypothesis being tested to be true, P, say. Thus, by redefining the test criteria the definition of the "most extreme test criterion" changes from being the greatest value, in the case of the χ^2 statistic, to the smallest value, in the case of the P statistic.

It is possible that two distribution functions may not differ by a large amount even though the shapes of the distributions may be quite different. The Kolmogorov test applied to Bartlett's data does not indicate significance at the 0·05 level (Bartlett, 1963, Hawkes's contribution to the discussion). However, the tendency for the inter-arrival, or gap, times to be longer or shorter than under the null hypothesis would be an indication of departure from this hypothesis (Durbin, 1961). As such an indication was obtained by examining the cumulative distributions of gap times for the observed data and each of the random samples of the reference set the following test criteria based on the distributions of gap times were considered.

Consider paired samples where one sample is Bartlett's data and the other is a sample from the reference set. A two-sample Kolmogorov–Smirnov test can be carried out on the distributions of inter-arrival times for these 19 paired samples. The minimum value of the maximum difference between the relative cumulative distributions of inter-arrival times for these 19 paired samples was found to be 0·18. Then the minimum value of λ, the test statistic for the Kolmogorov–Smirnov two-sample test (for the case when the sample sizes are large), is 1·44. Assuming the hypothesis of no bunching to be true and using a one-sided Kolmogorov–Smirnov two-sample test, the probability of getting a value of 1·44 or greater is less than 0·05. Thus it is concluded that the distribution of inter-arrival times for Bartlett's data is not the same as that for each of the simulated samples.

The 13 highest ranked (that is, the longest) inter-arrival times of each of the 19 random samples are less than the correspondingly ranked values of inter-arrival times of Bartlett's data. Thus, any sensible test criterion based on these values or any subset of them will be larger for the observed data than for any member of the reference set, leading to the conclusion that the process is not Poisson.

Should a clustering model be more appropriate one would expect to find a larger number of short gap times in Bartlett's data than in the simulated samples. Except for very small values, this is the case. It is not unreasonable to exclude these values if it is assumed that the cars are travelling at normal speeds for a two-lane rural road. Also, it is possible that there will be a larger number of these very short inter-arrival times in the simulated samples because under the Poisson model it is assumed that cars have no length. Any sensible test criterion based on the cumulative distribution of inter-arrival times between 0·0008 and 0·0100 will lead to the rejection of the null hypothesis because, in this range, the cumulative distribution of inter-arrival times for Bartlett's data is always greater than that of any of the members of the reference set.

Suppose that a cluster is defined as a group of 2 or more observations such that the inter-arrival, or gap, time between successive arrivals is less than or equal to ω. The value of ω should be relatively small. If ω is large it will often be possible to split the clusters into a number of closely grouped observations, or sub-clusters, separated by relatively large gaps. Taking ω to be 0·0020 the following test criteria were considered (although it would be sufficient to consider only one of them): the total number of clusters, the mean number of observations per cluster and the mean cluster width. For each test criterion the value based on Bartlett's data is greater than all the corresponding values obtained from the simulated samples forming the reference set. Again, it is concluded that there is evidence of a departure from the hypothesis of no clustering.

At the outset of this problem it was difficult to ascertain what statistics to use in testing the hypothesis of no bunching. After generating the random samples it became apparent that these samples differed greatly from the observed set of data. Consideration of several possible test criteria, discussed above, led to the same conclusion for each criterion. There is little doubt that Bartlett's data are not from a population having a Poisson distribution. This conclusion was not in question prior to this analysis—Bartlett's data were used only to illustrate the simplified Monte Carlo test procedure.

As has been shown, the simplified Monte Carlo procedure of significance testing is easy to apply. By keeping the reference set to a small number of random samples generated in accordance with the hypothesis being tested, comparison of these samples with the observed data is not as cumbersome as it is if the size of the reference set is large.

3. Power Functions and Some of their Properties

3.1. *Assumptions*

Assume the distribution function of the test criterion under all relevant hypotheses to be continuous. The random variable $u = F_0(t)$, where $F_0(t)$ is the distribution function of the test criterion under the null hypothesis, is a monotone increasing function of t and could be used as the test criterion. Then under the null hypothesis, u is uniformly distributed over the unit interval.

If the probability density of t is $f_0(t)$ under the null hypothesis and $f_1(t)$ under the alternative hypothesis, the probability density of u under the alternative hypothesis is

$$r(u) = f_1(t)/f_0(t).$$

The likelihood ratio $r(u)$ is called the power density function. It will be assumed that $r(u)$ is a monotone function of t and, in particular, a monotone increasing function of t. Then a uniformly most powerful test based on it is known to exist (Lehmann, 1959, p. 68).

3.2. *Power Functions*

Assuming the properties of Section 3.1 to hold, the power of the uniformly most powerful test at the α significance level is

$$P = \int_{1-\alpha}^{1} r(u)\,du = \int_0^1 H(u)\,r(u)\,du, \tag{3.1}$$

where

$$H(u) = \begin{cases} 0, & \text{if } 0 \leqslant u < 1-\alpha, \\ 1, & \text{if } 1-\alpha \leqslant u \leqslant 1, \end{cases} \tag{3.2}$$

and $0 < \alpha < 1$.

The power of the Monte Carlo procedure of significance testing at the same significance level, α, is

$$P_{mc(N)} = \int_0^1 M_N(u)\,r(u)\,du, \tag{3.3}$$

where

$$M_N(u) = \sum_{i=0}^{N\alpha-1} \binom{N-1}{i} u^{N-1-i}(1-u)^i, \tag{3.4}$$

$N\alpha$ and N are positive integers, the values of which are determined by the procedure adopted for judging significance, and α is a rational number in the range $(0,1)$. An alternative form of $M_N(u)$, convenient for use in the numerical evaluation of $P_{mc(N)}$ for given distributions, is

$$M_N(u) = \sum_{i=0}^{N\alpha-1}(-1)^{N\alpha-1-i}\binom{N-1}{i}\binom{N-2-i}{N\alpha-1-i}u^{N-1-i}.$$

3.3. *Increasing the Power of the Monte Carlo Test Procedure*

First, consider the trivial case $f_1(t) = f_0(t)$ for all t. Then $r(u) = 1$ for all u. Hence

$$P = \int_0^1 H(u)\,du = \alpha$$

and

$$P_{mc(N)} = \int_0^1 M_N(u)\,du = \sum_{i=0}^{N\alpha-1}\binom{N-1}{i}B(N-i, i+1) = \alpha$$

for all positive integral values of $N\alpha$ and N, where $B(a,b)$ is the beta function with parameters a and b. Therefore, the power of the Monte Carlo test procedure for all positive integral values of $N\alpha$ and N is equal to that of the corresponding uniformly most powerful test.

Now, consider the case where $f_1(t) \neq f_0(t)$ and $r(u)$ is a monotone increasing function of u. Suppose, for a given power density function, α is fixed and $N\alpha$ is increased to $N'\alpha$ (that is, N is increased to N'), where the positive integer N' is chosen such that $N'\alpha$ is also a positive integer.

For all integral values of $N\alpha$, $M_N(0) = 0$, $M_N(1) = 1$ and $M_N(u)$ is an increasing function of u. For small values of u, $M_{N'}(u) \leqslant M_N(u)$. But, since

$$\int_0^1 M_N(u)\,du = \alpha$$

for all integral $N\alpha$, $M_{N'}(u) \geqslant M_N(u)$ for large values of u. Because $r(u)$ is a monotone increasing function of u and these properties of the weight function $M_N(u)$,

$$P_{mc(N')} \geqslant P_{mc(N)}.$$

That is, for given $r(u)$ and α, the power of the Monte Carlo test procedure is increased by increasing the value of $N\alpha$ and, hence, by increasing the size of the reference set. An analytical proof of this heuristic argument follows.

Increase $N\alpha$ to $N'\alpha$, where $N\alpha$, $N'\alpha$ and $n\alpha = (N'-N)\alpha$ are positive integers and α is fixed. That is, N is increased to $N' = N+n$, where N and n are positive integers chosen such that $N\alpha$ and $N'\alpha$ are also positive integers. Let

$$\begin{aligned}D_n(u) &= M_{N'}(u) - M_N(u)\\ &= \sum_{i=0}^{N'\alpha-1}\binom{N'-1}{i}u^{N'-1-i}(1-u)^i - \sum_{i=0}^{N\alpha-1}\binom{N-1}{i}u^{N-1-i}(1-u)^i,\end{aligned}$$

where $N' = N+n$. Then

$$D_n(u) = u^{N-N\alpha}\left\{\sum_{i=0}^{N'\alpha-1}\binom{N'-1}{i}u^{N\alpha+n-1-i}(1-u)^i - \sum_{i=0}^{N\alpha-1}\binom{N-1}{i}u^{N\alpha-1-i}(1-u)^i\right\}.$$

By algebraic manipulation it can be seen that

$$D_n(u) = (1-u)^m u^{N(1-\alpha)} \left\{ \sum_{i=m}^{N'\alpha-1} \binom{N'-1}{i} u^{N\alpha+n-1-i}(1-u)^{i-m} \right.$$
$$\left. - \sum_{i=m}^{N\alpha-1} \binom{N-1}{i} u^{N\alpha-1-i}(1-u)^{i-m} - u^{N\alpha-m} \sum_{i=0}^{n-1} \binom{N-1+i}{m-1} u^i \right\},$$

where $1 \leqslant m \leqslant N-1$. Therefore, by further manipulation of terms, it can be seen that

$$D_n(u) = [(1-u)^\alpha u^{1-\alpha}]^N g_n(u),$$

where

$$g_n(u) = \sum_{i=0}^{n\alpha-1} C(i)\, u^{n-1-i} - \sum_{i=n\alpha}^{n-1} \binom{N+n-2-i}{N\alpha-1} u^{n-1-i}$$

and

$$C(i) = \sum_{j=0}^{N\alpha-1-i} (-1)^j \binom{i+j}{i} \binom{N+n-1}{N\alpha+i+j} - \binom{N+n-2-i}{N\alpha-1}.$$

Now, $D_n(u) = 0$ when $u = 0, 1$ and $g_n(u) = 0$. Since

$$g_n(0) = -\binom{N-1}{N\alpha-1} < 0 \quad \text{and} \quad g_n(1) = \binom{N-1}{N\alpha} > 0$$

there is at least one value of u in $(0, 1)$ for which $g_n(u) = 0$. Furthermore, if there is more than one positive root of $g_n(u) = 0$ in $(0, 1)$, there is an odd number of them altogether.

The polynomial $g_n(u)$ is of degree $(n-1)$ in decreasing powers of u with real coefficients. By inspection of the coefficients of u it can be seen that there are $n\alpha$ variations in sign. In accordance with Descartes' sign rule, there are at most $n\alpha$ positive roots of $g_n(u) = 0$. However, as the number of positive roots must be odd, there are at most $n\alpha$ positive roots if $n\alpha$ is odd and at most $(n\alpha-1)$ positive roots if $n\alpha$ is even.

Therefore, if $n\alpha = 1, 2$, $g_n(u) = 0$ has at most one positive root. Hence, for $n\alpha = 1, 2$, there is exactly one positive root of $g_n(u) = 0$ in the range (0,1). Suppose u_1 is the value of u, $0 < u_1 < 1$, such that $g_n(u_1) = 0$. Then

$$g_n(u) \leqslant 0 \quad \text{for} \quad u \leqslant u_1 \qquad \text{and} \qquad g_n(u) \geqslant 0 \quad \text{for} \quad u \geqslant u_1$$

and, hence,

$$D_n(u) \leqslant 0 \quad \text{for} \quad u \leqslant u_1 \qquad \text{and} \qquad D_n(u) \geqslant 0 \quad \text{for} \quad u \geqslant u_1.$$

Since $r(u)$ is a monotone increasing function of u

$$P_{mc(N')} - P_{mc(N)} = \int_0^1 \{M_{N'}(u) - M_N(u)\}\, r(u)\, du = \int_0^1 D_n(u)\, r(u)\, du$$
$$\geqslant \left\{ \max_{u \in (0, u_1)} r(u) \right\} \int_0^{u_1} D_n(u)\, du + \left\{ \min_{u \in (u_1, 1)} r(u) \right\} \int_{u_1}^1 D_n(u)\, du$$
$$= r(u_1) \int_0^1 D_n(u)\, du = 0,$$

where $N, N' = N+n$ and $N\alpha$ are positive integers and $n\alpha = 1, 2$. Therefore, if $n\alpha = 1, 2$, $P_{mc(N')} \geqslant P_{mc(N)}$ for all positive integral values of $N, N' = N+an$, where $N\alpha$, $N'\alpha$ and a are positive integers.

The most commonly used significance levels, $\alpha = 1/10$, $1/20$, $1/100$, $1/1000$, are of the form $\alpha = 1/n$ where n is a positive integer. For these cases, as has just been shown, the power of the Monte Carlo test procedure is a monotone increasing function of N where $N\alpha = a$ and a is a positive integer. This has also been shown to be true for the case where $N\alpha = 2a$ and a is a positive integer. The more general case, $N\alpha = x$ where x is any positive rational (that is, for any rational value of α in (0,1)), has proved to be mathematically intractable.

An equivalent expression for the power of the Monte Carlo test procedure is

$$P_{mc(N)} = \frac{1}{B\{N(1-\alpha), N\alpha\}} \int_0^1 u^{N(1-\alpha)-1} (1-u)^{N\alpha-1} \left\{ \int_u^1 r(x)\, dx \right\} du,$$

where $B(a, b)$ is the beta function with parameters a and b, N and $N\alpha$ are positive integers and $r(x)$ is the power density function. Letting

$$R(u) = \int_u^1 r(x)\, dx,$$

it can be seen that

$$P_{mc(N)} = \mathscr{E}\{R(u)\},$$

where the probability density of u is the beta distribution with parameters $N(1-\alpha)$ and $N\alpha$. For this distribution the expected value of u is

$$\mathscr{E}(u) = 1-\alpha$$

and the variance of u is

$$\sigma^2(u) = \frac{\alpha(1-\alpha)}{N+1}.$$

For given α, $\sigma^2(u) \to 0$ as $N \to \infty$.

Consider

$$P_{mc(N)} = \frac{1}{B\{N(1-\alpha), N\alpha\}} \left\{ \int_0^{l_1} u^{N(1-\alpha)-1} (1-u)^{N\alpha-1} R(u)\, du \right.$$
$$\left. + \int_{l_1}^{l_2} u^{N(1-\alpha)-1} (1-u)^{N\alpha-1} R(u)\, du + \int_{l_2}^1 u^{N(1-\alpha)-1} (1-u)^{N\alpha-1} R(u)\, du \right\},$$

where

$$l_1 = (1-\alpha) - KN^{\frac{1}{3}} \sigma(u),$$
$$l_2 = (1-\alpha) + KN^{\frac{1}{3}} \sigma(u)$$

and K is a constant. $R(u)$ is a monotone decreasing function of u over $(0, 1)$ and $R(0) = 1$, $R(1) = 0$. Then

$$R(l_2) P(l_1 \leqslant u \leqslant l_2) \leqslant P_{mc(N)} \leqslant \{1 - P(l_1 \leqslant u \leqslant l_2)\} + R(l_1) P(l_1 \leqslant u \leqslant l_2), \tag{3.5}$$

where

$$P(a \leqslant u \leqslant b) = \frac{1}{B\{(N(1-\alpha), N\alpha\}} \int_a^b u^{N(1-\alpha)-1} (1-u)^{N\alpha-1}\, du.$$

Applying Chebyshev's inequality,

$$1-P(l_1\leqslant u\leqslant l_2)=P\{|u-(1-\alpha)|\geqslant KN^{\frac{1}{3}}\sigma(u)\}\leqslant\frac{1}{K^2N^{\frac{2}{3}}}\to 0$$

as $N\to\infty$. Also, as $N\to\infty$, l_1 and l_2 approach $(1-\alpha)$. Then, assuming that $r(x)$ is continuous in the neighbourhood of $(1-\alpha)$, $R(l_1)$ and $R(l_2)$ tend to $R(1-\alpha)$ as $N\to\infty$. Therefore, as $N\to\infty$, it can be seen from inequality (3.5) that $P_{mc(N)}\to R(1-\alpha)$. But $R(1-\alpha)$ is the power of the uniformly most powerful test, P; that is, for given α, as the size of the reference set increases the power of the Monte Carlo test procedure approaches that of the uniformly most powerful test.

3.4. *Power Loss*

Referring to equations (3.1) and (3.3) it can be seen that

$$\begin{aligned}P-P_{mc(N)}&=\int_0^1\{H(u)-M_N(u)\}\,r(u)\,du\\&=-\int_0^{1-\alpha}M_N(u)\,r(u)\,du+\int_{1-\alpha}^1\{1-M_N(u)\}\,r(u)\,du.\end{aligned}$$

As $r(u)$ is a monotone increasing function of u,

$$\max_{u\in(0,1-\alpha)} r(u)=r(1-\alpha)\quad\text{and}\quad\min_{u\in(1-\alpha,1)} r(u)=r(1-\alpha).$$

Also, for $u\in(1-\alpha,1)$, $1-M_N(u)\geqslant 0$ and $1-M_N(u)=0$ only when $u=1$. Therefore,

$$\begin{aligned}P-P_{mc(N)}&\geqslant -r(1-\alpha)\int_0^{1-\alpha}M_N(u)\,du+r(1-\alpha)\int_{1-\alpha}^1\{1-M_N(u)\}\,du\\&=r(1-\alpha)\left\{\alpha-\int_0^1 M_N(u)\,du\right\}\\&=0.\end{aligned}$$

Thus $P\geqslant P_{mc(N)}$. That is, the power of the Monte Carlo test procedure is less than or equal to the power of the corresponding uniformly most powerful test procedure.

The extent of power loss depends on the form of the power density function as well as the procedure adopted for judging significance. In the previous section it has been shown that the power of the Monte Carlo test procedure, for a monotone increasing likelihood ratio and a given value of α, can be increased by increasing N and, in the limit, approaches the power of the uniformly most powerful test. Thus, the loss of power in using the Monte Carlo procedure instead of the uniformly most powerful test procedure is reduced by increasing the size of the reference set. The power loss will not exceed $\{1-M_N(1-\alpha)\}P$. In fact, the upper bound of the power loss will be considerably less than this and it is dependent on the likelihood ratio.

When the separation between the null and alternative distributions is small, $r(u)$ does not differ very much from unity for all u; hence, $P_{mc(N)}$ will be only slightly less than P. Also, as the separation between the two distributions becomes very wide $r(u)$ becomes a J-shaped distribution being approximately zero for most values of u and increasing rapidly for u in the neighbourhood of unity; and, again, the power loss would be very small, as expected. But the main concern is the extent of power loss in using the Monte Carlo test procedure instead of the corresponding uniformly most powerful test where the departure from the hypothesis being tested is moderate. To obtain some insight into this some examples are considered.

4. Comparison of Power Functions

4.1. *Power Functions for Normally Distributed Test Criteria*

Suppose a sample of M observations is selected at random from a normal population with unknown mean μ and unit variance. The null hypothesis is taken to be

$$H_0 :\cdot \mu = 0$$

while the alternative hypothesis is taken to be

$$H_1 : \mu > 0.$$

Then the test criterion, t, of the uniformly most powerful test is the sample mean,

$$\bar{x} = \frac{1}{M} \sum_{i=1}^{M} x_i.$$

The power of this test is

$$\mathrm{P}(\mu) = \int_{1-\alpha}^{1} r(u)\,du = \int_{t_\alpha}^{\infty} f_1(t)\,dt,$$

where

$$t_\alpha = \frac{1}{\sqrt{M}} \Phi^{-1}(1-\alpha), \quad f_1(t) = \left(\frac{M}{2\pi}\right)^{\frac{1}{2}} \exp\left[-\frac{M}{2}(t-\mu)^2\right]$$

and Φ is the standard normal distribution function. That is, $P(\gamma) = 1-\Phi(z_\alpha - \gamma)$, where $z_\alpha = \Phi^{-1}(1-\alpha)$ and $\gamma = \sqrt{(M)}\,\mu$.

The power of the Monte Carlo test procedure, at the α significance level, is

$$P_{mc(N)}(\gamma) = \sum_{i=0}^{N\alpha-1} (-1)^{N\alpha-1-i} \binom{N-1}{i} \binom{N-2-i}{N\alpha-1-i}$$
$$\times \int_{-\infty}^{\infty} [\Phi(x)]^{N-1-i} \frac{1}{\sqrt{(2\pi)}} \exp[-(x-\gamma)^2/2]\,dx$$

where $\gamma = \sqrt{(M)}\,\mu$.

Values of $P_{mc(N)}(\gamma)$ were obtained for $\alpha = 0{\cdot}05$, $N\alpha = 1, 2$ by numerical integration. Fig. 1 shows the power functions thus obtained, $P_{mc(20)}(\gamma)$ and $P_{mc(40)}(\gamma)$, and the power function of the corresponding uniformly most powerful test, $P(\gamma)$.

4.2. *Power Functions for Test Criteria with the t-Distribution*

Suppose that M observations are selected at random from a normal population with mean μ and the null and alternative hypotheses are the same as in the previous example; however, here no assumption is made about the population variance. Then, for the class of tests that are statistically independent of the unknown population variance, the test criterion of the uniformly most powerful test within this class is Student's t-statistic,

$$t(x) = \bar{x}/s_{\bar{x}},$$

where

$$s_{\bar{x}} = \left[\frac{\sum_{i=1}^{M}(x_i-\bar{x})^2}{M(M-1)}\right]^{\frac{1}{2}}.$$

Then the probability density of the test criterion under the null hypothesis, $f_0(t)$, is the central t-distribution with $m = M-1$ degrees of freedom and, under the alternative hypothesis, $f_1(t)$, is the non-central t-distribution with non-centrality parameter $\delta > 0$ and m degrees of freedom.

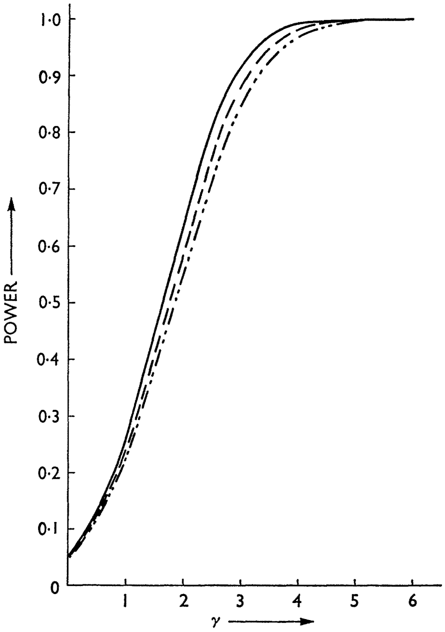

FIG. 1. Power functions for normally distributed test criterion. Significance level: $\alpha = 0{\cdot}05$. ———, Power of the uniformly most powerful test. – – – –, Power of the Monte Carlo test—19 random samples in the reference set. — — —, Power of the Monte Carlo Test—39 random samples in the reference set.

The power functions $P(\delta)$ and $P_{mc(N)}(\delta)$ are of the same form as those given in Section 4.1 except that (i) $f_0(t)$ and $f_1(t)$ are replaced by the central t-distribution with m degrees of freedom and the non-central t-distribution with non-centrality parameter $\delta > 0$ and m degrees of freedom, respectively,

$$\text{(ii)} \quad u = \int_{-\infty}^{t} f_0(y)\,dy = \theta(t)$$

and (iii) $t_\alpha = \theta^{-1}(1-\alpha)$. The power functions $P(\delta)$ and $P_{mc(N)}(\delta)$ were evaluated for $\alpha = 0{\cdot}05$, $N\alpha = 1, 2$ and different values of m and δ. Representative examples of these functions are given in Figs. 2 and 3 for $m = 3$ and $m = 15$, respectively.

As can be seen in these figures, the power of the Monte Carlo test increases considerably when $N\alpha$ goes from 1 to 2. However, the difference between the power of the uniformly most powerful test and that of the corresponding Monte Carlo test

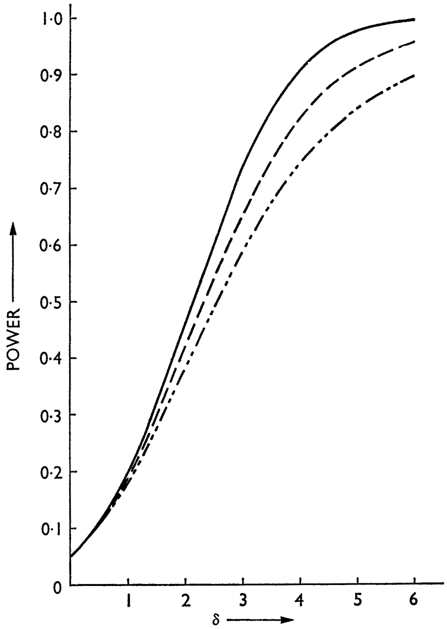

FIG. 2. Power functions when the test criterion is Student's *t*-statistic with 3 degrees of freedom. Significance level: $\alpha = 0{\cdot}05$. ———, Power of the uniformly most powerful test. — — — —, Power of the Monte Carlo test—19 random samples in the reference set. — — —, Power of the Monte Carlo test—39 random samples in the reference set.

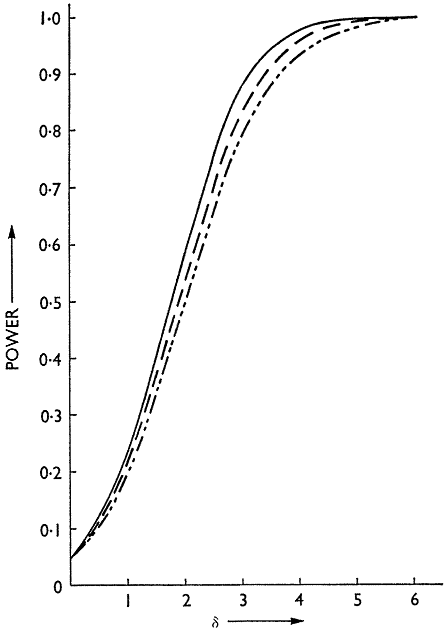

FIG. 3. Power functions when the test criterion is Student's *t*-statistic with 15 degrees of freedom. Significance level: $\alpha = 0{\cdot}05$. ———, Power of the uniformly most powerful test. — — — —, Power of the Monte Carlo test—19 random samples in the reference set. — — —, Power of the Monte Carlo Test—39 random samples in the reference set.

when using the t-statistic as the test criterion is considerably larger than when using the sample mean as the test criterion. As would be expected, as m (the number of degrees of freedom) increases the power of the Monte Carlo test procedure, when the t-statistic is the test criterion, approaches that for the case when the sample mean is the test criterion. For example, when $P = 0{\cdot}95$, $P_{mc(20)} = 0{\cdot}899$ when the mean is the test criterion, whereas $P_{mc(20)} = 0{\cdot}805$ for $m = 3$ and $P_{mc(20)} = 0{\cdot}889$ for $m = 15$ when the t-statistic is the test criterion. Graphs of the power density functions, $r(u)$, for the t and normal distributions having the same power for the uniformly most powerful test showed that $r(u)$ for the t-distribution gives more weight to the lower values of u and less to the higher values of u than does $r(u)$ for the normal distribution; thus the power of the latter would be expected to be greater than that of the former.

4.3. *Power Functions for Exponentially Distributed Test Criteria*

Suppose that the test criterion, t, is exponentially distributed with parameter δ and that the hypothesis

$$H_0\colon \delta = 1$$

is to be tested against the hypothesis

$$H_1\colon \delta = \delta_0, \quad 0 < \delta_0 \leqslant 1.$$

Then, using the same notation as in the previous sections, for $t \geqslant 0$,

$$f_0(t) = e^{-t}, \quad f_1(t) = \delta_0 e^{-\delta_0 t},$$

$$u = F_0(t) = 1 - e^{-t} \quad \text{and} \quad F_1(t) = 1 - e^{-\delta_0 t},$$

where

$$F_i(t) = \int_0^t f_i(x)\,dx, \quad i = 0, 1.$$

If t_α is the value of t such that

$$\alpha = 1 - F_0(t_\alpha) = e^{-t_\alpha},$$

$$P(\delta_0) = \int_{1-\alpha}^{1} r(u)\,du = \int_{t_\alpha}^{\infty} f_1(x)\,dx = e^{-\delta_0 t_\alpha} = \alpha^{\delta_0}.$$

For positive integral values of N and $N\alpha$,

$$\begin{aligned} P_{mc(N)}(\delta_0) &= \delta_0 \sum_{i=0}^{N\alpha-1} \binom{N-1}{i} \int_0^\infty (1-e^{-t})^{N-1-i} (e^{-t})^{\delta_0+i}\,dt \\ &= \delta_0 \sum_{i=0}^{N\alpha-1} \binom{N-1}{i} B(N-i, \delta_0+1+i) = \delta_0 \frac{\Gamma(N)}{\Gamma(\delta_0+N)} \sum_{i=0}^{N\alpha-1} \frac{\Gamma(\delta_0+i)}{\Gamma(1+i)}. \end{aligned}$$

Since it can be shown by mathematical induction that

$$\sum_{j=0}^{k} \frac{\Gamma(j+p)}{\Gamma(j+1)} = \frac{\Gamma(k+p+1)}{p\Gamma(k+1)}, \quad p > 0, \quad P_{mc(N)}(\delta_0) = \frac{\Gamma(N)}{\Gamma(\delta_0+N)} \frac{\Gamma(N\alpha+\delta_0)}{\Gamma(N\alpha)} = \prod_{i=N\alpha}^{N-1} \frac{i}{\delta_0+i}.$$

Figs. 4 and 5 give the graphs of $P(\delta_0)$ and $P_{mc(N)}(\delta_0)$ for $N\alpha = 1$ and $\alpha = 0{\cdot}05$ and $\alpha = 0{\cdot}01$, respectively. As can be seen, the power loss, which will decrease with the increase of N for given α, is exceedingly small when $N\alpha = 1$.

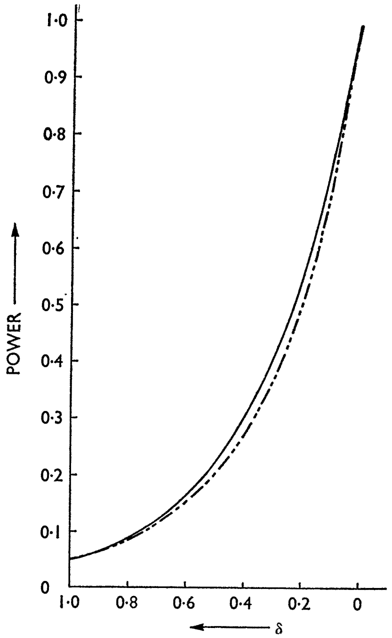

FIG. 4. Power functions for exponentially distributed test criterion. Significance level: $\alpha = 0{\cdot}05$. ———, Power of usual test. – – – –, Power of the Monte Carlo test—19 random samples in the reference set.

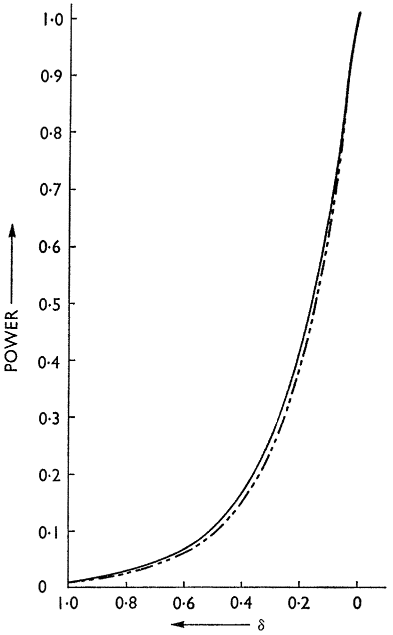

FIG. 5. Power functions for exponentially distributed test criterion. Significance level: $\alpha = 0{\cdot}01$. ———, Power of usual test. – – – –, Power of the Monte Carlo test—99 random samples in the reference set.

TABLE 1

Distribution of efficiency, $\beta(\gamma)$, of the Monte Carlo test to the uniformly most powerful test when the test criterion is $N(\gamma, 1)$. Significance level: $\alpha = 0{\cdot}05$

γ	$\beta(\gamma)$ $N\alpha = 1$ $N = 20$	 $N\alpha = 2$ $N = 40$
0·25	0·915	0·955
0·50	0·911	0·953
0·75	0·907	0·952
1·00	0·904	0·950
1·25	0·900	0·948
1·50	0·898	0·947
1·75	0·896	0·945
2·00	0·894	0·944
2·25	0·892	0·943
2·50	0·889	0·942
2·75	0·886	0·941
3·00	0·884	0·941
3·25	0·882	0·941
3·50	0·881	0·941
3·75	0·877	0·941
4·00	0·876	0·941
4·25	0·872	0·939
4·50	0·870	0·937
4·75	0·868	0·935
5·00	0·868	0·935

4.4. *Efficiency of the Monte Carlo Test Procedure*

For further comparison it would be useful to obtain a measure of efficiency of the Monte Carlo test to the corresponding uniformly most powerful test. To do so, the measure of efficiency of the Monte Carlo test to the uniformly most powerful test is taken as the ratio of the sample size of the uniformly most powerful test to that of the Monte Carlo test, where the uniformly most powerful test has the same power function as that of the Monte Carlo Test.

By direct comparison of the power curves of the examples considered in Sections 4.1 and 4.2 some indication of the efficiency of the Monte Carlo test to the corresponding uniformly most powerful test can be obtained. For the case when the test criterion is normally distributed and $\alpha = 0{\cdot}05$, estimates of the efficiency of the Monte Carlo test relative to the corresponding uniformly most powerful test are

$$\text{(i)} \quad 0{\cdot}8 \quad \text{when } N\alpha = 1,$$

$$\text{(ii)} \quad 0{\cdot}88 \quad \text{when } N\alpha = 2.$$

Testing at the 0·05 significance level, the estimates of efficiency of the Monte Carlo test to the uniformly most powerful test when the test criterion is Student's *t*-statistic are

$$\text{(i)} \quad 0{\cdot}3125 \quad \text{when } N\alpha = 1,$$

$$\text{(ii)} \quad 0{\cdot}5455 \quad \text{when } N\alpha = 2.$$

The values of efficiency for the case of the *t*-distribution indicate that it would be preferable to have a reference set consisting of more than 39 samples when $\alpha = 0{\cdot}05$ (that is, $N\alpha > 2$ for $\alpha = 0.05$). However, in the case of the normal distribution, the efficiency is relatively large even for $N\alpha = 1$, $\alpha = 0{\cdot}05$.

The case of the normally distributed test criterion will now be considered in more detail. On studying the graphs of the power functions $P(\gamma)$ and $P_{mc(N)}(\gamma)$ it seemed that a simple relation existed between these two functions. It was conjectured that there exists a positive real number, β, not greater than unity, such that, for each value of $N\alpha$ (α fixed),

$$P_{mc(N)}(\gamma) \approx P(\surd(\beta)\,\gamma).$$

To obtain an estimate of β, for given $N\alpha$ and α, the method of least squares is used. That is, it is necessary to find the value of β that minimizes

$$I(\beta) = \int_0^\infty [P\{\surd(\beta)\,\gamma\} - P_{mc(N)}(\gamma)]^2\,d\gamma,$$

where $P\{\surd(\beta)\,\gamma\} = 1 - \Phi\{t_\alpha - \surd(\beta)\,\gamma)$ and the other quantities are as defined in Section 4.1. The desired value of β is a solution of the following equation obtained by differentiating $I(\beta)$ with respect of β, setting the resulting expression equal to zero and carrying out integration and algebraic manipulation:

$$\sum_{i=0}^{N\alpha-1}(-1)^{N\alpha-1-i}\binom{N-1}{i}\binom{N-2-i}{N\alpha-1-i}\int_{-\infty}^{\infty}\{\Phi(x)\}^{N-1-i}\,\Phi(X)\,X\exp\left[\frac{-\{t_\alpha-\surd(\beta)\,x\}^2}{2(\beta+1)}\right]dx$$
$$-\left[\frac{\alpha}{\beta}\exp\left(-\frac{t^2{}_\alpha}{2}\right)+\left(1+\frac{1}{\beta}\right)\left\{\frac{\Phi\{\surd(2)\,t_\alpha\}}{\surd 2}+t_\alpha(1-\alpha^2)\left(\frac{\pi}{2}\right)^{\frac{1}{2}}\right\}\right]=0,$$

where $X = \{\sqrt{(\beta)}\,t_\alpha + x)/\sqrt{(\beta+1)}$, and $t_\alpha = \Phi^{-1}(1-\alpha)$. To solve this equation for β numerical procedures were used. For $\alpha = 0{\cdot}05$ and $N\alpha = 1$ the value of β obtained by this method is $\beta = 0{\cdot}797$. For $\alpha = 0{\cdot}05$ and $N\alpha = 2$ the value of β so obtained is $\beta = 0{\cdot}893$. Hence, when $\alpha = 0{\cdot}05$ and the test criterion is normally distributed with known variance, the Monte Carlo test is 0·797 efficient relative to the uniformly most powerful test for $N\alpha = 1$ and it is 0·893 efficient for $N\alpha = 2$. These values do not differ markedly from the measures of efficiency obtained by a direct comparison of the graphs.

More indicative of the situation than the least-squares estimates of efficiency is the distribution of efficiency, $\beta(\gamma)$, given in Table 1 for $N\alpha = 1, 2$ and $\alpha = 0{\cdot}05$ when the test criterion is $N(\gamma, 1)$. These values are all greater than the least-squares estimates of efficiency. The smaller least-squares estimates can be accounted for by the fact that the procedure used gives too large a weight to the contributions of the tail regions. In either case, the measure of efficiency is large even for $N\alpha = 1$.

4.5. *Power Loss when the Test Criterion is* $N(\gamma, 1)$

Assuming that there is a value of β, $0 < \beta \leqslant 1$, such that $P\{\sqrt{(\beta)}\,\gamma\}$ approximates $P_{mc(N)}(\gamma)$, $N\alpha = 1, 2, \ldots$, for all γ, the maximum power loss in using the Monte Carlo test procedure instead of the uniformly most powerful test when the test criterion is the sample mean can be obtained.

Let

$$\begin{aligned} L_N(\gamma) &= P(\gamma) - P_{mc(N)}(\gamma) \\ &\approx P(\gamma) - P\{\sqrt{(\beta)}\,\gamma\} = \Phi\{t_\alpha - \sqrt{(\beta)}\,\gamma\} - \Phi(t_\alpha - \gamma), \end{aligned}$$

where $t_\alpha = \Phi^{-1}(1-\alpha)$.

Possible extrema of $L_N(\gamma)$ exists for values of γ satisfying

$$\frac{d}{d\gamma} L_N(\gamma) = 0.$$

Since $\gamma > 0$, the only possible value is

$$\gamma_1 = \frac{t_\alpha(1-\sqrt{\beta}) + \{t_\alpha^2(1-\sqrt{\beta})^2 - (1-\beta)\ln\beta\}^{\frac{1}{2}}}{1-\beta}.$$

As

$$\frac{d^2}{d\gamma^2} L_N(\gamma_1) < 0$$

the maximum power loss is $L_N(\gamma_1)$.

Using the least-squares estimates of β given in Section 4.4, the maximum power loss occurs at $\gamma = 2{\cdot}237$ when $N\alpha = 1$ $(\alpha = 0{\cdot}05)$ and at $\gamma = 2{\cdot}178$ when $N\alpha = 2$ $(\alpha = 0{\cdot}05)$. Then for $\alpha = 0{\cdot}05$ the maximum power losses for $N\alpha = 1$ and $N\alpha = 2$ when the test criterion is $N(\gamma, 1)$ are 0·0855 and 0·0428, respectively. These values are in agreement with the actual results obtained.

ACKNOWLEDGEMENTS

The author wishes to thank Professor G. A. Barnard and Dr D. J. G. Farlie for their guidance.

The award of a research scholarship by the Royal Commission for the Exhibition of 1851 made it possible for the author to carry out this research at Imperial College —for this she wishes to express sincere gratitude.

REFERENCES

BARTLETT, M. S. (1963). The spectral analysis of point processes. *J. R. Statist. Soc.* B, **25**, 264–296.
DURBIN, J. (1961). Some methods of constructing exact tests. *Biometrika*, **48**, 41–55.
FISHER, R. A. (1935). *The Design of Exepriments.* Edinburgh: Oliver & Boyd.
LEHMANN, E. L. (1959). *Testing Statistical Hypotheses.* New York: Wiley.