On 2 % by Monte-Carlo

John E. B. Ponsonby

Emeritus, Jodrell Bank, UK
johnpon@supanet.com

1. Introduction

It seems that the Radio Astronomy Service (RAS) will accept, in certain circumstances, that 2% of all 2000 second integrations may be lost to interference. Furthermore it is proposed that whether or not a prospective new source of interference, a new satellite system say, will exceed the 2% level should be determined by Monte-Carlo simulation. The implication is that the new system will only be allowed if it causes interference below the 2% level.

This note ignores all the radio technical aspects of such simulations and simply assumes them to be perfect in all regards. It addresses only the statistical aspect of the problem.

The question that it is proposed the Monte-Carlo method should answer is:

"Will more than 2% of integrations suffer unacceptable interference?"

Unfortunately it is not a question to which the Monte-Carlo method can give a definitive answer! Because it is a statistical method it can only give a probable answer to such questions. One would obviously like the answer to have a "reasonably high" probability of being correct. That "reasonably high" probability needs to be stated in the question and agreed in advance by all concerned. For illustrative purposes I've chosen 90% confidence or a 10% risk that the true probability exceeds 2%. Thus the Monte-Carlo method can in principle answer the modified question:

"Can we be C% sure that no more than 2% of integrations will suffer unacceptable interference?"

(C%=90% for illustration) The Monte-Carlo method is inherently ill suited to determining the probability of unlikely events. Because they occur infrequently a great number of simulations must be carried out to build up significant statistics on their frequency. So it requires a very large number of simulated 2000 second integrations to get a reasonably authoritative answer to the modified question.

The modified question may be answered by the method of *Inverse Probability* and the theoretical framework of this method will be developed. It is concerned with the probability of the probability being less than 2%. A computer program (BORP4.EXE) has been developed in conjunction with this report and some numerical results obtained with it will be presented.

2. Rough statistics.

Let us suppose that each 2000 sec simulated integration can be assessed and labelled in a clear cut way, either as an A, meaning that the interference is acceptable, or as a U, meaning the interference is unacceptable. Then a long series of simulated integrations becomes a sequence of the form {AAAAAAAAUAAAAAAAAUAAU···}. If the true probability of getting a U is $p = 2\%$ then in a series of N integrations the *expectation value* for the number of U's is $0.02 \times N$. But this will seldom be the actual number of U's. The actual number is subject to statistical fluctuation of order $\sqrt{0.02N}$. For example in Table 1 I give the *expectation value* for the number of U events in N *trial* integrations. Each is subject to a fluctuation of order its square-root. The last column expresses the results as percentages. It is apparent that to get reasonably good statistics the number of simulated 2000 sec integrations must run into the thousands.

N	expectation value	fluctuation	percentages
200	4	2	2% ± 1%
450	9	3	2% ± 0.66%
800	16	4	2% ± 0.50%
1250	25	5	2% ± 0.4%
1800	36	6	2% ± 0.333%

Table 1. Statistics based on $p = 2\%$

3. The Bernoulli Distribution

A more precise approach is to compute the exact probability distributions. But this must be done with some care. The approach is based on the key assumption that each trial 2000 sec integration is totally independent of the rest.

This assumption would be violated if for instance a constellation of satellites were simulated continuously, so that 2000 sec integrations followed consecutively one after another. In that case the configuration of the system at the end of one integration would be the same as at the beginning of the next, and the trials would not then be statistically independent.

To develop the probability distribution let us start with the most elementary case. If only two trials ($N = 2$) are conducted then there are four possible sequences of events:

AA	AU UA	UU
$(1-p)^2$	$2p(1-p)$	p^2

p is the probability of getting a U

Since AU and UA both have the same number of A's and U's they represent indistinguishable cases. Thus there are only three possible outcomes:

Probability of getting two A's is $(1-p)^2$
Probability of getting one A and one U is $2p(1-p)$
Probability of getting two U's is p^2

If $p = 0.02$ the probabilities are: 0.9604, 0.0392 & 0.0004 respectively.

If there are three trials ($N = 3$) there are eight possible sequences of events and four possible outcomes:

AAA			
	AAU		
	AUA		
	UAA		
		AUU	
		UAU	
		UUA	
			UUU
$(1-p)^3$	$3p(1-p)^2$	$3p^2(1-p)$	p^3

If $p = 0.02$ the probabilities are 0.941192, 0.057624, 0.001176 & 0.000008.

If there are four trials ($N = 4$) there are sixteen sequences of events and five possible outcomes:

AAAA				
	AAAU			
	AAUA			
		AAUU		
	AUAA			
		AUAU		
		AUUA		
			AUUU	
	UAAA			
		UAAU		
		UAUA		
			UAUU	
		UUAA		
			UAUU	
			UUUA	
				UUUU
$(1-p)^4$	$4p(1-p)^3$	$6p^2(1-p)^2$	$4p^3(1-p)$	p^4

It is easy to see the emerging pattern. The numerical factors are Binomial coefficients, the powers of p ascend from the left and the powers of $(1-p)$ descend. There are 2^N sequences of events and $(N+1)$ possible outcomes. The n^{th} term is:

$$\frac{N!}{n!(N-n)!}p^n(1-p)^{N-n} \qquad n = 0, 1, 2, 3, ...(N-1), N \qquad (N+1) \text{ terms in all.}$$

and is the probability of getting n U's with $(N-n)$ A's in N *trial* integrations. The zeroth term, all A's, with $n=0$ is $(1-p)^N$ and the last term, the Nth, with $n=N$, corresponding to all U's, is p^N. This well known result is known as Bernoulli's Theorem. See Woodward (1953).

Using $P(x)$ to mean "the probability of x" rather than some specific function of x, the *Bernoulli Distribution* can be written as:

$$P_p(n) = \frac{N!}{n!(N-n)!}\, p^n(1-p)^{N-n} \tag{1}$$

meaning the probability of getting n U's given the specific value of p. It is a discrete distribution having value only at integer values of n. It is a proper probability distribution however in so far as it sums to unity.

$$\sum_{n=0}^{N} P_p(n) = 1 \tag{2}$$

However the problem that faces us is to determine the distribution

$$P_n(p)$$

which is the probability distribution for p (the probability of the probability p) given a specific value for n, which has been obtained as a result of N Monte-Carlo simulated integrations. This is the so called *"Inverse Probability"* problem.

4. Inverse Probability.

Again following Woodward the product law for probabilities is

$$P(x,y) \;=\; P(x)P_x(y) \;=\; P_y(x)P(y) \tag{3a, b}$$

Here $P(x,y)$ is the joint probability that two variables named X and Y have specified values x and y simultaneously. $P(x)$ is the probability of X alone having the specified value x. Likewise $P(y)$ is the probability of Y alone having the specified value y. $P_x(y)$ is the *conditional probability* that Y has the value y given that X has the specified value x, and $P_y(x)$ is the *conditional probability* that X has the value x given that Y has the specified value y. Should it happen that $P_x(y)$ is independent of x, then $P_x(y) = P(y)$ and the law reduces to the more familiar form

$$P(x,y) \;=\; P(x)P(y) \tag{4}$$

which applies when X and Y are statistically independent.

Taking equation (3b) and changing the notation to the case of interest we can write

$$P(p)P_p(n) \;=\; P_n(p)P(n) \tag{5}$$

or

$$P_n(p) \;=\; P(p)P_p(n)/P(n) \tag{6}$$

Now the result of our N trials is a specific value of n. Call it n'. So we want to find the probability distribution for p given n':

$$P_{n'}(p) \;=\; P(p)P_p(n')/P(n') \tag{7}$$

Since we know the specific n' the quantity $P(n')$ is of no consequence, it is so to speak history, so again following Woodward we write

$$P_{n'}(p) \;=\; k\, P(p) P_p(n') \tag{8}$$

where k is a constant to be determined retrospectively by normalization.

We must now consider the *prior* probability distribution for p: $P(p)$. Is there any *a priori* reason to expect any particular values of p to occur with greater or lesser probability than others? Certainly we expect $P(p)$ to be small for high values of p, unless the interference is quite appalling, but down in the few % region it is surely substantially constant. So it is fair to write

$$P_{n'}(p) \;=\; k\, P_p(n') \tag{9}$$

(If this is found unsatisfying there are whole treatises to be read on the *PRIOR*-problem.) Here the RHS is the discrete Bernoulli Distribution evaluated at $n = n'$ over a range of values of p. We can imagine in the (n, p) plane $P_p(n)$ as horizontal cuts at constant p, whilst $P_n(p)$ are vertical cuts at constant n. $P_{n'}(p)$, which is a continuous function of p, is one such specific vertical cut. Unfortunately the integral along one such vertical cut

$$\int_{p=0}^{p=1} P_p(n)\, dp \;=\; \int_{p=0}^{p=1} \frac{N!}{n!(N-n)!}\, p^n (1-p)^{N-n}\, dp \;\neq\; 1 \tag{10}$$

(for clarity I now drop the prime on n) so the vertical cuts are not properly normalized probability distributions. It can be shown however, see Appendix, that the integral evaluates to $1/(N+1)$. So

$$P_n(p) \;=\; \frac{(N+1)!}{n!(N-n)!}\, p^n (1-p)^{N-n} \tag{11}$$

is the properly normalized continuous probability distribution for the probability p given the single value n.

This probability distribution may be integrated numerically upwards from $p = 0$ so that the upper limit p_C may be identified such that it is C% certain that the true value of p lies below p_C according to:

$$\int_{p=0}^{p=p_C} P_n(p)\, dp \;=\; C \tag{12}$$

If $C = 0.9$ and $p_{90} \leq 0.02$ then the Monte-Carlo simulation answers the modified question with 90% confidence that the true value of p, the probability that one 2000 sec integration will suffer unacceptable interference, is less than 2%.

5. Computation.

The program BORP4.EXE requires as input three numbers, N the number of trials, n the number of U results and C% the desired confidence limit. It plots on screen the function $P_n(p)$ given by equation (11). At the same time it performs a numerical integration which is used to detect the value of p_C. The integration is continued and the total area under the curve given by the value of sigma. For all normal cases this should give the value 1.000 thus showing that the function is indeed correctly normalized. As the curve is plotted, passage of the $C\%$ point is indicated by a change of colour. The safe part of the curve, below the $C\%$ point, appears in green. The dangerous part, above the $C\%$ point, appears in red. When $n << N$ and both are small the curve is markedly asymmetric. When the *expectation value* for p, n/N, is neither close to 0 nor to 1, and N is large, then as expected the curve has Gaussian form. The expectation value and p_C are given numerically and as coloured marks on the p axis. Examination of the plotted curve gives one a feel for the statistics and allows one to judge just how big the true value of p might be should ill fortune have yielded a value for n subject to an unusually large negative fluctuation. The plotting and the numerical integration proceed in steps $\Delta p = 1/4000$. Consequently p is always given as a multiple of 0.025%.

6. Numerical Results.

A few interesting results are given in Table 2. which all give $p_{90} = 2\%$.

N	n	(n/N) %	$P_{90}\%$
191 thru 193	1	0.52%	2%
262 thru 265	2	0.76%	2%
330 thru 333	3	0.91%	2%
395 thru 399	4	1.01%	2%
458 thru 463	5	1.08%	2%
521 thru 525	6	1.14%	2%
582 thru 588	7	1.19%	2%
643 thru 650	8	1.23%	2%
97 thru 1008	14	1.39%	2%

Table 2. Some values of N and n that all give $P_{90} = 2\%$.

The reader is urged to obtain more by running the program.

7. Discussion.

This methodology always gives a value for p_C which is somewhat greater than the expectation value n/N. The difference between them naturally gets smaller as N increases but only very slowly. To be sure, with $C\%$ confidence, that the true probability p is less than 2% means that very often it will be considerably less, but there is no way this can be known. There is of course a $(100 - C)\%$ risk that it will be higher and this too cannot be known. Just how much above 2% it might be can be judged by examining the red portions of the curves.

Insofar as the true value of p will often be substantially less than p_C adoption of this methodology will be to the advantage of the RAS. But this is not a manifestation of any sort of hidden bias, it is simply a consequence of taking the modified question seriously.

In all the talk within SE21 and TG1/5 about Monte-Carlo simulation, I have never heard any estimates of how many simulations are necessary to obtain reasonably reliable results. This report and the accompanying program allows this question to be addressed.

A word of warning. It is most important that the number N be decided in advance. The methodology would be invalidated if simulations are run with an eye on how the expectation value n/N is getting on and stopping when is seems to be making a negative fluctuation. That would seriously prejudice the methodology.

8. Conclusion.

It has been pointed out that the question:

"Will more than 2% of integrations suffer unacceptable interference?"

cannot be answered by Monte-Carlo simulation. Only a modified question of the form:

"Can we be C% sure that no more than 2% of integrations will suffer unacceptable interference?"

can be addressed. I've adopted $C\% = 90\%$ for illustration. The RAS may decide on an alternative figure. Some say 95% would be more appropriate.

The methodology of *Inverse Probability* has been described and applied to a Bernoulli probability distribution. A program implementing it, BORP4.EXE, is presented which computes the probability of the probability p given n, and produces an answer to the modified question. It is clear that to get reasonably accurate results the number of independent 2000 sec integrations that must be simulated runs into the thousands.

Adoption of this interpretative methodology will be to the advantage of the Radio Astronomy Service but one must therefore expect it to be resisted by its opponents. However it is the only rational and honest approach I know to interpret the Monte-Carlo simulations.

Appendix.

The integral on the Left Hand Side of the inequality of equation(10) may be evaluated as follows. We are concerned to prove that

$$\frac{N!}{n!(N-n)!}\int_0^1 x^n(1-x)^{N-n}\,dx \;=\; \frac{1}{N+1} \tag{A1}$$

where N and n are positive integers such that $n < N$.

PROOF:

Write $(N-n) \;=\; m$ and define

$$I_{m,n} = \frac{(m+n)!}{m!n!}\int_0^1 x^n(1-x)^m\,dx \tag{A2}$$

Integrate by parts to get

$$I_{m,n} \;=\; \frac{(m+n)!}{m!n!}\left\{\left[\frac{(1-x)^m\,x^{n+1}}{n+1}\right]_0^1 - \int_0^1 \frac{(-1)\,x^{n+1}\,m(1-x)^{m-1}}{n+1}\,dx\right\} \tag{A3}$$

$$= \;\frac{(m+n)!}{(m-1)!(n+1)!}\int_0^1 x^{n+1}(1-x)^{m-1}\,dx \;=\; I_{m-1,n+1} \tag{A4}$$

since $\frac{m}{m!} \;=\; \frac{1}{(m-1)!}$ and $\frac{1}{(n+1)n!} \;=\; \frac{1}{(n+1)!}$.

We now see that

$$I_{m,n} = I_{m-1,n+1} = I_{m-2,n+2} = I_{m-3,n+3} = \cdots = I_{0,m+n} \tag{A5}$$

But

$$I_{0,n+m} = \frac{(m+n)!}{(m-m)!(n+m)!}\int_0^1 x^{n+m}\,dx = \left[\frac{x^{m+n+1}}{(m+n+1)}\right]_0^1 = \frac{1}{(m+n+1)} \tag{A6}$$

therefore

$$I_{m,n} = \frac{1}{(m+n+1)} = \frac{1}{N+1} \tag{A7}$$

QED.

Bibliography.

Woodward, P. M., *Probability and Information Theory with Applications to Radar*, Pergamon Press, London, 1953, 1964; McGraw-Hill, New York, 1953.
This is a much quoted classic and seminal work.