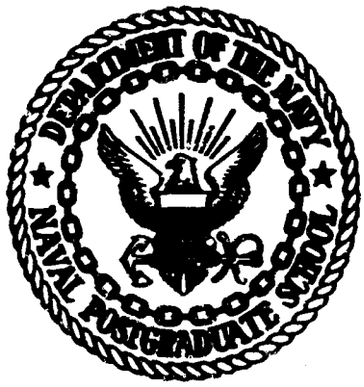
United States
Naval Postgraduate School



THESIS

PROCEDURES AND STATISTICAL

METHODOLOGY OF MODEL VALIDATION

by

James Matthew Arrison, III

December 1969

Procedures and Statistical Methodology of Model Validation

by

James Matthew Arrison, III
Lieutenant, United States Navy
B.S., United States Naval Academy, 1964

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
December 1969

Author _____

Approved by: _____
Thesis Advisor

_____
Chairman, Department of Operations Analysis

_____
Academic Dean

ABSTRACT

Determining the effectiveness of a computer simulation model in duplicating a desired real world phenomenon is an important unsolved problem. The purpose of this paper is to model the validation procedure in a broad context and develop a general methodology for the statistical part of validation. A procedure calling on utility, decision, simulation, and statistical theories is developed. The goals of statistical testing are presented, and the assumptions, properties, and results of several parametric and nonparametric tests are discussed and compared.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

With the advent of complex computer simulations of real world phenomena, a means of judging the worth or validity of the simulation has become very important, yet to associate with any simulation model a strict valid-invalid judgement is quite misleading. Models can be considered valid under certain circumstances and invalid under others, or when compared using different criteria they may be considered first valid then invalid.

In the past one of the largest problems of model validation has been the definition of the term. Normally what has been thought of as model validation is the statistical testing of collected data and the comparison of the results with a predetermined test criterion. For this reason validation has come under attack for being a form of statistical chicanery and merely a means to add credence to an already accepted model.

The purpose of this paper is to redefine the validation procedure in a larger context and to describe some of the problems, techniques, and assumptions associated with the statistical portion of validation. It is hoped that with this procedure model validation will become a more definitive process and that the mystical air normally associated with statistical testing procedures will be removed.

II. VALIDATION PHILOSOPHY

The present procedure of validation is basically as follows. After the requirement to validate a model has been given, agreement on a significance level for statistical testing is reached. The data is then given to a statistician to find an appropriate testing procedure. Using the predetermined level of significance it is then determined whether the difference between the real world and model data is significant. The decision-making procedure is quite simple, if the results of the test indicate a significant difference then the model is said to be invalid. If the difference is not significant then it is considered valid. This procedure is shown in Figure 1 and is the one used in a recent validation [12].*

This type of apparently straightforward validation has two basic problems. The decision rule while seemingly well defined is actually more complex and involves such things as cost and utility models as well as statistical theory. As an example, in a validation done by the Systems Analysis Group [23] the level of significance was set at a level of .5 in order to make the probability of accepting an invalid model small. This decision must have involved consideration of the costs of accepting an invalid model and of rejecting a valid model. After compiling these costs the principles of utility theory must have been used in arriving at a figure of .5 as the best level of significance. But, none of these considerations were mentioned in the report of the validation. So in the past, and even on present validation projects, the

* Number in brackets is reference number as listed on pages 50 and 51.

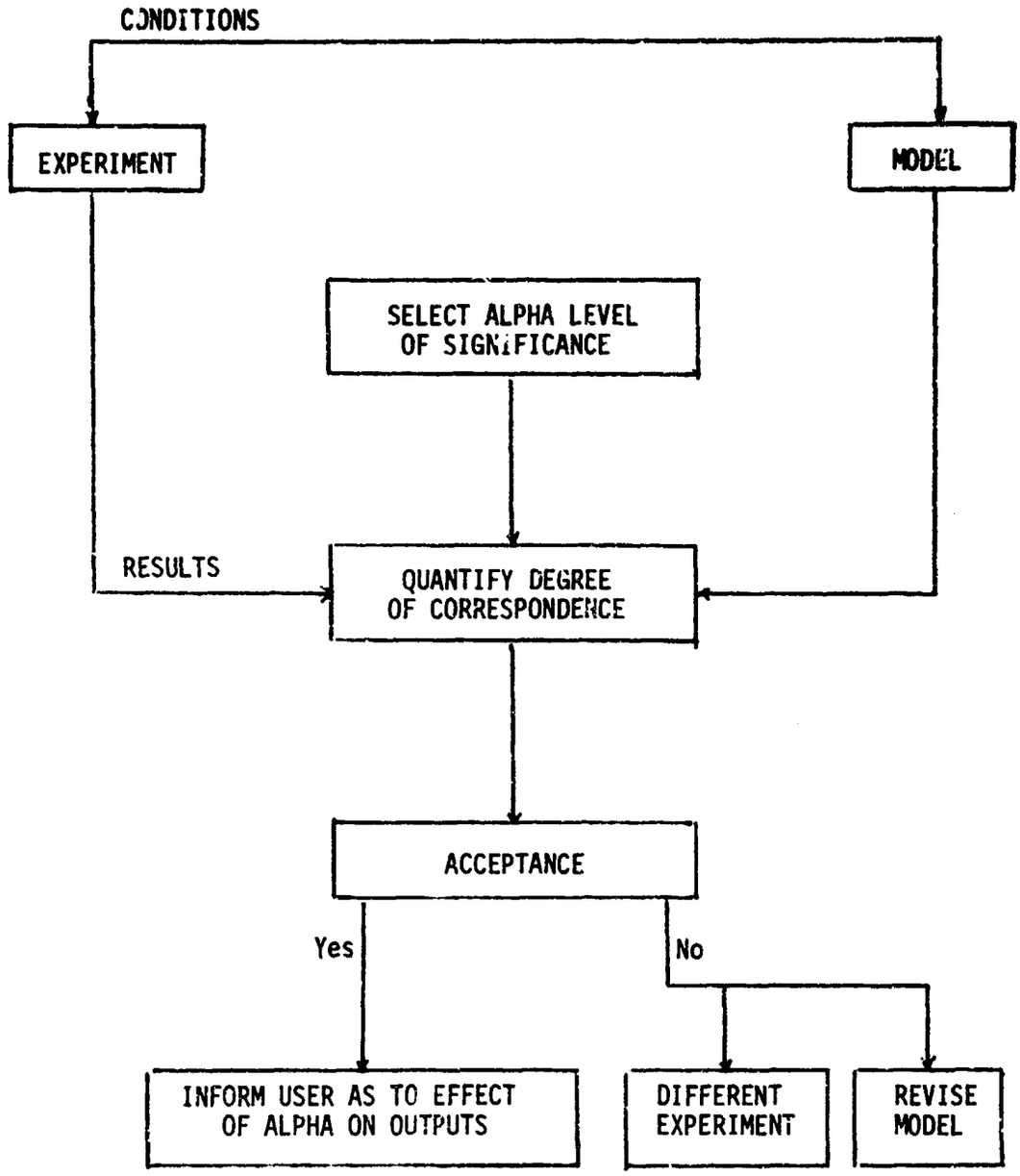


FIGURE 1

A PRESENTLY USED MODEL VALIDATION SCHEME

decision rules are not fully explained, but should be if meaningful validation is to be achieved. The second problem is that much confusion exists in the method of selection of a particular statistical test. In very few cases, if ever, will a test be perfect for the data. An assumption will therefore be relaxed slightly to make use of a strong property of a test, yet if another test were used that assumption might not have to be violated. The question of which assumptions and properties of a test are most important is very complex and the answers not clearly defined. Thus the properties and goals of statistical tests must be more completely defined. It must also be realized that while the passing or failing of a single test or at the most of a few tests constitutes a decision rule now, the results of these tests should only correspond to a single element of an n dimensional decision vector.

The philosophy of the present validation procedure is sound. What is proposed is a new procedure, rather than philosophy, directed at allowing the decision maker more flexibility. This involves describing decisions in terms of utility, simulation, and decision theory as well as just collected data and statistics.

The procedure can be thought of as the interchange of information between three modules, Simulation Theory, Statistical Theory, and Decision-Utility Theory. Within each module there are several nodes such as the testing node in the statistical module. Several nodes such as the criterion node share modules. In general the procedure would work as follows. Information concerning the model and real world, such as data, flow from the state of nature node through the validation information node and into the decision node. From the decision node several paths exist. The validation may be terminated, by accepting or rejecting

the model, the model may be temporarily rejected while further data is gathered or additional comparisons conducted, or the decision may be to compare the information by means of a statistical test. Regardless of the choice, if the validation is not terminated more information enters the validation information node as a result of the decision and the process will continue in a cycling manner until the decision to terminate the validation is given. Figure 2 illustrates the concept of modules and nodes interacting to form a validation procedure.

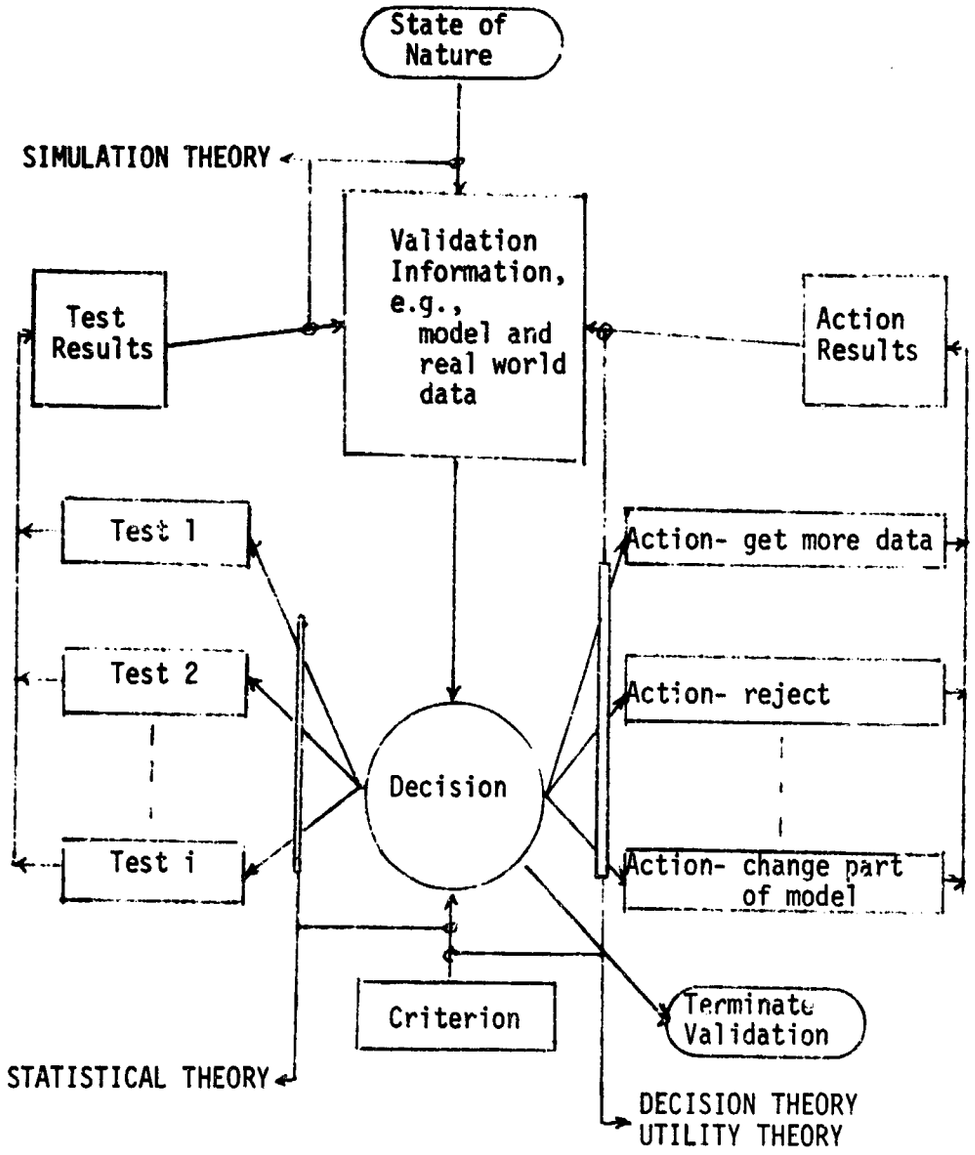


FIGURE 2

PROPOSED VALIDATION SCHEME

III. VALIDATION DECISIONS

There are four possible outcomes of any decision rule. These are based on the two decisions:

D_0 = Decision the model is valid

D_1 = Decision the model is invalid

and the two possible underlying states of nature:

S_0 = The model is valid

S_1 = The model is invalid .

One possible outcome is D_0S_1 or the incorrect decision that the model is valid when it is invalid. The other outcomes are D_0S_0, D_1S_0, and D_1S_1. Why the decision maker makes a particular decision depends upon the decision rule he is using. As an example of a simple decision rule consider a validation procedure which is similar to the one presently being used. It consists of one model, the two states of nature S_0 and S_1, and the two decisions D_0 and D_1. The statistical test is exact, that is:

$$P_r(X^* = x_0 | S_0) = 1$$

$$P_r(X^* = x_1 | S_1) = 1$$

or when the state of nature, S, is S_0 then the test statistic X^* is x_0 and similarly when $S = S_1$, then $X^* = x_1$. The simple decision rule is: if $X^* = x_0$ then $D = D_0$ and if $X^* = x_1$ then $D = D_1$. Again note that this is basically what is done in present validations. A test is performed and according to the results of the test alone the model is said to be valid or invalid.

Expanding the above, consider the following procedure consisting of the same model and states of nature. The test is no longer exact. Now, when $S = S_0$, X^* is a random variable with density function $p_0(x)$ and when

$S = S_1$ then X^* is a random variable with density function $p_1(x)$. X^* represents the test statistic whose range is the real line X. Since X^* is now a random variable the decision rule may become more complex but the possible results are still the same.

RESULTS

	S_0	S_1
D_0	Proper Acceptance	Improper Acceptance
D_1	Improper Rejection	Proper Rejection

PROBABILITIES OF VARIOUS RESULTS

	S_0	S_1
D_0	$1-\alpha$	β
D_1	α	$1-\beta$

FIGURE 3

RESULTS OF DECISION RULES AND THEIR PROBABILITIES

The probabilities of making the decisions are:

α = Probability of rejecting a valid model

β = Probability of accepting an invalid model

$1-\alpha$ = Probability of accepting a valid model

$1-\beta$ = Probability of rejecting an invalid model.

As an example of another decision rule consider a case where,

$p_0(x)$ is normal $(0,\sigma^2)$

$p_1(x)$ is normal (μ,σ^2) $\mu > 0$.

Let X_0 be that portion of the real line, X, such that all points in X_0 are less than or equal to X^+, while the other points constitute X_1, thus:

$$X_0 = \{x : x \leq X^+\}$$

$$X_1 = \{x : x > X^+\}$$

X^+ is an arbitrary point and can be determined either before or after data is collected depending upon the decision rule. If X^+ is determined before the test is performed then the decision rule might be that if X^* is in X_1 then D_1, otherwise D_0. With this decision rule the corresponding decision probabilities are:

$$p_r(D_0|S_1) = \beta = \int_{X_0} p_1(x)dx = \text{probability of accepting an invalid model}$$

$$p_r(D_1|S_0) = \alpha = \int_{X_1} p_0(x)dx = \text{probability of rejecting a valid model}$$

$$p_r(D_0|S_0) = 1-\alpha = \int_{X_0} p_0(x)dx = \text{probability of accepting a valid model}$$

$$p_r(D_1|S_1) = 1-\beta = \int_{X_1} p_1(x)dx = \text{probability of rejecting an invalid model}$$

Figure 4 shows the functions graphically with X^+ and the probabilities, α, and β.

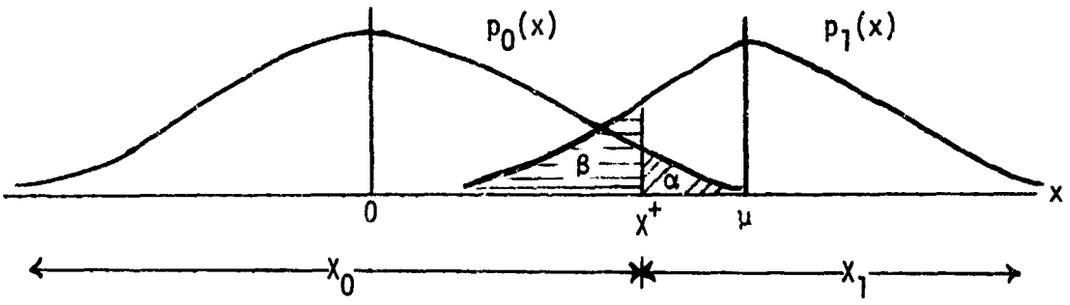


FIGURE 4

GRAPHIC DISPLAY OF POSSIBLE DISTRIBUTIONS OF TEST STATISTIC ASSUMING $p_0(x) = n(0,\sigma^2)$ AND $p_1(x) = n(\mu,\sigma^2)$

It should be noted that Figure 4 represents a much simplified pair of density functions. The nature of validation and the associated test statistics often prevents any knowledge of the exact distribution of $p_1(x)$. In only one of the tests performed in this paper can β be

readily determined. Another complicating feature of validation is that $p_1(x)$ usually flanks $p_0(x)$ or even overlaps $p_0(x)$ over its entire domain. Figure 5 shows these possibilities.

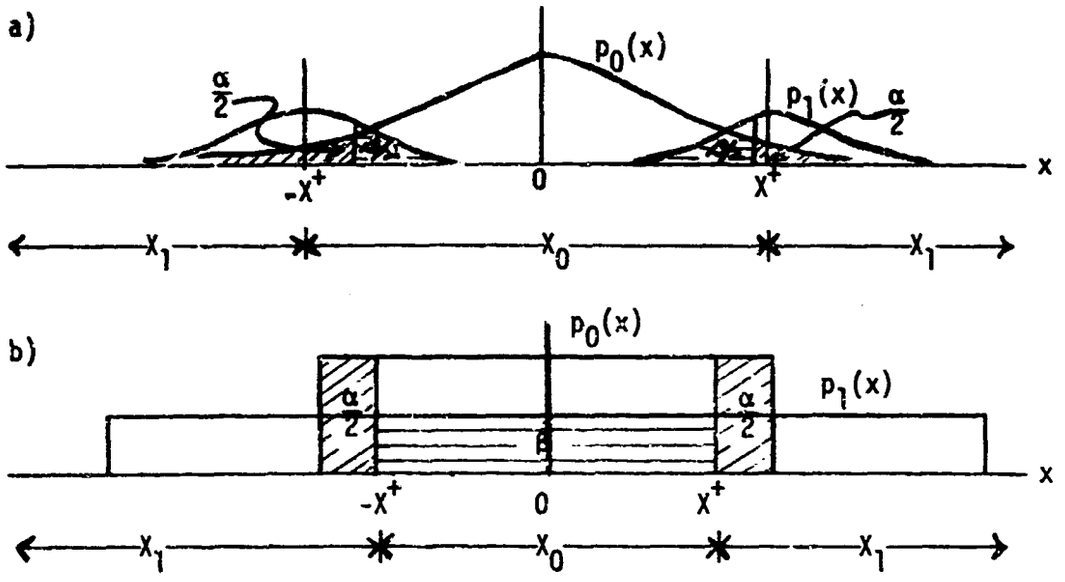


FIGURE 5

POSSIBLE INTERACTIONS BETWEEN $p_0(x)$ AND $p_1(x)$

In both parts a) and b) of Figure 5 the previous decision rule could have been used, but by associating costs with the various results of a decision rule, a payoff matrix can be formed and another type of decision rule employed.

Utility theory and cost structures will determine the values of the decision matrix but in general the S_1D_0 element will represent the highest cost for it represents the acceptance of an invalid model and thus all subsequent decisions based on the assumption that the model is valid will also be in error. S_0D_0 will usually cost the least but the ordering of S_1D_1 and S_0D_1 will vary depending on the costs of additional experimentation and realignment of the model.

With the costs of each decision result determined, and the decision maker willing to accept a priori knowledge of $P_r(S = S_0)$ and $P_r(S = S_1)$, then he can by using a rule such as minimax chose a decision which will minimize cost. If through information received from the statistical module he is willing to accept values of α and β then the costs of the decisions will become expected costs and by using the same minimax decision rule the minimum expected cost can be found. Thus with this decision rule, the relationship between the statistical and decision modules can be seen as one in which additional information received from the data is transmitted to the decision maker allowing him access to more information about the model and helping him to refine his decision.

The choice of which decision rule to use is a subject in itself and is left for future study. Regardless of the method chosen though, the value of statistical information from the data is apparent.

How to realign the simulation model, when the decision to reject it is made, is the subject of simulation theory. This also is a complex field and left for future study.

IV. STATISTICAL THEORY MODULE AND TESTS

Two measures of the probability of rejecting a valid model are α and P where P is determined by finding the largest value of α for which the null hypothesis that the model and real world are sampling from the same distribution can be accepted given the test statistic. Thus P represents the value of α at which the decision concerning the null hypothesis passes from acceptance to rejection and is determined after the test statistic is computed, whereas α is arbitrarily predetermined and used to compare with the results of the test. When comparing different tests, two approaches could be used. Either an α level of significance could be predetermined and each test receive a pass or fail rating, or a P value could be determined. For more sensitive comparisons P values will be determined when applying data to tests in the following sections.

The statistical module of model validation operates as follows. Real world data is compared with model simulation data by one of many statistical tests. For each test the $p_0(x)$ is known and the value of the test statistic computed. Given $p_0(x)$ and the test statistic the P value is determined along with β if $p_1(x)$ is known. This information is then passed into the test results node for further transmission into the validation information node. If α was predetermined and the test statistic X* fell into the critical region, i.e., P was less than α, then based on that particular test the decision that the model be accepted cannot be endorsed. The decision to reject or accept a model could be thought of as an n dimensional vector of which the test result is merely one component.

The types of statistical test previously used for model validation, and most likely to continue being used, are both parametric and nonparametric. Before describing several of these tests, a description of their inherent differences should be useful.

A. PARAMETRIC AND NONPARAMETRIC TESTS

A parametric statistical test is a test in which specific assumptions, such as $\mu = 0$ or $\mu_1 = \mu_2$, about the parameters of the sampled population are made, whereas a nonparametric test, as the name implies, makes no assumptions about the value of the parameters in the sampled population but rather assumes only that a distribution exists. Another term often used interchangeably with nonparametric is distribution free. A distribution free test differs from both the parametric and nonparametric tests in that it makes no assumptions about the form of the sampled distribution. In this paper the terms nonparametric and distribution free will be used interchangeably. More important than the difference in the definitions is the difference in the assumptions which must be made when testing parametrically vice nonparametrically. To determine critical values both tests require that the distribution of the test statistic be fully known. In the case of the parametric tests this often requires that the sample size be large so that the asymptotic distribution of the test statistic is known. The distribution, $p_0(x)$, of the test statistic in the nonparametric case is generally known precisely and need not be assumed. Other assumptions of the parametric tests may include independence of observations, underlying normal distribution of the sampled populations, homoscedasticity or at least, known ratio of variances among populations in the case of a multiple sample test, and that the data is measured in at least an interval

scale, meaning that operations with the data are isomorphic to arithmetic. The assumptions associated with nonparametric tests include only that sampled populations be continuous, and in some cases, be symetric or identical. As with parametric tests, the observations are assumed to be independent. For a more complete discussion of the assumptions see [2,21].

The more practical advantages of the nonparametric tests include their intuitive attraction, simplicity of derivation, and ability to be understood conceptually. They are often times easier to apply, but this quality deteriorates rapidly as the sample size increases past 30. Perhaps the largest advantage, however, is their statistical efficiency.* As Bradley explains [3]:

> When judged by the mathematical criterion of statistical efficiency, distribution-free tests are often superior to their most efficient parametric counterparts when both tests are applied under "nonparametric" conditions, i.e., conditions meeting all assumptions of the distribution-free test, but failing to meet some of the assumptions of the parametric test. When both tests are applied under "parametric" conditions, i.e., conditions meeting all assumptions of the parametric test, and therefore of both tests, distribution-free tests are usually very slightly less efficient at small sample sizes, becoming increasingly less efficient as sample size increases.

Thus with large samples the parametric tests are more powerful provided that their assumptions are met. This margin of power enjoyed by the parametric tests decreases with sample size until the sample size becomes small enough, 6-10, that the power differential is insignificant. On the other hand, when the parametric assumptions are falsely made, but the

---

* Power or statistical efficiency is defined as the ratio of the parametric test sample size to the nonparametric test sample size in order to make the power of the two tests equivalent. If the power efficiency of a nonparametric test is 96%, then if the more powerful parametric test has 10 samples the nonparametric test must have only 10/.96 = 10.4 samples to be of equal power.

nonparametric assumptions are not, the nonparametric tests are often more superior.

Since one of the underlying assumptions in validation is that the sample size of real world data will be quite small it is necessary to consider the effect of the parametric and nonparametric assumptions in terms of small samples, i.e., less than 10. Again according to Bradley, when the parametric assumptions are violated they have their most drastic effect and in addition are most unlikely to be detected due to the small sample size. If a parametric test can be used, even though it is more powerful, its advantage over the nonparametric test is slight due to the small sample size.

Because of the many facets of both types of tests it would be foolish to say that only tests of a single type should be used. Equally as foolish would be an attempt to categorize the types of data to be validated with specific statistical tests. This choice remains in the decision node of the validation procedure. So rather than attempt such a recipe it is beneficial to look at what has been done in several validations and what the differences in critical regions or P values are when tests requiring various assumptions are performed on the same data. In order to examine these differences, a sensitivity analysis on two sets of data with respect to statistical tests and their inherent assumptions was performed.

V. NONPARAMETRIC TESTS WITH SUBMARINE DATA BASE

The data used in the nonparametric tests of this section was obtained from a sequence of submarine exercises in which a submarine attempted to detect another submarine transitting through a defined region. If detection was made then both the range and aspect of the detected submarine were recorded. A stern aspect indicates a retreating contact and the corresponding range would be negative. A positive detection range indicates a bow aspect at initial detection of an incoming submarine. Thus for a given exercise the data might be: out of 10 possible detections initial detection was made at 8, 3, -6, and 1 miles. This should be interpreted as follows: the exercise was run 10 times and detection occurred on 4 runs. On these runs the initial detection was made when the transitting submarine was at a range of 8, 3, and 1 miles and closing, while on the fourth run detection was made at 6 miles but the range was opening. These exercises were simulated on the computer and similar results tabulated. Summarizing, the following constitutes the data base for this validation. The submarine model was tested under 10 various conditions such as speed and depth. Calling each set of conditions an input, there are 10 distribution functions each of which corresponds to an input. For each of the inputs there are several samples from the real world exercises and many, 100-120, samples from the computer simulation model. Two measures of effectiveness have been observed namely the frequency of detection and range of initial detection.

Once again, the goal of the statistical module is to determine $p_0(x)$, and $p_1(x)$, the size of the critical region or P value, and β.

Where the distribution of the test statistic under the null hypothesis that the real world and model are sampling from the same distribution is $p_0(x)$, and $p_1(x)$ is the distribution of the test statistic when the two distributions are not the same.

A. TESTS

The tests chosen to illustrate what might be done in validating the submarine model include the nonparametric Kolmogorov-Smirnov Test, the Fisher Exact Test, the Wilcoxon Test, and the tests used in the initial validation of this model [24].

B. KOLMOGOROV-SMIRNOV TEST

Perhaps the most heuristic of the statistical tests is the Kolmogorov-Smirnov Two Sample Test often referred to as the Smirnov Maximum Deviation Test. The test statistic is the maximum deviation between the two empirical cumulative distribution functions.

To compute the test statistic rank the n real world and m model observations and give each a subscript corresponding to its rank. For each possible rank i, i=1,...,n+m, calculate d_i. Where:

$$d_i = \frac{r_i}{n} - \frac{s_i}{m}$$

and

r_i = the number of real world observations less than the ith order statistic

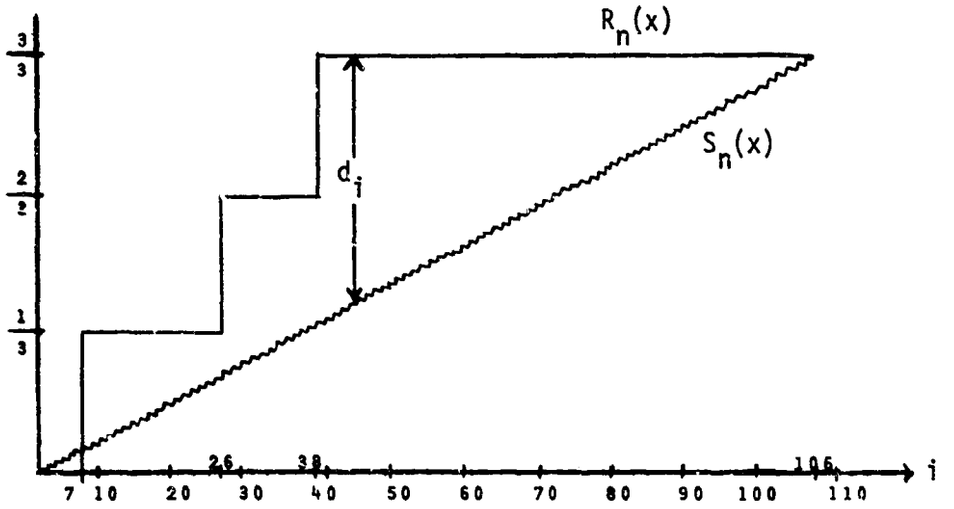
s_i = the number of simulated observations less than the ith order statistic.

The test statistic, D, is max $|d_i|$, i=1,...,n+m. Under the hypothesis that the observations came from the same distribution, the distribution of D is known and can be calculated for any combination of n+m [4,20].

As an illustration of this test consider the 10th input where n+m is 106, and n=3. The ranges of real world detections were ranked among the model detections and the values of i are i = 7,26,38. The two step functions are shown in Table I.

TABLE I

CUMULATIVE STEP FUNCTIONS IN KOLMOGOROV-SMIRNOV TEST WITH INPUT 10 OF SUBMARINE DATA



D occurs at i = 38 where

$$d_{38} = 3/3 - 35/106 \approx .67.$$

Using the approximation to $p_0(x)$, the P value is found to be .2024. Table II gives a summary of the P values when the test was applied in a similar fashion with the remaining 9 inputs.

This test has all the previously mentioned advantages of nonparametric statistics, especially intuitive appeal. It also has the advantage of testing for differences in the distributions caused by all the properties of the distribution function instead of just the differences in mean or variance.

A major restriction is placed on the validity of the results by using the approximation to $p_0(x)$. Hodges [15] has shown that as m and n

increase the approximation may differ significantly from $p_0(x)$, thus not only does power efficiency decrease with increase sample size but the approximation of $p_0(x)$ also becomes less valid. The effect on P of approximating the distribution function can be shown by comparing the results of this approximation with those of an exact test. This is shown in Table XVII, page 45.

TABLE II

SUMMARY OF RESULTS FOR KOLMOGOROV-SMIRNOV TEST WITH SUBMARINE MODEL DATA

INPUT	P VALUE
1	.485
2	.260
3	.980
4	.998
5	.941
6	.491
7	.922
8	.577
9	.792
10	.202

C. WILCOXON TEST IN THE ORIGINAL VALIDATION

In the original validation of the submarine encounter model and the associated data [26], the Wilcoxon Test was used with the initial range of detection data. In this test the observations from both sources for a given input are aggregated and ranked in order of magnitude. If both model and real world are sampling from the same distribution then all

combinations of ranks are equally likely. The smallest and largest possible rank sums make up the critical region since they represent the least likely results. For example, suppose the real world detections occurred at 1, 5, 8 miles out of 6 runs and out of 20 runs the simulation results had initial detection at 6, 10, 12, 14, 20 miles. The summed ranks for the real world would be 1+2+4=7 and at an alpha level of .109 the difference between real world and model results is significant.

Table III lists the acceptance regions in ranked sums for various model and real world outputs. In each case the level of significance is .109 and the real world ranks are to be summed. Exact P values were not found in the original validation.

There are two difficulties with the original use of the test however. In an apparent attempt to avoid the tedious counting procedures outlined in the next section, the model data was divided into sets of 20 then tested with the set of real world data. Each test was considered independently, thus the level of significance is considered to be $(1-.109)^6$ or .5. This is derived by the following argument. Since α was predetermined as $\alpha \geq .5$ then $(1-P_r)^n \leq .5$ where P_r is the probability of failing the test if the state of nature is S_0, and n is the number of tests. If n = 6 then P_r becomes .109. It seems very difficult to believe however that these tests are independent if the same real world observations are to be used in each test. The second difficulty is the method in which the rank sums were determined. Instead of considering an initial detection of 8 miles differently than one of -8 miles, both were given the same rank.

TABLE III

RANK-SUM ACCEPTANCE REGIONS WITH $\alpha = .109$*

R1 \ R2	2	3	4	5	6
2	-	7-12	11-18		
3	3-8	7-14	12-21		
4	3-10	8-17	12-23		
5	4-12	8-18	13-26		
6	4-13	9-21	14-29		
7	4-15	10-24	16-33	23-42	
8	5-17	10-26	16-35	24-46	
9	5-19	11-28	18-38	25-49	
10	5-20	12-31	19-41	27-53	
11	6-22	12-32	20-44	29-57	38-70
12	6-24	13-35	21-47	30-60	40-74
13	6-25	13-37	22-50	32-64	42-78
14	7-27	15-40	23-53	33-67	43-82
15	7-29	15-42	24-56	34-70	45-86
16	8-31	16-45	25-59	36-73	47-90
17	8-32	16-46	26-62	37-78	49-95
18	8-34	17-49	28-65	39-82	51-99
19	9-36	-	28-67	40-85	53-103
20	-	-	30-71	41-88	55-107

R_1 = No. of detections by model in sample size 20.

R_2 = No. of detections in real world exercise, with 6 runs.

* By permission from Submarine ASW Encounter Simulation Model Detection Validation (U) by Systems Analysis Office, ASW Systems Project Office (1967).

Both these difficulties are corrected and a more exact test made in the next test. Table IV presents a summary of results using this testing procedure.

TABLE IV

SUMMARY OF THE ORIGINAL WILCOXON TEST
RESULTS USING THE SUBMARINE DATA

RUN NUMBER	P VALUE
1	greater than .5
2	greater than .5
3	greater than .5
4	greater than .5
5	greater than .5
6	greater than .5
7	greater than .5
8	less than .5
9	greater than .5
10	less than .5

D. EXACT WILCOXON RANK SUM TEST

An improvement over the original use of the Wilcoxon Rank-Sum Test in validating this model can be made by not dividing the model data into subsets but rather considering it as a sample of size 120, and by determining the exact distribution of the Wilcoxon Test statistic. In order to determine the distribution of $p_0(x)$ it is necessary to compute such things as the number of possible ways 4 numbers can be sampled without replacement from the positive integers 1 through 124 such that their sum is always less than or equal to 165. A recursive counting

procedure for large numbers like 165 has been developed by Fix and Hodges [10] and was used in determining P values for 4 of the 10 inputs. As an approximation to the test it can be realized that the ranks form a finite population, thus the expected value and variance of the average rank sum can be determined exactly. The distribution of the average value of the observed ranks minus its expected value and divided by its standard deviation is approximately the unit normal. Kruskal and Wallis [16] have suggested an addition correction for continuity when using this approximation.

As an example of the effect of the normal approximation observe the summary of P values using the Wilcoxon Rank-Sum Test in Table V.

These are the first results based on the exact distribution of $p_0(x)$, but as shown by the results in Table V it is evident that the approximations are quite close. Again the inherent advantages of the nonparametric statistics are present but also is the lack of knowledge of $p_1(x)$. For a more complete discussion of the efficiency of this test see [5].

E. BINOMIAL TEST IN THE ORIGINAL VALIDATION

The other measure of effectiveness used in the statistical portion of the validation of this model is the probability of detection, P_d. In the original validation [25] a very inexact test was used. If m model runs and n real world runs were to be compared then the probabilities of each possible outcome were estimated. These probabilities are shown in Table VII for n equal 4 and m equal 20.

To see the inexactness of this test consider how the probabilities are determined. The probability of x detections in n runs is

$$\binom{n}{x} p_d^x (1-P_d)^{(n-x)}.$$

TABLE V

SUMMARY OF SUBMARINE MODEL P VALUES WITH
DETECTION RANGE DATA USING THE WILCOXON RANK-SUM TEST

INPUT	EXACT	APPROXIMATION
1		.3140
2	.0066	.009322
3		.7900
4		.7860
5	.8735	.8728
6	.2351	.2448
7		.5666
8		.3600
9		.2684
10	.0952	.0902

TABLE VI

SUMMARY OF P VALUES USING NONPARAMETRIC
TESTS ON SUBMARINE MODEL RANGE OF DETECTION DATA

INPUT	K-S TEST	WILCOXON RANK SUM TEST		ORIGINAL RANK SUM TEST
	APPROX.	EXACT	APPROX.	APPROX.
1	.485		.3140	> .5
2	.260	.0066	.009322	> .5
3	.980		.7900	> .5
4	.998		.7860	> .5
5	.941	.9735	.8728	> .5
6	.491	.2351	.2448	> .5
7	.922		.566	> .5
8	.577		.3600	< .5
9	.792		.2684	> .5
10	.202	.0952	.0902	< .5

If the null hypothesis is true and the model and real world runs are independent then the probability of observing y out of m detections from the model and x out of n detections from the real world is:

$$P_r(x,y) = \binom{m}{y} P_d^y (1-P_d)^{(m-y)} \binom{n}{x} P_d (1-P_d)^{(n-x)}$$

$P_r(x,y)$ is a concave function and by taking derivatives with respect to P_d, it can be shown that if $0 < x+y < m+n$ then:

$$P_r(x,y) \leq \binom{m}{y}\binom{n}{x}\left(\frac{x+y}{n+m}\right)^{(x+y)}\left(1 - \frac{x+y}{m+n}\right)^{(m+n)-(x+y)} \equiv \hat{P}(x,y).$$

Thus $\hat{P}(x,y)$ is an upper bound on the probability that x and y detections will occur. $\hat{P}(x,y)$ are the values listed in Table VII.

The data has again been divided into groups of 20 and thus the critical region reduced to .109 for each test. For the particular values of n and m the critical region has been partioned in Table VII. Should a pair (x,y) fall into this region for any of the six tests then the hypothesis is rejected at the .5 significance level.

The primary objection to the test besides the division of model observations into groups of 20 is the fact that each $\hat{P}(x,y)$ is equal to or larger than its exact value yet the size of the critical region is still assumed to be .109. This would seem to indicate that when the null hypothesis is accepted using this test, it might be rejected when using a more exact test. This is in fact the case as shown in Table XI, page 37.

F. FISHER EXACT TEST

When using the number of detections divided by the number of runs to test model validity, tests having more exact knowledge of $p_0(x)$ are also available. One such test is the Fisher Exact Test based on the hypergeometric distribution. The P value is determined by computing

TABLE VII

$\hat{P}$ (x,y) VALUES FOR n = 4 AND m = 20*

y \ x	0	1	2	3	4
0	-	.062623	.006144	.000474	.000021
1	.313114	.081919	.014194	.001611	.000093
2	.194556	.089891	.022945	.003524	.000262
3	.134835	.091778	.031708	.006277	.000583
4	.097514	.089873	.040012	.009900	.001125
5	.071870	.085359	.047517	.014391	.001973
6	.053350	.079194	.053965	.019721	.003230
7	.039598	.071953	.059163	.025835	.005023
8	.029231	.064093	.062972	.032647	.007509
9	.021365	.055975	.065294	.040045	.010883
10	.015393	.047883	.066074	.047883	.015393
11	.010883	.040045	.065294	.055975	.021365
12	.007509	.032647	.062972	.064093	.029231
13	.005023	.025835	.059163	.071953	.039598
14	.003230	.019721	.053965	.079194	.053350
15	.001973	.014391	.047517	.085359	.071870
16	.001125	.009900	.044012	.089837	.097514
17	.000583	.006277	.031708	.091778	.134835
18	.000262	.003524	.022945	.089891	.194556
19	.000093	.001611	.014194	.081919	.313114
20	.000021	.000474	.006144	.062623	-

NOTE: 1. Table entries represent the equation:

$$\hat{P}(x,y) = \binom{4}{x}\binom{20}{y}\left(\frac{x+y}{24}\right)^{x+y} \cdot \left(1 - \frac{x+y}{24}\right)^{24-(x+y)}$$

2. P (x,y) values lying between shaded (****) region define the acceptance region.

* By permission from Submarine ASW Encounter Simulation Model Detection Validation (U) by Systems Analysis Office, ASW Systems Project Office (1967).

TABLE VIII

SUMMARY OF THE ORIGINAL BINOMIAL TEST USING THE SUBMARINE P_d DATA

RUN NUMBER	P VALUE
1	greater than .5
2	greater than .5
3	greater than .5
4	greater than .5
5	greater than .5
6	greater than .5
7	greater than .5
8	less than .5
9	greater than .5
10	less than .5

the probability of receiving the exact combination of model and real world detections as well as any of the more extreme combinations.

Consider the data as presented in Table IX.

TABLE IX

FISHER EXACT TABLEAU WITH SUBMARINE DATA FROM INPUT 7

	NUMBER OF DETECTIONS	NUMBER OF NON-DETECTIONS	TOTAL
REAL WORLD	3	5	8
MODEL	80	40	120
TOTAL	83	45	128

The probability of receiving this combination of detections and non-detections is:

$$\frac{83!\ 45!\ 8!\ 120!}{128!\ 3!\ 5!\ 80!\ 40!} = .07905$$

For a proof see [6].

The more unlikely combinations, keeping the totals fixed, and their probabilities are listed in Table X.

TABLE X

MORE EXTREME TABLEAUS IN FISHER EXACT TEST FROM INPUT 7

	DET.	NONDET.	DET.	NONDET.	DET.	NONDET.
REAL WORLD	2	6	1	7	0	8
MODEL	81	39	82	38	83	37
PROBABILITY	.01953		.002653		.0001518	

The sum of all these probabilities is .10183; but this represents the critical region in only one tail of $p_0(x)$ and since the alternate hypothesis is compound the sum must be doubled. P is therefore .20277. The results of this test with all 10 inputs are listed in Table XI.

These exact probabilities can be very tedious to compute and again approximations are available. A normal approximation when the sample size in large is described by Brownlee [7], along with guidelines on when the approximation is valid.* Unfortunately none of the input results met the criterion but in three cases they came reasonably close. The results of using the approximations are listed in Table XI.

Along with the standard attributes of nonparametric statistics the Fisher Exact Test and its normal approximation both have well defined power, $1-\beta$, functions associated with them. $1-\beta$ for input 1 was computed using the methods suggested by Brownlee [8] and is listed in Table XI as .529.

---

* Working in reverse it has been shown by Tocher [27] that the Fisher Exact Test can be used when the conditions of the normal approximations do not hold.

TABLE XI

SUMMARY OF SUBMARINE MODEL P VALUES
USING THE FREQUENCY OF DETECTION DATA

INPUT	FISHER EXACT TEST	NORMAL APPROXIMATION	POWER	ORIGINAL BINONIAL
1	.954	.928	.529	> .5
2	.932			> .5
3	.7772	.668		> .5
4	.894			> .5
5	.968			> .5
6	1.000	.984		> .5
7	.203			> .5
8	.031			< .5
9	.266			> .5
10	.101			< .5

For information concerning the power function and the Fisher Exact Test see [1,14,19]. Thus for the first time in the tests described, $p_1(x)$ and $p_0(x)$ can be found.

G. SUMMARY OF TESTS WITH SUBMARINE DATA

This concludes a far from exhaustive presentation of possible statistical tests which could be used in validating the submarine model. Hopefully the types of assumptions that are necessary in nonparametric testing are reasonably clear. Brownlee, Bradley, and Seigel give a far more in-depth discussion of nonparametric statistics in their texts referenced in this section. For a more complete discussion of the power of nonparametric tests see [9] as well.

Before going on to a parametric test and one where dependence among samples is considered, examine the difference in the critical regions obtained by using various tests requiring slightly different assumptions of the distribution of the test statistic, Tables VI and XI, pp. 32 and 37.

While care must be used in explaining the cause of the differences, it is certainly safe to say that the assumptions of the Wilcoxon Rank Test are most closely adhered to while the ranking procedure of the original rank sum test and the approximation of $p_0(x)$ in the Kolmogorov-Smirnov Test would tend to discount the validity of their results.

VI PARAMETRIC AND NONPARAMETRIC TESTS WITH AIRCRAFT DATA BASE

The data to be used in the statistical tests of this chapter was obtained from eight independent aircraft-submarine exercises. In each exercise aircraft monitored a string of eight sonobuoys in an attempt to gain and maintain the detection of a transitting submarine. All exercises were made under similar conditions and therefore the conditions can be considered identical. The measure of effectiveness in these exercises is detection modulus or probability of detection. Detection modulus, D.M., is computed by dividing the total number of minutes detection was held by the total number of minutes detection could have been held.

$$D.M. = \frac{\text{TIME DETECTION WAS HELD}}{\text{TIME DETECTION COULD HAVE BEEN HELD}}$$

For each of the eight runs the range and aspect of initial detection for each buoy was tabulated. Also recorded were the range and aspect at the time of losing contact, and the same information in the event that contact was regained. Because of this extensive data base there are several random variables which might be tested. All of these fall into two categories; however, those in which the assumption is made that the samples are independent and identically distributed and those which assume only that the samples are identically distributed. The Paired t, Wilcoxon, and Kolmogorov-Smirnov Tests fall into the first category while the Davisson Test falls into the latter.

Thus using detection modulus as a measure of effectiveness the nonparametric, Wilcoxon and Kolmogorov-Smirnov Tests and parametric Paired t and Davisson Test will be used to demonstrate the spectrum of tests

and their characteristics that might be used in validating a model with this type of data base.

A. PAIRED t TEST

Since we are testing the hypothesis that the real world and the model are sampling from the same distribution it is only natural to compare the differences in their outputs. Let d_i represent the difference between the real world and detection moduli for run i, and let

$$D = \frac{1}{n} \sum_{i=1}^{n} d_i \quad i=1,\ldots,n .$$

If S^2 is defined as:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (d_i - D)^2$$

then it can be shown [13] that

$$t = \frac{\sqrt{n}\, D}{S}$$

is asymptotically t distributed with n-1 degrees of freedom.

Since this test assumes that the d_i are independent, the sample size is eight and each sample is the difference between the real world and model estimate of the true detection modulus computed with all eight buoys operating in concert. Table XII shows the actual data and part of the calculations. The corresponding P value is approximately .42.

The Paired t Test has the advantages of the parametric tests, and $p_0(x)$ is known exactly to be $t_{(n-1)}$ for large n. Since this distribution is well tabulated and the arithmetic is basic, the test is easy to perform. The test does make some very restrictive assumptions. The independence assumption forces the aggregation of the data to the extent that much information may be lost. The asymptotic property of the

distribution of the test statistic adds another degree of complexity for no longer is $p_0(x)$ known exactly. It is known only in the limit as n increases.

TABLE XII

INDEPENDENT AIRCRAFT DATA FOR PAIRED t TEST

RUN	REAL WORLD	MODEL	d_i	d_i-D
i	D.M.	D.M.		
1	.4865	.3478	.1387	.1748
2	.3587	.4023	-.0436	.0075
3	.2500	.4711	-.2211	.1850
4	.4134	.5034	-.0900	.0539
5	.1729	.2967	-.1238	.0877
6	.3884	.4126	-.0242	.0119
7	.2601	.2517	.0094	.0445
8	.3171	.2702	.0469	.0830

$$S^2 = \frac{1}{7} \sum_{i=1}^{8} (d_i - D)^2 = \frac{.08344}{7} = .012635$$

$$t = \frac{\sqrt{n}\, D}{S} = \frac{\sqrt{8}\ (-.0361)}{.1124}$$

$$t = -.9084$$

B. KOLMOGOROV-SMIRNOV TEST

Another way of comparing the differences in the samples is by the Kolmogorov-Smirnov Test. The relative merits of the test have been discussed, but this data presents an opportunity to compare the exact $p_0(x)$ for the Kolmogorov-Smirnov Test to the previously used approximation of $p_0(x)$. Using the data given in Table XII, the maximum deviation is .25, and by the previous approximation to $p_0(x)$ the corresponding P value is .9639 whereas by Massey's exact computation [18] the P value is .6602. This very large difference indicates the dangers in using this approximation to $p_0(x)$.

C. WILCOXON TEST WITH NORMAL APPROXIMATION

The Wilcoxon Test when performed on the data of Table XII and with use of the normal approximation to $p_0(x)$ yields a P value of .46. The relative merits of this test are the same as described previously, and the results are given only for comparative purposes.

D. DAVISSON TEST WITH DEPENDENCE

In the past three tests the sample size was eight due to the fact that independence among samples was required by each test. What if one wanted to compare the average detection modulus of each buoy on each run or perhaps the average detection modulus in each five mile range band from -50 to 50 miles for each buoy on each run? In these cases and the many others that might be considered the values are dependent on each other and thus none of the assumptions of the tests mentioned so far are completely satisfied.

Davisson has shown that by comparing certain differences between the real world and model results that the maximum likelihood ratio yields a statistic with a known distribution [11].

Since the test is very tedious only a relatively short comparison with the aircraft data will be given. Consider the detection moduli of buoys 3, 4, and 5 on each run. The random variable to be tested is the average detection modulus of each buoy. Thus the null hypothesis is that the average detection moduli for buoys 3, 4, and 5 are the same in the real world as they are in the model and that their interdependence is also identical.

The first step in determining the test statistic is to compute the variance-covariance matrix of the computer's average detection moduli.

To do this the average detection modulus for each run on a buoy is subtracted from the average for that buoy over all runs.

The results are shown in Tables XIII and XIV.

TABLE XIII

AVERAGE BUOY DETECTION MODULI FOR THE MODEL AND THEIR AVERAGES

		BUOY		
		3	4	5
RUN	1	.240755	.560000	.491032
	2	.365082	.457377	.595555
	3	.104921	.596825	.514203
	4	.584210	.941052	.782857
	5	.051273	.164909	.158269
	6	.496562	.375031	.181154
	7	.038730	.465397	.579434
	8	.209259	.563148	.468054
AVERAGE		.261349	.513217	.468054

TABLE XIV

AVERAGE MODEL DETECTION MODULI MINUS THEIR AVERAGES

		BUOY		
		3	4	5
RUN	1	-.020594	.046783	-.049022
	2	.103730	-.055840	.127501
	3	-.156428	.083608	.046149
	4	.322862	.427835	.314803
	5	-.210076	-.348308	-.309785
	6	.235214	-.156186	-.286900
	7	-.222619	-.047820	.111380
	8	-.052090	.049931	.045875

The transpose of the 8x3 matrix in Table XIV when multiplied by itself yields the variance-covariance matrix Q. The Q matrix is shown in Table XV.

TABLE XV

VARIANCE-COVARIANCE MATRIX Q

.036453	.020347	.0098831
.020347	.043229	.034850
.0098831	.034850	.039085

Now a difference vector M is computed with each component being equal to the difference between the average real world detection modulus overall eight runs and the corresponding results from the model.

TABLE XVI

DIFFERENCE VECTOR M

REAL WORLD AVERAGE	MODEL AVERAGE	M
.449466	.261349	.1881
.365698	.513217	-.1475
.539570	.468054	.0715

Davisson has stated [11] that the distribution of

$$M^T Q^{-1} M$$

is asymptotically chi-squared with N degrees of freedom where N is the dimension of Q. In this case $M^T Q^{-1} M$ is 9.7682 and the corresponding P value is .02.

As was the case with the Paired t Test, this test has the advantages of being parametric but the disadvantages of its asymptotic properties and lack of knowledge of $p_1(x)$. The main drawback of the Davisson Test is its computational difficulty. As the dimension of Q increases a large computer becomes necessary and the sorting of data becomes quite tedious. Care must also be taken that accuracy is not lost in the inversion of Q and that subsets are chosen such that Q is not singular. In spite of all these disadvantages, the relief from the independence

assumption is very advantageous. If tolerance of its assumptions permits its use, the Davisson Test will yield a more detailed validation test. It is now possible to reject part of the model while accepting the rest, thus allowing trouble-shooting for the simulation analysts. This feature was also possible with the submarine model but only because 10 different inputs were sampled and thus data collection had to be more extensive and also more costly.

E. SUMMARY OF TESTS WITH AIRCRAFT DATA

The P values corresponding to each of the four tests applied to the aircraft data are listed in Table XVII.

TABLE XVII

SUMMARY OF P VALUES USING PARAMETRIC AND NONPARAMETRIC TESTS ON THE AIRCRAFT MODEL DATA

PAIRED t	KOLMOGOROV-SMIRNOV		WILCOXON	DAVISSON
	EXACT	APPROX.		
.42	.6602	.9639	.46	.02

It is not appropriate to compare the results of the Davisson Test to those of the other tests due to its unique properties, nor is it feasible to pass judgement on the remaining tests solely on the results in Table XVII. It should be noted however that the Kolmogorov-Smirnov and Wilcoxon Test results are based on exact knowledge of $p_0(x)$ while the Paired t Test and the approximate Kolmogorov-Smirnov Test are not, and that no additional knowledge of $p_1(x)$ is obtained by using these approximations. While the distribution of the Davisson Test statistic is not exact nor is information about $p_1(x)$ available, it does allow a more localized validation thereby allowing "trouble-shooting" which the other tests do not permit.

As with the submarine data, these tests are far from an exhaustive set of all those possible. They were chosen to represent the range and spectrum of assumptions needed to perform the validation of this type model with its data base.

VII. SUMMARY AND CONCLUSIONS

This paper has investigated the most salient problems of present day validation procedures and alleviated them by enlarging the scope of validation and by describing what is needed and can be expected from a statistical test with "validation type" data. It was shown that decision theory and cost analysis while present in previous validations received no mention, and that statistical testing with its pass or fail results did not allow the decision maker much flexibility. While only two simple decision rules and one type of decision criterion were presented, it became obvious that by determining P values from several tests and by trying to do such things as minimizing expected cost, the decision maker could avail himself of more information and have the capability to change more elements in his decision rule.

A general methodology for the statistical testing of validation data was also discussed. Included in the methodology are the goals of a "validation test," the types of tests available with their inherent assumptions and properties, the need for multiple testing, and the pitfalls of relaxing assumptions within a test.

It was seen that while a myriad of possible tests exists, those having exact knowledge of $p_0(x)$ and $p_1(x)$ will be the best. But, since $p_1(x)$ is seldom known due to the nature of the alternate hypothesis and calculation procedures necessitate approximations to $p_0(x)$ in many cases, these desirable tests are not always available. Some tests are clearly better than others, but in general, it was seen that several tests using different assumptions should be used to achieve the most reliable information about P and β.

In conclusion the problems of validation are analogous to those of systems analysis and cost-effectiveness. The goal or criterion can be defined as minimization of expected cost for a fixed level of validity, yet the methods of exact determination are not as well defined and need to be considered in concert instead of individually. In the past, one of the methods was statistical testing. When used alone there existed reasons to criticize the validations but when used in the procedure as presented in this paper, the validator has more flexibility and is able to use more information from his data and other sources.

Another important advantage of this procedure is the increased ability to see the effects of changes in a decision rule. All that could be seen previously was that at a significance level of .6 the model was considered invalid but at a .4 level it was not. Now such things as the changes in a decision rule caused by refusing to accept a priori knowledge of the states of nature can be observed.

So just as was done with systems analysis a new approach or way of looking at a problem has been proposed. This time it is to help the decision maker with his important and complex problems of model validati.

VIII. AREAS FOR FUTURE STUDY

Since this paper represents a pilot study in the expansion of model validation, almost any facet of the paper could and should be expanded.

The area of simulation theory is normally not considered an O.R. problem at least in the context of calibrating the model. The search for more nearly perfect statistical tests is also considered as second in importance to the development of decision rules applicable to model validation.

After several decision rules have been presented then case studies similar to those of systems analysis will make a valuable contribution to the field of validation.

LIST OF REFERENCES

1. Bennett, B. M., and P. Hsu, "On the Power Function of the Exact Test for the 2x2 Contingency Table," Biometrika, 34 (1947), pp. 123-138.

2. Bradley, James V., Distribution-Free Statistical Tests. New Jersey: Prentice-Hall, Inc., 1968, pp. 15-44.

3. Ibid., p. 18.

4. Ibid., pp. 288-290.

5. Ibid., pp. 108-110.

6. Brownlee, K. A., Statistical Theory and Methodology in Science and Engineering, second edition. New York: John Wiley and Sons, Inc., 1965, pp. 163-164.

7. Ibid., pp. 150-154.

8. Ibid., p. 154.

9. Fisz, M., Probability Theory and Mathematical Statistics. New York: John Wiley and Sons, Inc., 1963, pp. 566-578.

10. Fix, Evelyn, and J. L. Hodges, Jr., "Significance Probabilities of the Wilcoxon Test," Annals of Mathematical Statistics, 26 (1955), pp. 301-302.

11. FIXHEX Model Validation Methodology Report. Pennsylvania: Naval Air Development Center, 1968, p. 17.

12. Ibid., p. 2.

13. Ibid., p. 7.

14. Harkness, D. L., and L. Katz, "Comparison of the Power Functions for the Test of Independence in 2x2 Contingency Tables," Annals of Mathematical Statistics, 35 (1964), pp. 1115-1127.

15. Hodges, J. L., Jr., "The Significance Probability of the Smirnov Two-Sample Test," Arkiv För Matematik, 3 (1957), pp. 469-486.

16. Kruskal, W. H., and W. A. Wallis, "Use of Ranks in One-Criterion Analysis of Variance," Journal of the American Statistical Association, 47 (1952), pp. 583-621.

17. Massey, F. J., Jr., "The Distribution of the Maximum Deviation Between Two Sample Cumulative Step Functions," Annals of Mathematical Statistics, 22 (1951), pp. 125-128.

18. Ibid.

19. Pearson, E. S., and Maxine Merrington, "2x2 Tables; The Power Function of the Test on a Randomized Experiment," Biometrika, 35 (1964), pp. 331-345.

20. Seigel, Sidney, Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill Book Co., Inc., 1956, pp. 127-130.

21. Ibid., pp. 18-34.

22. Smirnov, N., "Table for Estimating the Goodness of Fit of Empirical Distributions," Annals of Mathematical Statistics, 19 (1948), pp. 278-281.

23. Submarine ASW Encounter Simulation Model Detection Validation (U), ASW Systems Project Office, Systems Analysis Group (SECRET), 1967, p. A-2.

24. Ibid., pp. 12-14.

25. Ibid., pp. A-3 - A-6.

26. Ibid., pp. A-7 - A-10.

27. Tocher, K. D., "Extension of the Neyman-Pearson Theory of Tests to Discontinue Variates," Biometrika, 37 (1950), pp. 130-144.

Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Naval Postgraduate School Monterey, California 93940	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE

Procedures and Statistical Methodology of Model Validation

4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)

Master's Thesis; December 1969

5. AUTHOR(S) (First name, middle initial, last name)

James Matthew Arrison, III

6. REPORT DATE December 1969	7a. TOTAL NO. OF PAGES 55	7b. NO. OF REFS 27
8a. CONTRACT OR GRANT NO b. PROJECT NO. c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Naval Postgraduate School Monterey, California 93940

13. ABSTRACT

Determining the effectiveness of a computer simulation model in duplicating a desired real world phenomenon is an important unsolved problem. The purpose of this paper is to model the validation procedure in a broad context and develop a general methodology for the statistical part of validation. A procedure calling on utility, decision, simulation, and statistical theories is developed. The goals of statistical testing are presented, and the assumptions, properties, and results of several parametric and nonparametric tests are discussed and compared.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

Security Classification

A-31408

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Model validation						
Statistical methodology of model validation						
Validation procedures						

DD FORM 1 NOV 65 1473 (BACK)