

NRL Report 8427

Evaluating the Validation of a Monte Carlo Simulation of Binary Time Series

D. R. ROQUE

*Systems Research Branch
Space Systems Division*

September 19, 1980



NAVAL RESEARCH LABORATORY
Washington, D.C.

Approved for public release; distribution unlimited.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|-----------------------|---|
| 1. REPORT NUMBER NRL Report 8427 | 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle) EVALUATING THE VALIDATION OF A MONTE CARLO SIMULATION OF BINARY TIME SERIES | | 5. TYPE OF REPORT & PERIOD COVERED Final report |
| | | 6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s) Diego R. Roque | | 8. CONTRACT OR GRANT NUMBER(s) |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Research Laboratory Washington, DC 20375 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 79-0701-J-0 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Arlington, VA 22217 | | 12. REPORT DATE September 19, 1980 |
| | | 13. NUMBER OF PAGES 31 |
| 14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) | | 15. SECURITY CLASS. (of this report) UNCLASSIFIED |
| | | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |
| 16. DISTRIBUTION STATEMENT (of this Report) Approved for public release, distribution unlimited. | | |
| 17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report) | | |
| 18. SUPPLEMENTARY NOTES | | |
| 19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Monte Carlo evaluation Simulation validation Statistical tests Binary time series | | |
| 20. ABSTRACT (Continue on reverse side if necessary and identify by block number) A statistical technique has been developed for the validation of a Monte Carlo simulation process or other processes whose results can be reduced to a finite time sequence of equally spaced events with dichotomous outcomes. Essential to the technique is the Bahadur-Lazarsfeld representation of the probability distribution of the populations consisting of all binary vectors with a specific number of elements. This paper analyzes the properties of the test and the adequacy of the Bahadur-Lazarsfeld representation for practical application purposes. It examines the probability of (Continued) | | |

DD FORM 1 JAN 73 1473

EDITION OF 1 NOV 68 IS OBSOLETE
S/N 0102-014-6601

1

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT (Continued)

rejection which results when the significance level α is specified and computes the power of the test when the null hypothesis is tested against known alternatives. The evidence collected suggests that the statistical test is an adequate tool for the purposes for which it was designed.

CONTENTS

| | |
|--|----|
| INTRODUCTION | 1 |
| MAJOR FINDINGS | 2 |
| VALIDATION METHOD | 3 |
| A MONTE CARLO EVALUATION OF THE TEST | 3 |
| AN ANALYTICAL APPROACH | 7 |
| SUMMARY | 10 |
| REFERENCES | 10 |
| APPENDIX A – Validation Methodology | 11 |
| APPENDIX B – Monte Carlo Evaluation | 15 |
| APPENDIX C – Details of Analytical Formulation | 24 |

EVALUATING THE VALIDATION OF A MONTE CARLO SIMULATION OF BINARY TIME SERIES

INTRODUCTION

This report describes the results of an effort to characterize the properties of a statistical test developed by the Naval Research Laboratory (NRL) to validate the Antisubmarine Warfare Program Surveillance Model (APSURV) Mod 1.4 simulation model. The APSURV digital computer simulation models have been the Navy's approved undersea surveillance models since their initial development. They have been used to predict the performance of the Sound Surveillance System (SOSUS). The validation effort was one of the first attempts to assess the adequacy of the APSURV models in representing the detection performance of the SOSUS. This paper presents some preliminary results in evaluating the validation method itself. Its focus is on the properties of a statistical test developed by Johnson and Wiener [1] designed to conduct the validation effort. Although the evaluation has been limited by the large amount of computer time required to perform a complete analysis, the available evidence confirms the adequacy of the validation procedure and opens up the topic as an area for further research.

The term validation means determining whether, within the degree of assurance of statistical tests, the simulation represents the process being modeled. In [1] a quantitative procedure for a simulation model validation has been adopted and the required statistical tests developed to conduct the task. The validation technique developed can be briefly described as follows. Each Monte Carlo replication of the simulation model for any one sensor produces a vector of m binary elements. Based on a sample of these binary vectors, a representation is obtained of the probability distribution of the population of binary vectors from which the sample was drawn. Using this representation the likelihood of any vector of m binary elements may be computed under the hypothesis that it comes from the same statistical population as the vectors generated by the simulation model. In particular the method computes the likelihood of the binary vector resulting from an observed run of the actual process under the same conditions that are represented in the simulation. The question of the validity of the simulation model, at the significance level α , is then resolved by observing whether the likelihood of the experimentally obtained vector exceeds the α^{th} percentile of the likelihoods of the simulation generated vectors. The statistical technique is appropriate for the validation of a Monte Carlo simulation of processes such as detection processes, whose results can be reduced to a finite sequence of thresholding events having dichotomous, i.e., binary, outcomes. Moreover, the validation technique can be applied even when the process being simulated cannot be experimentally repeated. Thus the simulation's validity for the case examined can be determined, to within a specified level of statistical significance, with only a single observation of the real-world process being simulated. The statistical technique is nonparametric, it does not assume independence between events occurring at different times, and it does not require the assumption of any stationarity or steady state behavior of the process simulated. The simulation validation procedure can be summarized as follows.

As a particular target traverses the surveillance zone it generates a track history of detections for each sensor in the zone. For each sensor i there is an observed vector $\hat{x}_i = (\hat{x}_{i1}, \dots, \hat{x}_{im})$ of 0's and 1's from an unknown probability distribution p_o^i and the simulation generates n vectors $x = (x_1, \dots, x_m)$ of 0's and 1's from a probability distribution p_i . Using the n generated vectors the statistical technique obtains an estimate \hat{p}_i of p_i . The simulated data are then applied to \hat{p}_i to obtain the

sample distribution of the n vectors as an approximation to the population distribution. The test consists of determining whether the observed vector \hat{x}_i has \hat{p}_i -value in the upper $1 - \alpha$ region of the sample distribution. If it does then the hypothesis that \hat{x}_i comes from the distribution p_i is accepted. The test is applied for all i and many acceptances that \hat{x}_i is from p_i will confirm the null hypothesis that p_i is a good approximation to p_o^i , thereby validating the simulation. It is to be expected that due to statistical fluctuation some sensors will fail the test. Hence a distinction must be made between validation of the simulation model in general and specific statements about the simulation of the individual sensors. Statements about the latter are simply understood to carry the uncertainty inherent in the statistical test itself.

This paper examines the properties of the statistical test applied to validate the simulation. Of immediate interest is the question of whether specifying the α -th percentile of the sample distribution does result in a probability of false rejection equal to α . In addition there are such concerns as to how one should deal with alternative hypotheses and what kind of power the test exhibits in evaluating alternatives. Partial answers have been obtained to some of these questions and enough information has been learned to establish the adequacy of the test procedure. A full analysis of the properties of the test remains for further investigation.

Following the introduction in this report is a section on the major conclusions reached. It is followed by a brief statement on the validation method that was evaluated. The next two sections consist of a Monte Carlo study on the validation method and an analytical approach. Finally there is a summary and a series of appendices presenting details, as necessary, on the validation method and technical support material on the evaluation methods.

MAJOR FINDINGS

Each Monte Carlo replication of the simulation model for each sensor produces a vector of m binary elements. When the number of elements is small, say $m = 2$ or 3 , the statistical test at a significant level α does not result in a probability of false rejection equal to α . As m increases, for α small, the probability of false rejection approaches α from below and it is practically α for $m = 5$ or 6 .

Both a Monte Carlo evaluation and an analytical approach were taken to examine the power of the test. In the former two alternative hypothesis were tested against the same null hypothesis. Both alternatives yielded high power and the more dissimilar the alternative to the null hypothesis the higher the power of the test. In the analytical approach a simple structure was considered that allowed one to attain power curves for the cases $m = 1$ and $m = 2$. These cases shed some light into the behavior of the power function in general. Depending on the null hypothesis the range of possible values of the power of the test tends towards higher power as the two distribution become more dissimilar. This is not to say that minimum power is achieved whenever the null and alternative hypothesis coincide. It simply states that this is the case for the longer portion of the range of possible values of the alternative distribution. The number of exceptions gets small as α gets small and for small α in general the exceptions occur within a range that remains relatively close to the null hypothesis anyway. This analysis will be discussed later in more detail. Both approaches suggest however that the test performs adequately with respect to the power of the test.

The statistical test consists of testing the null hypothesis that the observed vector is a sample from the population from which the n vectors produced by the simulation were generated. In so doing one specifies a significance level α , in this case the α^{th} percentile of the sample distribution obtained from n replications of the simulation. If the likelihood of the observed vector falls below the α^{th} percentile of the likelihoods of the simulation generated vectors the null hypothesis is rejected.

In order to conduct the evaluation of the test a Monte Carlo approach was first taken. Sample distributions were obtained from simple known distributions. This was done by repeatedly generating n vectors as if by the simulation and then generating from the same distribution an $n + 1$ st or observed vector. In each instance the test was conducted. By repeating the test 20 times a proportion of rejections was computed. By further repetitions in blocks of 20 repetitions each, a sequence of independent, identically distributed random variables was generated. The Central Limit Theorem then allowed one to obtain confidence intervals on the resulting mean proportion of rejections, which should approximate the resulting probability of rejection when applying the test. The evidence gathered supports the conjecture that as the vector length m increases the probability of false rejection begins to approach α . For the cases considered the results indicate that the probability of rejection is actually less than α for $m = 2$ or 3, but it approaches α from below as m increases and is practically α for $m = 5$ and 6. By varying the above approach slightly the $n + 1$ st vector was then generated from an alternate known distribution different from the distribution of the first n vectors. The resultant proportion of rejections now estimates the power of the test. For the cases considered the test appears to discriminate very well.

To obtain the representation of the distribution of the population from which the n simulated vectors come, it is necessary to estimate the function $f(\underline{x})$ which measures the correlation effects among elements of the vector \underline{x} . Appendix A specifies the functional form of $f(\underline{x})$. An important consequence of the evaluation is the fact that it is necessary to pay attention to the vector sample size n to ascertain when the estimate $\hat{f}(\underline{x})$ is close enough to its theoretical value $f(\underline{x})$ for the representation of the distribution in question to be adequate.

VALIDATION METHOD

The simulation model in question, APSURV Mod 1.4, was designed to estimate the performance of an acoustic sensor in the detection of a target in the ocean. Specific details of the validation methodology for the simulation are given in Appendix A of this paper. The simulation is essentially based on Monte Carlo replications of time-phased detect/no detect events and the validation considers a real-world observed sequence of detect/no detect decisions made by the acoustic sensor. Table 1 illustrates the structure of the validation data. The simulation generates n replications of vectors with m elements and the test determines whether the random process generating the $n \times m$ matrix is equivalent to the random process generating the $(n + 1)$ -st or observed vector. The Method of Bahadur [2] and Lazarsfeld [3] is used to obtain the representation of the probability distribution of the population from which the n simulation replications come. From this is obtained for each vector $\underline{x} = (x_1, \dots, x_m)$ a likelihood value

$$p(\underline{x}) = p_{[1]}(\underline{x}) \cdot f(\underline{x})$$

where $p_{[1]}(\underline{x})$ is the probability of vector \underline{x} under the assumption of independent vector elements and $f(\underline{x})$ is a function which represents the degree of correlation along the time stream.

For an α -level test one checks whether the likelihood of the observed vector $p(\underline{x}_o)$ is above the first $n \cdot \alpha$ likelihoods from the simulated set. If so, the hypothesis that \underline{x}_o is a member of the same random process generating the n replications is not rejected.

A MONTE CARLO EVALUATION OF THE TEST

Of the various properties of the test one is interested in evaluating, the most immediate concern is whether specifying the α -th percentile of the sample distribution does result in a probability of false rejection equal to α . In addition there are such concerns as to how one should deal with alternative hypotheses and what kind of power the test exhibits in evaluating alternatives. Another important factor is the adequacy of the estimate of the probability distribution of the n simulated replications of a track history of detections. Of particular interest is whether the number of replications is sufficient for the estimate of the correlation function $f(\underline{x})$ to settle about its theoretical value.

Table 1 — Structure of the Validation Data

240-HOUR TRACK, 50 MODEL REPLICATIONS
1: DETECTING THE TARGET,
0: NOT DETECTING TARGET

| Elapsed Time (hours) | 1 | 2 | 3 | 4 | 5 | 6 | ... | 239 | 240 |
|---------------------------------|------|-----|-----|-----|-----|-----|-----|-----|-----|
| Observed Detection History | 0 | 0 | 1 | 1 | 1 | 0 | ... | 1 | 1 |
| Model Replication | | | | | | | | | |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | ... | 1 | 1 |
| 2 | 0 | 1 | 0 | 0 | 0 | 0 | ... | 0 | 1 |
| 3 | 0 | 0 | 1 | 1 | 0 | 0 | ... | 0 | 1 |
| . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . |
| 50 | 0 | 0 | 1 | 0 | 0 | 0 | ... | 0 | 0 |
| Predicted Detection Probability | 0.00 | .20 | .60 | .20 | .00 | .00 | ... | .00 | .60 |

Under the assumption that the simulation model is valid, the statistical test for a specific sensor-track combination would normally constitute a Bernoulli trial with rejection probability α and probability of no rejection $1 - \alpha$. Here the sample distribution (of the n replications) is used as an approximation to the theoretical distribution; hence the actual rejection probability may appear to be data-driven. However repetitions of the test itself under identical conditions are identically distributed and should generate a proportion of rejections approximating the probability of rejection as the number of repetitions increases. In general there are $n + 1$ vectors each of length m . Ideally the probability of rejection should equal α when the $n + 1$ st vector is from the same distribution as the first n , and it should equal the power of the test against a specified alternative distribution when the $n + 1$ st vector is from the alternative distribution. To make inferences about the probability of rejection successive repetitions of the same test may be generated. The repetitions should be grouped into subsets of equal size from each of which a proportion of rejections may be computed. The computed proportions are a sequence of independent, identically distributed random variables. Taking a large enough number of subsets one can then appeal to the Central Limit Theorem and use classical statistical techniques to make inferences on the probability of rejection that results when applying the given test at a specific level α .

A preliminary evaluation of the properties of the test has been conducted with the use of Monte Carlo methods. It was basically intended to determine an estimate of the level of significance obtained in the test. The approach taken used sets of $n + 1$ random binary vectors which were repeatedly generated from a known distribution. Several cases were considered by varying the vector length m and by varying n , the number of replications, for each m . The vector elements in all cases were independent Bernoulli random variables where the probability of a 1 varied by vector element. The various cases of m considered were 2, 3, 4, 5 and 6 vector elements. For the case $m = 6$ the probabilities of a 1 for all vector elements were (.2, .4, .6, .8, .8, .8) respectively. Similarly one used for $m = 5$ the probability vector (.2, .4, .6, .8, .8), for $m = 4$ (.2, .4, .6, .8), for $m = 3$ (.2, .4, .6) and for $m = 2$ (.2, .4). For all of these distributions the correlation function has a theoretical value of $f(x) = 1$. In a typical case, for example $m = 4$ and $n = 50$ with a random number routine, 51 vectors of length 4 are generated and the test is applied to determine whether the 51th vector does or does not belong to the same population from which the first 50 came. The decision is made whether or not to reject such a hypothesis. This procedure is repeated 20 times and the proportion of rejections is computed. The sets of size 20 are replicated 100 times. The one hundred computed proportion of rejections are then treated as 100

independent, identically distributed random variables whose mean is an estimate of the probability of false rejection resulting from the application of the test. The test itself consists of computing the likelihood q of the $n + 1$ -st vector and the set of likelihoods $\{p_g = p(\underline{x}_g): g = 1, \dots, n\}$. The test procedure is to reject the hypothesis of association if the observed value q falls below the α -th percentile of computed values p_g . Define N to be the number of elements in the set $\{g: p_g \leq q, 1 \leq g \leq n\}$. If $N \geq n \cdot \alpha$, the hypothesis of association is not rejected. The hypothesis is rejected at the significance level α if $N < n \cdot \alpha$.

In this evaluation the value of α was fixed at $\alpha = 0.30$ and for each value of m the number of replications (i.e., values of n) used were from the set $\{20, 50, 100, 200, 500, \text{ and } 800\}$. Not all values of n were used for all values of m . Only for $m = 5$ were all values of n used. In addition as m and n became larger both the number of repetitions in each subset from which the proportion of rejections were computed and the number of subsets itself were decreased due to limitations in computer time and cost.

For the cases considered the significance level achieved appears to approach α from below as m , the number of vector elements increases.

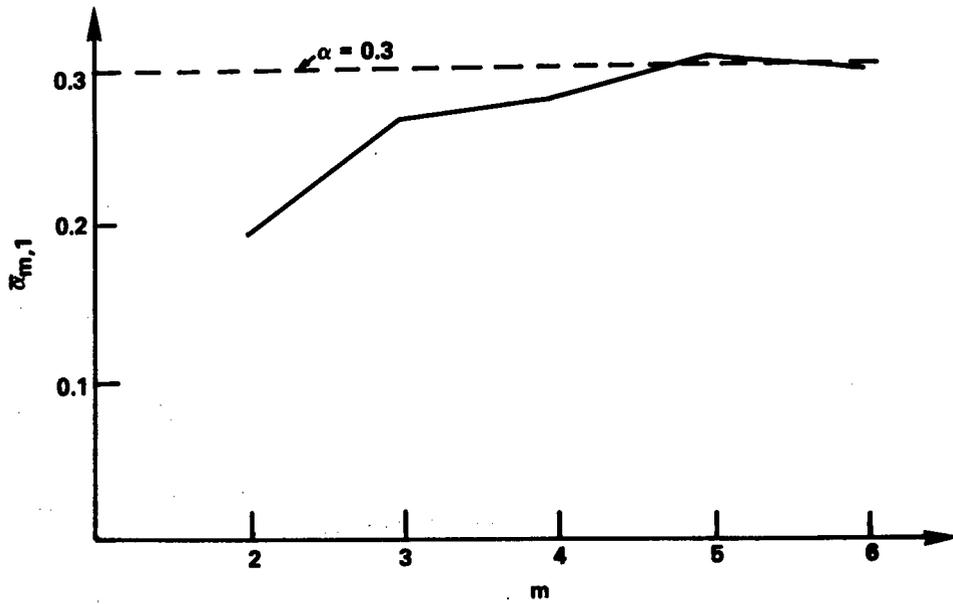
The analysis was also conducted to estimate the power of the test when the null hypothesis was tested against a specified alternative. For the case $m = 5$ tests were also conducted where the $n + 1$ -st vector was generated from a known distribution other than the distribution from which the first n vectors were generated. These alternate distributions consisted of independent, identically distributed vector elements where the probability of a 1 was given by $p = .1$ in one case and $p = .5$ in another. The theoretical value of the correlation function in this case is also $f(\underline{x}) = 1$. The proportion of rejections in this case estimates the power of the test, that is, the probability that the null hypothesis is correctly rejected. When comparing both alternatives to the null hypothesis the test appears to discriminate very well and the more dissimilar the alternative to the null hypothesis the higher the power of the test.

An interesting problem revealed in the evaluation is that vectors with negative likelihoods arise as a function of the fluctuation of computed values of $f(\underline{x})$ around its theoretical value. In the cases considered the theoretical value is $f(\underline{x}) = 1$. As n increases the distribution of the n values $\hat{f}(\underline{x})$ tends towards a spike at 1 and the number of vectors resulting in negative likelihoods (because $\hat{f}(\underline{x}) < 0$) decreases or disappears. To assess the effects of the estimates \hat{f} tests were performed first using the estimates \hat{f} and then replacing these estimates with the actual theoretical values $f(\underline{x}) = 1$.

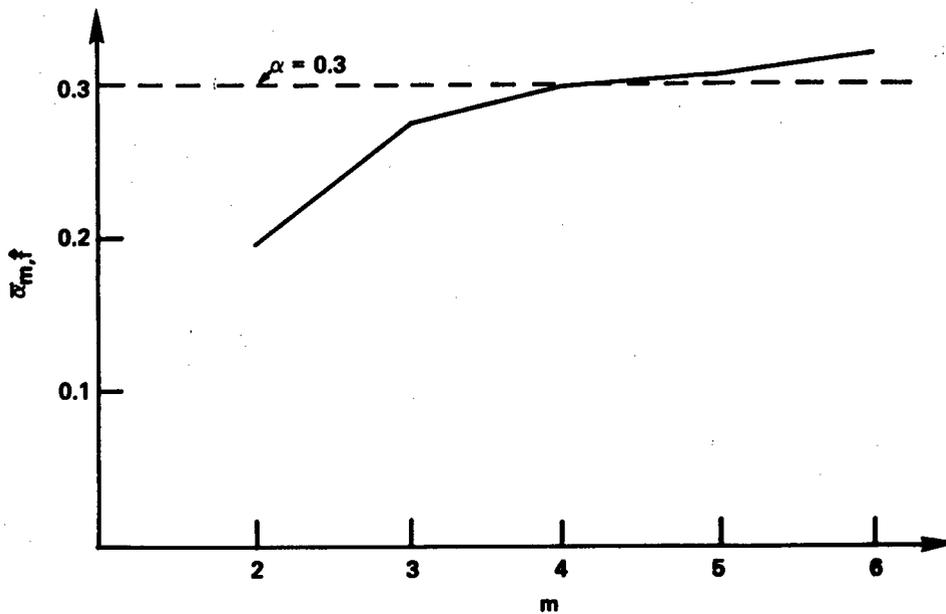
The data were collected by cases of m , n , and either $f(\underline{x}) = \hat{f}$ or $f(\underline{x}) = 1$. For each case labeled (m, n, \hat{f}) there corresponds a case $(m, n, 1)$. To obtain an overall measure of the resulting probability of false rejection when the test was applied at a significance level α for each m , the resulting estimates $\hat{\alpha}$ were averaged over all values of n used for that specific m . These resulting averages can be labeled $\bar{\alpha}_{m,\hat{f}}$ and $\bar{\alpha}_{m,1}$. To compute the $\bar{\alpha}_{m,\hat{f}}$ only those values of n were included for which the estimates \hat{f} were considered adequate based on a comparison with $f(\underline{x}) = 1$ cases. Figure 1 plots the overall average probability of false rejection resulting for each m . For the cases considered the results indicate that the probability of rejection is actually less than α for $m = 2$ or 3 but it increases to α as m increases and is practically α for $m = 5$ and 6.

In addition, in these cases where the theoretical correlation function is $f(\underline{x}) = 1$ some attention must be paid to having a large enough number of replications (n) to make certain that the computed values $\hat{f}(\underline{x})$ are an appropriate representation of the theoretical correlation function. An alternative would be to find a better estimator for $f(\underline{x})$.

The conclusions of these evaluations cannot be accepted as general because the analysis was limited to a few known distributions of simple structure. Enough evidence was collected, however, to establish confidence in the adequacy of the procedure developed by Johnson and Wiener [1].



(a)



(b)

Fig. 1 — Estimated probability of false rejection vs. values of m for $\alpha = 0.3$ (a) $f(x) = 1$ and (b) $f(x) = \hat{f}$

Appendix B contains more detailed information on the Monte Carlo evaluation including the data collected and tables and figures representing the information gathered from the evaluation.

AN ANALYTICAL APPROACH

The statistical test consists of computing the likelihood q of the $n + 1$ -st vector \underline{x}_0 and the set of likelihoods $\{p_g = p(\underline{x}_g): g = 1, \dots, n\}$, one likelihood for each of the n replication vectors $\{\underline{x}_g: g = 1, \dots, n\}$. The test procedure is to reject the hypothesis of association if the observed value q falls below the α -th percentile of the computed values p_g . Define N to be the number of elements in the set $\{g: p_g \leq q, 1 \leq g \leq n\}$. If $N \geq n \cdot \alpha$ the hypothesis of association is not rejected. The hypothesis is rejected at the significance level α if $N < n \cdot \alpha$.

As discussed in appendix A, the likelihood of any vector \underline{x} is given by its Bahadur representation.

$$p(\underline{x}) = p_{[1]}(\underline{x}) \cdot f(\underline{x})$$

where $p_{[1]}(\underline{x})$ is the probability of vector \underline{x} under the assumption of independent vector elements and $f(\underline{x})$ is a function which represents the degree of correlation between vector elements. In this section only the simplest cases will be considered. The $n \times m$ data matrix will consist of n independent vectors each of length m where the entries $x_{i,j}$ are independent, identically distributed Bernoulli random variables with probability of a 1 given by v . In this case for any vector \underline{x} , $f(\underline{x}) = 1$, and hence $p(\underline{x}) = p_{[1]}(\underline{x})$. Now let $E(\cdot)$ denote expected value, then

$$v = E(x_{i,j}) \quad i = 1, \dots, n; j = 1, \dots, m$$

where it is assumed $0 < v < 1$. In this case maximum likelihood estimates are used to obtain \hat{v} . Since this is a simplification of a slightly more general representation the estimates are obtained from the columns of the $n \times m$ matrix.

$$p_{[1]}(\underline{x}_g) = \prod_{i=1}^m \hat{v}_i^{x_{gi}} (1 - \hat{v}_i)^{1-x_{gi}}, \quad g = 1, 2, \dots, n$$

where

$$\hat{v}_i = \begin{cases} 1/2n & \text{if } \sum_{g=1}^n x_{gi} = 0 \\ 1 - (1/2n) & \text{if } \sum_{g=1}^n x_{gi} = n \\ (1/n) \sum_{g=1}^n x_{gi} & \text{otherwise for } i = 1, 2, \dots, m \end{cases}$$

Notice first that $v_i = v_j$ in our case for all $i, j = 1, \dots, m$ and secondly since the v_i 's are assumed to be neither 0 nor 1 a reasonable correction is made should the data seem to indicate they are when estimating them. Even though $v_i = v_j$ for all $i, j = 1, \dots, m$ it is likely that $\hat{v}_i \neq \hat{v}_j$ for most i, j pairs.

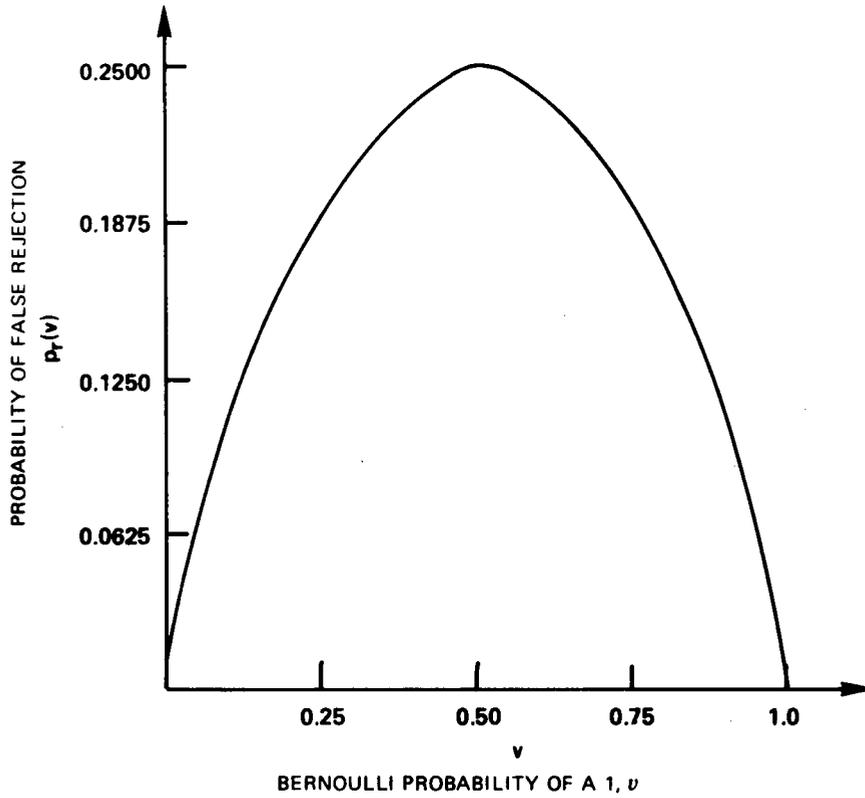
The particular data structure considered in this section shows that it may be the case that the probability of false rejection when applying the test approaches the significance level α as m , the vector length, increases. For any vector length m there are 2^m possible binary outcomes or members of the population of vectors of length m . The analysis consists of letting $n = 2^m$ where all binary vector elements consist of independent, identically distributed Bernoulli random variables with the probability of a 1 given by v . There are $(2^m)^{2^m}$ possible outcomes from which the Bahadur representation may be obtained. To each of these outcomes one may associate 2^m possible $n + 1$ -st or observed vectors yielding a total of $(2^m)^{2^m+1}$ elementary outcomes. Consider first the case where the observed vectors come

from the same distribution as the sets of first $n = 2^m$ vectors. Specifying α , evaluating N for each outcome, and applying the rules of the test one may establish an association between α and the actual probability of false rejection. Table 2 lists all possible outcomes when performing the test for the case $m = 1$. This table shows the 4 cases of "simulated" vectors that are possible, the estimates $\hat{\nu}$ of ν that each case generates, the computed vector likelihoods p_g for each case, the likelihoods g for each of the possible "observed" vectors, and the value of N for each combination of simulated vectors and observed vector. Here $n = 2^m = 2$, $(2^m)^{2^m} = 4$ and $(2^m)^{2^{m+1}} = 8$. In this case, the probability of false rejection is independent of a specified value of α since N assumes only two values either 0 or 2. For any $\alpha \in (0, 1)$ the null hypothesis is rejected only if $N = 0$; this occurs only whenever $x_1 = x_2 = 1$ and $x_0 = 0$ with probability $\nu^2(1 - \nu)$ or if $x_1 = x_2 = 0$ and $x_0 = 1$ with probability $(1 - \nu)^2\nu$. Hence, for any $\alpha \in (0, 1]$ the resulting probability of rejection is given by $p_r(\nu) = \nu^2(1 - \nu) + (1 - \nu)^2\nu = \nu(1 - \nu)$, depending only on ν . The function $p_r(\nu)$ has domain $[0, 1]$, is concave, and is symmetric around the point $\nu = 1/2$ where it achieves its maximum of $1/4$. If one considers next the case where the observed vector comes from a different distribution, say a Bernoulli random variable with the probability of a 1 given by t , then the power of the test is also independent of α and is given by $1 - \beta = \nu^2(1 - t) + (1 - \nu)^2t$, where β is the probability of a type II error (false acceptance). The power of the test depends only on the values of ν and t . Figure 2(a) shows $p_r(\nu)$ as a function of ν and Fig. 2(b) shows $1 - \beta$ as a function of t for various values of ν . For each value ν_0 the function $1 - \beta$ is a straight line in t with y intercept ν_0^2 and slope $(1 - 2\nu_0)$. Figure 2 illustrates the behavior of the power function. When $\nu < 1/2$, $1 - \beta$ increases as t increases for $t > \nu$ and $1 - \beta$ decreases as t decreases for $t < \nu$ hence the increase in power occurs as t and ν become more dissimilar for the larger portion of the range of t . When $\nu > 1/2$, $1 - \beta$ increases as t decreases for $t < \nu$ and $1 - \beta$ decreases as t increases for $t > \nu$ hence the increase in power occurs as t and ν become more dissimilar for the larger portion of the range of t . In either case as ν becomes small or large ($p_r(\nu)$ small) the range of t for which power decreases as t and ν become more dissimilar is very small. Naturally for $\nu = 1/2$ (all vectors equally likely) $1 - \beta = p_r(\nu) = 1/4$ for all values of t , that is, the power function is a constant equaling the probability of false rejection for all values of t .

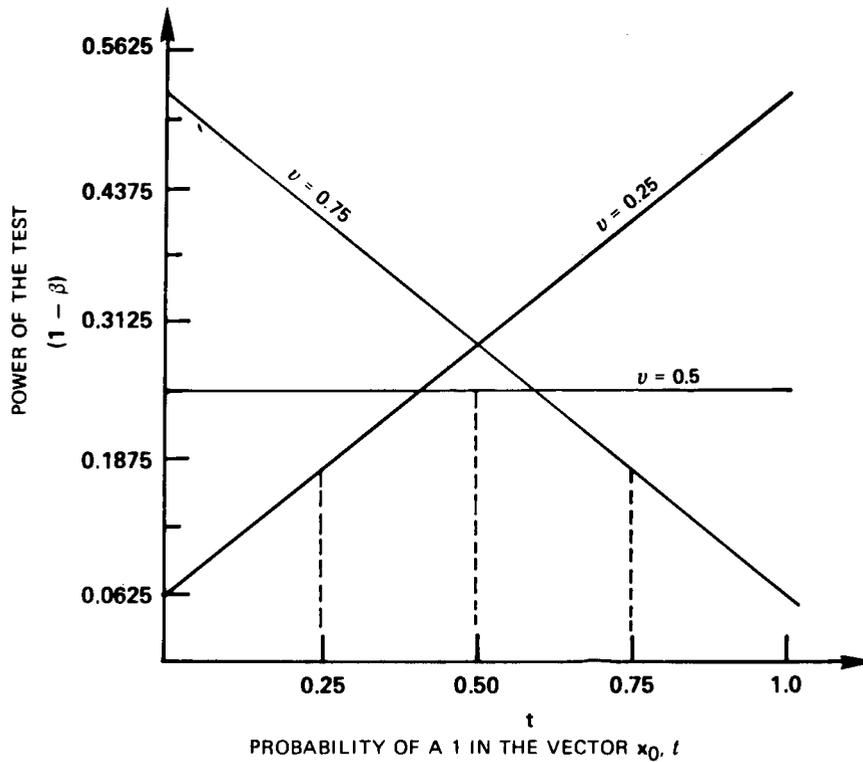
Table 2 — Test Outcomes for Case $m = 1$
Specified Data Structure

| Simulated vectors x_g | | Estimated value of ν | Likelihood of vector x_g $p_g(x)$ | | Likelihood of observed vector x_o q | | Number of cases less than q N | |
|----------------------------|---------|--------------------------|--|---------|--|-----------|--------------------------------------|-----------|
| $g = 1$ | $g = 2$ | $\hat{\nu}$ | $g = 1$ | $g = 2$ | $x_o = 0$ | $x_o = 1$ | $x_o = 0$ | $x_o = 1$ |
| 1 | 1 | 3/4 | 3/4 | 3/4 | 1/4 | 3/4 | 0 | 2 |
| 1 | 0 | 1/2 | 1/2 | 1/2 | 1/2 | 1/2 | 2 | 2 |
| 0 | 1 | 1/2 | 1/2 | 1/2 | 1/2 | 1/2 | 2 | 2 |
| 0 | 0 | 1/4 | 3/4 | 3/4 | 3/4 | 1/4 | 2 | 0 |

For larger values of m the number of elementary outcomes to consider increases very rapidly yet one can discern a pattern from the cases $m = 1, 2$, and 3. Table I listed all possible outcomes for the case $m = 1$. Similarly one may list all possible outcomes for the case $m = 2$. There are 1024 of these. Following the rules of the test one may evaluate the value of N for each outcome. N assumes only the integer values 0, 1, 2, and 4. For the case $m = 3$ enough outcomes were listed until it became obvious that N assumes only the integer values 0, 1, 2, 3, 4, 5, 6, and 8. The integral scale suggests a partition of $(0, 1)$ into overlapping subintervals where α may assume its values. For the case $m = 2$ the test indicates that for $\alpha \in (0, 1/4]$ one rejects the null hypothesis if and only if $N = 0$; for $\alpha \in (1/4, 1/2]$



(a)



(b)

Fig. 2 — Test characteristics for case $m = 1$. (a) Probability of false rejection.

(b) Power of the test for various values of v .

one rejects if and only if $N \leq 1$ ($N = 0$ or 1) and for $\alpha \in (1/2, 1)$ one rejects if and only if $M \leq 2$ ($M = 0, 1$ or 2). The case $\alpha = 1$ is not realistic and should be ignored as a trivality. For $m = 2$ one considers their 3 subintervals of $(0, 1)$ where α may assume its values. By listing the 1024 elementary outcomes and their corresponding values of N one may single out those events that correspond to a rejection of the null hypothesis for each of the 3 subintervals where α assumes its values. The events for which $N = 0$ are rejection events for $\alpha \in (0, 1/4]$, the events for which $N = 0$ or 1 are rejection events for $\alpha \in (1/4, 1/2]$ and the events for which $N = 0, 1$ or 2 are rejection events for $\alpha \in (1/2, 1)$. One associates to each event its corresponding probability and add the event probabilities corresponding to each subinterval where α assumes values. This means that for values of α in the intervals $I_1 = (0, 1/4]$, $I_2 = (1/4, 1/2]$ and $I_3 = (1/2, 1)$ there correspond three different probability of rejection functions $p_{r_1}(\nu)$, $p_{r_2}(\nu)$ and $p_{r_3}(\nu)$ depending only on ν . These functions have range $[0, h_1]$, $[0, h_2]$ and $[0, h_3]$ respectively where $h_1 \in I_1$, $h_2 \in I_2$ and $h_3 \in I_3$. The functions $p_{r_2}(\nu)$ and $p_{r_3}(\nu)$ corresponding to the larger values of α are concave and symmetric around $\nu = 1/2$ where they achieve their maxima h_2 and h_3 . As α gets smaller, in this case $\alpha \in I_1$, the function $p_r(\nu)$ begins to behave in a different manner. It remains symmetric around $\nu = 1/2$ but in this instance becomes bimodal. It jumps to a quicker maximum achieved at about $\nu = .3$ (also $\nu = .7$) and remains fairly close to its maximum for values of ν between $\nu = .3$ and $\nu = .7$. These functions are plotted in Appendix C.

For values of $m > 1$ the number of nonoverlapping subintervals covering $(0, 1]$ where α may be specified is given by $2^m - 1$. The rightmost subinterval always has length $1/2^{m-1}$. All others have length $1/2^m$. The length of the subintervals decreases rapidly and to each one there corresponds a unique probability of rejection function $p_r(\nu)$ with range $[0, h]$ where h is contained in the subinterval where α assumes its value. The functions $p_r(\nu)$ depend only on ν and are symmetric around $\nu = 1/2$. As m increases $h \rightarrow \alpha$ and it appears that for smaller α 's the functions $p_r(\nu)$ tend to achieve their maximum rapidly and remain close to this maximum for values of ν between the modes approximately a rectangular shape. This means that for large m and small α the probability of false rejection is approximately α for all values of ν except perhaps those close to 0 or 1. This behavior supports that observed from the computer evaluation. In this case as m becomes large and α becomes small the probability of rejection approaches α .

Appendix C contains plots of the different probability of rejection functions and the power functions for the case $m = 2$.

SUMMARY

Both the Monte Carlo evaluation and the analytical approach indicate that the statistical test developed by Johnson and Wiener is a valuable tool that accomplishes its objective as a technique for validating the simulation model. In evaluating the properties of the test several questions have arisen that remain open for further investigation. However, there is sufficient evidence to suggest that the test technique is a good discriminating tool that may be applicable to simulation models resulting in binary time series output in general.

REFERENCES

1. R.E. Johnson and H.L. Wiener, "Validation of a Monte Carlo Simulation of Binary Time Series," Unpublished, Systems Research Branch, NRL, October 1977.
2. R.R. Bahadur, "A Representation of the Joint Distribution of Responses to n Dichotomous Items," in *Studies in Item Analysis and Prediction*, ed. by H. Solomon (Stanford University Press, Palo Alto, Calif., 1961), pp. 158-168.
3. P.F. Lazarsfeld, *Some Observations on Dichotomous Systems*. Sociology Department Report (Columbia University, New York, 1956).

Appendix A

VALIDATION METHODOLOGY

The simulation model in question was designed to estimate the performance of an acoustic sensor in the detection of a target in the ocean. The target's acoustic characteristics and the target track, i.e., the history of the geographical positions of the target at each hour, are provided as input to the simulation model. The model operates by replicating this track many times. During a particular replication, at each hour, the model decides that the sensor either is detecting or is not detecting the target, based on factors such as sensor-target range and geometry; the acoustic properties of the ocean in the sensor-target vicinity; sensor alertment due to possible detections at previous hours of this replication; and the magnitude of a Gauss-Markov fluctuation approximated by summing three independent Ehrenfest random walk terms (Feller [A1]) at each hour. Thus for a target track lasting m hours, a single replication by the model produces a vector of m elements, each element being either 1 or 0, where 1 indicates the sensor is detecting and 0 not detecting the target at a particular hour. It is clear that a high degree of correlation exists between the detection events occurring at consecutive hours. This is because the probability of detection of a given target by a given sensor is a strong function of the target's geographical position. In addition there is an alertment effect, because of prior detections, that must be considered. Moreover, it would be most unusual for the same probability of detection to obtain at all points of a target track. Since a moving target is changing its position and aspect relative to the sensor as time progresses, no steady state will be reached in the finite duration of a target track.

The recorded history of detections of a given target by a given acoustic sensor must be considered as a sample of size one from a population with an unknown statistical distribution. It is a random binary vector of 0's and 1's with m elements. A factor affecting the selection of the validation technique is the lack of data concerning realizations of the simulated process. Multiple realizations of the same target track are impossible to obtain for most targets and sensors of interest. Hence one is testing the validity of the simulation model as a representation of the statistical structure underlying the recorded history of detections. That is, the hypothesis being tested is that the unknown statistical distributions of which the observed history of detections constitutes a single sample point is the same as the unknown statistical distribution of which the models replications constitute many sample points.

As a particular target traverses the surveillance zone it generates a track history of detections for each sensor in the zone. For each sensor i there is an observed vector $\hat{x}_i = (\hat{x}_{i1}, \dots, \hat{x}_{im})$ of 0's and 1's from an unknown probability distribution p_o^i and the simulation generates n vectors $x = (x_1, \dots, x_m)$ of 0's and 1's from a probability distribution p_i . Using the n generated vectors the statistical technique obtains an estimate \hat{p}_i of p_i . The simulated data are then applied to \hat{p}_i to obtain the sample distribution of the n vectors as an approximation to the population distribution. The test consists of determining whether the observed vector \hat{x}_i has \hat{p}_i -value in the upper $1 - \alpha$ region of the sample distribution. If it does then the hypothesis that \hat{x}_i comes from the distribution p_i is accepted. The test is applied for all i and many acceptances that \hat{x}_i is from p_i will confirm the null hypothesis that p_i is a good approximation to p_o^i , thereby validating the simulation. It is to be expected that due to statistical fluctuation some sensors will fail the test. Hence a distinction must be made between validation of the simulation model in general and specific statements about the simulation of the individual sensors. Statements about the latter are simply understood to carry the uncertainty inherent in the statistical test itself.

The statistical hypothesis test uses the representation by Bahadur [A2] and Lazarsfeld [A3] for the probability distribution underlying binary sequences, which is summarized as follows. Assume the target track is m hours long. Let X be the set of all points $\hat{x} = (x_1, x_2, \dots, x_m)$ with each $x_i = 0$ or 1,

and suppose $p(\underline{\hat{x}})$ is a probability distribution on the elements of X ; that is, $p(\underline{\hat{x}}) \geq 0$ for all $\underline{\hat{x}} \in X$ and $\sum_{\underline{\hat{x}} \in X} p(\underline{\hat{x}}) = 1$. Let $E_p(\cdot)$ denote the expected value of the expression in parentheses when the distribution p obtains. Then let

$$v_i = E_p(x_i) \quad 0 < v_i < 1; \quad i = 1, 2, \dots, m; \quad \text{and}$$

$$z_i = (x_i - v_i) / \sqrt{v_i \cdot (1 - v_i)} \quad i = 1, 2, \dots, m.$$

Next define the family

$$r_j = E_p(z_i \cdot z_j) \quad i < j;$$

$$r_{ijk} = E_p(z_i \cdot z_j \cdot z_k) \quad i < j < k;$$

$$r_{12\dots m} = E_p(z_1 \cdot z_2 \cdot \dots \cdot z_m).$$

For $\underline{\hat{x}} = (x_1, x_2, \dots, x_m)$ define

$$p_{[1]}(\underline{\hat{x}}) = \prod_{i=1}^m v_i^{x_i} (1 - v_i)^{1 - x_i}$$

and

$$f(\underline{\hat{x}}) = 1 + \sum_{i < j} r_{ij} \cdot z_i \cdot z_j + \sum_{i < j < k} r_{ijk} \cdot z_i \cdot z_j \cdot z_k \quad (1)$$

$$+ \dots + r_{12\dots m} \cdot z_1 \cdot z_2 \cdot \dots \cdot z_m.$$

Then for each $\underline{\hat{x}}$ in X ,

$$p(\underline{\hat{x}}) = p_{[1]}(\underline{\hat{x}}) f(\underline{\hat{x}}). \quad (2)$$

Thus $p_{[1]}(\underline{\hat{x}})$ denotes the joint probability distribution of the x_i 's under an assumption that the x_i 's are independently distributed, and $f(\underline{\hat{x}})$ represents the effects of correlation.

In this representation it is natural to refer to the parameters r_{ij} as second order correlations, to the parameters r_{ijk} as third order correlations, and so forth, culminating in $r_{12\dots m}$, the m -th order correlation. The distribution p then is said to have order s if one correlation of order s is non-zero and all correlations of order greater than s are equal to zero. If a distribution is known to be of a certain order s , then the representation (1) need only extend to correlations of order s or less.

In some applications either the nature of the situation being studied or computational problems might make it necessary to assume a specified order to the distribution, even though the value of that order cannot be known precisely. If the selection is in error, then the "truncated" form of expression (1) will be in error and so will the resulting values of $f(\underline{x})$ and $p(\underline{x})$. As defined by (6), the estimated $p(\underline{x})$ may not be a probability distribution and may assume negative values for some \underline{x} . This point is discussed by Bahadur [A2]. In addition, to obtain the estimate \hat{p} of (2) one must first obtain the estimate \hat{f} of (1). The statistical fluctuation associated with the values $\hat{f}(\underline{x})$ is another source of error leading to negative values of $\hat{f}(\underline{x})$ for some \underline{x} and consequently for $\hat{p}(\underline{x})$. Because \hat{p} may not be a probability distribution, values taken by \hat{p} are referred to as *likelihood* values.

In the application to the APSURV simulation several considerations led to truncating the form of (1). A fourth-order approximation to the Bahadur-Lazarsfeld representation has been employed, truncating the expression in (1) after the fourth-order correlations and using a time window of twelve hours, that is, assuming zero correlation between time steps more than twelve hours apart. From observations of the detection process it seemed reasonable to assume both that the correlation between epochs separated by more than twelve hours is negligible compared to the correlations between epochs closer together, and also that the contribution of correlations of order greater than four is relatively insignificant. The truncation has also been necessary in order to keep the computer costs within reason.

The observed realization compared with the simulation results consists of a track history of duration m hours together with the associated detection history. The simulation model is programmed to run with input parameters characterizing the observed situation. Then n replications of the model are run, each producing a time series (vector) $\hat{x}_g = (x_{g1}, x_{g2}, \dots, x_{gm})$, $g = 1, 2, \dots, n$, of m binary elements, where 1 denotes a detection and 0 no detection of the target by the acoustic sensor. Actual values of m and n used were $m = 240$ and $n = 50$. The n binary vectors are used to estimate the parameters in the Bahadur-Lazarsfeld representation of the probability distribution corresponding to the population from which the n sample vectors are generated. Simple unbiased estimators were chosen for all parameters. The estimators for the v_i 's are maximum likelihood when the distribution $p_{\{1\}}$ obtains, that is, when the x_{gi} 's are independently distributed. Since the v_i 's are assumed to be neither 0 nor 1 a reasonable correction is made should the data seem to indicate they are. The estimates are obtained by:

$$\tilde{v}_i = \begin{cases} 1/2n & \text{if } \sum_{g=1}^n x_{gi} = 0 \\ 1 - (1/2n) & \text{if } \sum_{g=1}^n x_{gi} = n \\ (1/n) \sum_{g=1}^n x_{gi} & \text{otherwise for } i = 1, 2, \dots, m; \end{cases} \quad (3)$$

$$z_{gi} = (x_{gi} - \tilde{v}_i) / \sqrt{\tilde{v}_i \cdot (1 - \tilde{v}_i)} \quad i = 1, 2, \dots, m, \quad (4)$$

$$g = 1, 2, \dots, n;$$

and

$$\begin{aligned} \tilde{r}_{ij} &= (1/n) \sum_{g=1}^n z_{gi} \cdot z_{gj}, & 1 \leq i < j \leq m; \\ \tilde{r}_{ijk} &= (1/n) \sum_{g=1}^n z_{gi} \cdot z_{gj} \cdot z_{gk}, & 1 \leq i < j < k \leq m; \\ &\dots & \\ \tilde{r}_{12\dots m} &= (1/n) \sum_{g=1}^n z_{g1} \cdot z_{g2} \cdot \dots \cdot z_{gm}. \end{aligned} \quad (5)$$

The likelihood p_g of the g -th model replication x_g is given by

$$p_g = p(\hat{x}_g) = f(\hat{x}_g) \cdot [\prod_{i=1}^m \tilde{v}_i^{x_{gi}} \cdot (1 - \tilde{v}_i)^{1-x_{gi}}], \quad g = 1, 2, \dots, n,$$

with $f(\hat{x}_g)$ as characterized by (1) and the z -values as given by (4). Then under the hypothesis that the random mechanism for the simulation is a model for the random mechanism underlying the observed sensor's detections, the recorded sequence of detections of that target by the specified sensor, the binary vector $\underline{x} = (x_1, x_2, \dots, x_m)$, has likelihood $q = p(\underline{x})$, relative to the Bahadur-Lazarsfeld representation of the n model replications given by (2) using the \tilde{v} 's and \tilde{r} 's computed by (3) and (5) from the model replications. Once the numbers $q = p(\hat{x})$ and $\{p_g = p(\hat{x}_g): g = 1, 2, \dots, n\}$ have been obtained, it can be determined whether to accept the simulation model as valid at a significance level α . The test procedure (a straightforward rank test) is to reject the hypothesis of association if the observed value q falls below the α -th percentile of computed values p_g . Define N to be the number of elements in the set $\{p_g: p_g \leq q, 1 \leq g \leq n\}$. Then if $N \geq n\alpha$, the simulation model is determined to be valid in predicting the performance of the specified acoustic sensor in detecting that target. The model is rejected as not valid at significance level α if $N < n\alpha$. The reason a one-sided test is used here is that the higher the likelihood of the recorded detection history relative to the Bahadur-Lazarsfeld representation of the model replications, the better the agreement is between the recorded detection history and the simulation output. Hence one need only be concerned with rejecting the simulation model if the likelihood of the recorded detection sequence is low relative to the likelihoods of the model replications.

Finally it should be mentioned that although Johnson and Wiener [A4] could not know the structure of the probability distribution they were working with, in particular the exact structure of the correlation function, they did address the problem of establishing the adequacy of the number of replications used (values of n) via a Smirnow two-sided goodness-of-fit test. The distribution of likelihoods

(relative to their respective Bahadur representations) of three independent sets of fifty replications for the same target and acoustic sensor were generated. The Smirnov two-sided test for goodness-of-fit (at the 0.20 significance level) indicated that the three samples could be assumed to have come from the same population. From this it was concluded that fifty replications of the model are sufficient for validation purposes. In a similar manner three independent sets of twenty replications each for the same target and acoustic sensor were generated. Although these samples passed the Smirnov two-sided goodness-of-fit test at the 0.10 significance level, at the 0.20 significance level the test indicated the three samples did not come from the same population. From this it was concluded that the variability between samples of only 20 replications was too great to permit their used in a validation.

REFERENCES

- A1. W. Feller, *An Introduction to Probability Theory and its Applications*, (Wiley, New York, 1968), Vol. I.
- A2. R.R. Bahadur, "A representation of the joint the responses to n dichotomous times," *Studies in Item Analysis and Prediction*, ed by H. Solomon, (Stanford University Press, Palo Alto, Calif., 1961), pp. 158-168.
- A3. P.F. Lazarsfeld, *Some Observations on Dichotomous Systems*. Sociology Department Report (Columbia University, New York, 1956).
- A4. R.E. Johnson and H.L. Wiener, "Validation of a Monte Carlo Simulation of Binary Time Series," Unpublished, Systems Research Branch, NRL, October 1977.

Appendix B

MONTE CARLO EVALUATION

In this appendix a detailed presentation is given of the data collected in performing an evaluation of the test technique developed by Johnson and Wiener to validate the APSURV MOD 1.4 simulation model.

Three factors were investigated, first the ability of the test to generate a probability of false rejection $\hat{\alpha}$ which approximate the assumed specified value α ; second the influence of computed values $\hat{f}(\underline{x})$ of the correlation term $f(\underline{x})$ in Eq. (6) of appendix A; and third the power of the test in correctly rejecting vectors which did not come from the same population as did the original set.

The approach taken used sets of $n + 1$ random binary vectors which were repeatedly generated from known distributions. Several cases were considered by varying the vector length m and by varying u for each m . The vector elements in all cases were independent Bernoulli random variables where the probability of a 1 varied by vector element. The various cases of m considered were 2, 3, 4, 5 and 6 vector elements. For the case $m = 6$ the probabilities of a 1 for all vector elements were (.2, .4, .6, .8, .8, .8) respectively. Similarly one used for $m = 5$ the probability vector (.2, .4, .6, .8, .8), for $m = 4$ (.2, .4, .6, .8), for $m = 3$ (.2, .4, .6) and for $m = 2$ (.2, .4). For all of these distributions the correlation function has a theoretical value of $f(\underline{x}) = 1$. Two sets of data were obtained. One set used the theoretical values $f(\underline{x}) = 1$ the other the computed estimates $\hat{f}(\underline{x})$. For the case $m = 5$ tests were also conducted where the $n+1$ -st vector was generated from a known distribution other than the distribution from which the first n vectors were generated. This alternate distributions consisted of independent, identically distributed vector elements where the probability of a 1 was given by $p = .1$ in one case and $p = .5$ in another.

In a typical case, for example $m = 3$ and $n = 20$, with a random number subroutine 21 vectors of length 3 are generated and the test applied to determine whether the 21-st vector belongs to the same population from which the first 20 come. The decision is made to reject or not to reject this hypothesis. The procedure is repeated 20 times and the proportion of rejections is computed. In turn the sets of 20 are repeated 100 times. The one hundred computed proportion of rejections are then treated as 100 independent, identically distributed random variables whose mean is an estimate of the probability of false rejection resulting from the application of the test. When the $n+1$ -st vector is generated from a specified alternate distribution the proportion of rejections in this case estimates the power of the test, that is, the probability that the null hypothesis is correctly rejected.

Tables B.1 and B.2 summarize the data collected in the Monte Carlo evaluation of the test. Table B.1 consists of values obtained with $f(\underline{x}) = 1$ and Table B.2 the values obtained with $f(\underline{x}) = \hat{f}(\underline{x})$, as computed from the data. As usual, m denotes the vector length and n the number of replications from which the Bahadur-Lazarsfeld representation is obtained for each distribution. Each case $(m, n, 1)$ or (m, n, \hat{f}) yields an estimate α . This estimate is the average proportion of rejections computed from k subsets each of size z . The product $k \times z$ represents the total number of times the statistical test is performed for each case $(m, n, 1)$ or (m, n, \hat{f}) . The computed sample standard deviation of the estimate is denoted s . The factor s/\sqrt{k} is used to obtain confidence intervals for α using percentage points of the Student-t distribution. For each vector length m the size of the population of such vectors is 2^m possible binary vectors. In order to obtain a uniform measure of the relative size of the number of replications from which the Bahadur-Lazarsfeld representation is obtained for each case (n, m) , the ratio $n/2^m$ is used. These ratios are listed along with the data in Tables B1 and B2. In Table B2 the

Table B1 — Test specified for $\alpha = 0.3$
 Data obtained with $f(x) = 1$

| Vector Length m | No. of Subsets size of each subset k/z | Number of replication per case n | $(n/2^m)$ | Estimate of α $\hat{\alpha}$ | Sample standard deviation s |
|----------------------|---|---|-----------|---|--|
| 2 | 100/20 | 20 | 5 | .1900 | .098 |
| 2 | 100/20 | 50 | 12.5 | .2015 | .0963 |
| 2 | 100/20 | 100 | 25 | .1940 | .0799 |
| 2 | 100/20 | 200 | 50 | .1910 | .0880 |
| 3 | 100/20 | 20 | 2.5 | .2655 | .0892 |
| 3 | 100/20 | 50 | 6.25 | .2765 | .1011 |
| 3 | 100/20 | 100 | 12.5 | .2795 | .1054 |
| 3 | 100/20 | 200 | 25 | .2770 | .0968 |
| 3 | 40/10 | 400 | 50 | .2300 | .1436 |
| 4 | 100/20 | 20 | 1.25 | .2935 | .1079 |
| 4 | 100/20 | 50 | 3.125 | .2770 | .0988 |
| 4 | 100/20 | 100 | 6.25 | .2670 | .0972 |
| 4 | 100/20 | 200 | 12.5 | .2920 | .1167 |
| 4 | 40/10 | 400 | 25 | .2725 | .1281 |
| 5 | 100/20 | 20 | .625 | .3395 | .1142 |
| 5 | 100/20 | 50 | 1.56 | .3090 | .1021 |
| 5 | 100/20 | 100 | 3.125 | .3280 | .1081 |
| 5 | 100/20 | 200 | 6.25 | .2985 | .0981 |
| 5 | 40/10 | 400 | 12.5 | .2800 | .1652 |
| 5 | 20/10 | 800 | 25 | .2900 | .1619 |
| 6 | 100/20 | 100 | 1.56 | .2975 | .1003 |
| 6 | 100/20 | 200 | 3.125 | .3095 | .1053 |
| 6 | 40/10 | 400 | 6.25 | .2700 | .1381 |
| 6 | 20/10 | 800 | 12.5 | .3150 | .1424 |

number of the vectors resulting with negative likelihoods for each run is denoted ω and the number of these which were the observed vector is denoted ω_0 . Figure B1 presents a comparison of estimates $\hat{\alpha}$ as a function of $(n/2^m)$ for $f(x) = 1$ and $f(x) = \hat{f}$ where $m = 5$. It can be observed that as $(n/2^5)$ increases the estimates $\hat{\alpha}$ for the case $f(x) = \hat{f}$ approach the estimates α for the case $f(x) = 1$, both approaching the region $\hat{\alpha} = .3$. In the case $f(x) = \hat{f}$ this is accompanied by a decrease in the resulting number of negative vectors, as is indicated in Table B2. Based on a comparison with the case $f(x) = 1$, the values $n/2^5$ where the estimates $\hat{\alpha}$ for the case $f(x) = \hat{f}$ are considered adequate are 6.25, 12.5 and 25. The resulting negative vectors in these cases were $\omega = 7$, $\omega = 4$ and $\omega = 1$ respectively as compared to $\omega = 273$, $\omega = 121$ and $\omega = 65$ for the cases $(n/2^5) = .625$, $(n/2^5) = 1.56$ and $(n/2^5) = 3.125$ respectively. It should be noted that for these last three cases the value of $k \times z$ is equal to 2000 that is, a total of 2000 sample points, whereas for the previous three cases there were only 1200, 400 and 200 sample points ($k \times z$) respectively. A simple calculation yields $7/1200 = .006$, $4/400 = .01$ and $1/200 = .005$ negative vectors per sample point. Multiplying by 2000 yields 11.67, 20 and 10 negative vectors respectively, still a small number compared to the other three cases hence the number of negative vectors does decrease as n increases.

Table B2 — Results of Simulation Experiments
Test specified for $\alpha = 0.3$
Data Obtained with $f(x) = \hat{f}$

| Vector Length m | No. of Subsets Size of each subset k/z | Number of replications per case n | $(n/2^m)$ | Estimate of α $\hat{\alpha}$ | Sample Standard Deviation s | Vectors with negative likelihood w | Number of Observed vectors with negative likelihood w_o |
|----------------------|--|--|-----------|--|----------------------------------|---|--|
| 2 | 100/20 | 20 | 5 | .1950 | .0973 | 11 | 11 |
| 2 | 100/20 | 50 | 12.5 | .2030 | .0982 | 0 | 0 |
| 2 | 100/20 | 100 | 25 | .1945 | .0794 | 0 | 0 |
| 2 | 100/20 | 200 | 50 | .1910 | .0880 | 0 | 0 |
| 3 | 100/20 | 20 | 2.5 | .2980 | .0951 | 18 | 18 |
| 3 | 100/20 | 50 | 6.25 | .2835 | .0959 | 0 | 0 |
| 3 | 100/20 | 100 | 12.5 | .2795 | .0985 | 0 | 0 |
| 3 | 100/20 | 200 | 25 | .2795 | .0980 | 0 | 0 |
| 3 | 40/10 | 400 | 50 | .2600 | .1464 | 0 | 0 |
| 4 | 100/20 | 20 | 1.25 | .3895 | .0936 | 69 | 69 |
| 4 | 100/20 | 50 | 3.125 | .3130 | .1058 | 0 | 0 |
| 4 | 100/20 | 100 | 6.25 | .2855 | .1003 | 0 | 0 |
| 4 | 100/20 | 200 | 12.5 | .3060 | .1162 | 0 | 0 |
| 4 | 40/10 | 400 | 25 | .2925 | .1403 | 0 | 0 |
| 5 | 100/20 | 20 | .625 | .4785 | .1231 | 273 | 273 |
| 5 | 100/20 | 50 | 1.56 | .3715 | .1179 | 121 | 121 |
| 5 | 100/20 | 100 | 3.125 | .3600 | .1094 | 65 | 65 |
| 5 | 60/20 | 200 | 6.25 | .3233 | .1006 | 7 | 7 |
| 5 | 40/10 | 400 | 12.5 | .3075 | .1607 | 4 | 1 |
| 5 | 20/10 | 800 | 25 | .2850 | .1631 | 1 | 1 |
| 6 | 100/20 | 100 | 1.56 | .3835 | .1010 | 113 | 113 |
| 6 | 50/20 | 200 | 3.125 | .3440 | .1067 | 31 | 27 |
| 6 | 40/10 | 400 | 6.25 | .3000 | .1468 | 14 | 4 |
| 6 | 10/10 | 800 | 12.5 | .3400 | .1578 | 9 | 1 |

Vectors with negative likelihoods indicate an unstable correlation function. As the number of rows (n) increases the correlation function begins to settle at about its theoretical value 1 and the number of negative likelihood vectors decreases. The larger the number of outcomes (2^m) of the population from which the distribution is estimated, the larger the number of rows (n) must be for the negative values to begin to disappear. The reason is illustrated as follows:

Consider the term

$$\sum_{i < j} r_{ij} z_i z_j.$$

The number of additions or the cardinality of the set

$$\{(i,j) : i < j, i = 1, \dots, m - 1; j = 2, \dots, m\}$$

increases as m increases. The same happens for higher order correlations. Even as $r_{ij} \rightarrow 0$ fast the increased number of additions counteracts this effect somewhat as m increases. Hence for larger m 's the rate of decrease of vectors with negative likelihoods decreases even though the actual number of negative vectors continues to decrease as n increases.

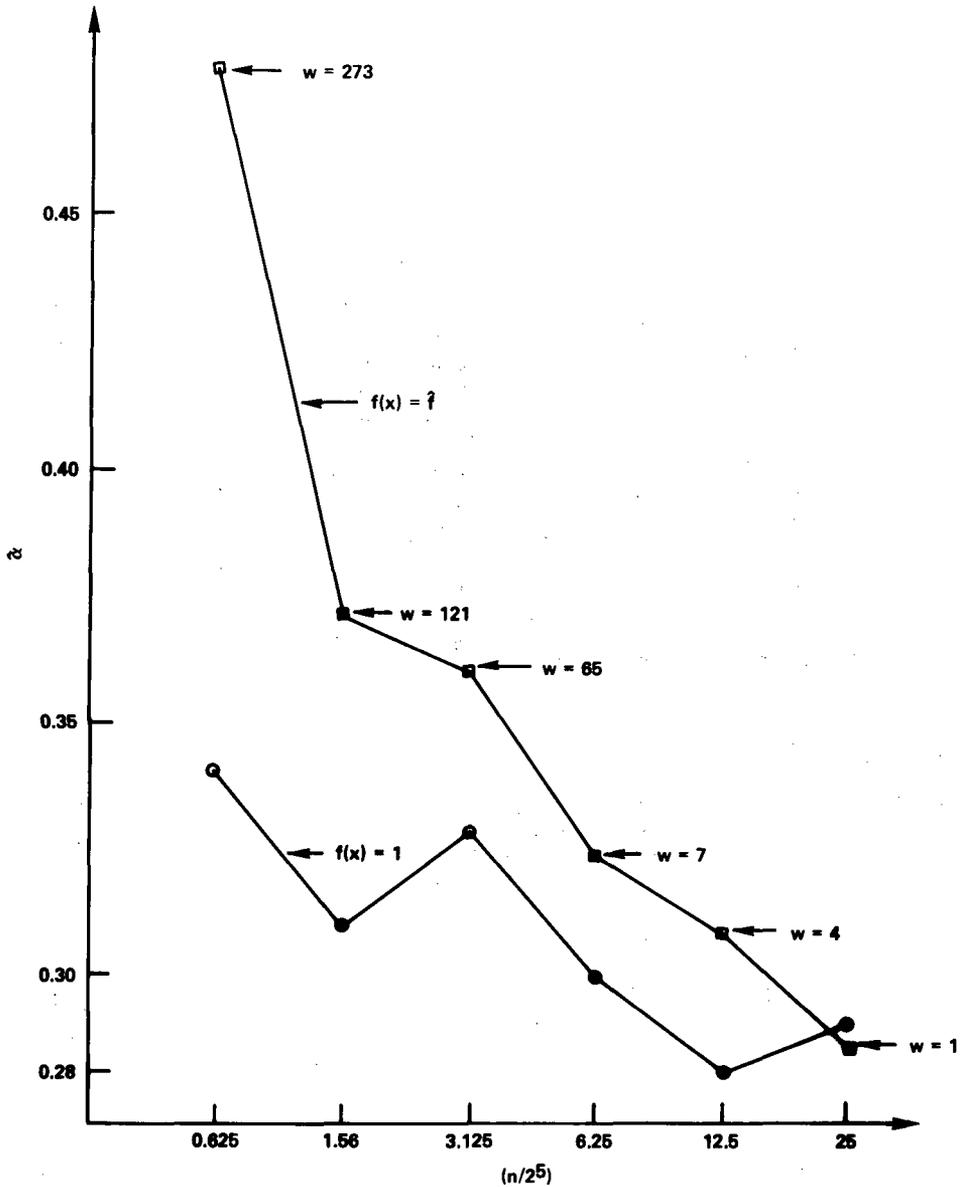


Fig. B1 — Estimates $\hat{\alpha}$ vs $(n/2^5)$ for $f(x) = 1$ and $f(x) = \hat{f}$; $m = 5$

In the procedure used whenever the observed $(n + 1)$ st vector had a negative likelihood it was assigned a very small probability and was automatically rejected. The overall rejection criterion was fixed in all cases at $\alpha = .3$. The sample estimate of the probability of rejection $\hat{\alpha}$ is very sensitive to the correlation function effects. Two effects can be distinguished: first, correlation term values that deviate significantly from $f(x) = 1$ tend to misrepresent the probability distribution; second, this same fluctuation generates a large number of negative observed vectors which inflates the proportion of rejections as such vectors are automatically rejected. Except for large ratios $(n/2^m)$ a large proportion of vectors with negative likelihood turn out to be the observed vector. This is the vector not included in obtaining the Bahadur representation. A possible explanation is that the Bahadur representation tends to represent its source. Independent of other properties of the test this is a desirable property in the sense that the test serves its purpose as a discriminating tool. All of these effects subside simultaneously as n or $(n/2^m)$ increase.

Figures B2 and B3 are plots representative of the density obtained in a Monte Carlo run for two cases of m and n estimating α . The two cases are $m = 6, n = 100, f(x) = 1$ and $m = 6, n = 400, f(x) = \hat{f}$. The respective values of k and z are 100 of 20 and 40 of 10, hence the densities are for 100 and 40 points respectively. The respective means are $\bar{\alpha} = .2975$ and $\bar{\alpha} = .3000$. The respective 95% confidence intervals using the factors s/\sqrt{k} as mentioned above, are $[\.2778, .3172]$ and $[\.2545, .3455]$. In both cases the confidence intervals contain the actual value $\alpha = 0.3$.

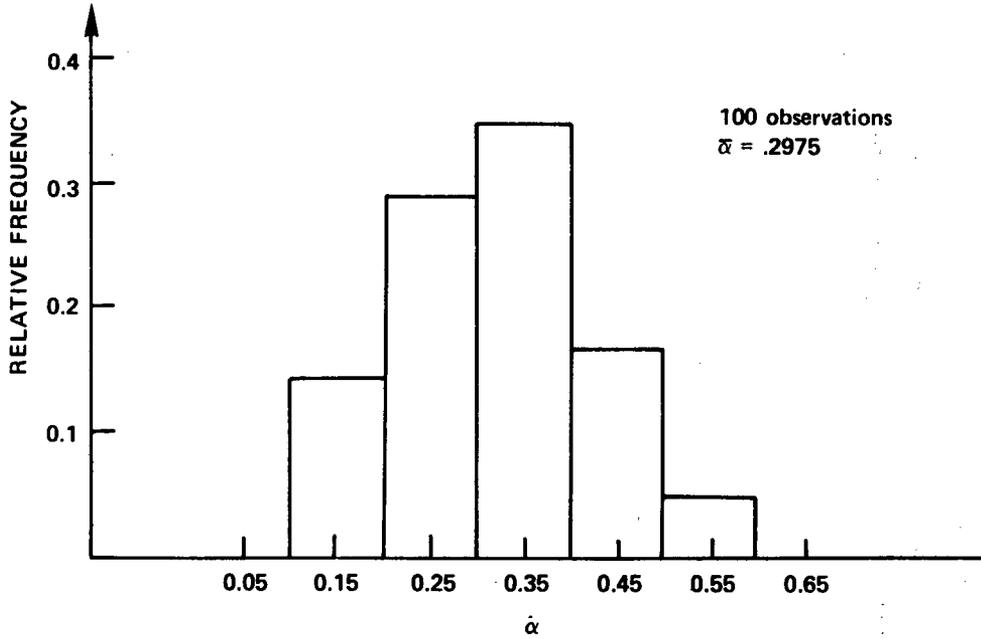


Fig. B2 — Sample density function for estimates of $\alpha, m = 6, n = 100, f(x) = 1$

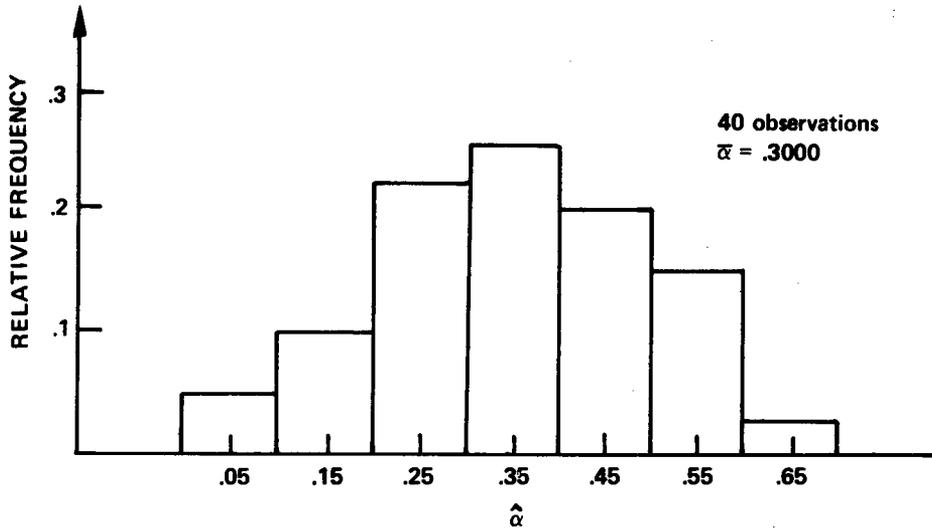


Fig. B3 — Sample density function for estimates of $\alpha, m = 6, n = 400, f(x) = \hat{f}$

In each Monte Carlo run to estimate α , the first $n \times m$ matrix of 0's and 1's was printed along with the corresponding value of $\hat{f}(\underline{x})$ for each row vector. Relative frequency histograms of the n values $\hat{f}(\underline{x})$ were plotted for each case (m, n). These plots revealed the expected result that as n increases for a specific m , the spread of values of $\hat{f}(\underline{x})$ go towards a spike at 1. An example is the case $m = 5$ and $n = 800$ plotted in Figure B4. The mean value is $\bar{f}(\underline{x}) = 1.037$ with standard deviation 0.251. Figure B5 shows a plot of values of $\bar{f}(\underline{x})$ for $m = 5$ as a function of values $n = 20, 50, 100, 200, 400$ and 800. The values $\bar{f}(\underline{x})$ consistently overshoot their theoretical value $f(\underline{x}) = 1$. For values of n greater than or equal to 200 the estimates are deemed close enough for an adequate approximation to the probability distribution they attempt to represent.

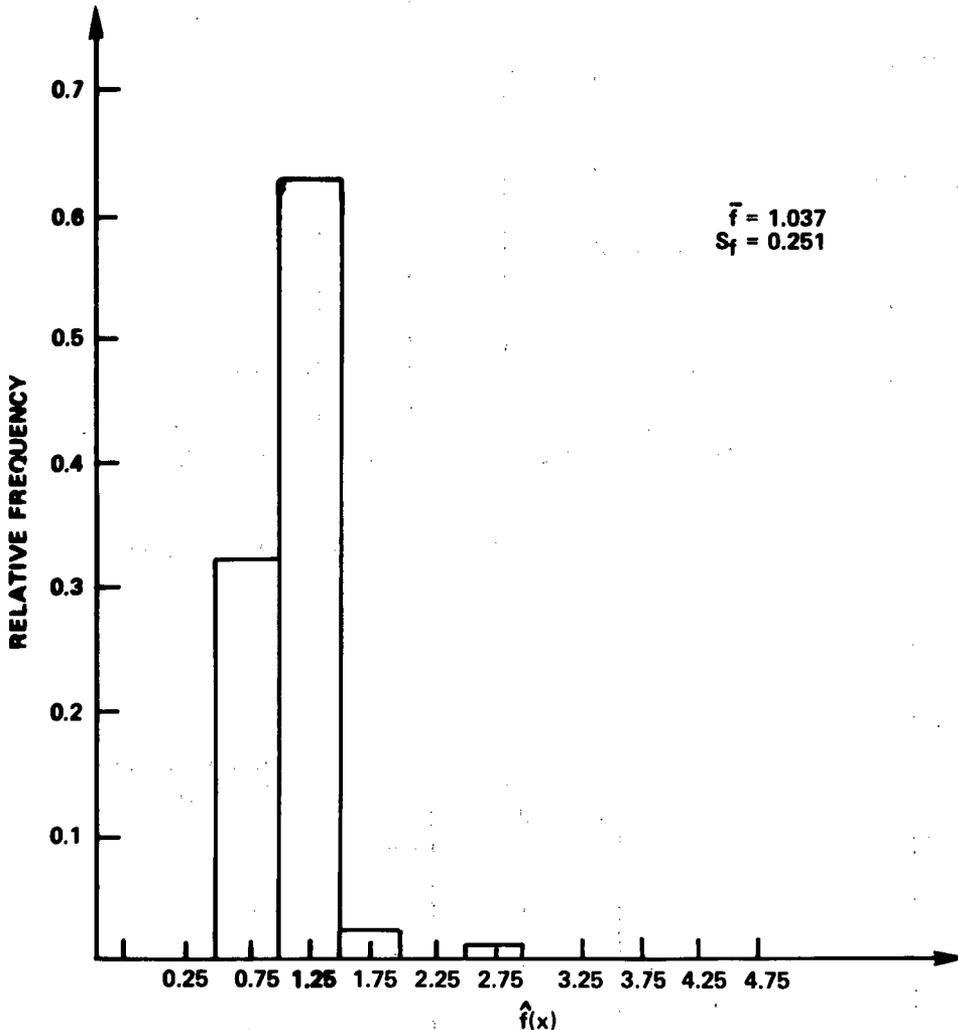


Fig. B4 — Histogram of values of $\hat{f}(\underline{x})$, $m = 5$, $n = 800$

Finally the power of the test was considered. The same Monte Carlo procedure was applied where the observed vector was now generated from an alternate distribution. Two alternate distributions were considered. The first one was one in which all vector elements were independent, identically distributed (iid), Bernoulli random variables with $p = .1$ and the second one was one where all vector elements were again iid, Bernoulli random variables with $p = .5$. For each of these distributions a Monte

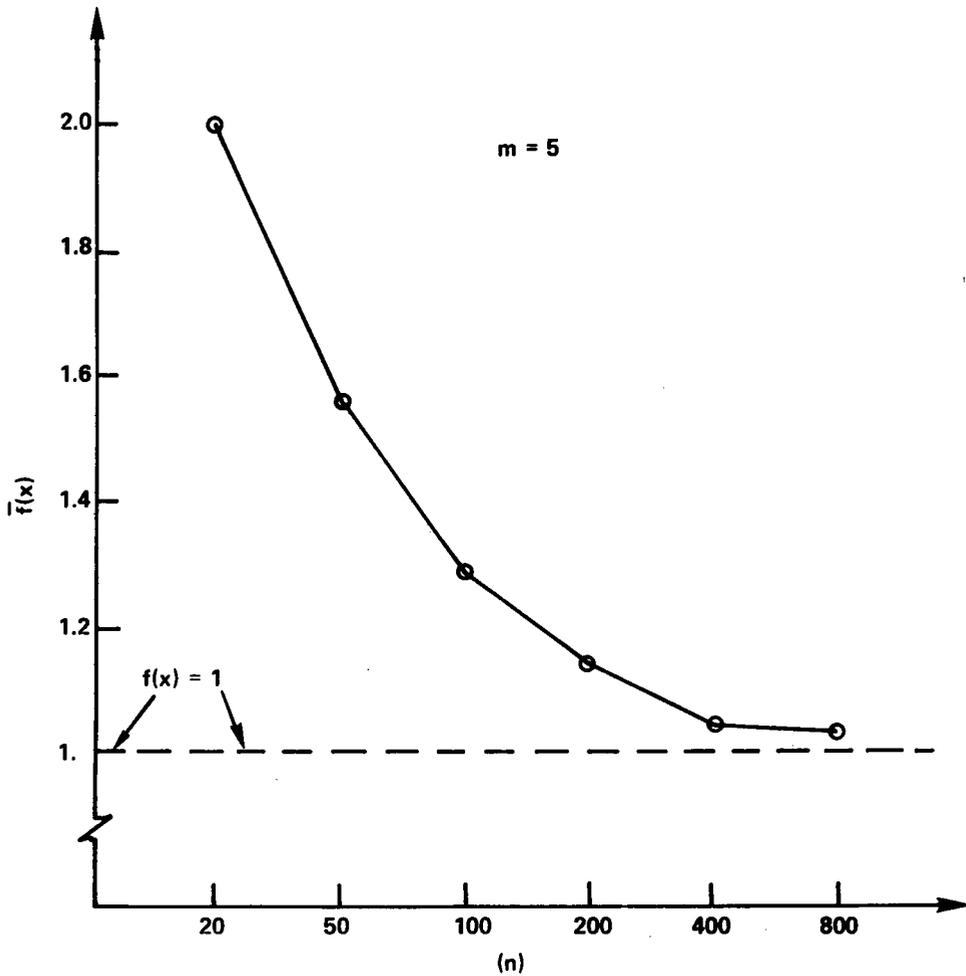


Fig. B5 - Mean values $\bar{f}(x)$ as a function of n

Carlo run was conducted first using $f(x) = 1$ and then using $f(x) = \hat{f}(x)$. The computed proportions now represent the power of the test. The case tried was $m = 5$ and $n = 200$. Not only did the test appear to discriminate well but also no appreciable difference was noted between the cases $f(x) = 1$ and $f(x) = \hat{f}(x)$. Figures B6, B7, B8 and B9 show plots of the relative frequency histograms resulting in estimating $1-\beta$ the power of the test for all cases considered, where β is the probability of a type II error, that is, of incorrectly accepting the $(n+1)$ -st vector. The figures show lower power for the case $p = 0.5$ as opposed to $p = 0.1$. For $p = 0.5$ the power is approximately 0.70; for $p = 0.1$ it is approximately 0.90. This is as it should be since $p = 0.1$ is "farther" from the simulated case than $p = 0.5$, which was closer to the simulated probability vector (0.2, 0.4, 0.6, 0.8, 0.8). The test for this two cases discriminated very well.

The Monte Carlo evaluation indicates that $\hat{\alpha}$ is approximately α for both m and n large, that \hat{f} gets close to $f(x) = 1$ as n increases and that the power of the test increases as "differences" between the base case and the $(n+1)$ -st vector increase.

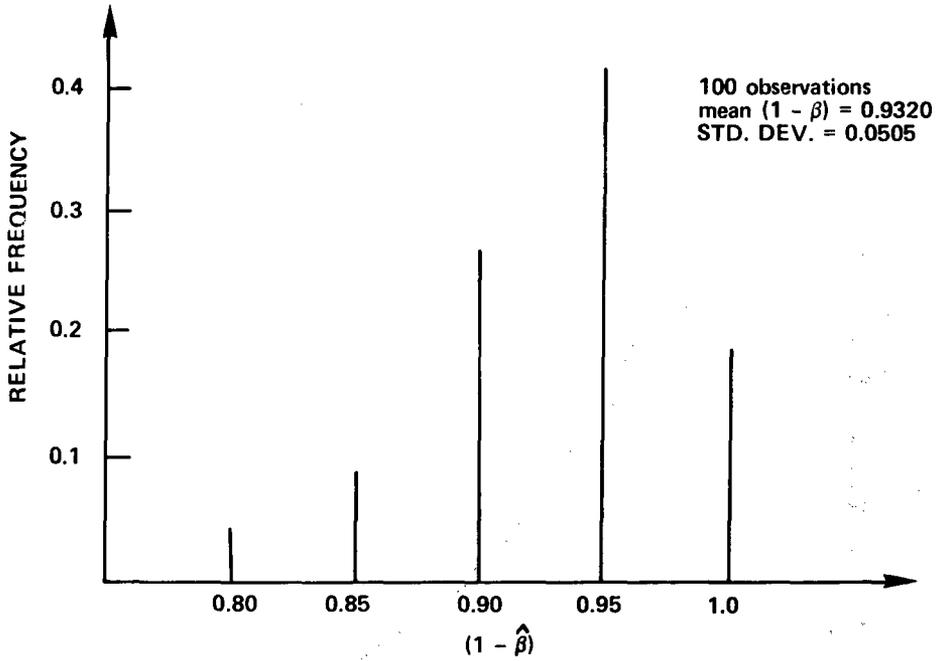


Fig. B6 — The power of the test for $m = 5$, $n = 200$, $f(\underline{x}) = 1$ and $p = .1$

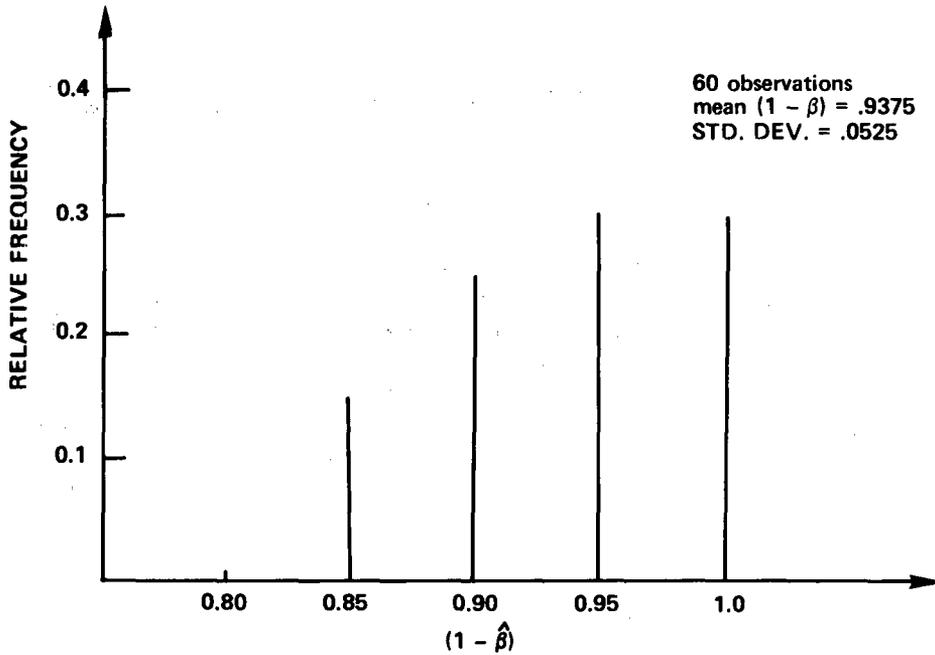


Fig. B7 — The power of the test for $m = 5$, $n = 200$, $f(\underline{x}) = \hat{f}$ and $p = .1$

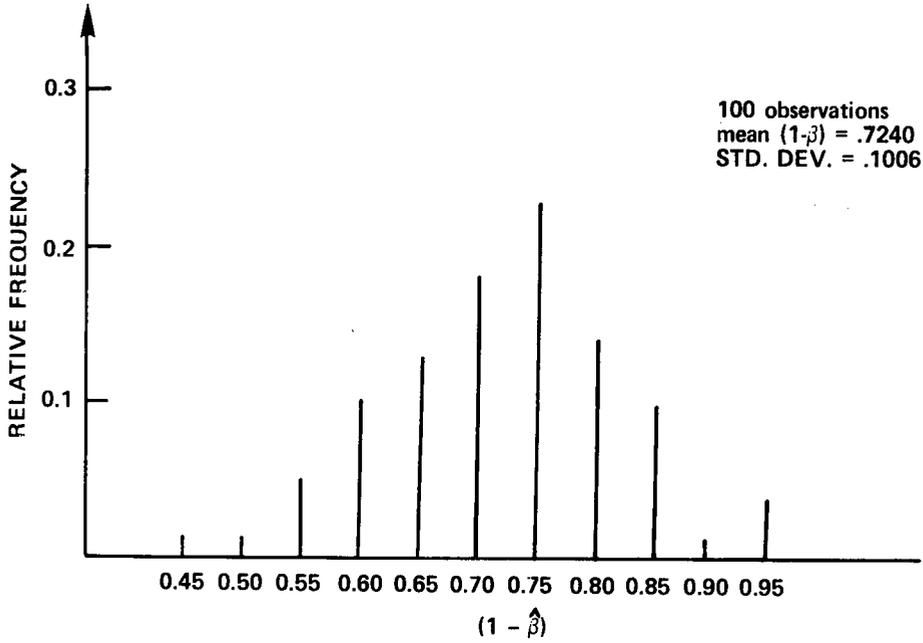


Fig. B8 — The power of the test for $m = 5$, $n = 200$, $f(\underline{x}) = 1$ and $p = .5$

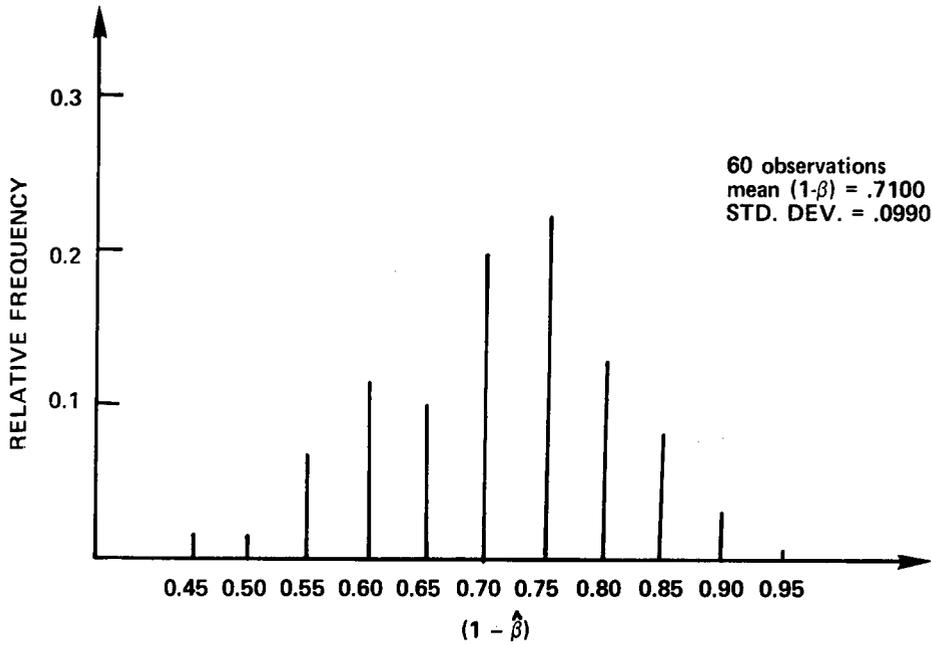


Fig. B9 — The power of the test for $m = 5$, $n = 200$, $f(\underline{x}) = 1$ and $p = .5$

Appendix C

DETAILS OF ANALYTICAL FORMULATION

The particular data structure considered in the analytical formulation of the problem shows that it may be the case that the probability of false rejection where applying the test approaches the significance level α as m the vector length, increases. For any vector length m there are 2^m possible binary outcomes or members of the population of vectors of length m . The analysis consists of letting $n = 2^m$ where all binary vector elements consist of independent, identically distributed Bernoulli random variables with the probability of a 1 given by v . There are $(2^m)^{2^m}$ possible outcomes from which the Bahadur representation may be obtained. To each of these outcomes one may associate 2^m possible $(n + 1)$ -st or observed vector yielding a total of $(2^m)^{2^{m+1}}$ elementary outcomes. The case where the observed vector comes from the same distribution as the set of first $n = 2^m$ vectors is considered first. Specifying α , evaluating N for each outcome, and applying the rules of the test one may establish an association between α and the actual probability of false rejection. Both the cases $m = 1$ and $m = 2$ were discussed in section 5. This appendix presents further results on the analytical approach. It deals specifically with the case $m = 2$. Applying the rules of the test by direct enumeration one obtains that the possible values of N are either 0, 1, 2, or 4. Here the test indicates that for $\alpha \in (0, 1/4]$ one rejects the null hypothesis if and only if $N = 0$, for $\alpha \in (1/4, 1/2]$ one rejects if and only if $N \leq 1$ ($N = 0$ or 1) and for $\alpha \in (1/2, 1)$ one rejects if and only if $N \leq 2$ ($N = 0, 1$ or 2). The case $\alpha = 1$ is not realistic and should be ignored as a triviality. By allowing the $(n + 1)$ -st vector to come from an alternate distribution in this case each vector element a Bernoulli distributed independent random variable with probability of a 1 given by t one may then analyze the resulting probability of rejecting the null hypothesis correctly, that is, the power of the test. Consider the analysis for the case $m = 2$. In this case there are 256 outcomes from which the Bahadur Lazarsfeld representation may be obtained. To each of these outcomes one may associate 4 possible observed vectors yielding a total of 1024 elementary outcomes to be considered. These were listed in detail and the functions $p_{r_1}(v)$, $p_{r_2}(v)$ and $p_{r_3}(v)$ were obtained along with their corresponding power functions $1 - \beta_1$, $1 - \beta_2$ and $1 - \beta_3$. The value of N may be 0, 1, 2 or 4, yielding three intervals in which α may assume values, $I_1 = (0, 1/4]$, $I_2 = (1/4, 1/2]$ and $I_3 = (1/2, 1]$. To each of these intervals correspond one $p_r(v)$ function and one $1 - \beta$ function as indicated by the respective indices. The functions $p_r(v)$ are of the form

$$p_r(v) = \sum_{j=1}^9 a_j v^{10-j}(1-v)^j$$

where the a_j 's are positive integers depending on the number of outcomes corresponding to each case or to possible values of N . The functions are symmetric such that for each term $a v^x(1-v)^y$, $x \neq y$ there corresponds a term $b v^y(1-v)^x$ where $a = b$. The probability of rejection functions and the power functions for the case $m = 2$ may be obtained by listing the 1024 elementary outcomes and their corresponding values of N . One may then single out those events that correspond to a rejection of the null hypothesis for each of the three subintervals where α assumes its values. The events for which $N = 0$ are rejection events for $\alpha \in (0, 1/4]$, the events for which $N = 0$ or 1 are rejection events for $\alpha \in (1/4, 1/2]$, and the events for which $N = 0, 1$ or 2 are rejection events for $\alpha \in (1/2, 1)$. One associates to each event its corresponding probability and add the event probabilities corresponding to each subinterval where α assumes values. Table C1 lists the probability of false rejection functions $p_{r_1}(v)$, $p_{r_2}(v)$ and $p_{r_3}(v)$. The first column lists the a_j accompanying each term for the function $p_{r_1}(v)$ which corresponds to all cases $N = 0$ or $\alpha \in (0, 1/4]$. The second column lists all numbers of terms that must be added to $p_{r_1}(v)$ to complete or form the function $p_{r_2}(v)$ which corresponds to the cases $N = 0$ or $N = 1$ (i.e., $\alpha \in (1/4, 1/2]$). The last column lists all number of terms that must be added to $p_{r_2}(v)$

Table C1 — Functions $p_{r_1}(v)$, $p_{r_2}(v)$ and $p_{r_3}(v)$

$$\text{Each } p_r(v) = \sum_{j=1}^9 a_j v^{10-j} \cdot (1-v)^j$$

| Term | a_j terms for $\alpha \in (0, 1/4]$ | additional a_j terms for $\alpha \in (1/4, 1/2]$ | additional a_j terms for $\alpha \in (1/2, 1]$ |
|--------------|---|---|---|
| $v^9(1-v)$ | 2 | — | — |
| $v^8(1-v)^2$ | 9 | 8 | — |
| $v^7(1-v)^3$ | 20 | 16 | 24 |
| $v^6(1-v)^4$ | 34 | 60 | 24 |
| $v^5(1-v)^5$ | 42 | 102 | — |
| $v^4(1-v)^6$ | 34 | 60 | 24 |
| $v^3(1-v)^7$ | 20 | 16 | 24 |
| $v^2(1-v)^8$ | 9 | 8 | — |
| $v(1-v)^9$ | 2 | — | — |

to complete or form the function $p_{r_3}(v)$ which corresponds to the cases $N = 0, 1$ or 2 (i.e., $\alpha \in (1/2, 1]$). Another way to interpret Table C1 is that column 1 consists of all probabilities for which $N = 0$ column 2 consists of all probabilities for which $N = 1$ and column 3 consists of all probabilities for which $N = 2$. Table C2 lists the power functions $1 - \beta_1$, $1 - \beta_2$ and $1 - \beta_3$ in a manner similar to Table C1. The power function $1 - \beta$ consists of sums of terms of the form

$$a v^x(1-v)^y t^r(1-t)^s$$

where $x + y = 8$ and $r + s = 2$; $x, y = 0, 1, 2, \dots, 8$ and $r, s = 0, 1, 2$

The function $p_{r_1}(v)$ has range $[0, .173]$, $p_{r_2}(v)$ has range $[0, .430]$ and $p_{r_3}(v)$ has range $[0, .523]$. Their curves are plotted in Fig. C1. The power curve $1 - \beta_1$, $1 - \beta_2$ and $1 - \beta_3$ are plotted in Fig. C2 where $v = .10$. Figure C2 indicates that the power of the test behaves as it is supposed to for most of its range. The more dissimilar the values of v and t the higher the power of the test. Figure C1 illustrates the beginnings of the conjectured behavior for $p_r(v)$ as m increases and α becomes small. The case $\alpha \in I_1$, should be of particular interested to the reader in relationship to the conclusions reached in the text section of this report.

Table C2 — Individual Terms for the
Power Functions $1-\beta_1$, $1-\beta_2$ and $1-\beta_3$.

| Term | a terms for $\alpha \in (0, 1/4]$ | additional a terms for $\alpha \in (1/4, 1/2]$ | additional a terms for $\alpha \in (1/2, 1]$ |
|---------------------|---|---|---|
| $v^8 t(1-t)$ | 2 | — | — |
| $v^8(1-t)^2$ | 1 | — | — |
| $v^7(1-v)t^2$ | 0 | — | — |
| $v^7(1-v)t(1-t)$ | 8 | 8 | — |
| $v^7(1-v)(1-t)^2$ | 8 | 0 | — |
| $v^6(1-v)^2 t^2$ | 0 | 0 | — |
| $v^6(1-v)^2 t(1-t)$ | 12 | 8 | 24 |
| $v^6(1-v)^2(1-t)^2$ | 24 | 4 | 0 |
| $v^5(1-v)^3 t^2$ | 0 | 8 | 0 |
| $v^5(1-v)^3 t(1-t)$ | 8 | 48 | 0 |
| $v^5(1-v)^3(1-t)^2$ | 8 | 47 | 0 |
| $v^4(1-v)^4 t^2$ | 2 | 8 | 24 |
| $v^4(1-v)^4 t(1-t)$ | 26 | 8 | 2 |
| $v^4(1-v)^4(1-t)^2$ | 2 | 8 | 24 |
| $v^3(1-v)^5 t^2$ | 8 | 47 | 0 |
| $v^3(1-v)^5 t(1-t)$ | 8 | 48 | 0 |
| $v^3(1-v)^5(1-t)^2$ | 0 | 8 | 0 |
| $v^2(1-v)^6 t^2$ | 24 | 4 | 0 |
| $v^2(1-v)^6 t(1-t)$ | 12 | 8 | 24 |
| $v^2(1-v)^6(1-t)^2$ | 0 | 0 | — |
| $v(1-v)^7 t^2$ | 8 | 0 | — |
| $v(1-v)^7 t(1-t)$ | 8 | 8 | — |
| $v(1-v)^7(1-t)^2$ | 0 | — | — |
| $(1-v)^8 t^2$ | 1 | — | — |
| $(1-v)^8 t(1-t)$ | 2 | — | — |

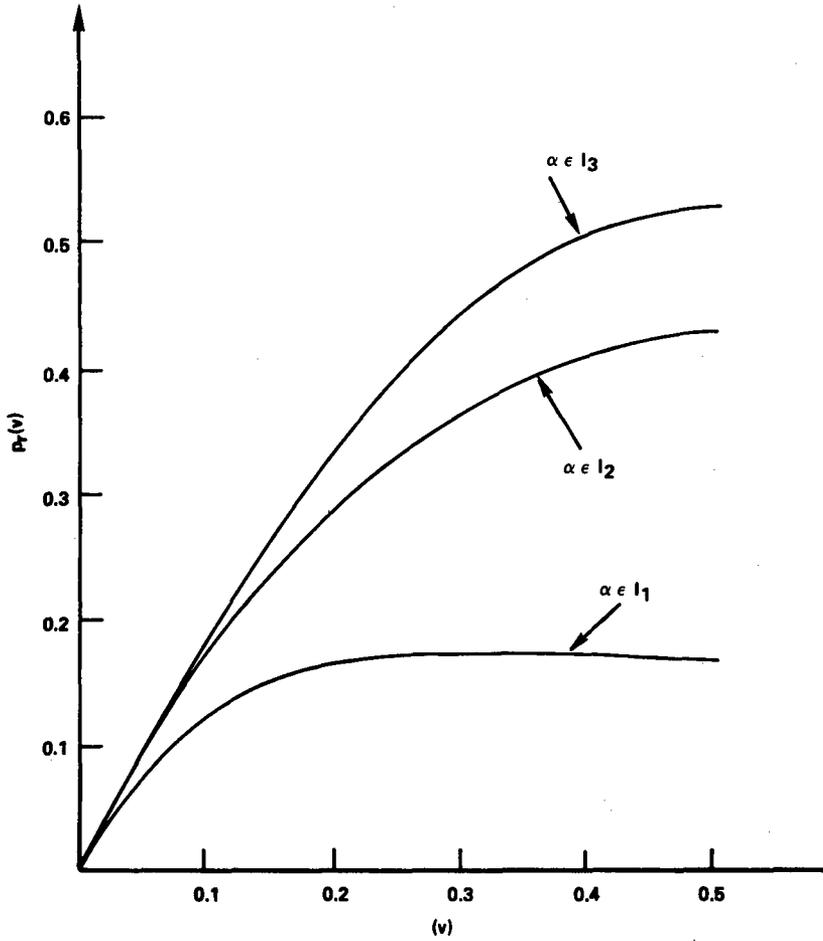


Fig. C1 — Functions $p_r(v)$ for the case $m = 2$

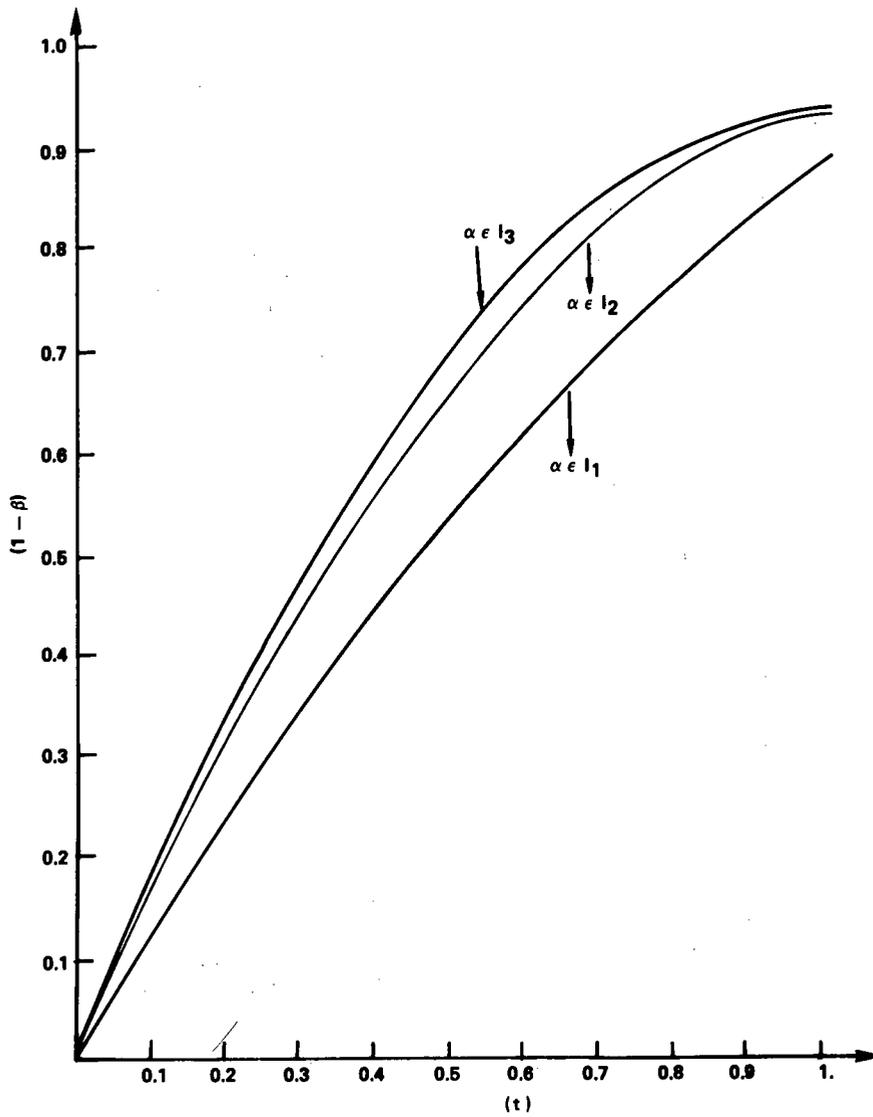


Fig. C2 - Power functions ($m = 2, v = .1$)