

DESIGN OF ENVIRONMENTAL EXPERIMENTS AND RELIABILITY PREDICTION

A. Bulfinch
Feltman Research and Engineering Laboratories
Picatinny Arsenal

PREFACE. This paper demonstrates how accepted statistical techniques stand to reduce the cost of testing missiles and missile components. These techniques are not treated in full; however, examples of their use and references are given.

The subject matter covered is oriented towards practising reliability engineers. It is hoped that some of their troublesome problems have been clarified.

ABSTRACT. Describes the need for, and use of factorial designs in surveying the separate effects of a large number of environmental treatments with maximum reliability and a minimum number of test specimens. Shows how information from the factorial experiment can be used to define reliability, and how this information can be used in tests of increased severity to predict "reliability-in-use."

SUMMARY. Reliability is defined.

The need for testing to failure is emphasized by comparing construction engineering problems with missile reliability problems. Methods are given to convert "safety margins" to measures of probability which can be used to predict "reliability-in-use."

The advantages of using factorial designs to survey the separate effects of a large number of environmental conditions are described. Information obtained from factorial experiments is used to formulate the definition of reliability in terms of the severest environments. These environments are used in tests of increased severity. Tests of increased severity are used to establish the relationship between use conditions and test conditions in order to predict "reliability-in-use."

Detailed examples are given of methods of predicting high "reliabilities-in-use" with small sample sizes.

CONCLUSIONS.

1. Any particular component can have many reliability values simultaneously. There is one reliability value associated with each possible combination of environmental condition and measurable functioning characteristic.

2. The results of factorially designed environmental experiments should be used in defining component reliability.

3. Tests of increased severity should be used in combination with factorial designs to predict "reliability-in-use" from test results.

4. Tests of increased severity can demonstrate high "reliabilities-in-use" with small sample sizes.

RECOMMENDATIONS.

1. Factorial designs should be used in combination with tests of increased severity to predict "reliability-in-use."

2. The test condition used should be experimentally determined.

3. Military standards should be revised to permit the experimental determination of the test conditions to be used.

4. The terms on which reliability is defined should be experimentally determined.

... ----- ...

I INTRODUCTION

A General

The statistical aspects of reliability are not new. All of the necessary concepts are adequately treated in modern statistical literature. The lack of information about measurable characteristics of the missile system and the environment it experiences in use, as well as the high cost of test specimens, have created the current problems.

Urgently needed are highly efficient, experimental techniques that can be uniformly applied by various segments of the same organization and by different organizations. Highly efficient techniques are required because of the need to demonstrate very high reliabilities with very small sample sizes. High reliabilities are required, of course, to assure successful functioning of complex systems composed of many components. Only small samples can be used because of the high cost and/or scarcity of test specimen. Uniform, standardized procedures are required for the collection of comparable data.

B Purpose

This paper purposes to do the following:

1. Define component reliability.

2. Describe how factorial designs can be used to survey the effects of several environmental conditions (with minimum sample size) preparatory to defining reliability in terms of these conditions.

3. Describe how "reliability-in-use" can be predicted from results of laboratory tests of increased severity.

4. Describe how tests of increased severity can use information from factorial experiments.

C Scope

The accepted statistical definition of reliability follows: Reliability is "the probability of successful functioning in use." This is a general definition that is applicable to any operating system. To define component reliability, the general definition must be modified to include:

a. Environmental conditions under which successful functioning took place.

b. The component characteristics that functioned successfully.

This means that every component can have as many reliabilities as the total number of possible combinations of environmental conditions and component characteristics. To have the weapon system reliability meaningful, the component reliabilities should be defined in terms of the most severe conditions so that the stated component reliability will be the minimum.

D Background

The techniques described, the most efficient known, are designed to maximize the amount of information obtainable from a given sample size. In addition, these techniques are definitive enough to serve as standard procedures throughout the same or different organizations over an extended period of time.

The uniform applicability of these techniques is as important as the efficiency. A large part of the value of experimentally determined reliability data is their scope of applicability. That is, reliability data collected by means of standardized procedures are cumulative in the mathematical sense. Hence, the precision with which reliability values are known can be improved with time as additional data are collected. This makes it possible to collect a reference file of reliability data on a variety of standard components.

II STATEMENT OF PROBLEM

In the usual case, the development engineer has one or more items to test under many different environmental conditions. The items to be tested may have two or more properties that must be evaluated.

Objectives of the usual reliability experiments for newly developed items follow:

1. Determine how well the engineer has succeeded in developing a reliable item.
2. Obtain an unbiased estimate of the "strength" (i.e., ability to withstand stress) of the item with minimum cost.
3. Determine the separate and combined effects of the environmental treatments on the reliability of each property measured.
4. Determine the effect of the length of time under the separate and combined effects of the environmental treatment on the reliability of each property measured.
5. Predict the "reliability-in-use" from the test results.

Objective 4 requires life-testing techniques which have been treated extensively elsewhere. This paper describes other techniques for increasing test severity.

The multiple properties of an item can be measured simultaneously. This poses no particular problem. The real problem confronting the engineer derives from his having different components that must be treated with a large number of different environments on a very limited budget.

Since most components are unique, they must be treated as separate problems. There is no known way to combine different kinds of components into a single integrated experiment. However, a single integrated experiment can be designed for several kinds of environments. As a result, the problem is one of designing the most efficient experiment for a single type of item and repeating the process for each type.

It is assumed that in every missile component there exists a true but unknown "strength," created by the particular design developed and used by the engineer in building the component. It is further assumed that the true "strength" is a constant and not a random variable for any particular design over short periods of time.

The present practice of designing components to pass the current military standards without failure does not attain the intended objectives for the following reasons:

1. An unbiased estimate of the true strength of the item cannot be obtained unless some items fail.
2. Cost prohibits testing all items at the same level of severity.
3. Reliabilities are demonstrated only in proportion to the number of items tested.

Testing without failure all items at the same level of severity can lead to completely erroneous conclusions. For example, in comparing two designs, there may be available more test specimens of the poorer design. Tests by military standards result in an equal number of failures for the two designs. Under these conditions, experimental evidence favors the poorer design.

Instead of subjecting all test specimens to the same test conditions regardless of their intended use, the level of severity of the test should be progressively increased until a failure is obtained. This procedure will lead directly to an unbiased estimate of the true strength with a minimum number of test specimens and establish the correct level of severity for each type of component at which the failure rate should be determined. This procedure will correctly differentiate between different designs.

Experience in the construction engineering field has shown that assurance that an item will not fail in use requires a large "safety factor" to be built into the item. To determine the "safety margin," the load applied must be increased until the test item breaks, or fails in some other manner. This is, in effect, a test of increased severity that leads directly to an estimate of the true breaking load the engineer is seeking to determine. The average, and standard deviation of only a few (3-6) such results are all that are required, because each value so obtained is an estimate of the true value. The difference between the "observed average breaking load" and the "load expected in use" divided by the "load expected in use" is called the "safety margin" or margin of safety. The larger this value is, the "safer" the engineer feels in predicting that the item will do the intended job without failing.

The construction engineer could have elected to load each test specimen with only the load expected in use, but because he has designed the item to withstand the load expected in use, this procedure tells him nothing about the true breaking load. All he learns is what he already knows--that it will not fail! Now if he wants to "feel safe" in predicting invariably successful functioning of the item, he must test many items. Asked to conduct his test in this manner, the engineer would rebel because he knows, as we do, that--it is far too costly!

In missile component testing, we should simulate the procedure used by the construction engineer—"load the item until it breaks" and then calculate the "safety margin." To do this, we must shift our attention from finding the number of items to be tested without failure at a single level of severity to finding the level of severity that will cause failure, and then finding the failure rate at that level. That is, to find the reliability (the probability of success) we must first find the failure rate (the probability of failure). To do this with small sample sizes, we must use a test of increased severity to find the level that will cause failure.

Robert Lusser, formerly of Redstone Arsenal, has advocated this approach of "testing to failure" for some time (Ref. 1). However, he neither showed how to "load" missile components until they fail, nor

predicted the "reliability-in-use" by means of the laws of probability. He was satisfied with using "margins of safety" (Ref. 2).

III APPLICATION OF PROBABILITY LAWS

An item will not fail until the applied stress exceeds the item's "strength." If the "strength" is much greater than the stress expected to be experienced in use, the chance (probability) of failure in use is very small, and the chance of success (reliability) is very high. It is in this sense that "high reliability" is defined. That is, high reliability means high probability of successful functioning under actual use conditions; it does not mean high reliability under the test conditions.

To translate the reliability demonstrated under test conditions to a "reliability-in-use" value, the relation between the "use" and "test" conditions must be known. Experience has shown that this relationship can be adequately represented by frequency distributions. This places the relationship on a probabilistic basis, and makes possible the use of the laws of probability.

When the average of the conditions in use is known, the level of severity required at which items must be tested to demonstrate any given reliability can be calculated in advance. As a result, reliabilities can be correctly predicted with small sample sizes without testing to failure. Alternatively, when the ultimate strength of the item is desired, the first failure method described below can be used. Both of these procedures predict "reliabilities-in-use" with small sample sizes through the use of the multiplication law which states that the probability of simultaneous occurrence of two independent events equals the product of the probabilities of separate occurrences of the events. Examples of both of these procedures are given below.

In reliability testing, the two simultaneous events referred to are test specimen failure and test condition causing that failure. Both the "failure rate" and the "chance" of the test condition's occurrence in "use" can be considered probabilities. By the above law, the predicted failure rate in use will be the product of the failure rate obtained in testing and of the chance that the test conditions could occur in use. The predicted "reliability-in-use" will then be equal to one minus this product.

When nothing is known about the environmental conditions expected in use, or when these conditions may vary in an unpredictable way, no prediction can be made about a unique "reliability-in-use." However, these methods of testing to failure can still be used to advantage. Knowing how much punishment components can take before failing reduces the number of unknowns, can be valuable in deciding how, and under what conditions, a particular missile can be used. This information can be useful in choosing between missiles for particular purposes. Moreover, where nothing is known about the conditions in use, the "most severe use condition" the item can be subjected to for any specified reliability can be calculated. An example is given below.

The methods described below show how missile components can be loaded (tested) to failure through the use of tests of increased severity. These methods also show how safety margins can be used to predict "reliability-in-use."

IV LABORATORY TEST METHODS

It is assumed in these methods that the test item can fail in but one way; that is, the binomial distribution is applicable.

A Factorial Designs

Plans should be made to conduct the laboratory experiments in two stages. First, survey the separate effects of the several environmental conditions of interest in one integrated experiment. The two-to-the-nth factorial designs or their optimized modifications are the most efficient for this purpose. These designs can be used to select the treatments causing the highest failure rates. These treatments can then be used to define the reliabilities of the test item. If the reliabilities determined in terms of these treatments are acceptable, the reliabilities of the test item in terms of any of the other treatments will also be acceptable. This procedure will reduce the magnitude and complexity of the experiments conducted to determine and predict reliability. More importantly, component reliabilities obtained in this manner will furnish a more realistic basis for calculating systems reliability.

See references 5 and 6 for available designs. These designs are the most efficient known. Experiments based on these designs may be conducted without changing the treatment procedure except to arrange for the test specimens to receive the number and kind of treatments required by the particular design used. However, the best differentiation among treatments is obtained when the level of severity used will cause 50 percent of the test specimens to fail.

For the purpose of this application, the two levels of each treatment can be the presence and absence of the treatment. Alternatively, any two levels of the treatments can be used.

The number of test specimens required in the optimized designs is one more than the total number of treatments used (Ref. 5). The more versatile fractional factorial designs (Ref. 6) require at least 16 items for experiments containing from five through eight treatments, and at least 32 items for nine through 13 treatments. With twice these numbers of items, the latter type designs can also measure interactions--how the effect of any one environment depends upon the others. Interactions among treatments cannot be measured by any design except the factorial.

Factorial designs permit a type of statistical analysis that distinguishes between variations due to chance and variations having assignable causes, and produces more information from a given number of items than any other known procedure. These designs actually

increase the effective sample size by making it possible to use each observation (or measurement) for more than one purpose. In fact, each treatment effect is determined as though the entire experiment is conducted to determine that particular treatment effect alone. As a result, the reliability with which each treatment effect is determined can be based on the total number of items used in the experiment. The three-treatment-design example described below demonstrates this point.

Further advantages in using factorial designs in environmental testing experiments follow:

a. No control groups are required.

b. Each treatment effect can be determined independently of all the others. That is, unambiguous conclusions can be drawn about each treatment effect.

c. Complex experiments involving a large number of treatments can be easily handled with the factorial procedures.

d. This is the only experimental design in which the relationship among the treatments can be measured. That is, the factorial design can determine whether the effect of one environmental treatment depends upon any of the others. These effects are called interactions.

e. The probability of being right or wrong can be controlled.

f. When the number of treatments used becomes large (three or more), only a fraction ($1/2$, $1/4$, $1/8$, etc.) of the total number of combinations in the factorial need be used.

When multiple replications cannot be used and only attribute (go, no-go) data are available, these designs can still be used to take advantage of their efficiency. However, in cases of this kind, the usual analysis of variance cannot be made. Instead, the usual summations are made to obtain and compare two binomial proportions (by the Fisher exact method) to determine the effect of each treatment. See Example No. 1 below.

Results of factorial designs are used as a guide in determining how to define reliability prior to conducting the test of increased severity. That is, the factorial experiment surveys all of the environmental treatments of interest (with a minimum number of test specimens) to determine the difference, if any, among the environmental effects. A decision is then made whether to redesign the item. If the item is considered acceptable at this time, reliability is defined in terms of the environmental treatment or treatments found to be most severe. If no differences are found among the effects, reliability can be defined in terms of a combination of several of the treatments considered most important from an engineering point of view. If reliability is defined in terms of the most severe treatments, the reliability values obtained will be smaller than those obtained with the other treatments. This is a necessary condition if the system's reliability derived from the component's reliabilities is to have meaning.

Tests of Increased Severity

Results obtained from the factorial experiments can be used to determine which of the environmental treatments will be used in the following procedures to predict "reliability-in-use."

B First-Failure Method (Single Factor)

Increase the level of severity after each test result is obtained until the test item fails. If the test destroys the item, increase the level of severity used with each succeeding item tested until a failure is obtained.

The level of severity can be increased in a variety of ways, such as the following:

1. Using more extreme treatments (e.g., higher temperatures or higher G-values).
2. Using two or more treatments on each test specimen.
3. Repeating the same treatment or set of treatments on the same item.

NOTE: Increasing the length of time an item is subjected to a particular treatment is not used here as a means of increasing the degree of severity.

By starting at, or near the level of severity expected in use, a failure should be obtained within five or six trials (or items). After the level of severity has been found that will cause failure, three or more items should be tested to estimate the failure rate at this level.

To determine the predicted "reliability-in-use," find the probability of occurrence in use of the test condition (at which the failure rate was measured) from a table of individual terms of the Poisson distribution (such as Table 39 of reference 9), where " m " is the expected use condition used. This in effect determines the probability associated with the "safety margin." The product of this probability value and the failure rate found under the test condition is the predicted failure rate in use. The predicted "reliability-in-use" is equal to one minus this product.

V EXAMPLES**A Example No. 1**

This example demonstrates how factorial designs can be used in combination with tests of increased severity. A simple three-treatment-experiment example is given below. The treatments used in this example are identified and defined as follows:

Design of Experiments

Identification

A

B

C

Treatment

Trans. Vib.

Flight Shock

Waterproofness

For purposes of the factorial design, each treatment is considered to have two levels:

1. Lower level or absence of the treatment (designated by subscript 1).
2. Higher level or presence of the treatment (designated by subscript 2).

The total number of possible combinations of three treatments, each at two levels, is two cubed or 8. These 8 combinations can be written in the following pattern:

		<u>A₁</u>			<u>A₂</u>		
		<u>B₁</u>	<u>B₂</u>			<u>B₁</u>	<u>B₂</u>
<u>C₁</u>	(1)	b		a		a + b	
<u>C₂</u>	c	b + c		a + c		a + b + c	

A minimum of 8 items would be required for this plan, each receiving different treatment combinations as follows:

<u>Item Number</u>	<u>Treatment Combinations</u>
1	None (1)
2	B only
3	A only
4	A + B
5	C only
6	B + C
7	A + C
8	A + B + C

By using the same letters (A, B, and C) and symbol (1) to represent the results obtained from testing the eight items, it can be shown symbolically that the treatment effects can be independently determined, using the total number of items in the entire experiment for each treatment as follows:

EFFECT OF TREATMENT A

$$A + (A + b) + (A + c) + (A + b + c) - \\ [(1) + b + c + (b + c)] = 4A$$

EFFECT OF TREATMENT B

$$B + (B + c) + (a + B) + (a + B + c) - \\ [(1) + c + a + (a + c)] = 4B$$

EFFECT OF TREATMENT C

$$C + (b + C) + (a + C) + (a + b + C) - \\ [(1) + b + a + (a + b)] = 4C$$

One-fourth of these differences equals the average effect of the respective treatments. From the above equations, it can be seen that the results obtained from the eight items have been used three times--once for each treatment. This produces an effective sample size equal to 3×8 , or 24 items; yet, each treatment has been determined independently of the others.

The above three-factor factorial can be used as an example of a fractional factorial design as follows:

<u>A₁</u>		<u>A₂</u>	
<u>B₁</u>	<u>B₂</u>	<u>B₁</u>	<u>B₂</u>
C ₁	b	a	-
C ₂	c	-	a + b + c

A minimum of four items is required in this design. As before, the separate effects can be determined by a process of summation and subtraction as follows:

EFFECT OF TREATMENT A

$$A + (A + b + c) - (b + c) = 2 A$$

EFFECT OF TREATMENT B

$$B + (a + B + c) - (a + c) = 2 B$$

EFFECT OF TREATMENT C

$$C + (a + b + C) - (a + b) = 2 C$$

One-half of these differences equals the average effect of the respective treatments.

When there is only one item available for each treatment combination, and only success and failure data are available, the usual analysis of variance cannot be used. However, the above differences, which will be binomial proportions in this case, can be compared by the Fisher exact method for 2 x 2 tables (Ref. 7) to determine the treatment effects. A very convenient set of tables for this purpose can be found in Ref. 8, which contains tables of minimum contrasts based on Fisher's exact method.

When it can be determined, from the results of the factorial experiment, which environmental conditions will be used to define reliability, the level (or severity) of the condition required to demonstrate a given "reliability-in-use," with a small sample, can be calculated in advance of testing, on the assumption that no failures will be obtained; if the average condition in use is known.

The test conditions required can be calculated as follows:

$$R = 1 - P_t \text{ (UCL)}$$

where:

R = the specified "reliability-in-use."

P_t = the probability of test conditions^t occurring in use.

UCL = the upper confidence limit (associated with the specified confidence level) of the failure rate expected under the test conditions to be calculated below.

When R and UCL are known, P_t can be calculated from the above formula. Given P_t (the probability) and the average use condition (m), the required test condition (i) can be found from Table 39 of reference 9, or from the following formula:

$$P_t = m^i e^{-m} / i !$$

where (e) is the base of natural logarithms.

1 Sample Calculations

Using the same three-factor-experiment example as above gives the following typical set of results, when one is entered as a "failure" and zero is entered as a "success." It is assumed that a knowledge of the item being tested has led to the decision that transportation vibration, flight shock, and waterproofness, in that order, are the three environmental conditions most likely to affect the important functioning characteristic of this item; this characteristic is contact resistance.

The treatment procedure and work-sheet (to record results) for this experiment would be the following two-entry table. An "X" in the item column means that the item receives the corresponding treatment, while a blank means that the item does not receive the treatment.

Treatment Procedure

Order of Treatment	Item No.							
	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32
Trans. Vib. (A)			X	X			X	X
Flight Shock (B)		X		X		X		X
Waterproofness (C)					X	X	X	X
Results: Replication 1	0	0	1	1	1	1	0	1
	2	0	0	1	0	1	0	0
	3	0	1	1	1	1	0	1
	4	0	1	0	1	0	0	0
Totals	0	2	3	3	3	4	0	2

The results of one complete replication should be obtained under a single set of controlled conditions (e.g., in the same day, same operators, same instruments, etc.), before going to the next replication. This will make it possible mathematically to subtract out of the results the effect of changing conditions.

By placing these results in the usual factorial matrix, the following table would be obtained:

	<u>A₁</u>		<u>A₂</u>	
	<u>B₁</u>	<u>B₂</u>	<u>B₁</u>	<u>B₂</u>
C ₁	0	0	1	1
	0	0	1	0
	0	1	1	1
	0	1	0	1
	0	2	3	3
C ₂	1	1	0	1
	1	1	0	0
	1	1	0	1
	0	1	0	0
	3	4	0	2

In preparation for analyzing these results, the usual summing process would give the following series of two-factor tables:

Summing over A

	<u>B₁</u>	<u>B₂</u>	<u>Row Totals</u>
C ₁	3	5	8
C ₂	<u>3</u>	<u>6</u>	<u>9</u>
Column Totals	6	11	17

Summing over B

	<u>A₁</u>	<u>A₂</u>	<u>Row Totals</u>
C ₁	2	6	8
C ₂	<u>7</u>	<u>2</u>	<u>9</u>
Column Totals	9	8	17

Summing over C

	<u>A₁</u>	<u>A₂</u>	<u>Row Totals</u>
B ₁	3	3	6
B ₂	<u>6</u>	<u>5</u>	<u>11</u>
Column Totals	9	8	17

Each one of the marginal totals is the sum of 16 observations. The results can now be analyzed and interpreted as follows:

Source	Effects	Test of Significance ^a
<u>Main Effects</u>		
Trans. Vib. (A)	9/16 vs 8/16	non-significant
Flight Shock (B)	9/16 vs 11/16	non-significant
Waterproofness (C)	8/16 vs 9/16	non-significant
<u>Replication</u>		
1	5/8	non-significant
2	3/8	non-significant
3	6/8	non-significant
4	3/8	non-significant
<u>Interactions</u>		
A x B	8/16 vs 9/16	non-significant
A x C	4/16 vs 13/16	significant
B x C	8/16 vs 9/16	non-significant
A x B x C	7/16 vs 10/16	non-significant

^a By the Fisher exact method for the 95% (two-sided) confidence level

2 Interpretation (when the above order is used)

(a) None of the effects is significant except the AC interaction. This means that items which have received transportation vibration treatment are significantly less waterproof than those not receiving transportation vibration.

(b) None of the treatments taken alone is significant, although the flight shock effect approaches significance. This result suggests the need for additional flight-shock tests if this treatment is considered important from an engineering point of view.

(c) The fact that the three-factor (ABC) interaction is not significant shows the following:

(1) Waterproofness does not change the effect of transportation vibration on flight shock (AB interaction).

(2) Flight shock does not change the effect of transportation vibration on waterproofness (AC interaction).

(3) Transportation vibration does not change the effect of flight shock on waterproofness (BC interaction).

(d) The fact that replication is not significant means that conditions were under a state of control throughout the experiment.

These results show clearly that the effect of transportation vibration on waterproofness is the most severe combination. It would appear from the results that a decision to improve the waterproofness characteristics is required. After this has been done, reliability must be defined. The results of this experiment show that reliability should be defined in terms of contact resistance (the functioning characteristic of interest) under the following environmental conditions:

(1) Transportation vibration followed by waterproofness (since these two conditions interact).

(2) Flight shock (since this treatment effect approaches significance).

If the reliability of the contact resistance under these conditions is acceptable, the reliability of the contact resistance under the other conditions will also be acceptable.

If the average transportation vibration condition in use is assumed to be 5 G's and the required reliability is 0.995, the test condition required to demonstrate this reliability with a sample of 5 test specimens can be calculated as follows:

$$\text{when: } R = 1 - P_t \text{ (UCL)}$$

$$\text{then: } P_t = \frac{1 - R}{\text{UCL}}$$

$$\text{when: } R = .995$$

$\text{UCL} = .52$ (the upper confidence limit at the two-sided 95% confidence level for testing five items and obtaining no failures)

$$\text{then: } P_t = \frac{1 - .995}{.52} = .0096$$

From Table 39 of reference 9 the test condition (i) associated with a use condition (m) of 5 G's and a probability (P_t) of .0096 is found to be equal to 10.9 G's. This is the level of transportation vibration required, followed by the waterproofness test to demonstrate a contact resistance reliability of 0.995 if no failures are obtained. If failures are obtained, the test conditions required to demonstrate a reliability of 0.995 with a sample of five test specimens are as follows:

<u>Test Specimens</u>	<u>% Failure</u>	<u>Test Condition Required, in G's</u>
0	0	10.9
1	20	11.3
2	40	11.5
4	80	11.7

NOTE: It is evident from the above sample calculations that, for small-sample sizes, the difference in test conditions between zero and anything less than 100% failures is insignificant. This means that the required test condition can be conservatively estimated by expecting a high failure rate.

B Example No. 2

When the average of the conditions in use is known, very high values for the "reliability-in-use" can be correctly predicted with very small sample sizes, if the level of severity is increased until a failure is obtained:

Given:

Average use conditions (m) = 5 G's

Found (Using First - Failure Method):

Number of items used to find test condition to cause first failure	5
--	---

Test condition found (i)	18 G's
--------------------------	--------

Number tested at 18 G's	5
-------------------------	---

Number of failures at 18 G's	2
------------------------------	---

The probability (P_t) of the test condition's occurring in use, from Table 39 of reference 9, when $m = 5$ and $i = 18$, is found to be $P_t = 0.000004$. The upper confidence limit (UCL) of the observed failure rate (2/5) for the two-sided 95% confidence level, from Table V of reference 8, is found to be $UCL = 0.8534$.

Since:

$$R = 1 - P_t \text{ (UCL)}$$

Then:

$$R = 1 - (0.000004) (0.8534) = 0.9999966, \text{ which is the predicted "reliability-in-use."}$$

C Example No. 3

When nothing is known about the expected conditions in use, the "most severe condition in use" an item can withstand can be calculated:

Given:

Required "reliability-in-use" = 0.9999

Test condition used (i) = 10 G's

Number of items tested = 10

Number of failures obtained = none

Since:

$$R = 1 - P_t \text{ (UCL)}$$

$$P_t = \frac{1 - R}{\text{UCL}}$$

From Table IX of reference 8, UCL for no failures in 10 trials equals 0.3085, for the two-sided 95% confidence level.

Then:

$$P_t = \frac{1 - 0.9999}{0.3085} = 0.000324$$

From Table 39 of reference 9, the "most severe condition in use" (m) for $i = 10$ and $P_t = 0.000324$ is found to be:

$$m = 2.6 \text{ G's}$$

With the given test result and test condition, this is the "most severe condition in use" under which the item will have 0.9999 reliability.

REFERENCES

1. Lusser, Robert, "Unreliability of Electronics - Cause and Cure," Redstone Arsenal, Huntsville, Ala., Nov. 1957.
2. Lusser, Robert, "Reliability Through Safety Margins," Redstone Arsenal, Huntsville, Ala., Oct. 1958.
3. Dixon and Massey, "Introduction to Statistical Analysis," McGraw-Hill Book Co., Inc., NYC, 1957, Second Edition.
4. Davies, O. L. (Editor), "Statistical Methods in Research and Production," Hafner Publishing Co., NYC, 1957, Third Edition.
5. Plackett and Burman, "The Design of Optimum Multifactorial Experiments," Biometrika, Vol 33, 1946, page 305.
6. "Fractional Factorial Experiments Designs for Factors at Two Levels," NBS Applied Math. Series No. 48.
7. Fisher, R. A., "Statistical Methods for Research Workers," Hafner Publishing Co., NYC, 1950.
8. Mainland, Herrera, and Sutcliffe, "Tables for Use With Binomial Samples," Department of Medical Statistics, NY University College of Medicine, NYC, 1956.
9. Pearson, E. S., and Hartley, H. O., "Biometrika Tables for Statisticians," Volume I, Cambridge University Press, Cambridge, England, First Edition, 1956.

MULTI-DIMENSIONAL STAIRCASE DESIGNS FOR RELIABILITY STUDIES

David R. Howes
U. S. Army Chemical Corps Engineering Command

This paper suggests the need for a sequential staircase procedure whereby a given contour of response could be traced experimentally without necessarily defining the entire response surface. Such a method would have important application in reliability studies, in design and engineering, etc. where the intent is to hold malfunctioning of some type at some predetermined level.

As an example, suppose that an artillery shell is to be filled with poison gas and closed with a burster tube (Figure 1, on the next page). It has been found that the leakage of these shells is affected by two variables, the interference of the burster tube (Variable A) and a variable B which is a structural characteristic of the shell body.

Although it may be possible to specify levels of A and B which will be satisfactory, it is also necessary to know the threshold of leakage in order to set manufacturing tolerances, filling procedures, etc. This involves the response surface generated in leakage in the A-B space.

Possible Methods

1. Factorial

It is possible to fit a response surface to the results of the experiment shown in Figure 2 (see next page) by well-known methods. The sample size, 1800, may seem excessive for accuracy which cannot exceed 2%.

2. Confined Factorial and Staircase Method

Staircase methods have been described which permit the isolation of percentage points of a problem with only one variable:^{1,2,3} For a two variable case it would be possible to treat one variable by a staircase method, and the other factorially. (See Figure 3).

3. Multi-Dimensional Staircase Method

An extension of the staircase method to N variables is possible, although the methods have not been produced yet, since the theory doesn't exist. We would assume that a smooth response surface existed and that it would be possible, staircase-wise to follow some response contour on that surface say 90%, 93%, 99%, etc. Interaction of the variables is also a possibility.

The factorial approach may be inefficient here since it collects data not needed merely to trace a contour. It may also be wasteful of time, materials, and manpower, since the experiment cannot proceed until a large predetermined number of items are available. It should also be mentioned that it runs against the strong desire usually found

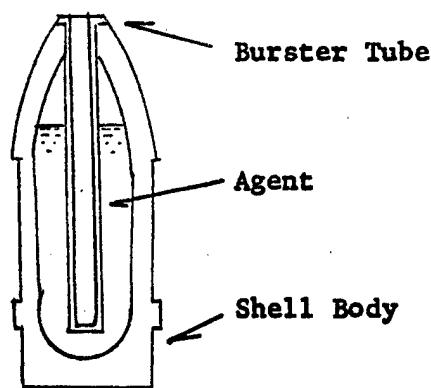


Figure 1.

Agent Filled Shell

		Number of Shells					
		1	2	3	4	5	6
Variable A	Level	50	50	50	50	50	50
	6						
	5	50	50	50	50	50	50
	4	50	50	50	50	50	50
	3	50	50	50	50	50	50
	2	50	50	50	50	50	50
	1	50	50	50	50	50	50

Figure 2.

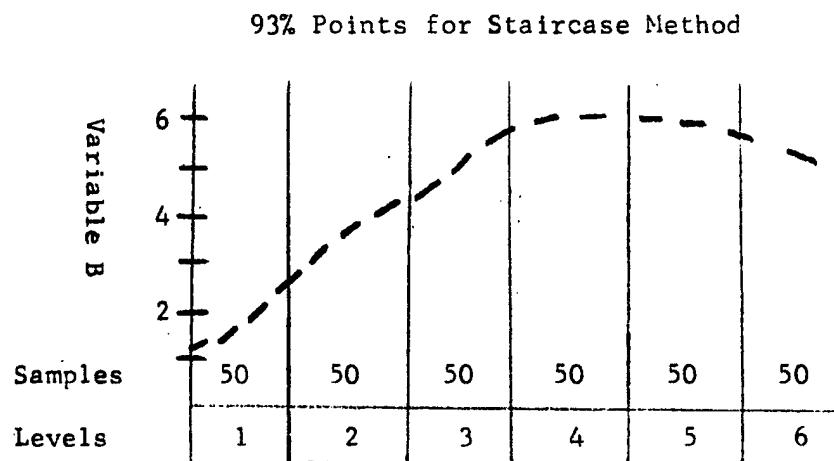


Figure 3.

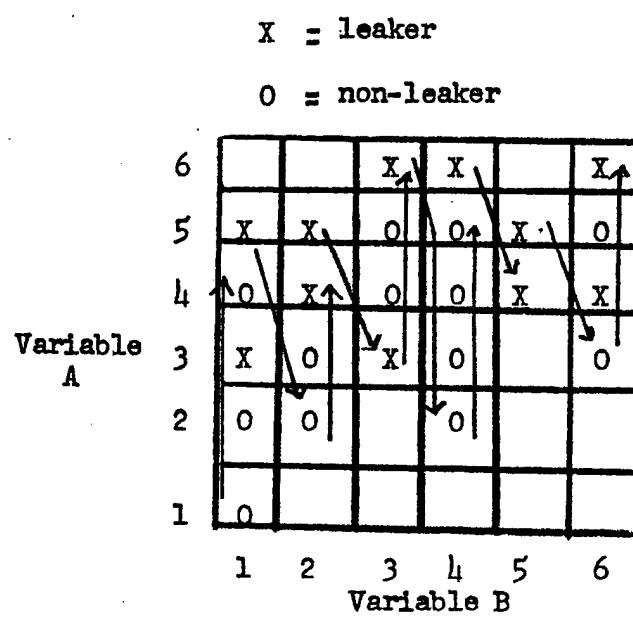


Figure 4.

among reliability engineers to try something, then try something else. Since the statistician must live with this, it would seem most desirable to adopt procedures which resemble to as great an extent feasible, those used by the reliability and test engineers. The sequential nature of the procedures has also the advantage of allowing experimentation to continue, while new experimental vehicles are being fabricated.

Example of a Multi-Dimensional Method:

In this procedure we might try a single sample at a time, and move over the response surface in accordance with certain rules based on the previous test result. (See Figure 4).

In this design, we proceed upward one step at a time in A in level No. 1 until two leakers have been found, then proceed to the next higher level of B at a level of A, one level below that on which the first leaker occurred. Then proceed upward in A until again two leakers have been found and proceed as before. Take the mid-point on A between the two leakers at each level of B as a point on the 50% leakage contour. Repeat this experiment as many times as necessary to get the desired precision of estimate. Instead of taking a single sample, we might take a sample of n and consider it "reliable" if no more than c leakers are found. This would, I suppose, lead to the tracing of percentage lines. Using a table of random numbers, I was able to get fairly good results in tracing a 10% contour over a bi-variate surface; good enough results to suggest the desirability of answering the following questions:

(1) What are the most effective rules to follow when traversing the response surface?

(2) What confidence can be placed in the results of K trials, using a sample size n in each trial, with an allowance of c defects as an estimate of the c/n fraction line?

(3) What method of computation or statistic would be used to obtain the c/n fraction line estimate from the data?

REFERENCES

1. National Defense Research Council Applied Mathematics Panel Report No. 101.1
2. Bureau of Ordnance, U. S. Navy, NAVORD Report No. 65-46, 21 March 46 - Anderson, McCarthy, & Tukey
3. Dixon, W. J. and Mood, A. M., "A Method of Obtaining and Analyzing Sensitivity Data" Journal of the American Statistical Association, Vol 43, No 241 (1948)

A PROPOSED RESEARCH PROGRAM FOR PROVIDING
A QUANTITATIVE BASIS FOR PREVENTIVE MAINTENANCE POLICIES
ON ORDNANCE EQUIPMENT

Walton M. Hancock
and
Randall E. Cline

This talk contains an outline of a proposed program which can be used to aid in the establishment of preventive maintenance policies. The program represents a general approach applicable to both existing systems and systems yet to become operational. It is anticipated that this approach should ultimately lead to the simplification of the maintenance of Ordnance equipment.

Since the Ordnance Corps has many different types of equipment, which vary both in complexity and density, no one PM policy can be applicable to all types of equipment. The research effort must therefore be oriented towards developing the proper general approach to the establishment of preventive maintenance policies for a variety of weapons systems. The applicability and usefulness of the approach, then, can be demonstrated by selecting a limited number of weapons systems and evaluating different preventive maintenance policies for them.

The talk is presented as follows: Part 1 contains a discussion of work done by others and comments on methods used to develop general solutions. Part 2 presents a general approach to the preventive maintenance problem, and Part 3 is a mathematical formulation which has been used to present, in a compact form, the ideas developed in the general approach. The fourth part contains a proposed program which will provide information for the evaluation of preventive maintenance policies on specific weapons systems in order to serve as examples of the application of this general approach to establishing preventive maintenance policies for other weapons systems.

1. CURRENT STATUS OF RESEARCH ACTIVITY CONCERNING MAINTENANCE POLICIES. An extensive library search and a number of visits have been made in order to take advantage of work performed by other research groups in the maintenance area.¹ One finds that there has been quite a bit of effort put into the specific details of establishing preventive maintenance policies for a particular type of equipment; there are, however, relatively few people that are concerned with a general approach.

The areas in which much work has been done include the military electronics field and the civilian trucking industry. An examination of the approaches used in these areas has proved quite helpful.

In the military electronics field, investigations have been made into the problems of reliability. The reliability of equipment is directly related to the amount and type of maintenance. Methods of analysis used in evaluating and improving reliability can also be used in developing maintenance programs. Briefly, the methods have been as follows:

1. A list of the most pertinent books and articles is given in the bibliography.

- a. Rather than conduct broad scale studies, particular units have been selected for detailed study.
- b. Only new equipment or equipment put through major overhaul is issued to the units to be studied.
- c. Detailed records are kept on the life history of the equipment. The amount of scheduled and unscheduled maintenance, the time required for repairs, basic causes of failure, parts usage and the time interval between failures are all carefully documented.
- d. The reliability of the equipment is related to the mission that is to be performed, and the mean time to failure is frequently used in deriving an expression of reliability. Since the missions are usually expressed in terms of the number of hours of use per mission, then the probability that the equipment will perform the expected mission can be predicted.
- e. Emphasis has been placed in the classification of the types of failures by basic components of the system such as by tube types, types of resistors, and capacitors. The reliability of the equipment is then expressed in terms of the reliability of its components.

Considerable success has been attained by the use of the above methods. The most notable of these have been realized by classifying the basic causes of failure for design purposes.

Since maintenance is one of the principal costs in the civilian trucking industry, effort has gone into the solution of their maintenance problems. The following represents the general methods used by the industry:

- a. A detailed life history is kept on each vehicle. These data contain a record of all repairs, including parts usage and costs, and a record of all maintenance performed. Incidence of breakdowns, associated costs, and the mileage of the vehicle are recorded for such events in the life history.
- b. By analysis of the life histories, norms are established for the expected life of each major component. Careful investigations are then made to determine the causes of failures which seem premature. Failures are also classified as to design deficiencies, improper maintenance or poor driving. Those that are in the design deficiency category are used to change the specifications on new equipment. Those caused by improper maintenance are used to modify the amount, kind, and time interval of scheduled maintenance. The failures caused by poor driving are analyzed for improved driver education.
- c. The "cost of maintenance per mile" is also derived from the detailed life history. This is a control technique used by

administrative personnel to see if the total cost of maintenance is kept within prescribed limits.

Proper feedback of information and the proper analysis of this information is considered by most trucking firms to be an absolute necessity. This applies to commercial trucking fleets, bus fleets, truck and car rental fleets, and users of off-the-road equipment. The same approach prevails regardless of the actual use to which the equipment is put.

2. A GENERAL APPROACH TO THE MAINTENANCE PROBLEM. In order to develop a general approach that can be applied to all types of weapons systems, it is first necessary to classify weapons systems in such a way that their common characteristics as well as their differing characteristics are evident. Preventive maintenance policies will then be related to these characteristics. It appears that all weapons systems can be classified (for purposes of establishing maintenance policies) in terms of the following basic parameters:

- a. Complexity
- b. Density
- c. Mission

Each of these three basic classifications is a vector quantity, or stated more simply, each can be described in terms of a number of factors. For instance, the complexity of a weapons system may be defined in terms of the crew requirements, the number of components, the total cost of the system, the average amount of time required to locate troubles, the ratio of time to check out the system compared to the time to complete a mission, etc. Similarly, density may be expressed in terms of geographical dispersion of equipment, travel time from support unit to supported units, total number of units in the field, etc. Missions can be defined in terms of the time equipment is required to be operable, the movements of operations which must be accomplished, the precision with which operations must be performed, etc.

Examples of the way these classifications are related to maintenance policies are as follows: Experience has shown that for electronic equipment an increase in complexity increases the maintenance requirements. The density of weapons systems affects the organization of maintenance crews and supporting test equipment. The mission also affects the type of maintenance. Since many weapons systems are required to perform a number of different missions, to achieve simplicity of maintenance at a minimum cost, the maintenance requirements may also vary. For example: trucks that are used on hard surface roads will require different maintenance than trucks used off the road. One trucking firm that was visited had different maintenance schedules for long distance vehicles than for local haul equipment, because the cost of a breakdown of a vehicle some distance from maintenance support was many times higher than for a vehicle used locally.

3. MATHEMATICAL FORMULATION OF A GENERAL APPROACH. Using the classification of weapons systems introduced in Part 2, a general approach to establishing preventive maintenance policies will now be considered. In the analysis of failure data for electronic equipment, the term reliability of a system or a component of a system has been generally used to denote the probability that a system or component will perform its required mission under given conditions for a specified operating time. The reliability of Ordnance equipment can be defined similarly. In developing a model which relates maintenance and reliability, the following assumptions are made:

- a. Reliability is dependent upon the age of the equipment. For example, tanks, trucks, missiles, etc., tend to fail more frequently as the equipment becomes older.
- b. The reliability is also dependent upon past usage of the equipment. For example, it is expected that the number of failures in trucks increase as the number of miles traveled increases. Similarly, the number of times a missile is checked is believed to affect the probability that the missile will fire.

Since both age and past usage are assumed to affect the reliability of equipment, it is useful to redefine reliability. Consequently, the following notation will be introduced. For any given system, let t = calendar age of the system, i.e., the number of years since the equipment was issued (new) to the user, and let $x(t)$ = usage of the system prior to time t , i.e., $x(t) = \int_0^t f(\tau)d\tau$, where $f(\tau)$ is some measure of usage.

The reliability of the system will be defined as follows:

The reliability of a system is the probability that the system will perform its required mission (which includes a specified operating time), given that the age of the system, t , and the past usage, $x(t)$, and the preventive maintenance policies are known. The reliability of a system will be designated by the symbol $r[t, x(t)]$.

It has been suggested that for wheeled and tracked vehicles, $x(t)$ = mileage, and for missiles, $x(t)$ = number of times certain checks have been made on the system.

Both scheduled and unscheduled maintenance are related to the reliability of a system. The purpose of maintenance is to increase $r[t, x(t)]$, and hence to maintain the reliability above some predetermined minimal level. The policies regarding scheduled and unscheduled maintenance are expected to be influenced by the complexity and by the density of the system. Visual inspections and operational check-out procedures are designed to ascertain whether the reliability of the system is above this predetermined level. Since a weapons system may be required to perform several different missions, we must consider all operations which the system may be required to perform and the associated performance times. Classify these missions (that is operation-time combinations) in groups in such a way that all missions in any given group are roughly equivalent

in terms of requirements on the system. Such a group of missions will be called a task. Now order these tasks in such a way that if the system can perform any given task, then it can also perform all simpler tasks. Designate these tasks by T_1, \dots, T_ℓ , ($\ell \geq 2$), where T_1 indicates the simplest task and T_ℓ corresponds to the most difficult task. Having specified a given task, say T_k , there exists a corresponding probability that the system can perform this task. Denote this probability by $r_k[t, x(t)]$.

Then for fixed t and $x(t)$, and all $k = 2, \dots, \ell$,

$$r_k[t, x(t)] \geq r_{k+1}[t, x(t)].$$

Now for each task T_k and any given maintenance policy, there exists a surface $r_k[t, x(t)]$ which represents the reliability of the system relative to T_k . Such a surface is illustrated in Figure 1.

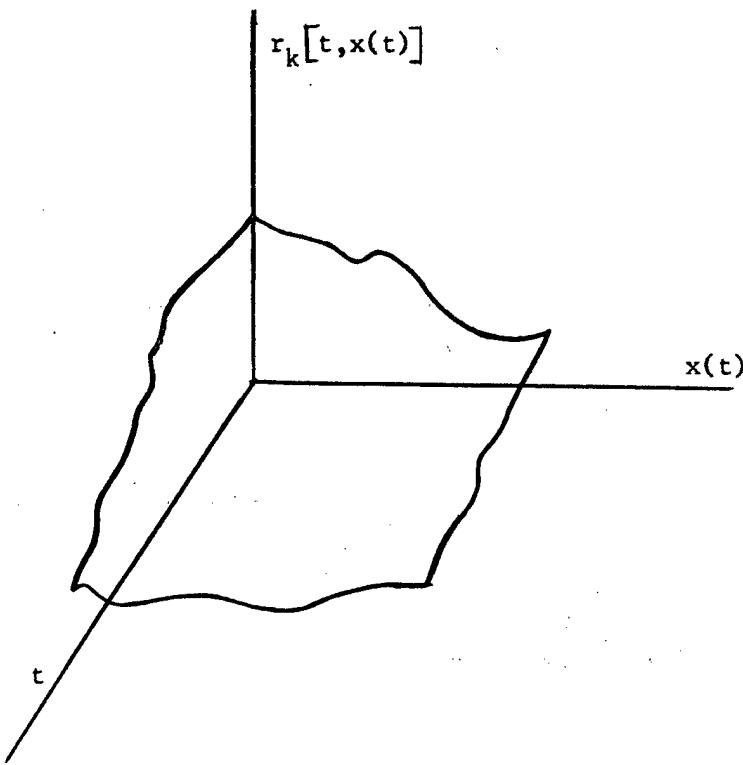


Figure 1

Theoretically, a maintenance policy should be designed in such a way that for any value of t , $0 \leq t \leq t'$, where t' is that time at which the system is discarded, and for any usage $x(t)$, the reliability of the system should be maintained in such a way that for some α , $0 < \alpha \leq 1$,

$$(1) \quad r_k[t, x(t)] \geq \alpha$$

for some task T_k . Since it is assumed that the reliability of a system decreases with both time and usage, then maintenance must be performed in an attempt to satisfy equation (1). The physical situation is illustrated in Figure 2 in which the lower surface represents the actual system reliability and the upper surface represents the desired goal of maintenance. It is to be noted that $r_k[t, x(t)]$ may actually exceed α . Conversely, the effect of inappropriate design on the system may be such that $r_k[t, x(t)]$ never attains the desired goal. Finally, it is observed that the goal may not be constant over the entire expected life of the system and may be lowered as equipment is phased out.

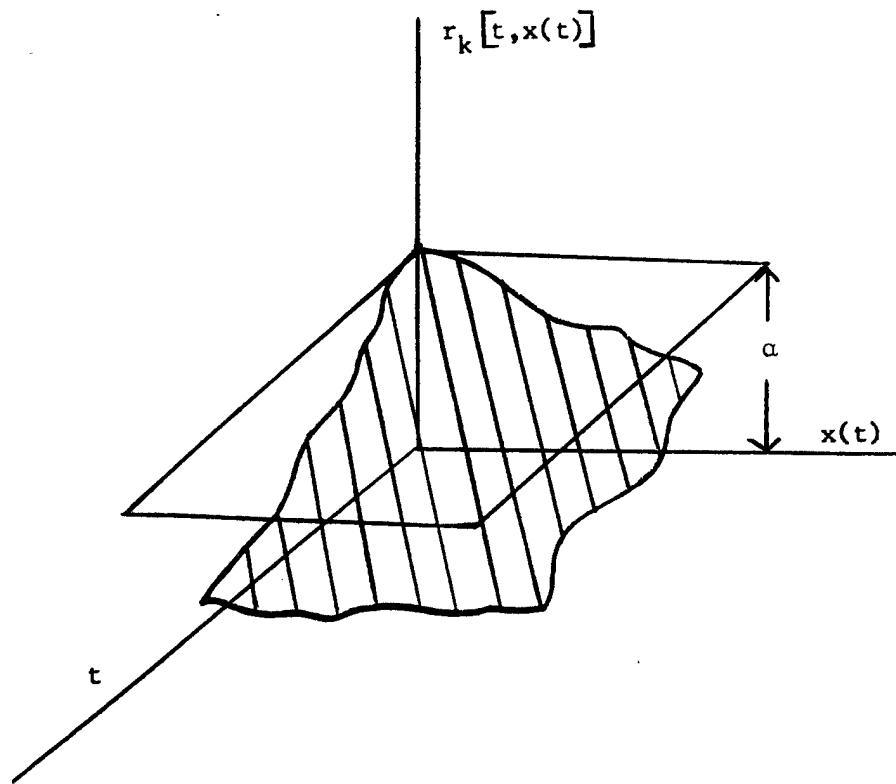


Figure 2

The life history of a system can be represented by a curve in the $t, x(t)$ plane. Associated with this curve is the corresponding reliability. Such a curve is shown in Figure 3.

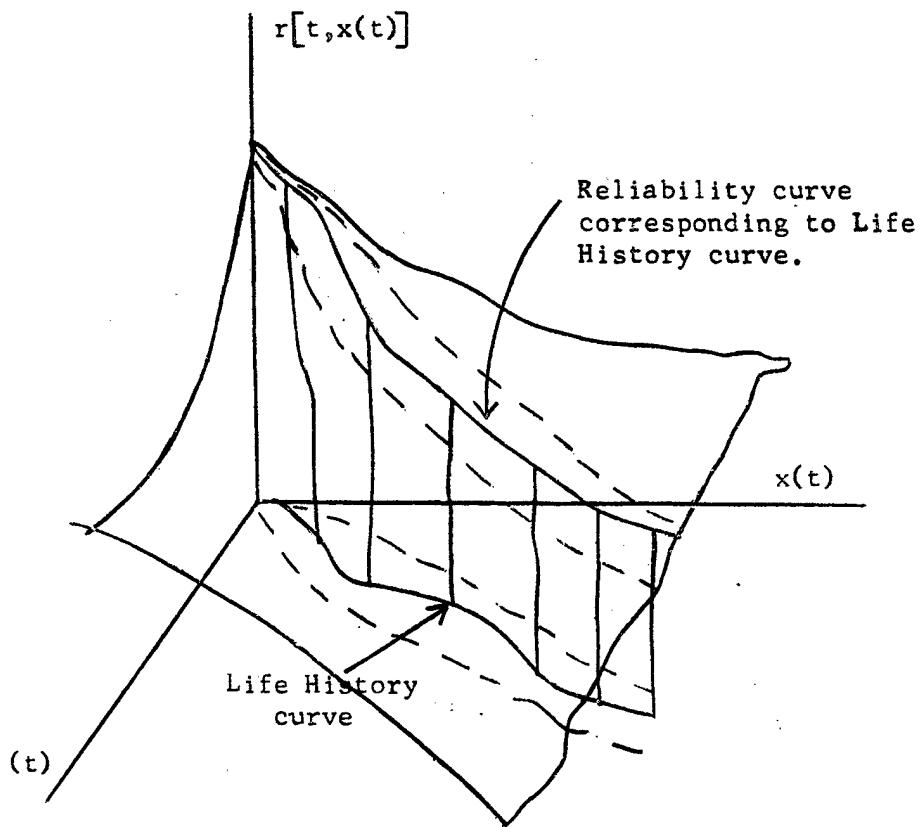


Figure 3

Observe now that since a collection of similar weapons systems in an Army unit will not be used in identical amounts, then observations taken from a particular unit will form a wedge as illustrated by the broken lines in Figure 3.

As mentioned above, maintenance of any type is intended to increase the reliability of the system. Thus, for any maintenance performed at a given point $[t, x(t)]$, let:

$$\Delta r_k [t, x(t)]$$

denote the change in $r_k [t, x(t)]$ obtained by performing the maintenance. The amount of this change is a random variable dependent upon the type of maintenance performed, the level of skill of the technician performing it, and the tools or test equipment available to him. Graphically, this may be illustrated as in Figure 4 for a particular system having maintenance performed at points $[t, x(t)]$, $[t_2, x(t_2)]$, ..., $[t_n, x(t_n)]$, where the jumps in the curve indicate the corresponding changes in reliability resulting from the maintenance.

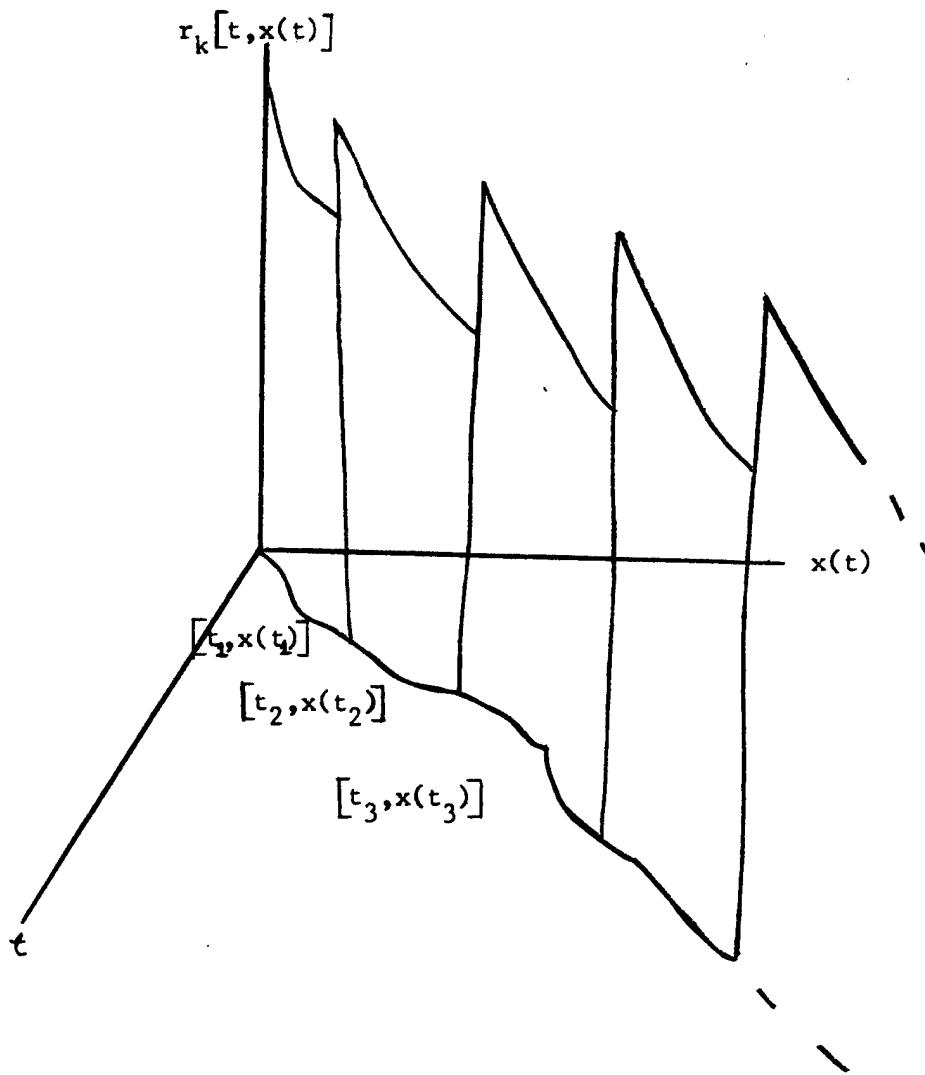


Figure 4

It is to be observed that if not only $r_k[t, x(t)]$ and $\Delta r_k[t, x(t)]$ are known, but also various costs of maintenance associated with $\Delta r_k[t, x(t)]$, then measures can be developed which relate costs of maintaining equipment to the expected loss from having equipment inoperative. Consequently, to use this approach for evaluating the effects of various preventive maintenance policies in terms of the change of the reliability of the system, it will be necessary to develop techniques for estimating the forms of $r_k[t, x(t)]$ and $\Delta r_k[t, x(t)]$ for various tasks and maintenance practices.

To aid in the estimation of these functions, it is useful to consider the effects of various components of the system on overall system performance. Now any given system can be represented as a collection of major components (or subsystems). Associated with each component is a

corresponding reliability surface again a function of t and $x(t)$. Assuming these major components are serially connected and statistically independent, then at any point $t, x(t)$,

$$r_k[t, x(t)] = \prod_{j=1}^n r_k^{(j)}[t, x(t)]$$

where there are n major components with reliabilities $r_k^{(j)}[t, x(t)]$, ($j=1, \dots, n$). By estimating not only $r_k[t, x(t)]$ and $\Delta r_k[t, x(t)]$, but also the corresponding quantities for major components, those components requiring the most maintenance will be apparent. Thus, concurrent with the collection of data to be used in estimating the overall system reliability, data on various components will also be collected. It is to be observed that this additional component analysis is essential for systems in which major components have been replaced.

Continued work on the general formulation will be directed toward relating the function $r_k[t, x(t)]$ and $\Delta r_k[t, x(t)]$ both to the organization of maintenance in terms of costs, skills, tools, etc., and to the complexity, density and task classifications of weapons systems.

4. PROPOSED PROGRAM FOR FIELD OBSERVATION AND DATA ANALYSIS. To obtain the information required to develop the curves and surfaces discussed in Part 3, a field observation program will be necessary. It is necessary that at least two, and preferably three, weapons systems be selected that represent different points in the complexity-density range. Since life histories are to be collected, new weapons systems should be selected whenever practicable. The specific weapons systems that are to be studied have not been chosen at the present time. A wheeled or tracked vehicle, a missile, and possibly a hand weapon will probably be selected. In selecting the weapons system and the units to be observed, proper consideration also must be given to the missions that are being performed.

Since a main purpose of the research is to quantitatively evaluate the effect of preventive maintenance practices to permit determination of the proper amount of scheduled maintenance that should be performed, it will be necessary, after an initial observation period, to slightly modify the existing maintenance practices and observe their effects.

Specifically, we will attempt to collect the following information:

A. Scheduled Maintenance

For each scheduled maintenance or operational check performed, the following information is desired:

1. Type of scheduled maintenance
2. Frequency of scheduled maintenance
3. Equipment usage between scheduled maintenance periods.

4. Deficiencies found during scheduled maintenance.
5. Parts replaced during scheduled maintenance.
6. Time to repair deficiencies found during scheduled maintenance periods.
7. Time to perform scheduled maintenance.
8. The echelons that perform the scheduled maintenance.

B. Unscheduled Maintenance

For each failure requiring unscheduled maintenance the following information is desired:

1. Frequency of failure
2. Basic cause of failure
3. Elapsed time since specific scheduled maintenance.
4. Usage since specific scheduled maintenance.
5. Parts needed for repair of failure
6. Parts available
7. Time to repair failure
8. Echelon performing the repair

C. Maintenance Organization

For the unit being observed the following information is desired:

1. Skills and equipment available at using unit and supporting units.
2. Inspection criteria at each echelon
3. Work load at each echelon.

We plan to initially place technically qualified field engineers on a full time basis with the units to assist user personnel in recording the above information. It is hoped that after approximately three to six months the cooperating units will be able to provide the necessary information, and the field staff will be required only to monitor the data collection program on a part-time basis. This would also free the field staff to initiate a similar observation program with an additional using unit.

As discussed in Part 3, the data collection program will be initiated concurrently with the continuation of the development of the general approach. As the program proceeds, the level, type and method of data collection may, of course, require modification. The initial information obtained will aid in selecting the most pertinent of many possible characteristics of weapons systems for first consideration in the mathematical formulation, and in defining the groups of missions required of the different types of systems studied. Such effort will be needed to further refine the general formulation to insure that the results will have practical significance to the particular weapons systems under consideration.

BIBLIOGRAPHY FOR APPENDIX III

1. "Effects of Maintenance on System Reliability," ARINC Research Corporation, Publication No. 101-16-144, Sept., 1959.
2. "An Analysis of the Effects of Maintenance on Parts Replacements," Ralph L. Madison, AERO. Radio Inc., Publication No. 112, Jan. 6, 1958.
3. "A Survey of the Current Status of the Electronic Reliability Program," R. R. Carhart, The Rand Corporation, RM 1131, Aug., 1953.
4. "Maximizing Expected Machine Up Time," Jay B. Heyne, Systems Development Laboratory, Hughes Aircraft Co., July, 1959.
5. "Evaluation and Prediction of Circuit Performance by Statistical Techniques," J. Marini, H. Brown, R. Williams, ARINC Inc., Publication No. 113, Feb., 1958.
6. "The Condition of 60 Farm Tractors Point Ways to Improve Maintenance," J. A. Weber, Society of Automotive Engineers, Paper No. 181, Sept., 1957.
7. "Problems Relating to Systems Checkout and Final Acceptance of Production Missiles," M. A. Stephens, Society of Automotive Engineers, Paper No. 48B, April, 1958.
8. "A Bibliography of Life Testing and Related Topics," William Mendenhall, Biometrika, Vol. 45, Parts 3 and 4, pp. 521-543, Dec., 1958.
9. "Maintenance Management--Operational Readiness Plan for the Redstone Missile System," Lt. Col. Brian O. Montgomery, Redstone Arsenal, Alabama, Sept., 1958.
10. "Optimum Component Redundancy for Maximum System Reliability," Robert Gordon, Operations Research Society Journal, Volume 5, No. 2, April, 1957. (pp. 229-243).
11. "Reliability of Military Electronic Equipment," The Advisory Group on Reliability of Electronic Equipment, Office of the Assistant Secretary of Defense (Research and Engineering), 4 June, 1957.
12. "Queues, Inventories, and Maintenance," Philip M. Morse, John Wiley and Sons, 1957.
13. "Reliability Factors for Ground Electronic Equipment," Henney, Lopatin, Zimmer, Poler, and Naresky, McGraw Hill, New York, 1956.
14. "A Good Preventive Maintenance Program," SAE Journal, Vol. 65, p. 88, July, 1957.
15. "Automatic Transmissions Make Some Truck Operator's Happy, Others Sad," SAE Journal, Vol. 65 pp. 56-58, Spetember, 1957.

16. "Simplicity - Key to Military Design," SAE Journal, Vol. 67, pp. 34 and 35, July, 1959.
17. "Air Force Model Shop Sets New PM Pace," Commercial Car Journal, Vol. 90, pp. 74-75, December, 1955.
18. "Maintenance Yardsticks Show Bus PM Efficiency," Commercial Car Journal, R. W. Ziffle, Vol. 90, pp. 68, 128-132, December, 1955.
19. "Systems Checkout Insure Missile Reliability," M. A. Stevens, SAE Journal, Vol. 66, pp. 80-81, August, 1958.
20. "Do Road Calls Measure PM Efficiency?" Commercial Car Journal, Vol. 90, pp. 71 and 186, September, 1955.

STATISTICAL ANALYSIS OF VARIOUS PARAMETERS
AFFECTING THE BURNING CHARACTERISTICS OF FLARE SYSTEMS

Bossie Jackson
Pyrotechnics Chemical Research Section
Pyrotechnics Laboratory
Picatinny Arsenal

The development of pyrotechnic flare compositions involves the investigation of a number of different variables. From the many investigations conducted previously numerous hypotheses were formed concerning the relationship of such variables as candlepower and burning rate with flare case coating, loading pressure, and the particle size of the fuel employed in the system. A better knowledge of the basic factors governing the burning characteristics of solid mixtures was desired. It was expected that the results of this investigation would tend to substantiate the various hypotheses.

Previous data tends to show that candlepower and burning rate are depended upon particle size, loading pressure, and are not affected by flare case coating. The analysis of these relationships was based on data obtained from this study using two statistical methods, the test for least significant differences, and in particular the analysis of variance. The experimental design for these studies is given in Figure 1. (See next page.) This configuration was used for four (4) levels of magnesium, M_1 , M_2 , M_3 , and M_4 . The flare case coatings are shown by C_1 , C_2 , C_3 , and C_4 . Seven (7) levels of loading pressure were studied ranging from P_1 to P_7 . Five samples were utilized for each combination of pressure, case coating, and magnesium particle size.

A standard flare composition (Table I) [Tables can be found at the end of this article] was used for this experiment. This composition contains 48% magnesium, 42% sodium nitrate, 2% polyvinyl chloride, and 8% Laminac resin. Flare compositions are consolidated in a variety of cylindrical cases under a specified pressure to obtain a cigarette-type propagating composition rather than one which flashes or explodes. The use of self-hardening polyester resins in flare compositions eliminated the need for consolidating them at very high pressures. Most flare compositions are presently loaded at pressures ranging from 4000 pounds per square inch (psi) to 10,000 psi.

Standard flare compositions each containing different mesh sizes of magnesium were evaluated in this study. These magnesium granulations together with their average particle diameter are given in Table II. The mesh sizes of magnesium are 20/50, 30/50, 50/100, and 100/200 with particle sizes varying from 437 microns to 110 microns. The compositions were consolidated at 2000, 4000, 7000, 10,000, 15,000, 20,000 and 25,000 pounds per square inch.

The effect of loading pressure on candlepower can be observed in Figure II. The mean candlepower values vary from 201,000 to 223,000 which approximates an eleven (11) percent change. It is apparent from

Design of Experiments

this graph that a definite trend exists with candlepower increasing with increased loading pressure. Table III summarizes the luminous intensity values observed at each pressure level. The least significant difference value at the ninety-five (95) percent confidence level is also given for these values. Despite the apparent trend of candlepower values, it will be observed that the difference in light output from 4,000 psi to 25,000 psi are not large enough to be significant based on the least significant difference value. This conclusion also holds for the candlepower at 2,000 and 4,000 pounds per square inch. However, the differences between the values at 2,000 psi and those at 7,000 psi and above are large enough to be termed significant. The appearance of this trend may be attributed to the relationship between porosity and heat conduction. A more porous column will conduct heat at a slower rate as a result of the air pocket acting as insulators giving slow burning rates and low candlepower. Conversely, the less porous column will conduct heat at a faster rate giving higher candlepower and burning rates.

Higher candlepower values are obtained from smaller particle diameters of magnesium as shown in Figure III. Candlepower plotted as a function of average particle diameter decreased with increasing particle size. The higher candlepower value obtained for the 168 micron magnesium compared to the finer 110 micron magnesium may be attributed to the distributional effect of particle size. As evidenced by the low average particle diameter, the 50/100 mesh magnesium contained a large percentage of fines which placed the material in the upper range of the finer 100/200 mesh magnesium which may account for the higher light output. The candlepower values for each mesh size magnesium are tabulated in Table IV. It is immediately apparent that these intensity values are significantly different from each other on application of the least significant difference value. It is also observed that the 50/100 mesh magnesium gave significantly higher candlepower than the 100/200 mesh fuel. This result reflects the necessity for reducing the tolerance limits of particle size for the different mesh sizes of magnesium. It is believed that data can be accumulated from studies conducted previously to show that candlepower definitely decreases with increasing particle size.

It was previously mentioned that any of the investigators in the field of pyrotechnics believed that candlepower was unaffected by case coatings. This was verified by the results obtained from this experiment as shown in Table V. The flare case coatings studied were Amberlac 292, Laminac resin 4116, Polyethylene 617, and paraffin wax. The candlepower values vary from 210 to 218 and are essentially the same based on the least significant difference value of 8.8. Flare case coatings are especially necessitated with compositions containing self-hardening resins. Since these resins undergo considerable shrinkage on curing, voids and air pockets are created as a result of the composition separating from the flare case wall. Such a condition gives rise to possible detonations or increased burning rates.

Just as candlepower shows an insignificant trend resulting from increased loading pressure, burning rate values show a parallel trend. Figure IV illustrates this trend as the loading pressure is increased

from 2,000 psi to 25,000 psi. The burning rate values show a trend towards reduction with increasing loading pressure. Table VI tabulates the mean burning rate values obtained at each pressure level. It also summarizes that the differences in burning rates are not large enough to be significant. Based on this method of analysis, it can be concluded that burning rate is not affected by increasing loading pressure. As a result of the oppositely parallel trends shown by candlepower and burning rate, it can be hypothesized that these two variables are interrelated. This hypothesis is corroborated and borne out when considering the effect of particle size of magnesium on burning rate.

Figure V shows burning rate plotted as a function of average particle diameter. It can be seen that burning rate decreases with increasing average particle size. As shown on Table VIII the differences between these values were found to be significant based on the least significant difference value of 0.05. It was previously observed that a corresponding effect was obtained with candlepower values, except in the result for 50.100 mesh magnesium. To further complicate the picture it was observed that significant differences in burning rates existed for the various flare case coatings. The results are given in Table VIII. The burning rates vary from a slow 3.61 inches per minute for the polyethylene to a fast 4.52 inches per minute for paraffin wax with Amberlac and Laminac resins yielding values in the middle. These differences based on the test for least significant difference indicate that the minimum and maximum values here are significantly different from the intermediate ones which are essentially the same. The significant effects of flare case coatings are undoubtedly due to their variation in binding strength, rate of thermal degradation, and end-products produced on combustion.

Burning rates of pyrotechnic compositions are also derived from the weight composition undergoing combustion per unit time. Figure VI illustrates the effect of loading pressure on the grams of composition per second from 2,000 pounds per square inch to 25,000 pounds per square inch. It is shown that grams of composition per second tends to increase with increasing loading pressure. The change in pressed density from 2,000 psi to 25,000 psi approximated twenty (20) percent. As shown in Table IX the differences in grams of composition per second are not significant based on the least significant difference value of 0.60. It was previously observed that the linear burning rate was not significantly affected by loading pressure.

In direct contrast to the above result, it was determined that the average particle diameter of magnesium had a significant effect on the weight of composition consumed per unit time. Based on the previous effects of particle size on both candlepower and burning rate this result could be anticipated. Figure VII shows grams of composition per second as a function of average particle diameter. It is observed that the number of grams burned per unit time varies inversely as the average particle diameter. The mean burning rate values tabulated in Table X are shown to be significantly different from each other as a result of the test for least significant difference.

Design of Experiments

By observing Table XI, it can be seen that flare case coatings do not significantly affect the grams of composition per second. This conclusion results in the fact that the average particle diameter of magnesium is the only parameter that significantly affects candlepower, burning rate (inches per minute), and burning rate (gram per second). The only other parameter contributing a significant effect was flare case coating on the linear burning rate.

Summarizing the results previously discussed, the analysis of variance table for candlepower is observed (Table XIII). This table gives the main effects, first and second order interactions of the parameters under consideration. Where the calculated F-ratio exceeds the critical F-ratio the effect of the parameter is considered significant. The main effects of magnesium particle size and case coatings are in accord with the results based on the test for least significant difference. The main effect from loading pressure observed here is significant as opposed to its insignificant effect based on the least significant difference value. Considering that such large changes in loading pressure (92%) results in very small changes in light output, it may indicate that the increasing trend of candlepower is insignificant. The first order interactions of magnesium loading pressure and case coating-loading pressure show a significant effect on candlepower, while magnesium-case coating is insignificant. Evidently, the flare case coating cancels out the effect of the magnesium particle size. The second order interactions of these parameters are shown to be insignificant.

Table VIII outlines the analysis of variance table for burning rate (inches) per minute. It is observed that main effects resulting from magnesium and case coating are significant and loading pressure insignificant. This corroborates the results based on the test for least significant difference. The first order interactions of magnesium-case coating and magnesium-loading pressure are shown to be significant, while case coating-loading pressure is insignificant. The second order interactions are significant.

The results of the analysis of variance for burning rate (grams per second) is given in Table XIV. Except for the fact that flare case coating is shown to be insignificant here, the results parallel exactly those obtained for the analysis of variance on the linear burning rate. This table also substantiates the results from the test for least significant difference.

TABLE I
COMPOSITIONS EVALUATED

Ingredients	Percent by Weight			
Magnesium, Atomized, 20/50, 437 microns	48			
Magnesium, Atomized, 30/50, 322 microns		48		
Magnesium, Atomized, 50/100, 168 microns			48	
Magnesium, Atomized, 100/200, 110 microns				48
Sodium Nitrate, 34 microns	42	42	42	42
Polyvinyl Chloride, 27 microns	2	2	2	2
Laminac Resin 4116*	8	8	8	8
* Laminac Resin 4116 - 98.5%				
Lupersol ddm	- 1.0%			
Nuodex	- 0.5%			

TABLE II

MAGNESIUM GRANULATIONS

<u>Mesh Size</u>	<u>Tapped Density, gm./cc.</u>	<u>Average Particle Diameter, Microns</u>
20/50	1.07	437
30/50	1.08	322
50/100	1.14	168
100/200	1.08	110

TABLE III

TESTS FOR LEAST SIGNIFICANT DIFFERENCE
OF LOADING PRESSURE VS. AVERAGE CANDLEPOWER

Least Significant Difference - 13.8

Level of Confidence, % - 95.0

<u>Loading Pressure Psi</u>	<u>Average Candlepower</u>
2,000	201.0
4,000	213.0
7,000	215.0
10,000	217.0
15,000	218.0
20,000	222.0
25,000	223.0

TABLE IV

TESTS FOR LEAST SIGNIFICANT DIFFERENCE
OF MAGNESIUM MESH SIZE VS. AVERAGE CANDLEPOWER

Least Significant Difference - 6.6

Level of Confidence, α - 95.0

<u>Magnesium Mesh Size</u>	<u>Trapped Density gms/cc</u>	<u>Average Particle Size, Microns</u>	<u>Average Candlepower</u>
20/50	1.07	437	130.0
30/50	1.08	322	154.0
50/100	1.14	168	293.0
100/200	1.08	110	285.0

TABLE VTESTS FOR LEAST SIGNIFICANT DIFFERENCE
OF FLARE CASE COATING VS. AVERAGE CANDLEPOWER

Least Significant Difference - 8.8

Level of Confidence, % - 95.0

<u>Flare Case Coating</u>	<u>Average Candlepower</u>
Amberlac Resin 292	218.0
Laminac Resin 4116	215.0
Polyethylene 617	218.0
Paraffin Wax	210.0

TABLE VI

TESTS FOR LEAST SIGNIFICANT DIFFERENCE

Least Significant Difference - 0.61

Level of Confidence, % - 95.00

<u>Loading Pressure, Psi</u>	<u>Average Burning Rate, Inches Per Minute</u>
2,000	4.41
4,000	4.54
7,000	4.29
10,000	4.23
15,000	4.03
20,000	4.17
25,000	4.06

TABLE VII

TESTS FOR LEAST SIGNIFICANT DIFFERENCE

Least Significant Difference - 0.05

Level of Confidence, % - 95.00

<u>Magnesium Mesh Size</u>	<u>Average Burning Rate, Inches Per Minute</u>
20/50	2.62
30/50	3.01
50/100	5.66
100/200	5.84

TABLE VIII

TESTS FOR LEAST SIGNIFICANT DIFFERENCE

Least Significant Difference - 0.14

Level of Confidence, % - 95.00

<u>Flare Case Coating</u>		<u>Average Burning Rate, Inches Per Minute</u>
Amberlac		4.13
Resin	292	
Laminac		4.15
Resin	4116	
Polyethylene	617	3.61
Paraffin Wax		4.52

TABLE IX

TESTS FOR LEAST SIGNIFICANT DIFFERENCE

Least Significant Difference - 0.60

Level of Confidence, % - 95.00

<u>Loading Pressure Psi</u>	<u>Burning Rate Grams Composi- tion/second</u>
2,000	5.06
4,000	5.37
7,000	5.38
10,000	5.49
15,000	5.48
20,000	5.58
25,000	5.57

TABLE X

TESTS FOR LEAST SIGNIFICANT DIFFERENCE

Least Significant Difference - 0.06

Level of Confidence, % - 95.00

<u>Magnesium Mesh Size</u>	<u>Burning Rate Grams Composi- tion per second</u>
20/50	2.41
30/50	2.82
50/100	6.11
100/200	6.33

TABLE XI

TESTS FOR LEAST SIGNIFICANT DIFFERENCE

Least Significant Difference - 0.23

Level of Confidence, % - 95.00

<u>Flare Case Coating</u>	<u>Burning Rate Grams Composi- tion/second</u>
Amberlac Resin 292	4.36
Laminac Resin 4116	4.32
Polyethylene 617	4.50
Paraffin Wax	4.50

TABLE XII
ANALYSIS OF VARIANCE TABLE FOR CANDLEPOWER VALUES

ANALYSIS OF VARIANCE

<u>Source</u>	<u>Main Effects</u>	<u>Sum of Squares</u>	<u>Degree of Freedom</u>	<u>Mean Squares</u>	<u>F - Ratios Calculated</u>	<u>Critical F - Ratios 95% Level of Confidence</u>
Magnesium (M)		3,076,232.87	3	1,025,410.956	2220	2.60
Pressure (P)		26,818.30	6	4,469.716	5.18	2.66
Coating (C)		5,592.08	3	1,864.0266	2.29	3.86
<u>Interactions,</u>						
<u>First Order</u>						
M + C		7,334.05	9	814.8944	1.77	1.88
M + P		15,535.53	18	863.085	1.88	1.57
C + P		9,805.12	18	544.7289	1.90	1.66
<u>Interactions,</u>						
<u>Second Order</u>						
M + P + C		15,415.97	54	285.6661	<1	1.32
<u>Residual Error</u>						
<u>Total Sum of Squares</u>						
		206,239.40	448	460.3558		
<u>Pooled Variance</u>						
		3,362,972.94	559			
<u>Standard Deviation</u>						
		460,355,800				
		$\sqrt{460,355,800} = 21,455$				

TABLE XIII

ANALYSIS OF VARIANCE TABLE FOR BURNING RATE (INCHES PER MINUTE) VALUES

BURNING RATE, INCHES PER MINUTE

<u>Source</u>	<u>Sum of Squares</u>	<u>Degrees of Freedom</u>	<u>Mean Square</u>	<u>F - Ratios Calculated</u>	<u>Critical F - Ratios 95% Level of Confidence</u>
<u>Main Effects</u>					
Magnesium (M)	12,718,235	3	4,272,745	19,600	2.60
Pressure (P)	160,662	6	26,777	1.94	2.66
Coating (C)	61,301	3	20,651	15.3	3.86
<u>Interactions</u>					
<u>First Order</u>					
M + C	12,163	9	1,351	6.2	1.88
M + P	248,338	18	13,797	63.1	1.57
C + P	30,731	18	1,707	1.3	1.66
<u>Second Order</u>					
M + P + C	70,522	54	1,306	6.0	1.32
<u>Residual Error</u>					
Total Sum of Squares	97,992	448	219		
Pooled Variance	13,399,945	559			
Standard Deviation	0.0219				
	$\sqrt{0.0219} = 0.15$				

TABLE VIANALYSIS OF VARIANCE TABLE FOR BURNING RATE (GRAMS PER SECOND) VALUESAnalysis of VarianceBurning Rate, Grams of Composition/Sec.

<u>Main Effects</u>	<u>Source</u>	<u>Sum of Squares</u>	<u>Degree of Freedom</u>	<u>Mean Square</u>	<u>F - Ratio Calculated</u>	<u>Critical F - Ratio 95% Level of Confidence</u>
Magnesium (M)	Magnesium (M)	18,333.445	3	6,111.148	176,000	2.60
Pressure (P)	Pressure (P)	150.184	6	25,031	1.89	2.66
Coating (C)	Coating (C)	35.199	3	11.733	2.13	3.86
<u>Interactions, First Order</u>						
M + C	M + C	34,800	9	3,866	11.2	1.88
M + P	M + P	238.236	18	13.235	38.4	1.57
C + P	C + P	24,573	18	1,363	1.07	1.66
<u>Interaction, Second Order</u>						
M + P + C	M + P + C	68,748	54	1,273	369	1.32
<u>Residual Error</u>						
Total Sum of Squares	Total Sum of Squares	154,606	448	345		
Pooled Variance	Pooled Variance	19,039.724	559			
Standard Deviation	Standard Deviation	0.0345		$\sqrt{0.0345} = 0.19$		

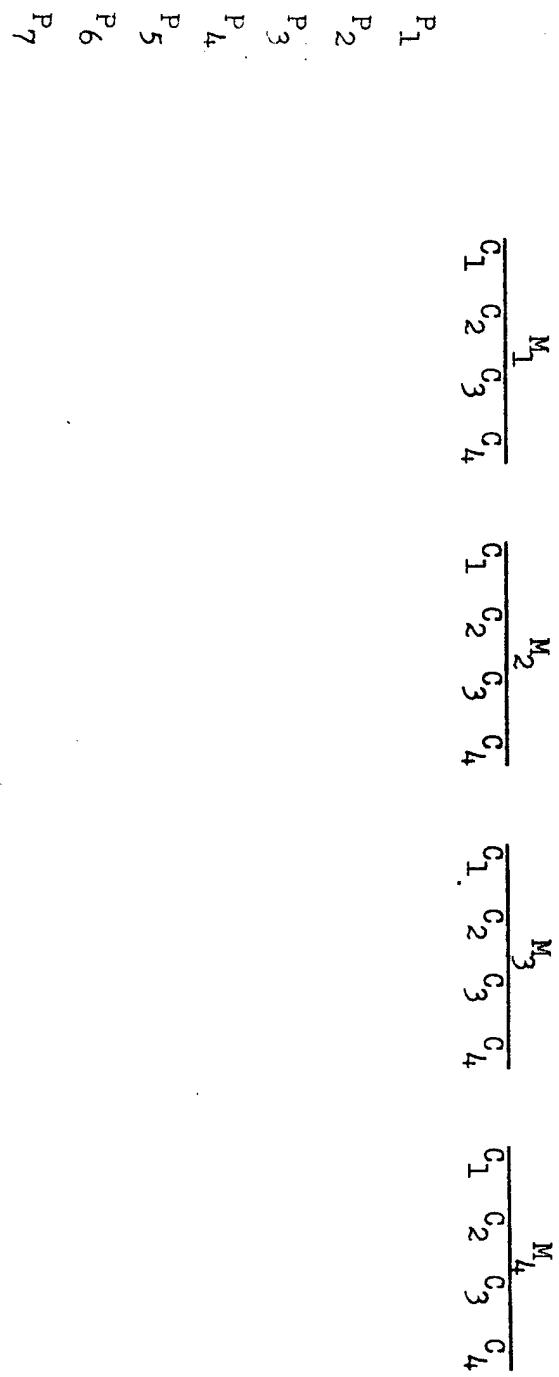
Figure IEXPERIMENTAL DESIGN

Figure II

232

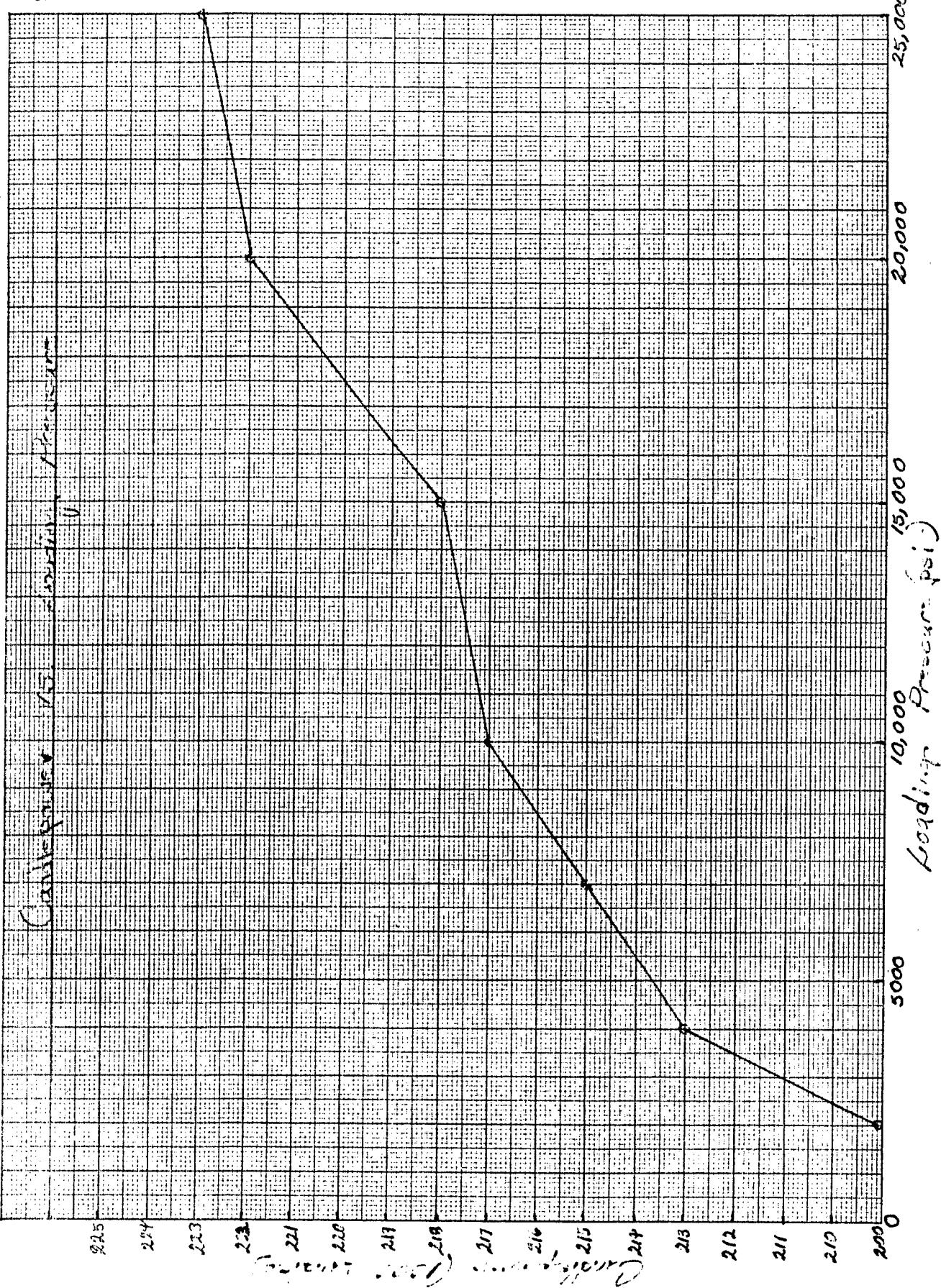


Figure III

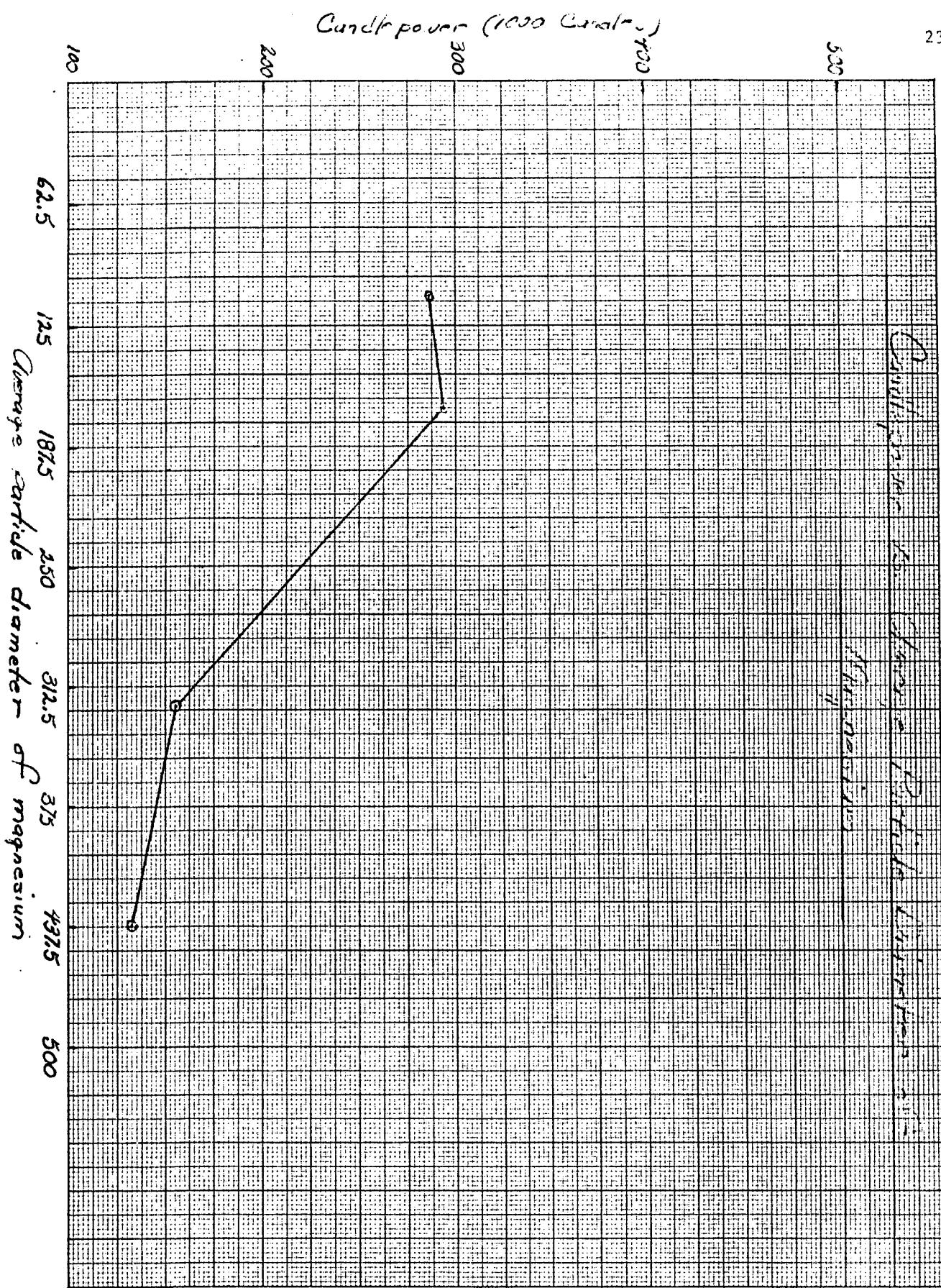


Figure IV

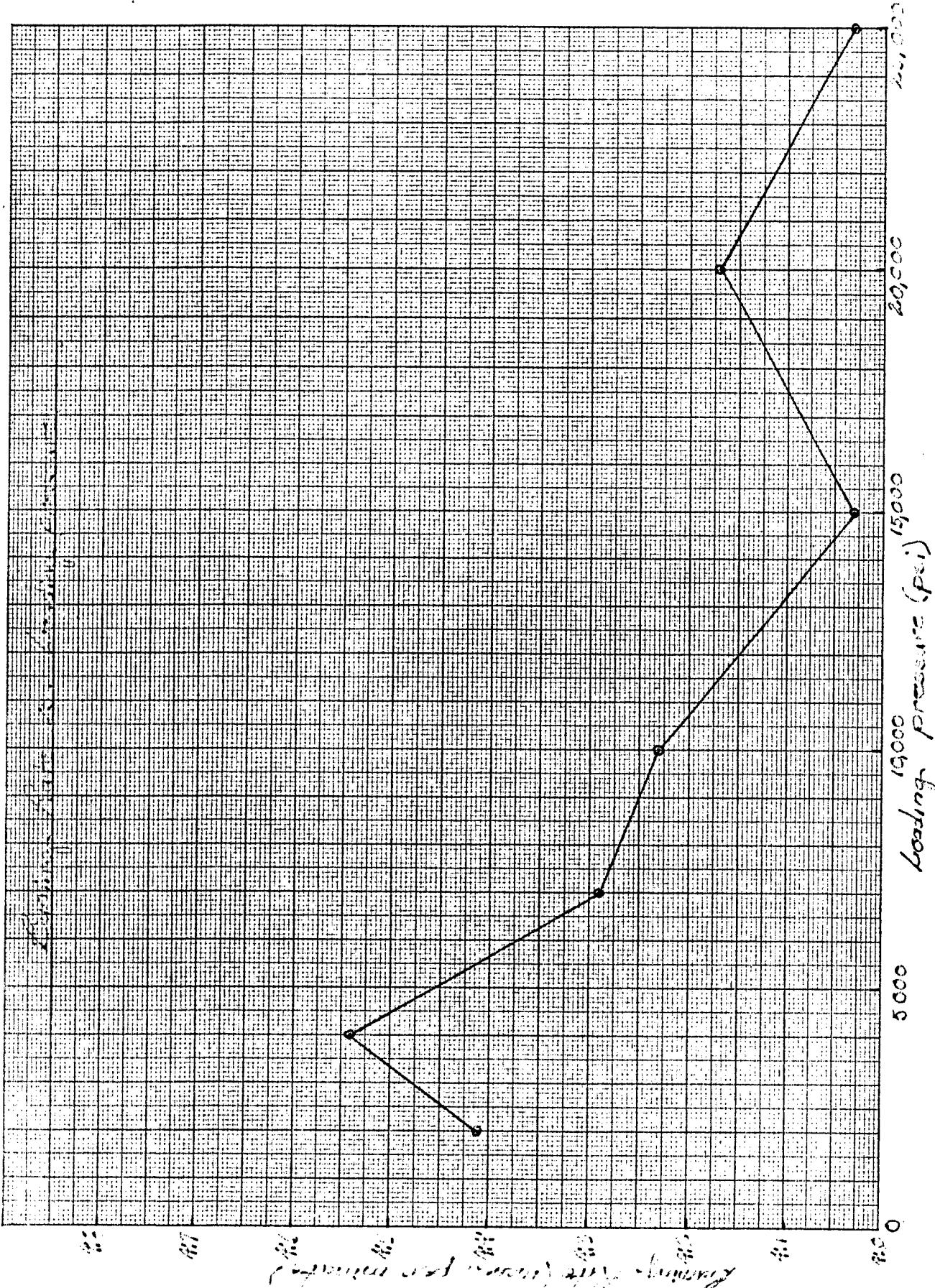


Figure V

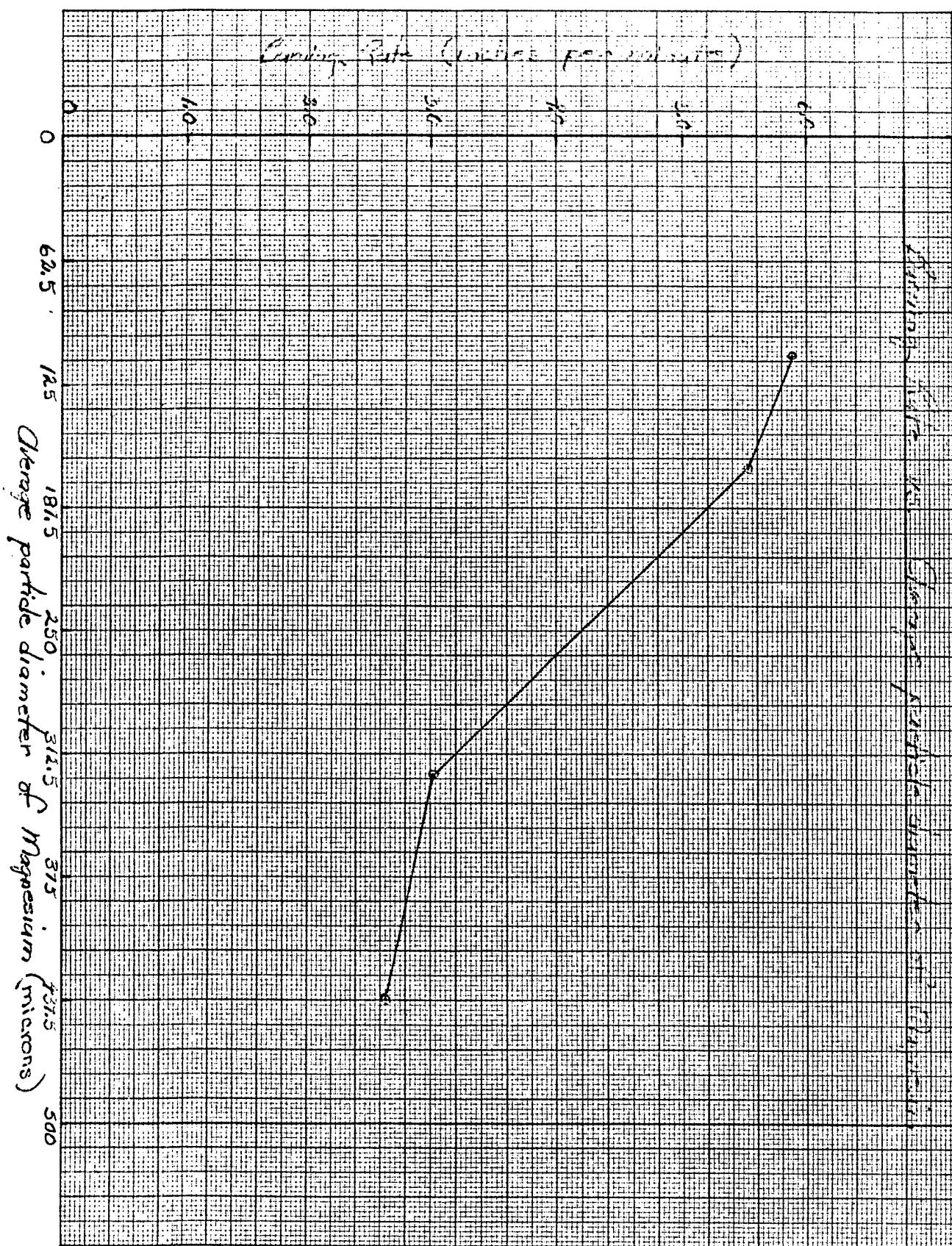


Figure VI
Cyclic Loading Test
Second Test

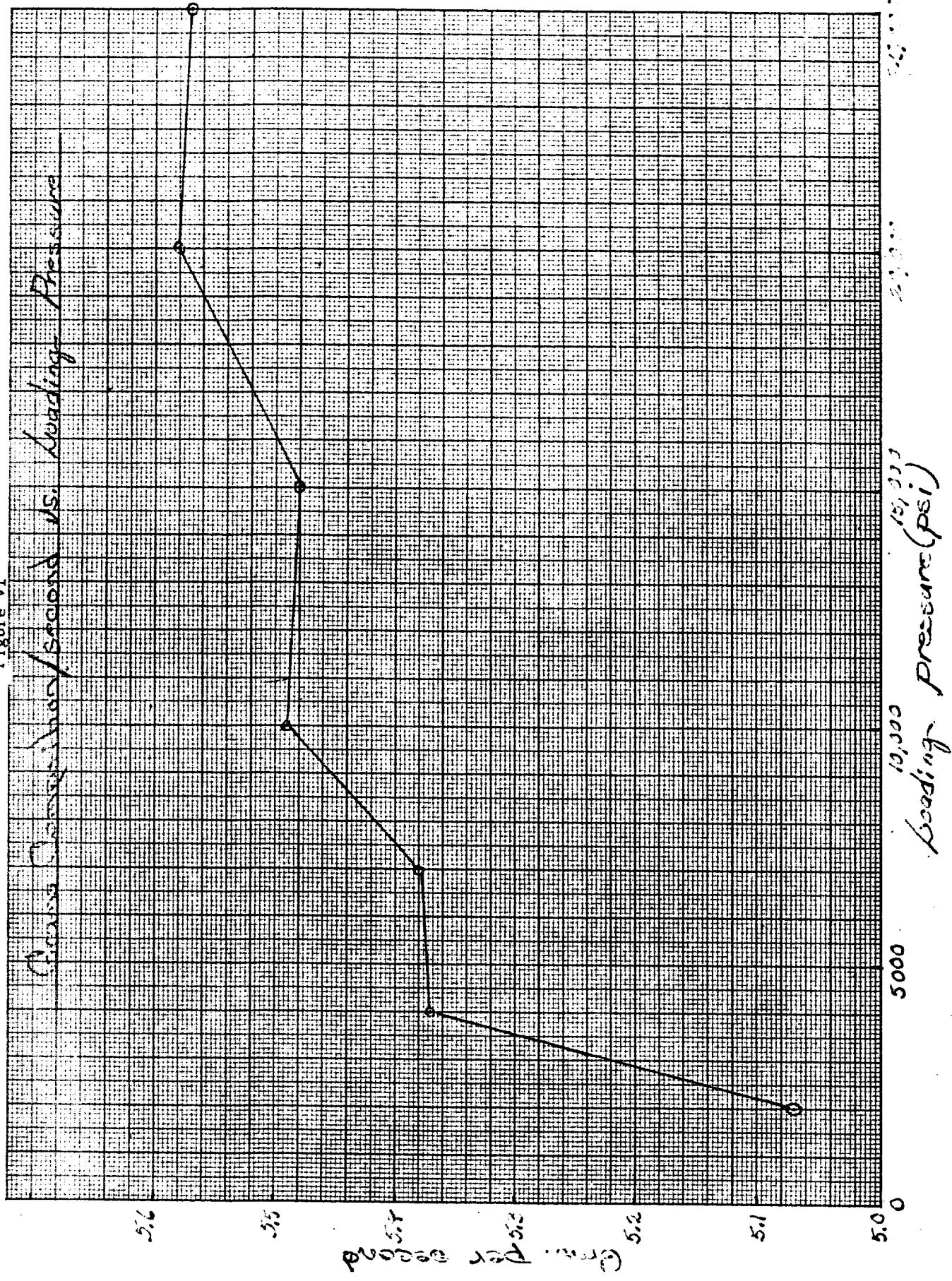
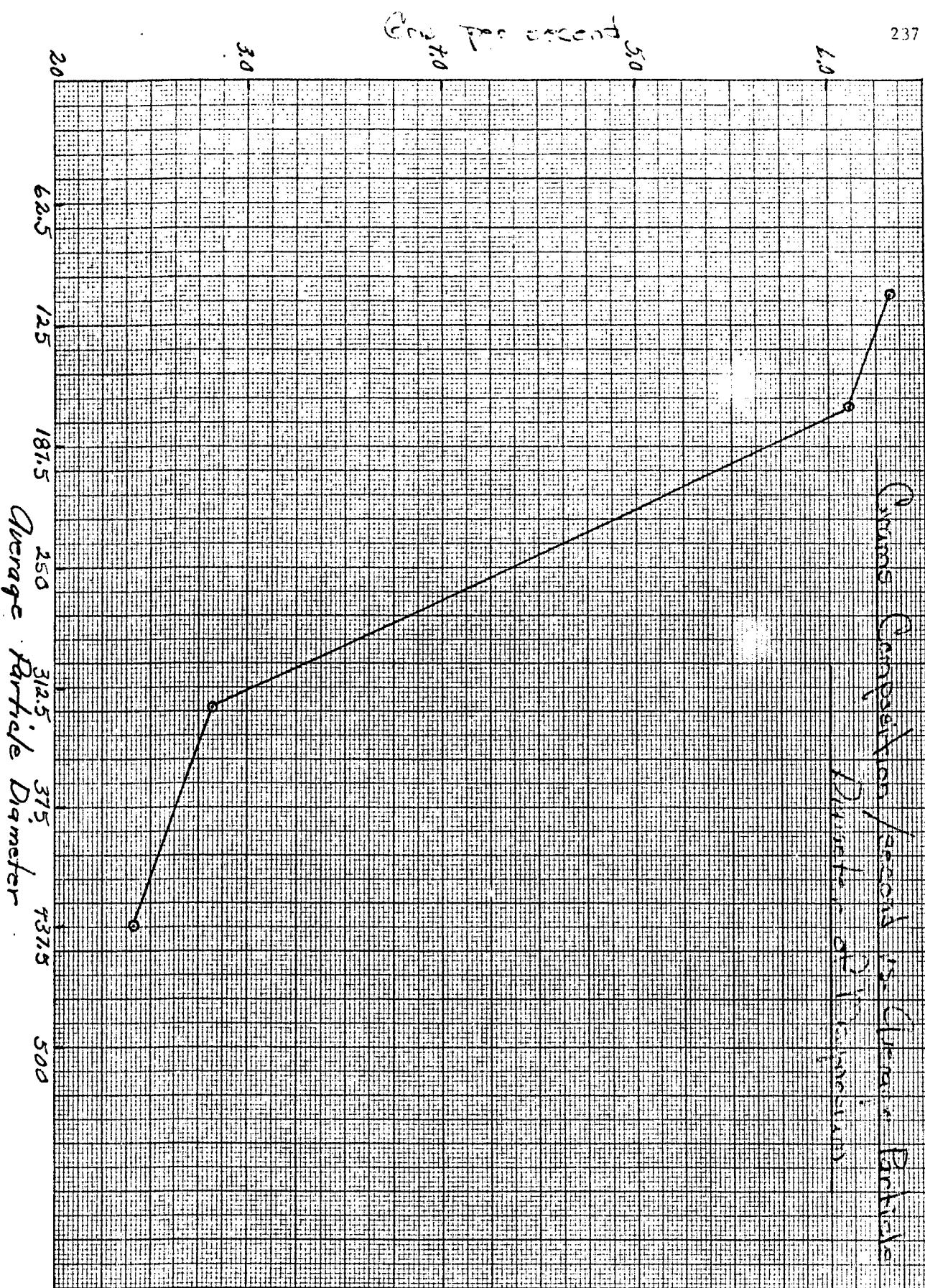


Figure VII

237



A STATISTICAL EVALUATION OF THE
PYROTECHNICS ELECTROSTATIC SENSITIVITY TESTER*

Everett Crane, Chester Smith, and Alonzo Bulfinch

...
OBJECT.

(a) To establish optimum operating conditions for the electrostatic sensitivity tester by determining statistically which design factors contribute most significantly to its performance.

(b) To determine whether electric spark sensitivity results obtained through use of this instrument on samples of fine (29 micron) magnesium powder are reproducible.

SUMMARY. An electrostatic sensitivity tester developed at Picatinny Arsenal was evaluated statistically. The factors found to contribute most significantly to optimum instrument operating conditions were resistance, humidity, energy, and the relationship of energy to resistance. The electrostatic sensitivity results obtained with fine magnesium powder specimens were found to be reproducible. It was concluded that further work should be conducted on a variety of samples to determine the effect of various characteristics of the circuit and the maximum energy input which will produce no burning in a specified number of trials. A method for measuring this can be developed by studying the lower tails of the spark sensitivity curves. Deviations in the lower tails of the curves, which are unique for each material, are the best indicators of the materials' sensitivity characteristics.

...
INTRODUCTION. Previously constructed electrostatic sensitivity testers were found to have one major shortcoming. The energy delivered to the sample was inconsistent because of losses within the system, and reproducibility of results was therefore erratic. An investigation of electrostatic sensitivity testers in use by the Bureau of Mines, the Naval Ordnance Laboratory, and the British armed forces was undertaken (Refs 1, 2, 3, and 4), and a modified apparatus was constructed in an attempt to eliminate this deficiency.

The action of the pyrotechnics electrostatic sensitivity tester developed at Picatinny (Fig 1)** is extremely simple. A sample is placed in the sample holder and a movable probe having a sharp point

* This paper appeared July 1959 as Dept. of the Army Project 504-01-027 report issued at Picatinny Arsenal.

** The figures have been placed at the end of this article.

is raised above it. The apparatus is then set at the desired voltage and R-C resistance. A chosen capacitor (charged to the desired voltage) is connected between the probe and the sample holder base. The capacitor is discharged by allowing the probe to fall to a fixed distance above the sample. The operator then observes and records the resulting action.

This is a classical experiment, as many such devices have been used in the past. However, despite its apparent simplicity, it has not, in past work, consistently produced satisfactory results. Because it has a built-in resistance, capacitance, and probe-down-time mechanism (Fig 2), the new device offers better opportunity for consistent results. One unfortunate difficulty, however, is that the probe (Fig 3) tends to become loosened by vibration, causing the operator to lose time in re-setting it. After preliminary tests have been conducted, improvements to eliminate this fault will be made.

Because a large volume of data has been collected in determining optimum instrument operating conditions, it was considered desirable to issue a report on this phase of the investigation. Electric spark sensitivity data on various pyrotechnic, propellant, and explosive materials will be included in subsequent reports.

Difficulties inherent in the study of this instrument are:

1. Only attribute (Go, No go) type data can be obtained. This type of data yields only a small amount of information per observation.

2. The property of the materials to be tested is sensitivity to electric spark. This property requires a test of increased severity which is a type of test that yields little information per observation.

3. The effects of a large number of variables are determined simultaneously.

4. The spark sensitivity of a large number of materials must be evaluated. The input energy and the effect of instrument variables for any given material are of little value in the study of spark sensitivity of other materials.

5. Because of the nature of the data, non-parametric methods of analysis must be used. These methods are less efficient than parametric methods of analysis.

To reduce these difficulties to a minimum and extract the maximum amount of information possible, statistically designed experiments called factorial experiments were used. This type of designed experiment is the most efficient known. It is possible in factorial experiments to study more than one variable at a time. In general, the efficiency of the experiment is increased when a greater number of variables are studied simultaneously (Refs 5 and 8).

EXPERIMENTAL DESIGN AND ANALYSIS. Since the equipment used in this experiment was new, little was known at the outset concerning either the magnitude of the input energy required to cause burning or the effects of such other variables as might be present in the system. Therefore, a sequential approach to the problem was adopted. In this manner, something was learned about the magnitude of the input required, and it was possible to examine the results of small experiments before doing further work. The results of these exploratory experiments were not included in this report because their contribution was mainly to eliminate "rough spots" in the apparatus.

The data was analyzed by the Kruskal-Wallis rank-sum test, sometimes called the H-Test. In determining the significance of the main effects, this test was used in the usual way (Ref 5), to determine differences among means. In determining the significance of the first-order interactions, the appropriate main effects were subtracted from each total interaction effect.

In these exploratory studies, fine (29 micron average particle size) magnesium powder¹ was used, since it was a convenient homogeneous material.

Experiment 1 (Energy Changes)

To obtain a first estimate of the input energy required, tests of increased severity were conducted using the run-down method (Refs 6 and 7). In these tests, all variables were held constant at convenient levels, except energy (in joules), which was varied by varying the capacitance. When the results were plotted on probability paper (Figs 4 through 9), they yielded essentially straight lines, which indicated that the data could be considered, for all practical purposes, to be normally distributed. This was an important finding since it simplified interpretation of the results. The average values from these graphs (the 50% points in terms of energy) were helpful in establishing the input energy level used as a standard in subsequent experiments.

Experiment 2 (Gap Length, Humidity, Voltage, and Resistance)

The results of Experiment 1 were as follows:

1. The effects of sample size were insignificant.
2. Only inconclusive data was obtained on the effects of gap length and humidity.
3. The data obtained indicated that more should be known about the effects of voltage and resistance.

On the basis of the above findings, Experiment 2 was designed as a 4-factor complete factorial experiment to determine the effects of humidity, gap length, voltage, and resistance. The energy level was adjusted to 0.100 joule, to provide a usable distribution of successes and failures. The experiment was repeated 5 times (Tables 1, 2, and 3).

¹ Sample 142, barrel No. 30, Golwynne Chemical Company

Experiment 3 (Energy, Capacitance, and Voltage)

It was clear from the 4-factor experiment that the greatest number of ignitions were being obtained by eliminating the resistance (which is connected in series between the capacitor and the probe). It now appeared desirable to determine the effect of voltage at different energy levels. For this purpose, a 3-factor factorial experiment was designed (Tables 4 and 5) involving 3 levels of voltage, 6 levels of energy, and 2 levels of resistance. Resistance was included to confirm the conclusions reached in the 4-factor experiment regarding the effect of resistance.

RESULTS.

Experiment No. 1

The tests of increased severity showed averages (50% ignitions) and standard deviations (slopes), in joules, as follows:

	Average	Std Dev
Figure 5	0.100	0.075
Figure 7	0.134	0.055
Figure 9	0.144	0.064

Experiment No. 2

The results of the 4-factorial statistical analysis detailed in Tables 1, 2, and 3 were:

Main Effects ^a	Effect
Voltage (V)	Not Significant
Resistance (R)	Significant ^b
Gap Length (G)	Not Significant
Humidity (H)	Significant ^b

^aTaken from the Analysis of Variance in Table 3

^bSignificant at the 95% confidence level

Interactions ^c	Effect
V x G	Not Significant
R x G	Not Significant
V x H	Not Significant

(contd.)

Interactions ^c	Effect
R x H	Not Significant
G x H	Not Significant
V x R	Significant

^cVery highly significant, beyond the 99.9% level

Experiment No. 3

Figure 10 represents percentage of hits (burnings) versus volts versus joules and Figure 11 shows percentage of hits versus joules for 3000, 4000, and 5000 volts. The curve in Figure 12 is a composite of the 3 curves in Figure 11. Tables 4 and 5 show that, while resistance (R) and energy (E) are both very highly significant, voltage (V) is not significant. Figure 12 shows the average to be 0.062 joule and the standard deviation to be 0.019 joule over the three voltage levels used.

DISCUSSION OF RESULTS. Elimination of the danger of accidental electrostatic initiation is a major reason for measuring the electric spark sensitivity of pyrotechnics, explosives, propellants, and other materials. For this purpose, instrument operating conditions that will produce the maximum burning rate at all energy levels can be considered optimum.

From Tables 1 and 4, it is clear that removing all resistance from the system produces a significantly greater burning rate at all energy levels. Zero resistance can therefore be considered the optimum resistance condition for magnesium powder..

The data in Tables 4 and 5 and Figure 11 shows that, for zero resistance, the effect of changing the voltage from 3000 to 5000 volts is not significant. The effective sample size for evaluating the effect of voltage is 30 trials at each voltage level. Hence, the conclusion that the effect of voltage at zero resistance is insignificant at all energy levels is based on a sample size sufficient to give very good precision.

The data (Tables 4 and 5 and Figure 12) also makes evident a correlation between increasing percentages of burnings and increasing energy (joules).

Information on gap length and humidity is given in Table 1. This table shows that, over the 5 resistance levels, the effect of changing the gap length from 0.01 to 0.02 inch is nil and the effect of changing the humidity from 30% to 80% is significant. The results shown in this

table are considered to be reliable because they meet the effective sample size requirement for gap length and humidity, which is 250 trials at each level.

Additional work should be done to define the electric spark sensitivity of pyrotechnics, explosives, propellants, and other materials in terms of the characteristics of the electric circuit used and the maximum energy input which produces no burning in a specified number of trials. Once this definition has been developed through experience with representative materials, a method for measuring this property can be developed. This can be done by studying the lower tail of each sensitivity curve shown as a broken line in Figure 12. Since errors in this portion of the curve are rather large, it is dangerous to extrapolate from present data. In addition, significant deviations from normality can be expected. These deviations cannot be predicted by any known means. However, past experience with the impact sensitivity of explosives has shown that these deviations in the lower tail of the sensitivity curve are unique for each material and are the best indicators of sensitivity characteristics.

Work should also be carried out to determine optimum instrument conditions for pyrotechnics, explosives, propellants, and other materials. It may be possible to classify most materials into a few general types for this purpose, so that only a few instrument settings will be required. If this is not possible, then a rapid method should be developed for determining optimum conditions for new materials.

CONCLUSIONS.

1. The maximum burning rate of magnesium powder cannot be obtained over the range of energy levels surveyed if resistance is added in series between the capacitor and the probe. Varying the voltage between 3000 and 5000 volts has no effect on the number of ignitions of magnesium powder at any energy level when the resistance level is held constant.

2. Ignition is dependent on the energy released by the electrostatic sensitivity apparatus. For magnesium powder, the percentage of burnings increases with increasing energy (joules).

3. There is highly significant interaction between resistance and voltage, that is, the effect of voltage is dependent upon the level of resistance employed. Thus, any statement concerning the effect of voltage on burnings must specify the level of resistance.

4. The electrostatic sensitivity results obtained for 29-micron-average-particle-size magnesium powder are reproducible.

5. Additional work will be needed to evaluate the effect of gap length and humidity at zero resistance and to determine the electric spark sensitivity of a wide range of pyrotechnics, explosives, and propellants.

TABLE I
Four-Factor Factorial Electrostatic Sensitivity Experiment^a
(Experiment 2) for 29-Micron Magnesium Powder

Relative Humidity	Gap Length, inches	Resistance, kilo ohms ^b	E_1		E_2		E_3		E_4		Total Hits
			.0222 mfd ^c	3000 volts	.0163 mfd	3500 volts	.0125 mfd	4000 volts	.0099 mfd	.0080 mfd	
25-40%	.021	0	11111	11111	11111	11111	01111	11111	11111	11111	24
"	"	90	10011	10110	11000	01011	10110	10110	10110	10110	14
"	"	170	11001	01011	11100	10010	10010	10110	10110	10110	14
"	"	260	01011	10101	11100	10110	01110	01110	01110	01110	15
"	"	350	11011	00111	10011	10101	01100	01100	01100	01100	15
"	".010	0	11111	11111	11111	11111	11111	11111	11111	11111	25
"	"	90	11101	11001	00011	00101	10010	10010	10010	10010	13
"	"	170	01100	00101	10010	01110	01110	01110	01110	01110	12
"	"	260	11010	01110	10101	01111	01111	11101	11101	11101	17
"	"	350	10010	01001	00101	01111	01111	00111	00111	00111	13
75-95	.021	0	11111	11111	11111	11111	11111	11111	11111	11111	25
"	"	90	11110	10111	00110	10110	10110	10110	10110	10110	16
"	"	170	11011	01111	10110	11110	11110	11110	01011	01011	18
"	"	260	11110	01101	10011	11010	11010	11010	10111	10111	17
"	"	350	01111	11010	10110	11101	11101	11101	11010	11010	17
"	".010	0	11111	11111	11111	11111	11111	11111	11111	11111	25
"	"	90	10111	11101	10011	01010	01010	11110	11110	11110	17
"	"	170	11101	11111	10001	10101	10101	11010	11010	11010	17
"	"	260	10110	10101	11010	11010	10001	10001	00111	00111	14
"	"	350	11011	11100	11101	11101	11101	11100	11100	11100	19

^aEnergy, $E = \frac{1}{2} C V^2 = 0.100$ joule at every level; probe dwell-time 2.5 seconds; 2 standard scoop quantities. 0 = No Reaction; 1 = Reaction.

^bFrom R-C resistance (See Table 2, p 8).

^cCapacity

^dVoltage

TABLE 2
Summary of Table 1 Data

	Trials	Hits	Misses	
Capacitance and Voltage*				
E ₁	100	75	25	
E ₂	100	71	29	
E ₃	100	63	37	
E ₄	100	69	31	
E ₅	100	69	31	
Resistance, ohms				
0	100	99	1	
90,000	100	60	40	
170,000	100	61	39	
200,000	100	63	37	
350,000	100	64	36	
Gap Length, inches				
.021	250	175	75	
.010	250	172	78	
Humidity, %				
25 to 40	250	162	88	
75 to 95	250	185	65	
	E ₁	E ₂	E ₃	
	E ₄	E ₅		
*Capacitance, mfd	.0222	.0163	.0125	.0099
Voltage	3000	3500	4000	4500
				5000

Energy was in all cases .100 joule.

TABLE 3
Non-Parametric Analysis of Variance of Table 1 Data

	Calculated H-value ^a	Degrees of Freedom	Critical Chi-Square
MAIN EFFECTS			
Voltage (V)	3.3	4	9.49
Resistance (R)	11.7 ^b	4	9.49
Gap Length (G)	0.0 ^b	1	3.84
Humidity (H)	4.8	1	3.84
INTERACTIONS			
V × G	2.3	9	16.92
R × G	12.5	9	16.92
V × H	0.0	9	16.92
R × H	14.5	9	16.92
G × H	2.0 ^c	3	7.81
V × R	85.9	24	36.42

^a $H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{(R_i)^2}{n_i} - 3(N+1)$. This H-test is the Kruskal-Wallis rank-sum non-parametric

test for the difference among means of counted data where H has a Chi-square distribution and N = Total number of determinations in all groups ($\sum n_i = N$)

k = Number of groups

n_i = Number of determinations in an individual group

R_i = Sum of the ranks in an individual group.

^b Significant at the 98% level

^c Very highly significant

TABLE 4

Three-Factor Factorial Electrostatic Sensitivity Experiment^{a,b} for 29-Micron Magnesium Powder

Energy, joules	Capacitance, microfarads	Voltage, kilovolts	Trials ^c	Total Hits
0.10	0.0222	3	1111111111	10
	.0125	4	1111111111	10
	.0080	5	1111111011	9
.08	.0178	3	1111111011	9
	.0100	4	1011101111	8
	.0064	5	0101111111	8
.07	.0155	3	11110111011	8
	.0088	4	0111101111	7
	.0056	5	1111010101	7
.06	.0133	3	11010101010	5
	.0075	4	0100010111	5
	.0048	5	01010101010	4
.05	.0111	3	0010010110	4
	.0063	4	0010010001	3
	.0040	5	000001000	1
.04	.0089	3	0001010001	3
	.0050	4	000000000	0
	.0032	5	0000000000	0

^aThis experiment was repeated for 10,000 ohms R-C resistance with 100% failures (No reactions). See Table 5 (p 11).

^bProbe dwell-time 2.5 secs, R-C resistance = 0 ohms, Gap length 0.01 to .02 in.; R. H. 25 - 35%

^c0 = No reaction; 1 = Reaction

TABLE 5
Summary of Table 4 Data (See also Figs 8 and 9)

Energy, joules	Voltage	% Hits	
		Zero Resistance	10,000 ohms Resistance
.10	3000	100	20
.10	4000	100	0
.10	5000	90	0
.08	3000	90	0
.08	4000	80	0
.08	5000	80	0
.07	3000	80	0
.07	4000	70	0
.07	5000	70	0
.06	3000	50	0
.06	4000	50	0
.06	5000	40	0
.05	3000	40	0
.05	4000	30	0
.05	5000	10	0
.04	3000	30	0
.04	4000	0	0
.04	5000	0	0

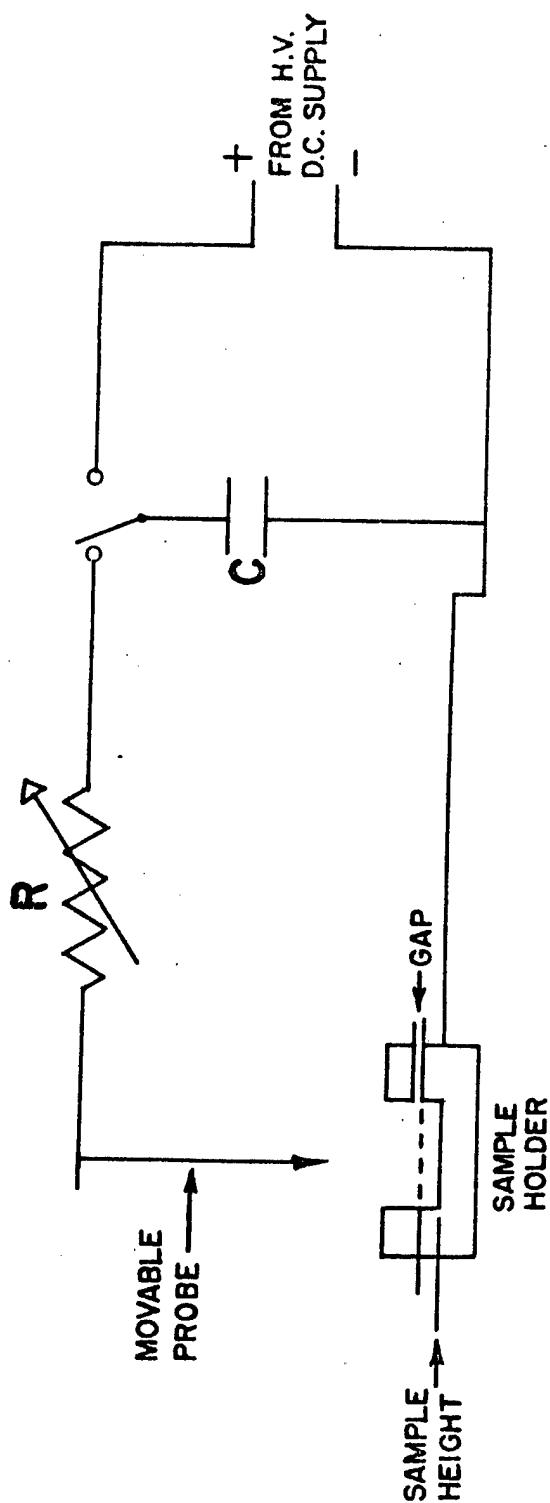


Fig 1 Schematic of Electrostatic Sensitivity Test Apparatus

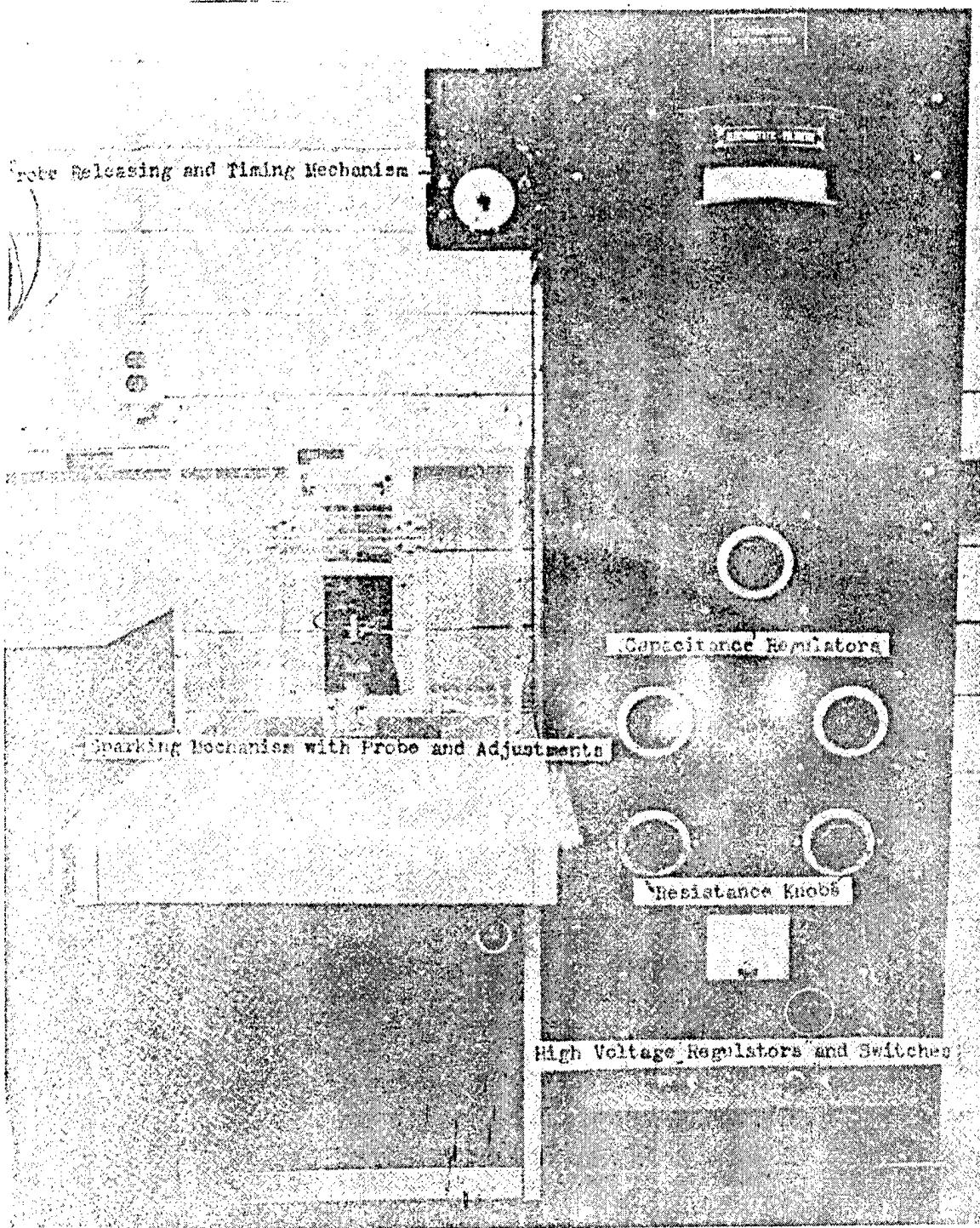


Fig 2 Pyrotechnic Electrostatic Sensitivity Tester

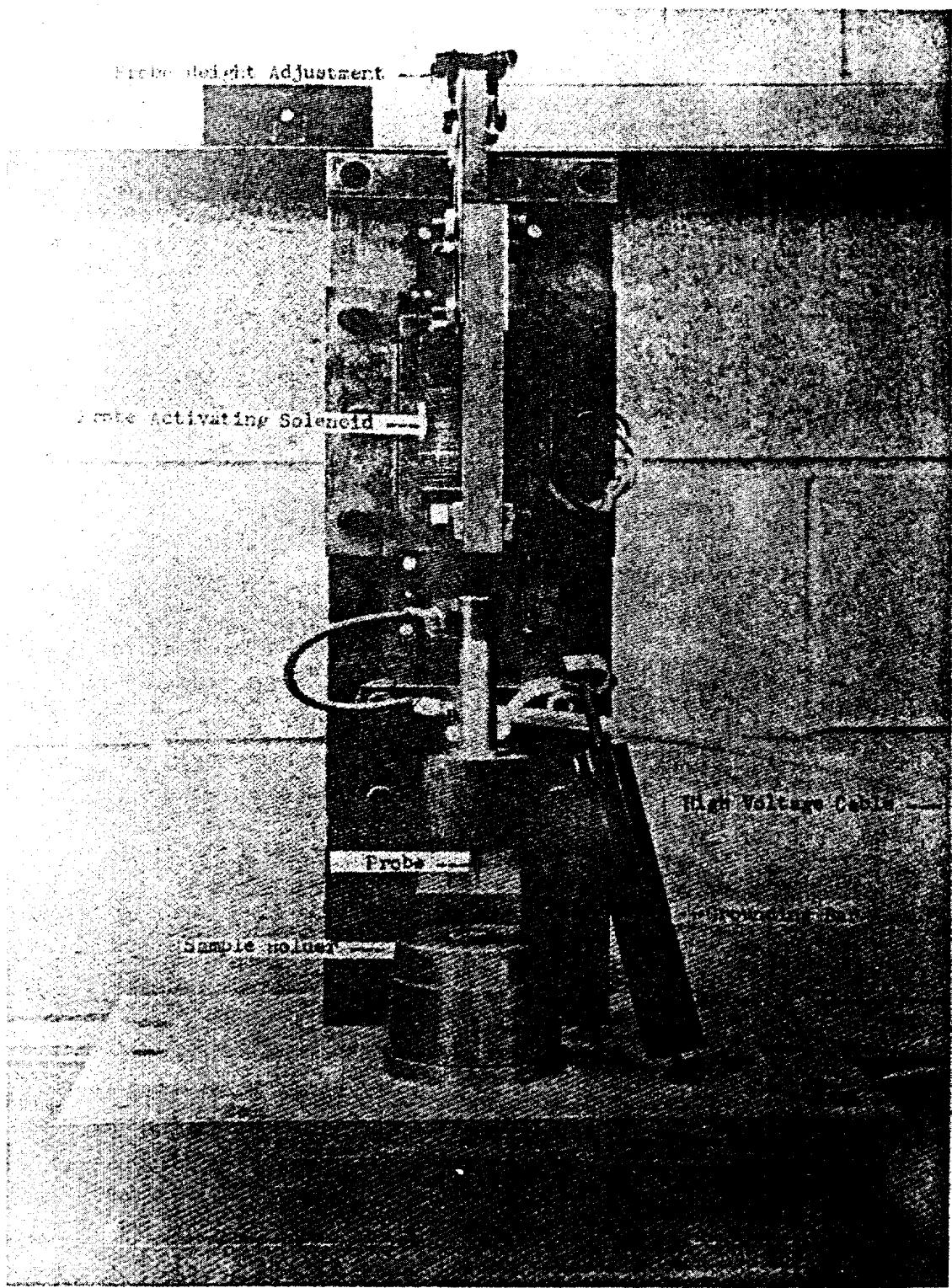


Fig 3 Sparking Mechanism, Probe, and Adjustments

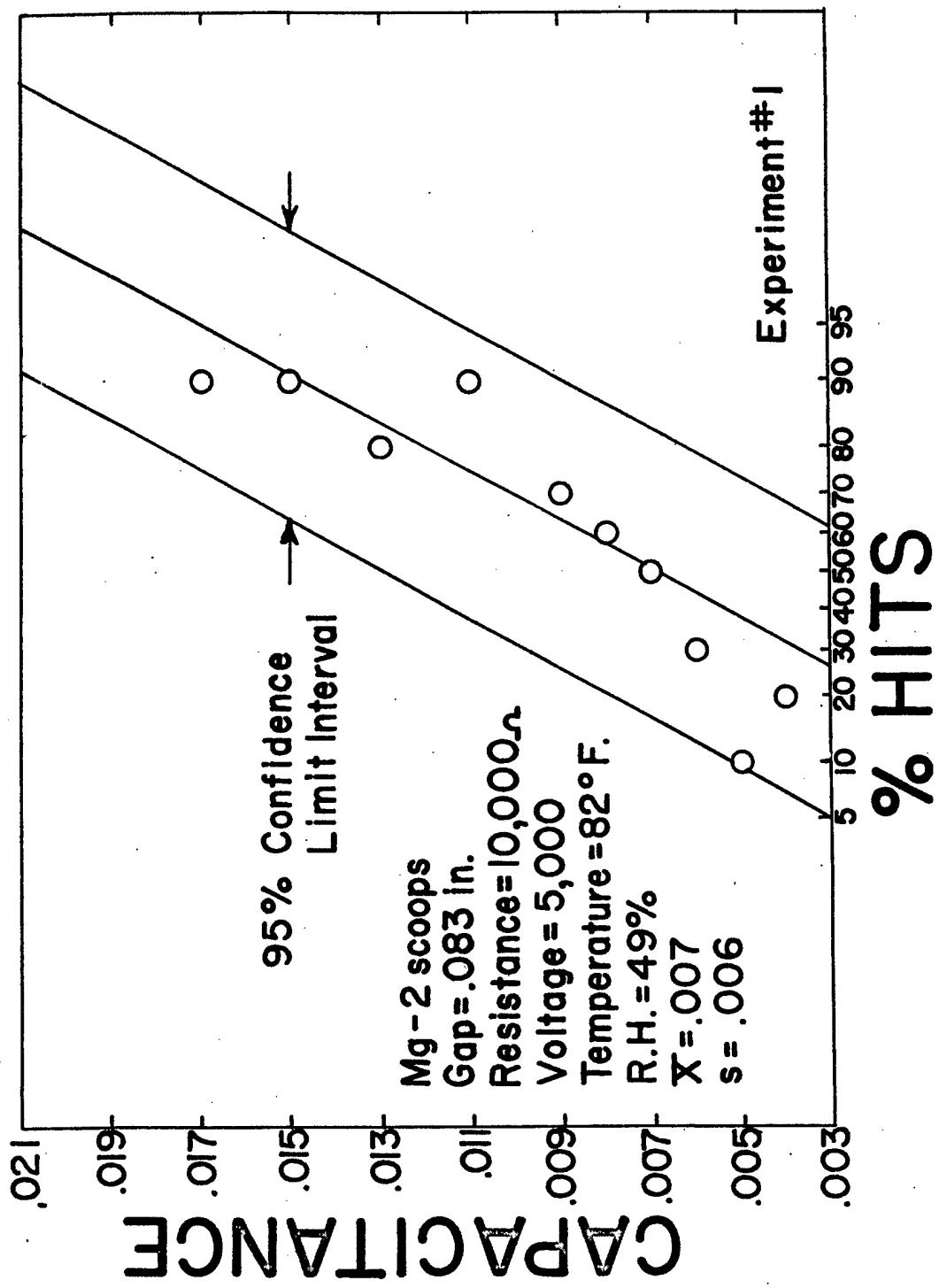


Fig 4 Normality of Distribution of Experiment 1 Capacitance Data

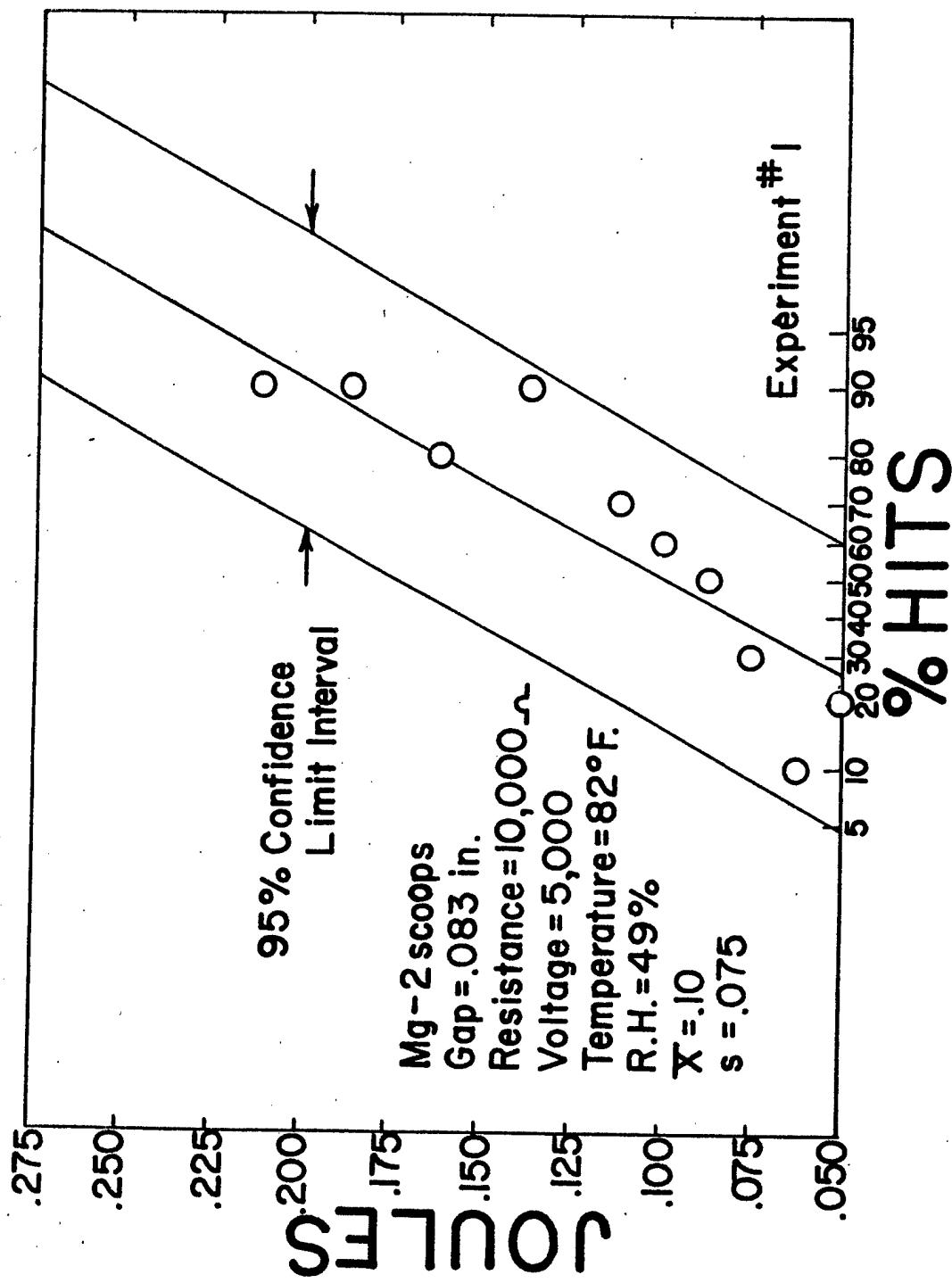


Fig 5 Normality of Distribution of Experiment 1 Energy Data

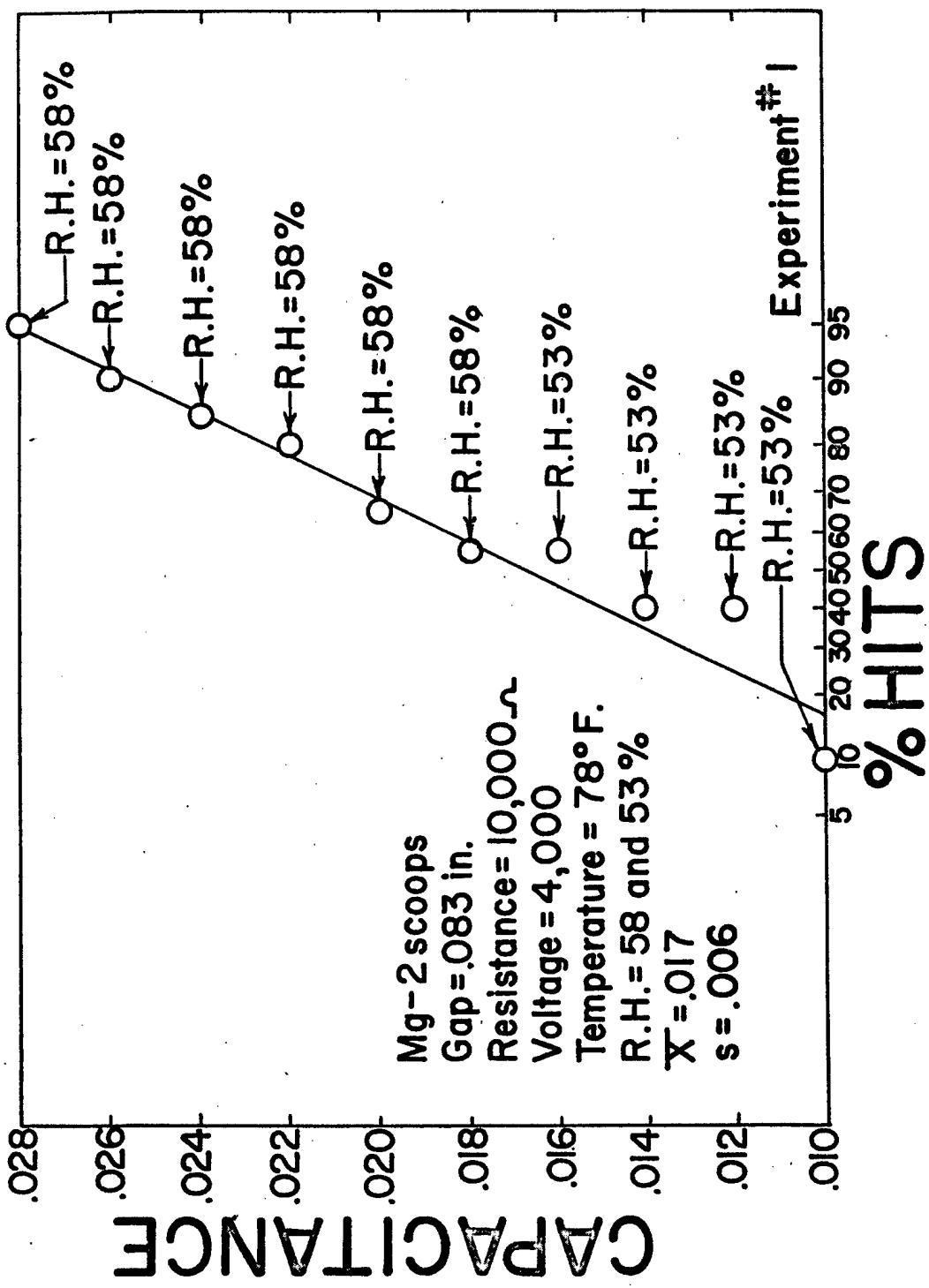


Fig 6 Preliminary Data on Effect of Humidity on Capacitance

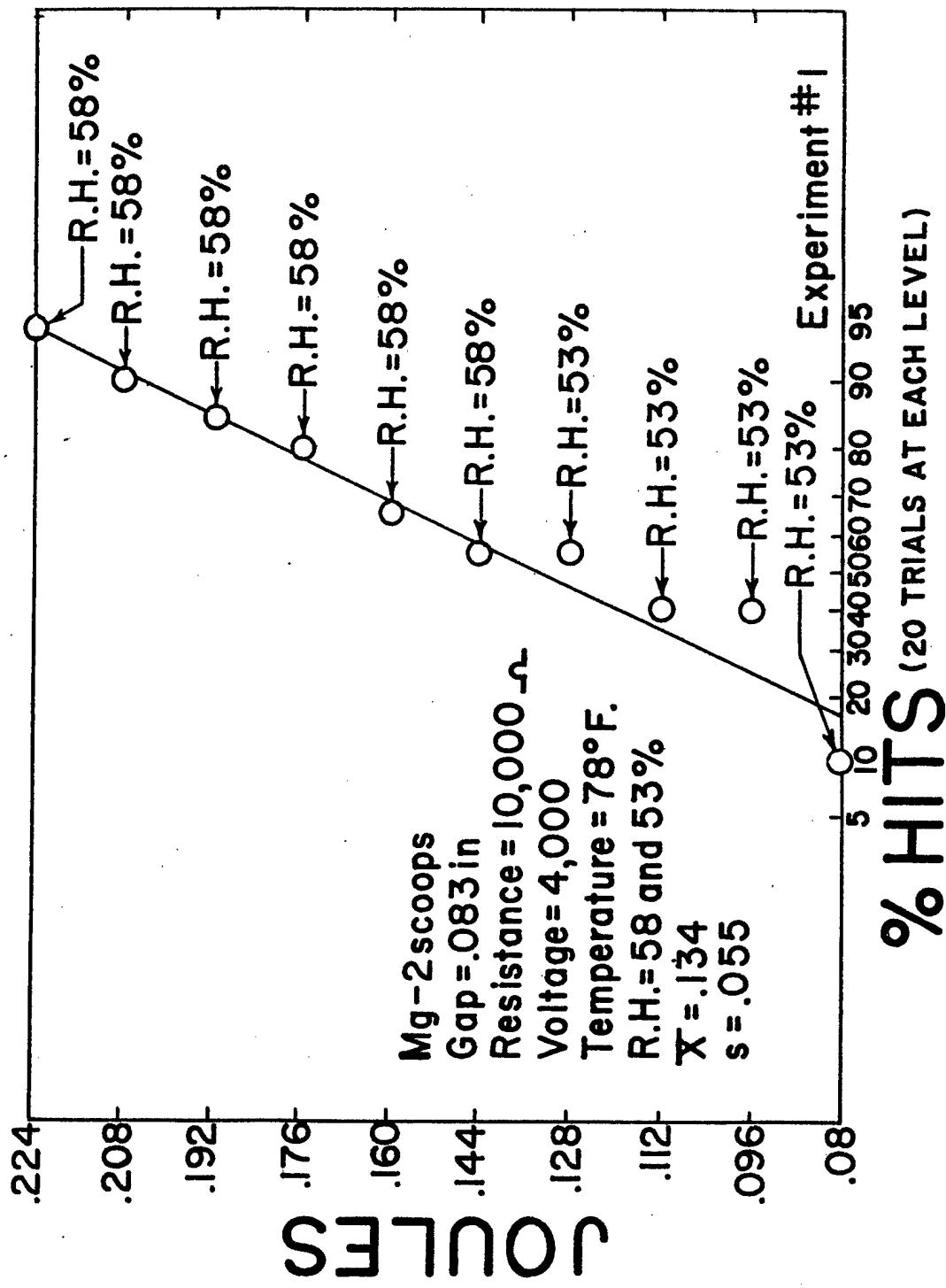


Fig 7 Preliminary Data on Effect of Humidity on Energy

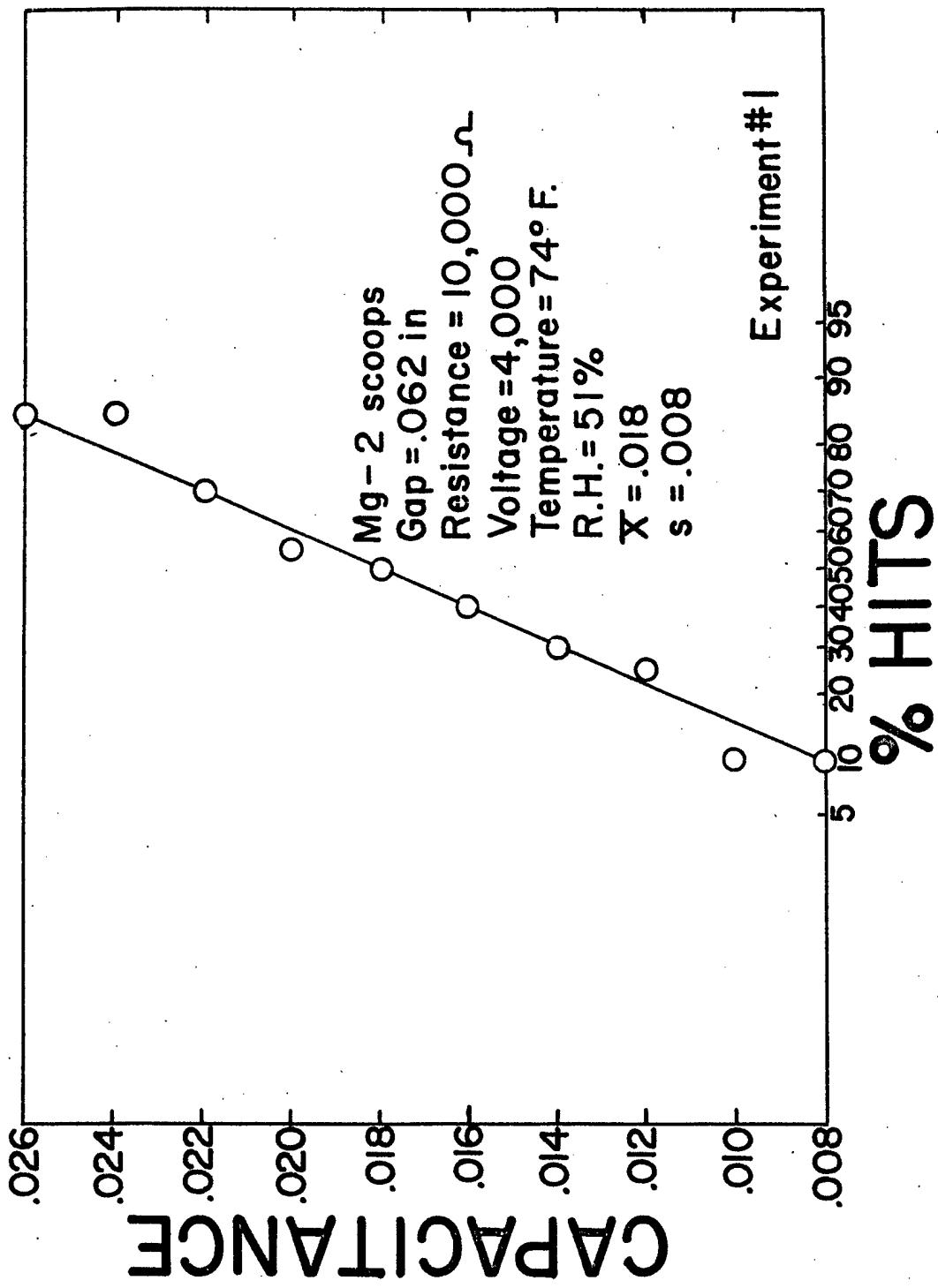


Fig 8 Confirmation of Figures 4 and 6 at Controlled Temperature and Humidity

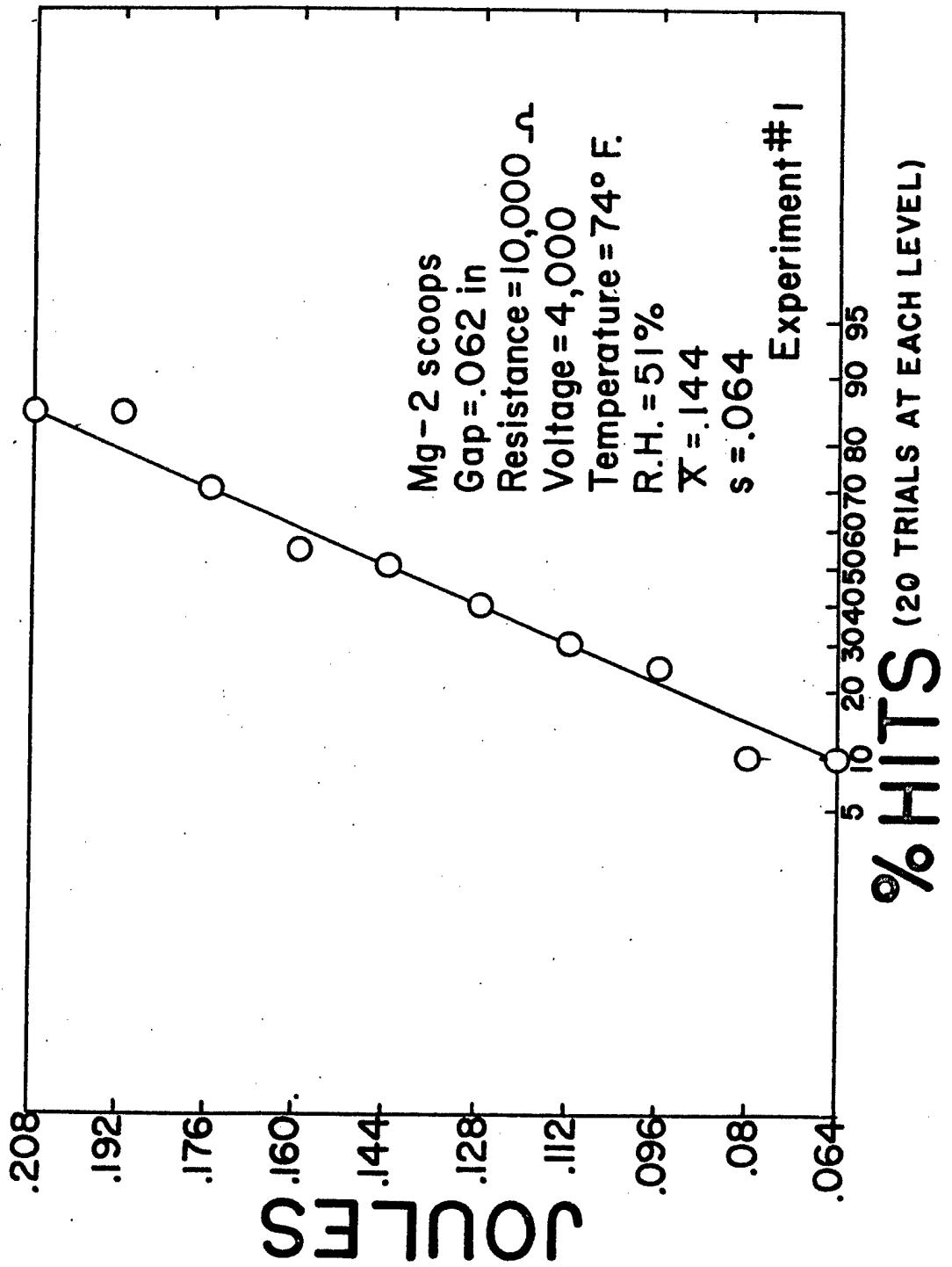


Fig 9 Confirmation of Figures 5 and 7 at Controlled Temperature and Humidity

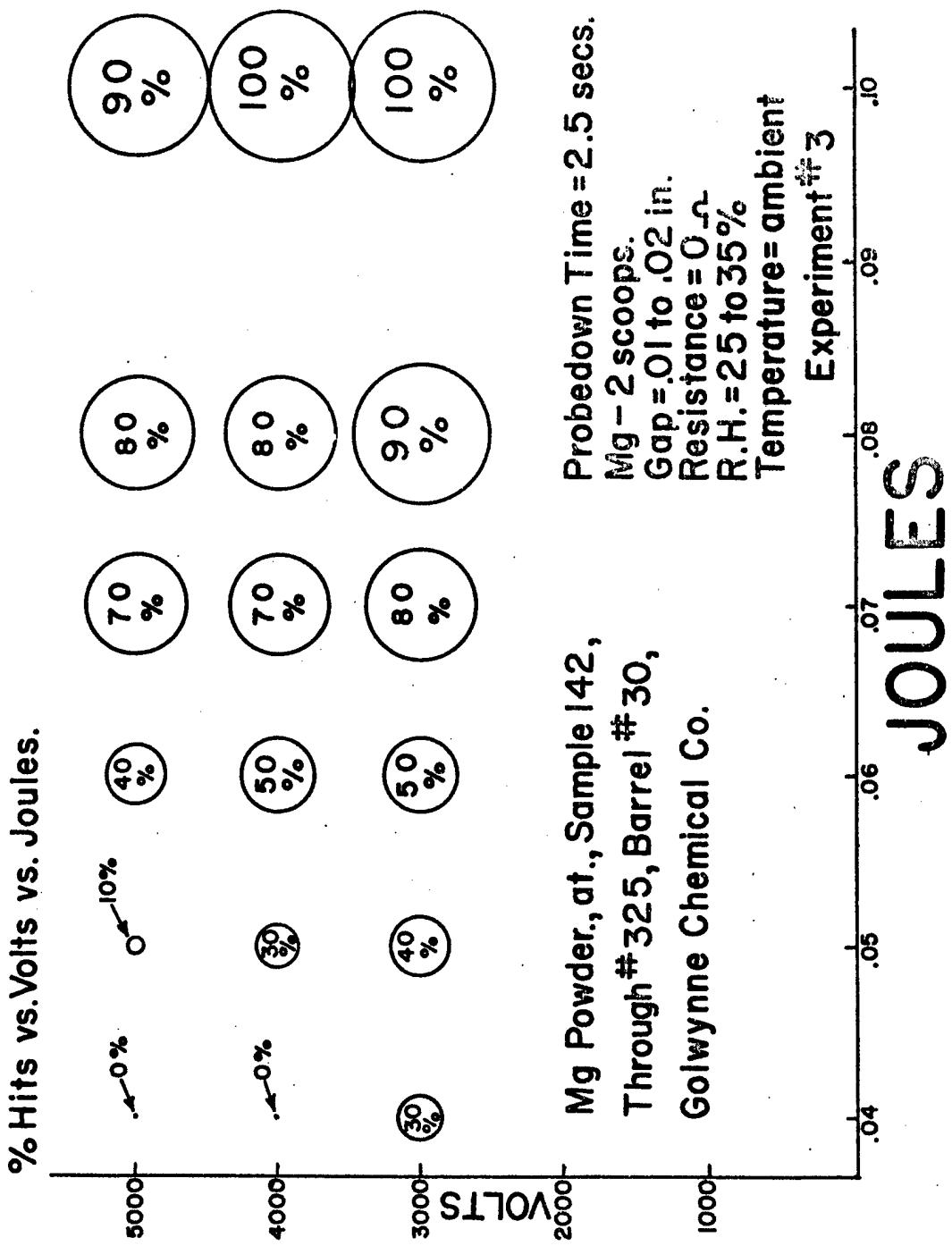


Fig 10 Area Graph Showing Interaction Between Voltage and Energy

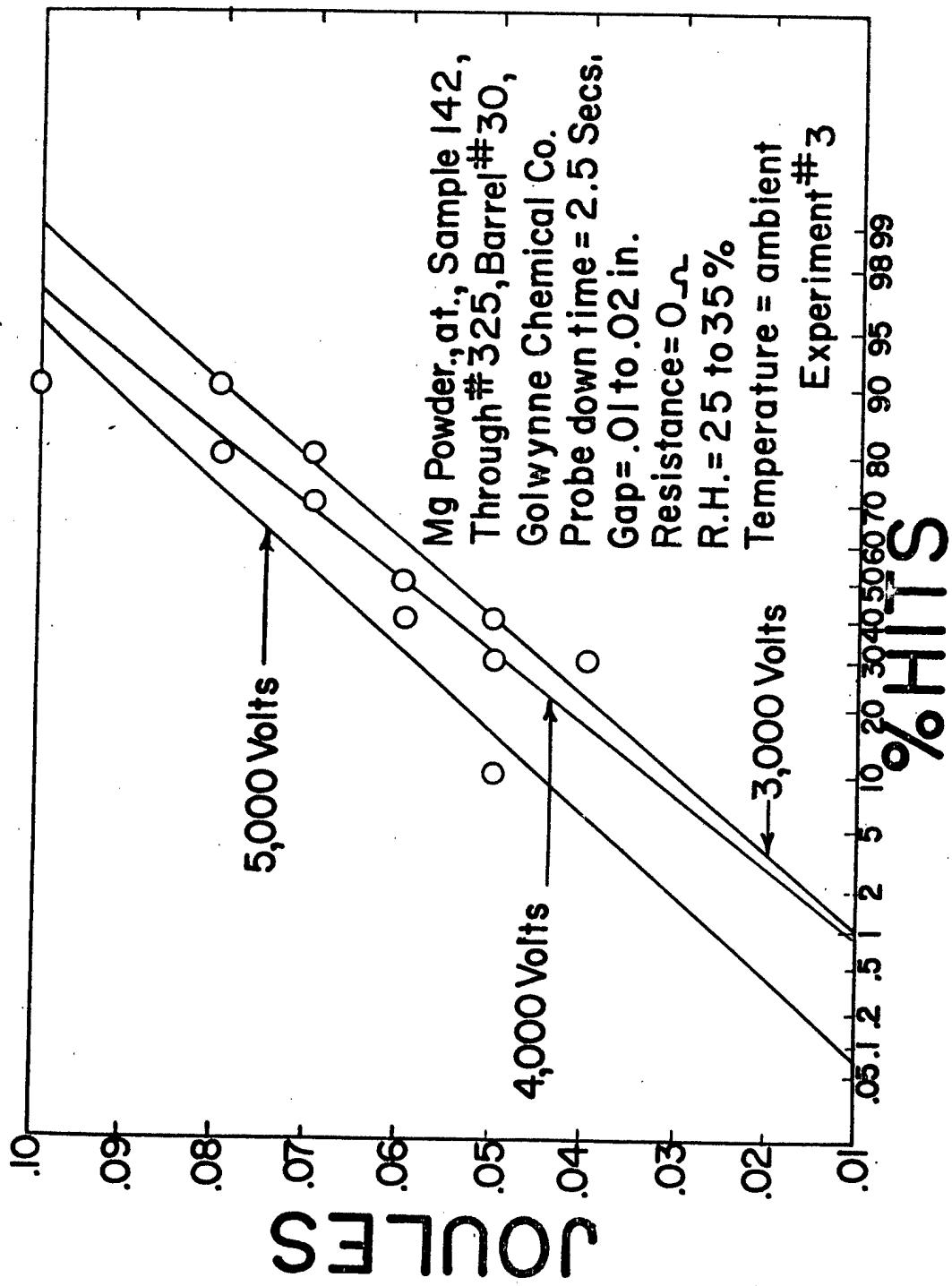


Fig 11 Line Graph Showing Interaction Between Voltage and Energy

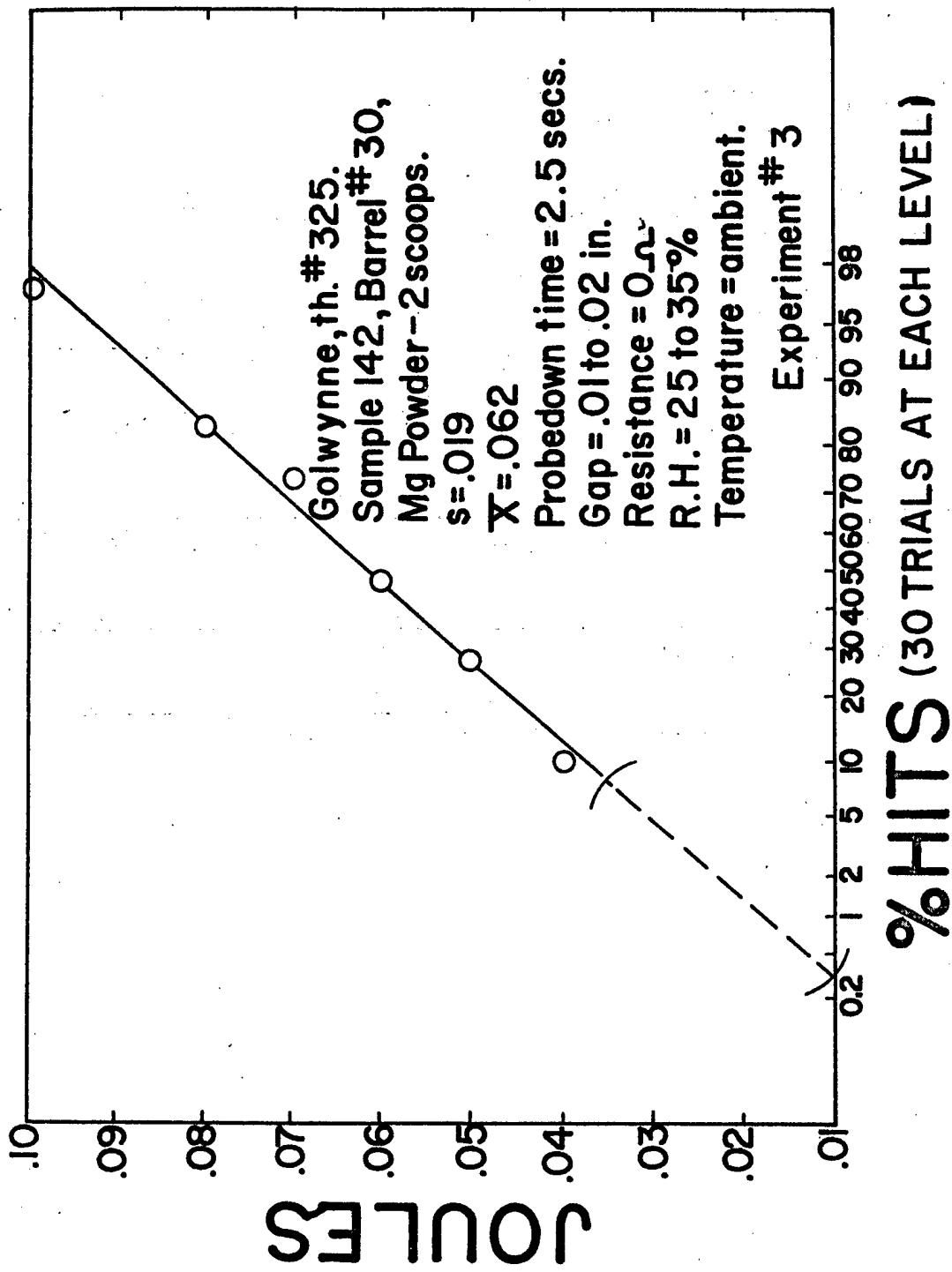


Fig 12 Composite of Figure 11 Curves

REFERENCES

1. I. Hartmann, J. Nagy, and H. R. Brown, Inflammability and Explosibility of Metal Powders, Bureau of Mines, R. I., 3722, October 1943
2. J. N. Ayres, The Design, Assembly, and Operation of the Explosive Electrostatic Sensitivity Tester, Naval Ordnance Laboratory Memo 9959, 7 Feb 1949
3. F. W. Brown, D. J. Kusler, and F. C. Gibson, Sensitivity of Explosives to Initiation by Electrostatic Discharges, Bureau of Mines Report 5002, September 1953
4. P. W. J. Moore, J. F. Sumner, and R. M. H. Wyatt, The Electrostatic Spark Sensitiveness of Initiators: Part 2 - Ignition by Contact and Gaseous Electrical Discharges, C35838(10), May 1956
5. W. Dixon and F. Massey, Introduction to Statistical Analysis, 2nd Edition, McGraw Hill Book Co., Inc., New York City, 1957, page 290
6. C. W. Churchman, Theory and Application of Sensitivity Curves of Small Arms Primers, as Determined by the Standard Drop Test Machine, Frankford Arsenal Report R-259, December 1942
7. C. W. Churchman, Manual for Proposed Acceptance Test for Sensitivity of Percussion Primers, Frankford Arsenal Report R-259A, January 1943
8. O. L. Davies, The Design and Analysis of Industrial Experiments, Hafner Publishing Co., New York City, 1954

ABSTRACT

EXPERIMENTAL DETERMINATION OF "BEST" COMPONENT LEVELS IN THERMAL POWER SUPPLIES

Sheldon G. Levin
Diamond Ordnance Fuze Laboratories

The paper presented the results of a study conducted by the Power Supply Branch of the Diamond Ordnance Fuze Laboratories. The primary purpose of the experiment was to find the combination of component weights (levels) which would give optimum performance of a particular battery under any use (test) condition. Three factors A, B, C, were considered at four levels each, and the details leading up to the selection of a 3/4 replicate were given. There were four responses: life, activation time, noise level, and peak volts, each examined at four test conditions. It was necessary to establish criteria of goodness and state the objectives in mathematical form.

There was no exact solution to this multivariate problem and the fact that the two different approaches to the analysis gave very similar results was very encouraging. The first explored each of the response surfaces and selected the "best" region by ordering the standardized form of the responses. The second method obtained a linear function of the factors which gave a single continuous variable $f(x,u,z) = f$. The responses of life, activation time, peak volts and noise level were then fitted as quadratic functions of f . The set of values of f which corresponded to the satisfactory (with regard to end-use requirements) values of each response variate was obtained. The intersection of these sets constituted the satisfactory, and in this case best region.

STATISTICS IN MEDICAL RESEARCH

W. J. Dixon
University of California, Los Angeles

I shall not attempt an introduction to the statistical methods used in the medical field. I shall attempt to discuss some of the special conditions which the nature of the field imposes on the design situation. As in any experiment we need to know something about the observer. I started out in mathematics, specialized in mathematical statistics, then worked with applications in engineering and social science, and now work in the biomedical and health sciences.

What is the professional environment of medical research? Who are the people conducting this research? They are:

1. Ph.D.'s from physical sciences working in biophysics in instrumentation and simulation research
2. Ph.D.'s from chemistry and physiology doing experimentation in pharmacology, psychiatry and chemotherapy.
3. M.D.'s of all specialties carrying out animal and plant work as well as clinical trials on humans

What problems are investigated? Problems studied in the medical research environment range from the most basic kinds of research of the type carried out in physics, chemistry and physiology and as carried out in psychology and sociology to mere collections of observations-- perhaps, somewhat objective--on standard treatments in standard medical practice. Some form of what we call "statistics" is present in all of these types of research.

What are examples of differences in medical applications? First, consider survey research. Surveys involving records only may present only the usual difficulties in probability sampling, stratification, etc.* But a serious difference may arise if direct interviews or questionnaires concerning health or previous medical history are undertaken. Here, additional problems of non-response and language arises (the National Health Survey is developing techniques in this area). If we seek mental health status, we find problems which at present seem insurmountable. (Nothing but plans that I know of). However, some of the most important differences and differences which force some adjustment arise from the fact that people are involved.

How are people involved?

1. Directly as experimental subjects

* There is an excellent article discussing these applications in Am. J. Ph H. 44 (1954) pp. 719-740, On the Use of Sampling in the Field of Public Health.

Or we may be involved

2. Only slightly less directly if the research outcome has an immediate effect on our own disease or on the diseases of those close to us or on a disease we might expect to contract.
 3. We feel involved as a member of the human community seeking a cure or being concerned about experimentation on humans.
 4. Even if there were no personal concern about ultimate results there are laws governing the treatment of human beings. Roughly stated the law defines the correct treatment as the treatment in general use.
1. Each individual is concerned about his α and β risks (even though they may be poorly formed in his mind) when he thinks about being included in a study. How can one obtain a random sample of objects when they can choose to be unsampled?
 2. If an individual's own life is to be greatly affected in a different way depending on what a study shows, how can he be objective?

How can one do careful and well planned research when so much pressure for time is brought to bear on laboratories connected with medical research? Even though work is still at the stage where it is being carried out on animals, or biological systems other than man, this research is often considered only an early step to later research of direct relevance to people--thus the pressure is on at all stages. Conclusions must be obtained quickly before assumed knowledge is prevalent--there may be no possibility for slow efficient sequential experimentation, particularly on man, because in the medical profession information moves rapidly from mouth to mouth, and if a treatment becomes generally accepted, whether rightly or wrongly, research cannot continue. In the urge for speed we may find the use of a control resisted due to the feeling: "If the drug works we can try it on twice as many people during the same period." As an example, radiation is regularly used as a supporting treatment to surgery for certain cancers. This is not experimental, but accepted. It was not given large scale scientifically designed trials before becoming accepted. This would now be very difficult to do. Since chemotherapy in addition to surgery was not accepted previously, experimentation is possible. Such research is in progress.

Another point which needs careful statement concerns the clinical result versus scientific or statistical result. At least at first, the clinician shows little interest in separation of two groups in mean even though the separation may be precise and real; in clinical work the immediate concern is with each single person. The question is usually asked: Is this result of clinical significance? This may mean can you classify individuals into one of several categories with minimum error?

In a very similar way difficulties arise in the choice and definition of measurement to be used in an investigation. In the basic sciences there is early attention to refinement of criteria, measurement such as refined weighing scales, an accurate radiation counter or some accurate electrical measurement. In the behavioral sciences there is often extensive development of test or questionnaire or interview procedure to gain adequate reproducibility of measurement. In medical research the same care must be taken. You may find that a technique for measuring blood pressure as carried out in office practice (which is entirely adequate for deciding whether a patient's blood pressure is closer to 250 than it is to 120) may immediately be used without refinement in a research study to investigate changes over short periods of time caused by small doses of tranquilizers. Of course, the measurement will almost surely fail to detect these differences. There is a common rationalization supporting the use of unrefined measures in medicine which arises from the confusion of the discovery of a new result and the later developmental problem of making a result of practical importance. The clinician may state: "If the result is not observable with the usual techniques, it will not be of practical value."

In addition to errors in the measuring instrument itself, which are often large, it may be necessary to refine the measurement to some basal state for the individual. One may be able to design a study with sufficient replication of measurement on enough individuals so that one need not hold fixed some of the greatest contributions to variation, but this is usually not the case. In the case of blood pressure for the basal state the recommended technique requires twelve hours of the patient's time and two or three hours of the clinician's time. On the other hand, rapid reading may have tremendous variation. Research is needed in developing for many measurements compromises which will result in more accuracy without too much cost. For the example of blood pressure, some workers are investigating the use of a reading taken a short time after giving a tranquilizer. Development in this direction, of course, introduces a new problem. For example, the use of a tranquilizer may change blood pressure differentially for the type of patient you wish to discriminate so that he is not separated from others.

It must be generally recognized that the type of measurement used should be chosen on the basis of the research goals. However, some research workers with a clinical background may understand that research will require different measurements than are used in clinical practice, but may seek more and more accuracy, when they may need only precision. In some cases precision may be available by a presently known or easily developed technique. If they are investigating changes only, precision may suffice. For example, they may need to observe only an upward or downward shift and be little concerned with absolute level.

Present Medical Record System Creates Problems for Research:
The collection and storage of certain measurements and observations on patients is required for various legal and accreditation requirements. In many places the clinician knows that these records are never used in more than a superficial way. Even though certain data must be coded

accurately, much data need not be. Much of his experience in recording data is with these medical records. Now, suppose you organize a study and require from a certain physician measurements on a patient. His first inclination will be to supply information in the form which appears in the medical records. This information may be of very little value for the research study either because of the use of very gross categories or because the desired information may be included for only certain types of cases. In some cases it can be important to note the presence or absence of the information rather than the size of the measurement if actually recorded. There are many record systems in hospitals and clinics. Records are kept by admissions, departments, by nurses, operating room staff, clinical laboratories, individual physicians. They are kept for medication, infections, special procedures; etc. When a research problem is instituted and a new form is introduced it may receive the easy cavalier verbal fluency of some present records rather than the persevering scrupulous accuracy necessary for good research.

Difficulty of Measuring One Component or Holding Other Variables

Fixed: Much of medical research is done on living organisms of a complex nature, man, for example. Any measurement may affect the individual so that immediate replication is often impossible. The body has many compensating systems so that measurement of only one characteristic is of little value. The component cannot usually be measured except when coupled to all other components of the body. Selective assembly of components is usually not possible. Progress may be made for large or specific types of response with few measurements, but frequently the only approach is through a multivariate analysis.

To what question does the medical research worker seek answers? He asks what to measure. The answers should not only give consideration to the accuracy of each particular measurement but the choice of which to use. It is recognized that many measurements will be required, but one cannot measure everything for reasons of time and money alone. Therefore, studies or redundancy of observations may be required. A component analysis of basic measurements may be helpful. A component analysis may also be helpful in the reduction of the number of variables which must be considered in solving the diagnosis problem.* Regression analyses on basic variates or on components may allow a reduction in the number of variates of importance for certain types of research problem. In other words, we may be able to help in determining which independent variables are of importance for different dependent variables.

There is the question: How to measure? For measurements on a continuous scale the effectiveness of a particular measurement in the analysis of an experiment may be increased by the use of a

* Discussions of these techniques are given by:

M. G. Kendall, *A Course in Multivariate Analysis*, Hafner, New York, 1957
S. S. Wilks, *Mathematical Statistics*, Princeton, N. J., 1943

transformation.* Many observations made in medical research are ordered but without a natural scale. For example, severity of response, answers to history questions, symptoms, laboratory findings, response to treatment. The effective use of such variables will usually require the assignment of a numerical scale. Decision will be required as to criteria underlying the scale to be chosen. One may choose a scale which will optimize a certain regression relationship or one may choose a scale to minimize interaction with another observation being made in the analysis.** The literature dealing with problems of scaling is certainly not complete, but see Torgerson.***

In contrast with this discussion consider clinical trials. Here one may attempt to control many variables not by multivariate analysis but by randomization. For example, one can accept cases as they arise and randomly assign them to treatment or control categories and trust the randomization to effect a balance on the many other related conditions. If knowledge of the status can be kept from the patient and from the doctor these studies are called "double blind." Since the treatment often involves a specific act by the doctor a placebo or dummy treatment is often used.

Examples: Heart surgery, chemotherapy, cold treatment, etc.

What does medical research need (from statisticians):

1. Improvements in basic measurements including scaling methods.
2. Computer programs to ease the pain of multivariate analysis.
3. Further developments in analysis of multivariate measurements including those made continuously in time, e.g., spectral analysis.

How to Use Statistical Methods in Medical Research?

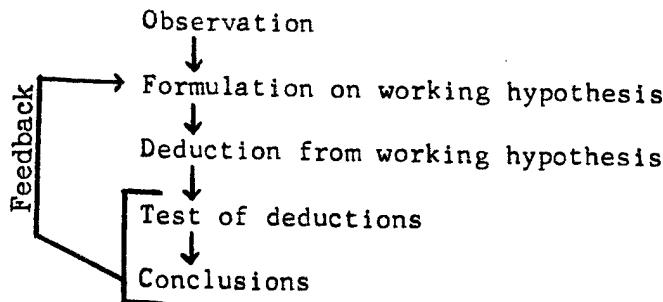
As can easily be seen from reading many current periodicals in the medical field, a great many papers use statistical findings. One can also note that the words probability, confidence and significance are used, but what has been attempted by the statistical analysis often could be called resurrection or sanctification. Since there is a tendency for statistical aid to be requested for the poorer study at the wrong time, it is important to show good research workers how modern statistical methods can be a part of their entire "grand strategy" of the use of the scientific method.

*For example, see John W. Tukey, On the Comparative Anatomy of Transformations, Annals of Math. Stat., Vol. 28 (1957), p. 602

**An interesting application of this criterion for scaling severity of response appeared in a paper by P. J. Claringbold and W. R. Sobey, Studies on Anaphylaxis, Australian J. of Biological Sciences, 10 (1957), 360-364.

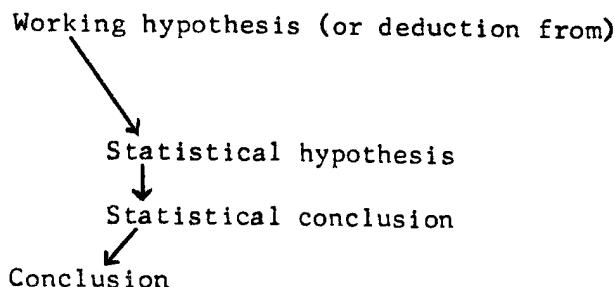
***Warren S. Torgerson, Theory and Methods of Scaling, John Wiley & Sons, New York, 1958.

If we think of the outline



we can perhaps show the research worker that the use of statistical methods can make more efficient at least one stage in the above diagram enabling the experimenter to move to the next stage of his "grand strategy" with some guideposts which will assist him in deciding with what certainty the last point has been established.

In the above outline we may insert a step where statistical models can be of assistance.



This new portion may be carried out by a mathematical model which may approximate reality sufficiently closely to be very useful, particularly if the mathematical model allows for individual variation, i.e., a statistical model. Differential equations have often been used if individuals show little variation so that a representation by mean values is sufficient.

When the statistical model is used to assist the research there are additional feedbacks to the future working hypotheses both from estimates obtained incidental to the main study and from side investigations designed to aid in constructing the statistical model itself.

For a statistical model to be of real assistance it should fit as closely as possibly to the known characteristics of the experimental situation. The research worker will gain confidence in the statistician if the statistician is interested in knowing what type of observations he has because the type of observations available will affect the kinds of hypotheses and certainly the kinds of analyses (or models) which are appropriate. Classifications often used are:

Name	Type	Example
Nominal scale	Identification	White or black
Ordinal scale	Order	Test tubes ordered on color intensity
Interval scale	Intervals are equated	(Most common measurements)
Ratio scale	Zero defined	Distance

How Is a Statistical Model (Including Design) Specified?

Type of population (whether individual or measurements)

Sampling method

Definition of measurement

Hypothesis

What is "different"

What is known from previous experiments (may specify some parameters for this experiment)

Form of distribution of observations (or at least kind)

What next step is anticipated in line of research (estimate additional parameters for future design)

Statement of risks

α risk of rejecting hypothesis when true

β risk of accepting hypothesis when difference = d

Estimate of sample size (or sequential plan)

Expected troubles and controls or safeguards

It is well for the research worker to know what kind of problems a statistician considers simple and what kind are more difficult.

A. Examples of simple problems

One variable

Does a change occur in one measurement when a treatment is given?

Do two treatments differ in their effect on one characteristic of a patient?

Do several treatments differ in their effect on one characteristic of a patient?

Two variables

Are two measurements associated in the same way for two types of material?

B. Examples of problems less simple

Which one of 7 treatments is really different from the rest?

Does a change occur when a treatment is given when change may be reflected in any of several variables or combinations of them?

Which group of 4 or 5 measurements out of 28 measurements is best for assigning an individual to Group A and Group B?

Example: from amino acid measurements assign to group with or without hepatitis

Example: from physical and social variables assign mentally defective children to groups according to ability to profit from training

If a variable y is affected by changes in any of four x variables what is a good strategy for selecting the particular x values to optimize y ?

The man developing the statistical theory will seek answers for the above questions, answers which will specify the appropriate model.

What Sort of Models Will the Statistician Give the Research Worker?

Most problems may be attacked by certain general statistical models. Or, a new model may be developed which follows more precisely the assumptions of the particular experiment under investigation.

The situation is somewhat like the tailor who has a shop with a number of ready-made suits. They can be used with little delay and sometimes slight alteration may be made. A tailor-made suit may require a great deal of time to construct, perhaps by many tailors and will certainly cost more. The ready-made suit will not fit everyone, but may fit a great many fairly well-some, very well.

What are some of the ready-made suits?

<u>Scale</u>	<u>Statistics</u>
Nominal	Binomial, multinomial, χ^2
Ordinal	Order statistics, non-parametric statistics, median percentiles
Interval	Mean, standard deviation, standard test, t-tests, correlation, regression, analysis of variance
Ratio	Little different from interval scale except some parameters may be specified and not estimated

For analysis of observations on the interval scale most ready-made suits assume normality. How does one think about satisfying the assumption of normality?

- a) Is it known?
- b) Is it to be verified?
- c) Will it be produced? (transformation--including averaging)
(here we change shape of the man to fit the suit; he may have to wear a transformation before the suit will fit)
- d) Will we show special caution in conclusions?
- e) Will the assumption be avoided, e.g., by use of non-parametric methods? (this may be like using a suit that is too big--it may not pinch, but it may not fit closely anywhere)
- f) Will a theory be developed for the appropriate distribution?

Another assumption which is made by many tailor made suits for analyzing observations on the interval scale is homoscedasticity (the assumption of equal variances). The same considerations can be listed for satisfying this assumption that were mentioned in the discussion of normality.

The research worker may be interested to know the criteria statisticians use in constructing models.

There is the goal of efficiency. Minimization of the number of observations for fixed risks α and β or minimization of one risk when the other risk and the sample size is fixed.

There is the goal of unbiasedness or accuracy. A statistical measure should be correct on the average.

There is the goal of minimum variance or precision. A statistical measure should have minimum variability as an estimate.

There is the goal of maximum power in the test of a statistical hypothesis. If the hypothesis is not true we should have a good chance of discovering this fact.

We must also tell the research worker that the particular statistic which will satisfy these goals depends on the type of observations which he makes. There is no universal answer.

It may be news to the person contacting the statistician for purposes of sanctification that the field of statistics also concerns itself with the construction of designs for the experimental attack itself. There are some answers to the questions:

Which cases should I select?

Which dose should be given next?

Which variables should I measure?

What combinations of treatments should be investigated together?

Do we make more measurements with no change in conditions or observe under more different conditions?

Some of these questions may come under the heading of the strategy of replication.

A chemist is quoted as follows, "I don't believe in replication. If you measure it once you know what you've got. If you measure it again and don't get the same answer, you don't know where you are."

Replication is often considered to occur only when several observations are made under identical conditions. The use of appropriately balanced designs can yield the advantages of replication at the same time one does experimentation over a wider range of other variables.

If we consider the example of investigating the effect of three variables at each of three levels, we make 27 observations all under different conditions if an observation is made for each combination of levels. But, since the experiment is performed in a balanced way we investigate the effects of changes in several variables simultaneously and can investigate the individual effects of each variable alone, the interaction of one variable with another and estimate the replication or measurement error. The randomized blocks and factorial designs can be used to advantage in medical research. However, we can also ask:

Is it essential to measure all possible combinations?

If not all are required we can perhaps suggest a carefully balanced subset which will still provide answers to the important questions. Such designs are latin squares or other fractional factorials.

My comments this evening are not intended to be comprehensive but only to indicate by examples the importance of both the statistician and the research worker continually educating each other and to list a few of the important points about which education should take place.

I wish to close with two comments which may be classified as philosophical.

First, there is a great concern in medicine for arriving at conclusions which will state a cause and effect relationship rather than an association. I have found it necessary to offer the information that statistical analyses in general only demonstrate association.

Second, the comparatively recent attention to Type II error in statistics may be at fault for its neglect in many scientific reports at the present time. Its continued neglect, however, may be tied to the notion of conservatism since the Type I or α -error is controlled. In diagnostic situations this is often the risk of challenge to authority. There seems to be less interest in the risk of continued acceptance of authority which is not correct.

SAMPLING IN BIOLOGICAL POPULATIONS

D. B. DeLury
University of Toronto

It is not my wish to burden you with the practical details of methods that are used to estimate the vital statistics of biological populations. I propose only to skim lightly over some of the more interesting methods and to take advantage of the occasion to preach a sermon on the sin of non-randomness.

We have today an elaborate, well-developed Theory of Sampling, aimed chiefly at human populations and the things they do. The place where we get a grip on these problems is the fact that these populations are fixed geographically or in some other way that provides a basis for effective stratification. Furthermore, these populations have the property that we can, in principle at least, get at all the individuals in them and therefore we have a basis for a positive randomization procedure to select samples within strata.

We have also today a well filled-out discipline called the Design of Experiments, and this too has meaning only in circumstances in which effective randomization is possible.

Not everybody today is as convinced of the importance of randomization as he should be and consequently some of our investigations fall short on the score of randomness. Perhaps a look at the antics that people engage in when randomness is not possible will point up the essential role that randomness plays.

Some biological problems, of course, fall nicely within the scope of standard sampling procedures, e.g., a study of a population of nesting birds, or a beaver population, even though it might be difficult enough to carry them out. These things have, in fact, been done a few times, but the job of carrying out the dictates of randomness is formidable indeed.

It is not questions of this kind that I want to talk about, but those in which random selection of samples is truly impossible. I shall speak particularly of sampling populations of fish as, perhaps, the most striking instance of this. The impossibility of doing anything positive to ensure randomness in samples of fish is obvious enough, and those negative steps we might take to avoid the most unpleasant consequences of non-randomness are unknown or are known only in a qualitative way. Furthermore, those features of the populations which can upset our procedures are known only qualitatively.

Fish are generally, though not always, highly mobile. They tend to stratify according to age or size and in other ways as well. Every method of capture we have devised is biased with respect to size and doubtless other features too. Indeed, the conviction is growing that the probability that a fish will be caught by any particular device varies widely from one individual fish to another.

In any event, let us look at some of the things people do, to try to estimate the size of a population of fishes, keeping in mind the difficulties I have been talking about. Nothing here is at all new; this is simply an account of things that have been done.

TAGGING. One of the older tricks is to catch a sample of fish, put tags on them or otherwise mutilate them so they can be recognized, release them and watch for their reappearance in subsequent catches. Then, if we can assume that the proportion of marked fish in the catch is equal to the proportion in the population, apart from sampling error, we get at once an estimate of the size of the population. This device goes back to Laplace, I believe, but he did not use it on fish..

To put this into symbols, using t to denote the time since the tagged specimens were released:

	Tagged	Total
Population	X_t	N_t
Sample	x_t	n_t

$$\frac{x_t}{n_t} = \frac{X_t}{N_t} = \frac{X_o}{N_o}, \quad N_o = \frac{n_t X_o}{x_t},$$

valid as long as any depletion affects tagged and untagged equally.

While there are many things one can think of which would upset this estimate, the weakest spot is surely the supposition that the proportions in the sample and in the population are equal -- i.e.,

$$\frac{x_t}{n_t} = \frac{X_t}{N_t},$$

or to put it otherwise, tagged and untagged are equally catchable. This is the kind of thing we look to randomness to ensure, the kind of thing we expect to be met a priori through the way we select our samples. If we were dealing with beads in a box, instead of fish, we would mix them thoroughly after introducing the marked beads and then draw out our samples. With fishes, what can we do? Perhaps not very much, but we are coming to know some of the pitfalls. To illustrate one of the common ones, I can tell you about a tagging experiment carried out some years ago to ascertain the number of black bass in a rather large lake. The bass were captured in traps, put here and there around the shore and left there the whole summer. Many bass were tagged and recaptures were numerous but the estimate of the total number of bass, based on these recaptures, was something like 500, a completely fatuous figure. Now it happened that good records had been obtained from the anglers in the lake, both numbers caught and numbers tagged. An estimate based on these records was around 30,000. What was happening here? A run through the records disclosed that, with

practically no exceptions, every bass recaptured was taken in the same trap in which he was first caught. We can see, then, what happened. As far as the traps were concerned certain individuals were highly exposed to recapture, over and over, while others had little chance of capture at all. The same effect has been seen in other populations, for perhaps different reasons. One thing to be feared, then, is that fish captured in any particular manner are more likely than the others to be captured again in the same manner -- i.e., tagged and untagged are not equally catchable. This is particularly dangerous when repeated recaptures are used. It seems prudent, then, to recapture by a method that is different from the one used in placing the tags or, better still, use several methods for both and keep such records as are needed to keep track of the methods by which each fish is captured.

Whatever we may do, however, the tagging procedure provides no real check on the crucial assumption that tagged and untagged are equally likely to be caught. Information of another kind is needed if we are to do so.

Tagging procedures have been extended far beyond the simple one described here. We can, for example, tag and sample simultaneously in various ways. Such procedures, in which we can no longer treat as constant the proportion tagged in the population, will be distorted by the operation of an appreciable mortality during the sampling period, which adds another source of uncertainty. The fact that they are so distorted means, of course, that they contain information about mortality and a few schemes for extracting estimates of the mortality rate from them have been proposed. For most part, they depend on repeated recaptures and, for this reason, our inability to sample randomly strikes them particularly hard. We can side-step this dependence on repeated recaptures, but the methods we must then use are very weak.

Perhaps you will permit me to take off here on another tack to speak of a question that has vexed me for some time, because I see no good way of getting a grip on it. It is a point, though, that may have some importance outside the immediate context. One of these tagging plans furnishes a good vehicle for the discussion.

Let us suppose that we tag and sample simultaneously, perhaps by tagging and releasing all untagged members of each sample and releasing also those already bearing tags. To keep the discussion free from difficulties that are here irrelevant, let us say that the size of the population is constant throughout the sampling period, i.e., no mortality, immigration and so on. Let N stand for the size. Then, using the notation written down earlier, and making the reasonable assumption that each sample is small compared to the size of the population; i.e., the sampling is effectively binomial, we can write the probability of getting x_t tags in a sample of n_t :

$$\binom{n_t}{x_t} \left(\frac{x_t}{N}\right)^{x_t} \left(1 - \frac{x_t}{N}\right)^{n_t - x_t} = f(t), \text{ say, and the likelihood}$$

is

$$L = \prod f(t).$$

A direct maximum-likelihood calculation yields, after some algebraic rearrangement, the estimating equation

$$\sum \frac{x_t - n_t X_t / N}{1 - X_t / N} = 0.$$

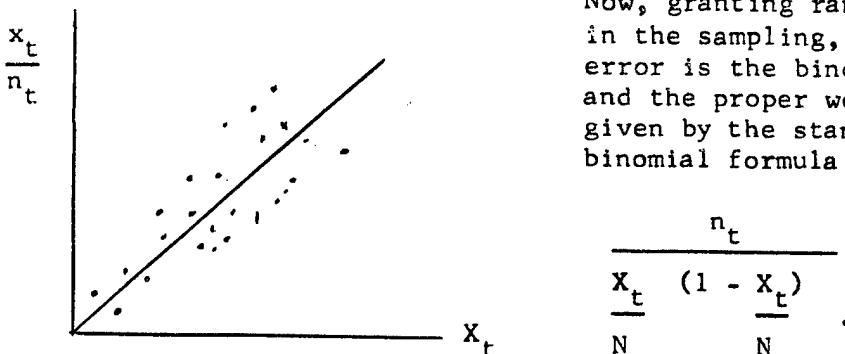
This equation can be solved for N by numerical methods, but this does not concern us here.

We can take another view of this question. Since $\sum \frac{x_t}{n_t} = \frac{X_t}{N}$,

we might think of plotting $\frac{x_t}{n_t}$ against X_t and fitting by least squares

a line passing through the origin. The slope of this line estimates $\frac{1}{N}$.

Now, granting randomness in the sampling, the only error is the binomial error and the proper weights are given by the standard binomial formula



Using these weights, the least squares equation can be rearranged into precisely the M.L. Equation, as might, perhaps, be expected. The point here is, then, that the M.L. estimate uses as weights functions of the population proportions $\frac{x_t}{N}$. Now if the sampling is, in fact, random,

this is presumably the most effective weighting. However, in these circumstances, it seems reasonable to question this weighting, because, when any sample is taken, the proportion available to it may be grossly different from $\frac{x_t}{N}$. It seems to me, therefore, safer to weight by sample size, n_t ,

which incidentally leads to a simple estimating equation. The general question, then, is: when our sampling may fall short with respect to randomness, are our so-called efficient statistics even as good as other, less efficient ones? I do not think any general answer is possible here. The question itself is not precise and cannot be accurately formulated. I have, however, compared the two approaches outlined above on artificial data in which randomness is assured and can perceive no persistent difference between them.

One could spend several hours on ramifications of the tagging method, but I must pass on to a procedure of a different kind.

CATCH-EFFORT. This method rests on a common observation that, as a population becomes depleted, the return from a given amount of sampling effort declines. In order to make quantitative use of this fact, some assumptions are necessary and data of a specific kind must be collected. Let us say, then, that for each of a number of consecutive intervals, we have a record of the catch - $c(t)$ and the effort expended - $e(t)$. $e(t)$ might be measured in net-nights, boat-hours, etc. Then, we can calculate for each interval the catch per unit of effort, $C(t) = c(t)/e(t)$ (t refers to the t^{th} interval). We can calculate also the accumulated catch and the total effort expended up to the t^{th} interval.

$$\begin{aligned} K(t) &= c(1) + c(2) + \dots + c(t-1), \\ E(t) &= e(1) + e(2) + \dots + e(t-1). \end{aligned}$$

Now, let us make the simplest assumptions we can think of, say that the population is closed and that a unit of effort takes always the same fraction k of the population. k has been called by some the "rate of fishing" and by others the "catchability" -- in any event, k is measured in (units of effort) $^{-1}$.

We have, now, from the definitions and assumptions:
 $c(t) = k e(t)N(t)$, $N(t)$ being the size of the population at time t , or, dividing by $e(t)$:

$$C(t) = kN(t) .$$

Also, because fish can leave the population only through being caught,

$$N(t) = N(1) - K(t) .$$

$$\text{Hence } C(t) = kN(1) - kK(t) .$$

Now C and K are observable quantities. We can plot them, and if reasonable straightness results we can fit the line and so estimate $kN(1)$ and k , hence $N(1)$.

From the same assumptions, we can derive another relation,

$$\log C(t) = \log [kN(1)] - kE(t).$$

While these two relations are equivalent in a mathematical way, as soon as they are embedded in a statistical setting, they show up quite different. The first has only to do with the conditional behavior of $C(t)$ given $K(t)$, which depends only on $K(t)$, whereas the second depends on the whole stochastic process up to the t^{th} interval. As far as I know, this has never been elucidated and we know less about how to fit it than we do for the first.

We may properly be concerned about the rather restricted assumptions under which these relations have been derived and indeed we may be sure that frequently they will be violated. On the other hand, there is a certain amount that can be done to ensure that they will hold reasonably

well. The one assumption that is largely outside our control is the one which asserts that a unit of effort always takes the same fraction of the population -- i.e., that the catchability is constant throughout the sampling period. Here again is an assumption that we would look to randomness to take care of. Everyone knows that this parameter is bound to fluctuate, perhaps widely, from day to day, but this is not as serious as the possibility of persistent trends. We may expect trends in the catchability to distort our graphs out of straightness, but it must be granted that day-to-day fluctuations are usually large enough to make decisions on this question rather difficult. Furthermore, other failures in our assumptions can produce curvature.

On the whole, then, we find ourselves with two methods, each dependent on a crucial assumption which cannot be tested with the kind of data usually collected in applying this method. Each of them is the kind of thing that we usually look to random sampling to take care of. In this case, the best we can hope for is to try to find independent checks on the critical assumptions.

A little reflection will show that each of these methods contains information that can be used to test the assumption on which the other depends, provided that the two methods, tagging and catch-effort, are applied to the population simultaneously.

The catch-effort method requires that catchability remain constant and the tagged portion of the population provides a population of known size, for which direct estimates of the catchability can be made from the sequence of returns. These estimates can then be inspected for trends. The tagging method, on the other hand, requires that tagged and untagged be equally catchable. A catch-effort analysis, conducted separately on the tagged and untagged parts of the population, puts these two catchabilities directly in evidence, so that a comparison can be made.

Investigations of this kind are necessarily elaborate and expensive and, as far as I know, this combined tagging and catch-effort study has not often been attempted. Tagging alone has been used many times, but it is likely that most of the population estimates so reached are seriously wrong. The catch-effort method is newer and has been used much less. We need more experience with it, but we can say that, in some instances at least, it works pretty well. When the sampling is extensive enough and intense enough to bring about heavy depletion in a rather short time, we may expect the catch-effort method to function fairly well. This does happen not infrequently in commercial fisheries and some sports fisheries.

I have not spoken of the elegant and theoretically powerful methods developed during the past few years by a number of people, P. H. Leslie and D. G. Chapman, to mention two. These methods depend on tagging and in some instances, on catch-effort as well, and they purport to yield estimates of all the vital statistics of a population, birth and death rates and so on, as well as population size. Not only do these methods rest heavily on random selection of samples, depending as they do on repeated recaptures, but they also involve actuarial notions such as mortality rate, presumably

constant and stable. Now these notions work well enough in human populations, because we have pretty well eliminated the catastrophe from our system of causes. Likewise, actuarial methods, i.e., methods based on mortality tables, may well be applicable in some controlled populations, e.g., laboratory populations. In natural populations, however, the most prominent effects are environmental, fluctuations are wide and capricious, masking entirely the built-in cause-and-effect mechanisms without which no population can survive. Only when the populations are extraordinarily dense or sparse do these mechanisms play a predominant role. For these reasons, I think we have to keep our sights fairly low in dealing with natural populations; in particular, we should not use methods which depend on the kind of stability we have become used to in human populations.

THE APPLICATION OF FRACTIONAL FACTORIALS IN MISSILE TEST PROGRAMS

Paul C. Cox
Ordnance Mission, White Sands Missile Range

I wish to begin by acknowledging the assistance I have received from Dr. Boyd Harshbarger in developing and applying the specific fractional factorial which I plan to discuss. I understand he had an entire class at VPI work with this design, and the fact that it is a design which is presently being used for a specific missile test program proved to be both stimulating and motivating to the class.

The statistical designing of a missile test plan is usually plagued with numerous serious problems. These include: (1) There are usually a large number of different types of treatments and at several levels which must be evaluated; (2) the sample size is almost always quite small; (3) there are quite often, physical limitations which will place severe restrictions on the design; (4) the test must be designed with the realization that the statistical evaluation is but one of many requirements to be satisfied by the test; (5) most missile test plans require several changes after testing has started; and (6) there will invariably be a few test failures which will either require substitute firings or will result in a loss of data.

I will restrict my discussion to the first two of these problems the large number of treatments required and the restricted sample size. Because of these characteristics, fractional factorials can sometimes be a useful tool in designing missile test plans. Fractional factorials are valuable because they provide for a reduction in sample size. This reduction is bought at a price and the price should be carefully studied before using a particular fractional factorial for a specific missile test plan. Actually, we at White Sands Missile Range consider fractional factorials frequently, but because of the limitations, use them rarely.

I will illustrate these points by discussing a problem we were recently faced with and how a fractional factorial appeared to be the best solution of the design problem.

It was required that a missile system be flight tested to determine its effectiveness under the following combinations of test conditions:

Type of Warhead	W_1	W_2	W_3
Propellant Temp.	T_1	T_2	T_3
Launcher-Target Range	R_1	R_2	R_3
Launcher Emplacement	E_1	E_2	E_3
Launcher	L_1	L_2	

(Two identical launchers).

It was decided that the missiles should be fired from the two launchers in pairs. In this way, the error mean square will contain only the

variability due to the missile system and no variability due to changes in weather, firing personnel, etc.

Ideally one should use a factorial design with a minimum of four rounds tested under every set of conditions. (One pair at one time and the second pair at some other time). This would require a total of $3 \times 3 \times 3 \times 3 \times 4 = 324$ rounds; a figure which is entirely unrealistic.

Since a factorial design could not be used, the obvious procedure was to study the physical characteristics of the test and try to develop a fractional factorial which would be suitable. It was determined from the study of the physical characteristics of both the system and the test that emplacements would probably not interact with any of the other test conditions. Consequently a fractional factorial was designed by setting up R, W and T as a factorial design; then for R_1 , E was introduced in the form of a Latin Square and similarly for R_2 and R_3 except the rows were permuted. (1) The same design is used for both L_1 and L_2 since the rounds are fired in pairs from the two launchers. The design is given in table 1. This is referred to as a 3^4 confounded fractional factorial in blocks of 27 units for which 1/3 of a replicate is given, and it requires a total of 54 rounds plus spares, a figure which is realistic in this particular case. A design similar to this one is plan No. 6A.18, P 290, Cochran and Cox. (2)

		L_1 and L_2		
		W_1	W_2	W_3
R_1	T_1	E_1	E_3	E_2
	T_2	E_3	E_2	E_1
	T_3	E_2	E_1	E_3
R_2	T_1	E_2	E_1	E_3
	T_2	E_1	E_3	E_2
	T_3	E_3	E_2	E_1
R_3	T_1	E_3	E_2	E_1
	T_2	E_2	E_1	E_3
	T_3	E_1	E_3	E_2

Table 1. A 3^4 , 1/3 replicate, confounded Fractional Factorial in Blocks of 27 units.

(1) It would have been possible to use Graeco-Latin Squares, and thus test some other treatment at 3 levels, providing certain assumptions could be made about this new treatment.

(2) "Experimental Designs," Second Edition, W. G. Cochran and G. M. Cox, 1957, John Wiley and Sons, Inc., New York

The precautions to be observed in this design are: (1) The RT, TW, and RW interactions are aliased with the E interactions; (2) The RWT interaction is aliased with the E Main Effect; and (3) It is doubtful if a suitable comparison can be made between the variability in performance due to the missile and the variability due to the missile plus metric conditions.

These difficulties were all studied carefully and it was agreed that the design in figure one was appropriate as far as our problem is concerned. It was felt that the likelihood of E interacting with R, W or T is negligible, and that the RWT interaction is probably negligible, also. The remainder of the discussion is based upon the hypothesis that the above assumptions are all correct.

Table 2 shows two analysis of variance tables. One is for the fractional factorial design described in table one and the other is for a 3^4 factorial with 2 replications.

Sources of Variance	Fractional Factorial	Factorial degrees of Freedom
W	2	2
T	2	2
R	2	2
E	(WRT) 2	2
WR	4	4
WT	4	4
RT	4	4
WE		4
RE		4
TE		4
WRT		8
WRE		8
TRE		8
WTE		8
WRTE		16
Interactions associated with E	6	
Error	27	81
Total	53	161
Sample Size	54	162

Table 2. Degrees of freedom associated with a 3^4 factorial with 2 replications and a 3^4 fractional factorial showing 1/3 of a replicate (repeated).

From table 2 it is clear that, if the assumption that the E interactions and the WRT interactions are negligible, we are buying almost as much from the fractional factorial with 54 rounds as from the complete factorial with 162 rounds. It is true the degrees of freedom for error are 27 as compared to 81 for the complete factorial, but the increase in power which results from using 81 rather than 27 degrees of freedom is usually unimportant.

The computation of sums of squares is extremely simple for this design. The procedure is the same as for a complete factorial design except the error term must be computed by taking one half the sum of squares of the difference between pairs. Then the term which we describe as interactions associated with E is obtained by subtraction.

This design will be illustrated by the data given in table 3 and the analysis of variance is given in table 4. The data of table 3 is fictitious, such data was chosen for two reasons: (1) to keep this presentation unclassified; and (2) the actual study is not far enough along to provide actual data. The data listed in table 3 is radial miss distance which has been transformed in order that an analysis of variance would be appropriate. The important thing to note is that the smaller the value the better the weapon has performed.

		L_1		L_2			
		W_1	W_2	W_3	W_1	W_2	W_3
R_1	T_1	36 E ₁	38 E ₃	31 E ₂	36 E ₁	46 E ₃	15 E ₂
	T_2	14 E ₃	15 E ₂	20 E ₁	20 E ₃	24 E ₂	35 E ₁
	T_3	32 E ₂	33 E ₁	25 E ₃	29 E ₂	26 E ₁	24 E ₃
R_2	T_1	21 E ₂	29 E ₁	32 E ₃	20 E ₂	26 E ₁	22 E ₃
	T_2	8 E ₁	6 E ₃	16 E ₂	4 E ₁	13 E ₃	11 E ₂
	T_3	28 E ₃	14 E ₂	13 E ₁	22 E ₃	12 E ₂	22 E ₁
R_3	T_1	45 E ₃	33 E ₂	32 E ₁	42 E ₃	37 E ₂	27 E ₁
	T_2	16 E ₂	33 E ₁	32 E ₃	15 E ₂	34 E ₁	32 E ₃
	T_3	33 E ₁	45 E ₃	25 E ₂	40 E ₁	42 E ₃	35 E ₂

Table 3. Results of a firing test using a 3^4 fractional factorial and two launchers.

Sources of Variation	S.S.	D of F	M.S.	F Ratio
W	100	2	50	2.17
T	1409	2	704	30.61 **
R	2223	2	1111	48.30 **
E (WRT)	466	2	233	10.13 **
WT	695	4	174	7.56 **
WR	167	4	42	1.83
TR	109	4	27	1.17
Due to E Interaction	156	6	26	1.13
Error	626	27	23	
Total	5951	53		

Table 4. Analysis of Variance for data in Table 3. (** Indicates significance at the 1 percent level).

Table 4 indicates that all main effects except for warhead type, have a significant influence upon missile accuracy, with range having the greatest influence of any. WT is the only significant second order interaction. At this time, it is desirable to investigate the mean square which is attributed to the various E interactions. If this were not small, one might have some doubt about whether the assumption concerning the E interactions being negligible was really sound. If we were absolutely certain that E interactions are not possible, this mean square term could be used for another purpose. If there had only been one replication (27 rounds) it would have been necessary to use this term as the error term. But since the rounds were fired in pairs we may now consider that the mean square attributed to E interactions is an estimate of the variability which is due both to the missile and to the day to day variation, while the mean square for error term is an estimate of the missile variability after stripping out the day by day variation. By comparison it is seen that the difference between the two estimates is negligible from whence it might be implied that day by day variation appears to be well under control.

Returning to the results of table 4, the overall mean is found to be 26.22 and the mean values for the various levels for the main effects are given by table 5 below.

Treatments \ Levels	1	2	3
W	25.61	28.11	24.94
T	31.55	19.33	27.78
R	27.72	17.72	33.22
E	27.05	22.28	29.33

Table 5. Mean Values for Each Level of W, T, R and E.

From table 5, it appears that warhead No. 3 results in the smallest while warhead No. 2 causes the greatest miss distance. However, the effects of warheads are not significant and there is no reason to believe that one warhead will cause a larger miss distance than another. Ambient temperature results in the smallest while low temperatures cause the greatest mean miss distance. Medium ranges have the smallest, while long ranges have the greatest mean miss distance. Finally it may be seen that launchers emplaced on level ground will result in the smallest mean miss distance while an emplacement on the fore side of a hill will result in the greatest mean miss distance. It would be very much in order to study the WT interactions, but this will be omitted, largely because the main effect W is not significant.

Conclusions: When testing missile systems it is usually the case that many levels of treatments must be studied from the data obtained from a limited sample size. Fractional factorials are frequently a very useful tool for designing such tests. There are many pitfalls to watch for when using fractional factorials, but many times this technique appears to give results nearly as good as those obtained from a complete factorial and by using a much smaller sample size.

THE DESIGN & REDESIGN OF AN EXPERIMENT

C. W. Mullis
Integrated Range Mission, White Sands Missile Range

INTRODUCTION. In May 1958, White Sands Missile Range undertook an evaluation which was particularly amenable to optimization of the experiment. An experiment was designed and data collection began. Unanticipated field problems required approximately four times the effort predicted in order to fill enough points in the design matrix to permit reasonable analyses. The end result was that an experiment expected to take six months, extended over a period of sixteen months and is just now nearing completion. In view of these circumstances, this paper might better be titled by using the often quoted expression "The best laid plans of mice and men often go astray."

THE PROBLEM. White Sands has been employing cinetheodolites manufactured by Askania-Werke A. G. since the range was established in 1945. Today there are approximately sixty instruments in regular use. In recent years a new instrument has appeared on the market which is purported to represent the state-of-the-art in cinetheodolite type instruments. This instrument, manufactured by Contraves A. G., Zurich, Switzerland is known as the Contraves EOTS. During 1958, the J. W. Fecker Division of the American Optical Company became the United States distributor and was desirous of obtaining information on the comparable performance of their new product and existing equipment in the field. They also wished to demonstrate that the Contraves was dynamically accurate to better than 5 sec of arc (one part in 250,000). Due to the potential market, the varied nature of the missile firing workload (approximately ten missiles of varying types fired each work day) and a unique capability to install instruments in dual installations side by side, Fecker proposed to White Sands that evaluation of comparison of the Contraves instrument to the existing Askania instruments be performed. Previously the only available comparison was Contraves at Eglin to Askania at White Sands or similar cases which left much room to challenge validity. White Sands welcomed the opportunity to settle the argument, and obtain first hand information on the new instrument.

CONSIDERATIONS. By utilizing the dual installations, operating on the same missions from the same timing distribution and control network, reading and reducing the data in the same plant, and comparing data taken at the same instant in time, it was felt that a valid comparison of the instruments in question would be forth-coming. The design of the experiment then centered around the features of the instruments and their deployment.

The first question was "How many to use." Fecker proposed to furnish two Contraves. It is a well known fact that although two station triangulation meets the mathematical requirement for a solution, the accuracy of the final computed data increases as the number of stations is increased. Consequently use of three instruments for each system was agreed upon.

The next question was "where to put them." Figure 1 shows the deployment of the existing Askanias. The nature of the missions can be summarized as follows: Ballistic type missiles are fired from the Small Missile Range near "N" Station, and from the launching strip extending

eastward from the Army Missile Test Center. Those missiles may impact in the 30, 50, 70 or 90 mile impact areas. Air to air missions are conducted above the four major impact areas. Surface to air missiles are launched from the launching strip and may intercept anywhere from a few miles north of the launch point to the northern boundary of the range. It was desired to locate the instruments such that as many of the various types of missions as possible could be included in the experiment.

Another factor influencing the placement of the instruments was intersection geometry. Needless to say, the accuracy of final data is quite sensitive to the angle of intersection of the lines of sight from the instruments to the target. Tracking capabilities should be taxed to the maximum but the capability of the instruments was not to be exceeded. An arbitration of these factors resulted in the deployment shown in Figure 2. This deployment had good geometry for at least two of the instruments on missiles launched from any launch area. Data could be obtained on all but air to air missions. The instruments experience varying modes of tracking severity depending on the launch point. All have wire lines for communication and timing distribution and they are all "close in."

The next consideration was the physical installation of the instruments. Figure 3 shows an actual installation with the Askania on the left and the Contraves on the right. The Askania is mounted on a hydraulic hoist which elevates it through a hole in the roof of the monolithic concrete building. The kinematic platform on top of the hoist is then rotated, locked into the building roof, and the hoist lowered leaving the instrument "sitting" on the building. Obviously, the Contraves would not pass through the hole. Therefore, a one inch thick steel plate was fabricated to cover the hole and the Contraves mounted to the plate. In essence then both instruments were using the same pedestal, namely the building.

Since all film was to be automatically processed by the high speed continuous processing machines, it was felt that no special consideration should be given this phase of the test.

When the film reached the data reduction portion of the system, other questions arose. For instance, what about human error in film measurement? In order to minimize this it was decided to read each target board frame five times and each data from three times and use the averages thus obtained as the reading.

Registration in the reader was another questionable area. Since the Contraves mechanism is pin registered and the Askania mechanism is not, it was decided to check the registration for both by resetting on each frame. In the Contraves, the frame is rotated through the elevation angle. The center of the recorded frame then moves according to the eccentricity of the elevation axis. Therefore the center of the frame was rechecked for each frame. For the Askania, the fiducials were checked as usual, each frame.

CALIBRATIONS. Calibrations are the measurements made on the instrument to determine the bias errors which exist in its various parts. A correction thus obtained is applied in the data reduction process so that biases do not

appear as errors in the end result data. These measurements may occur at very infrequent intervals or may be made each time the instrument is operated. Figure 4 shows the calibrations performed on the two types of instruments for this test. Lens sag does not exist for the Contraves and circle eccentricities are eliminated by diametrically opposed scale readings. However, these were measured and accounted for in the Askania. Although lens distortion was measured for both instruments and for the projection lenses of the film readers, it was not necessary to use it since the tracking was good. That is to say the target was always near enough the center of the frame that distortion could be neglected. To insure that no appreciable error was contributed by the film reading machines, the measuring cross hair digitizers were calibrated. The measurements thus obtained were used to correct the final data which were used for computation.

HYPOTHESES. As previously stated, Fecker desired to prove that the Contraves was accurate to 5 sec of arc or better. The Government wished to compare the performance of the two systems. An attempt was therefore made to design an experiment which would achieve both goals. The null hypotheses were stated as follows:

TEST 1

Null Hypothesis

The Contraves EOTS cinetheodolite does not exhibit smaller random errors than the Askania Kth 53 cinetheodolite.

Alternate Hypothesis 1

The Contraves EOTS cinetheodolite exhibits random errors less than 1/2 the value of the random errors of the Askania Kth 53 cinetheodolite.

Alternate Hypothesis 2

The Contraves EOTS cinetheodolite exhibits random errors less than 1/4 the value of the random errors of the Askania Kth 53 cinetheodolite.

TEST 2

Null Hypothesis

The Contraves EOTS cinetheodolite system (three stations) does not exhibit smaller random errors than the Askania Kth 53 cinetheodolite system (three stations).

Alternate Hypothesis 1

The Contraves EOTS cinetheodolite system exhibits random errors less than 1/2 the value of the random errors of the Askania Kth 53 cinetheodolite system.

Alternate Hypothesis 2

The Contraves EOTS cinetheodolite system exhibits random errors less than 1/4 the value of the random errors of the Askania Kth 53 cine-theodolite system.

These hypotheses are based on the Askania as a standard. By testing the Askania random errors against, the 5 second figure, it was felt than an absolute value for the accuracy of the two systems would be obtained.

To test these hypotheses, two methods were to be employed. For Test 1 the Variate Differences 1 techniques was to be used. For Test 2 a statistical analysis of the residuals from a standard Davis solution 2, 3 was to be used. Test 1 would give a comparison of instrument precision, where Test 2 would give system accuracy including biases. Since, by the nature of the techniques Test 2 required considerably more data than Test 1, Test 2 governed the determination of the data to be collected. The experiment then resolved itself into a study of two variables, one containing six parameters, the other two parameters to be investigated on three levels. A further division occurs when it is considered that each system containing three instruments is treated as a single variable. The design matrix is shown in Figure 5.

DATA COLLECTION. From the design matrix, it was planned to collect data on eighteen missions. This was to be accomplished in one month to six weeks. In actuality data were collected over a period of three months on eighty seven missions. From this we were able to select ten missions on which all six instruments (three Askanias and three Contraves) functioned properly and collected sufficient data for analysis. This is not to say that the instruments are that unreliable. For proper application of the variate differences analysis it is necessary to have 100 consecutive points. Loss of a single point due to a condition of mistrack eliminated numerous records from consideration. Loss of points due to the missile passing behind clouds had the same effect. Another difficulty was the fact that although 100 consecutive points may have existed for the various instruments, they did not overlap enough in time to allow a valid comparison.

FIELD PROBLEMS. As data collection progressed operational observations lead to questioning of the assumptions. Primary among these was the rigidity of the building. A rather long term effect was observed in the tilting of the building due to differential heating. On one particular day this amounted to approximately 40 seconds of arc over a period of 6 hours. This was not considered to be significant in terms of the effect during a particular mission since the mislevel of the instrument is read before and after each mission. However, one observer noted that there was a sudden shift of approximately 10 seconds of arc which occurred in a period of a few minutes. The reason for this shift could not be explained.

As a result of the uncertainty in the stability of the support structure, a method was devised whereby the movement of the building could be measured during a mission. This consisted of checking the movement of the

instrument base against a pendulum mirror which was referenced to earth's gravity. [4] Only a few measurements were made but the results showed that no significant movement of the building occurred during the mission.

Another assumption which is made in the usual operation of the system is that the target boards which are used to determine the orientation and mislevel of the instrument are stable. The azimuth stability of the target boards was noted in several spot checks using the Contraves film information. In most cases, the correspondence was within 3.6 seconds of arc. The elevation position of the target boards was not checked as accurately as this since turbulence research at White Sands has indicated that there is much less vertical deterioration than horizontal deterioration due to turbulence. However, refraction in the vertical direction may be significant.

Of the three stations used in the field, one station, "N", did not "fit in" with the other two for either the Contraves or the Askania. This means that the line of sight from that station did not pass through the intersection of the lines of sight from the other two stations within the prescribed limits of the data reduction procedure. Thus far, no explanation has been found for the large discrepancy observed. Surveys have been rechecked. Timing distribution delays were measured to be less than 100 micro-seconds. All known possibilities have been exhaustively checked with no answer being found.

ANALYSIS. Due to their simplicity, the variate differences calculations were started first. Although they have not yet been completed, a sample tabulation of some of the data from three missions is given in Figure 6. From this tabulation one can see the nature of the information which is being obtained. For instance one can make instrument to instrument comparisons in azimuth and/or elevation; compare the azimuth or elevation performance of like instruments; compare azimuth to elevation performance of a single instrument, etc. By performing a statistical analysis of the data for several missions, generalized statements relative to the characteristics of the instruments can be made. Once the data have been calculated for all the missions these analyses will be completed.

The analysis for Test 2 does not present so bright a picture. With the exception of one mission, it has been impossible to get the data from "N" station to work in a three station trajectory computation. Forced solutions have had residuals far beyond the rejection limits. On one mission the data would work in a two station solution with either of the other stations but three stations would not run. To date, this problem is still under consideration.

CONCLUSIONS. Time does not permit an excursion into the details of system improvements which have been effected as a result of the deficiencies discovered during this evaluation. One significant course of action as far as the experiment is concerned has been decided upon. Fecker is planning to take the raw data from some of the missions, read the data on their own equipment and perform an independent analysis. All the procedures will be similar except for the computer program, which will account for an assumption made in the Davis solution to linearize the equations.

White Sands Missile Range has proceeded with two station solutions using "C" and "T" stations and is preparing a statistical analysis of these data for Test II.

In summary, it may seem strange that this discussion has been presented to this group. The intent was to emphasize the physical problems encountered and the requirement for flexibility in the statistical design of experiments of this nature.

REFERENCES

1. M. G. Kendall, "The Advanced Theory of Statistics," Vol. II, P387-394, Hafner & Co., New York.
2. R. C. Davis, "Techniques for the Statistical Analysis of Cinetheodolite Data," NAVORD Rpt. 1299, NOTS 369, Naval Ordnance Test Station, China Lake, California, 22 March 1951.
3. F. P. Apostalas and J. B. Gose, "Askania Cinetheodolite Procedure Technical Memorandum 446, White Sands Missile Range, New Mexico, August 1957.
4. E. C. Schluter, "Cinetheodolite Dynamic Accuracy," Instrumentation Services Interim Report: April 1959, Measurements Division, Integrated Range Mission, White Sands Missile Range, New Mexico.

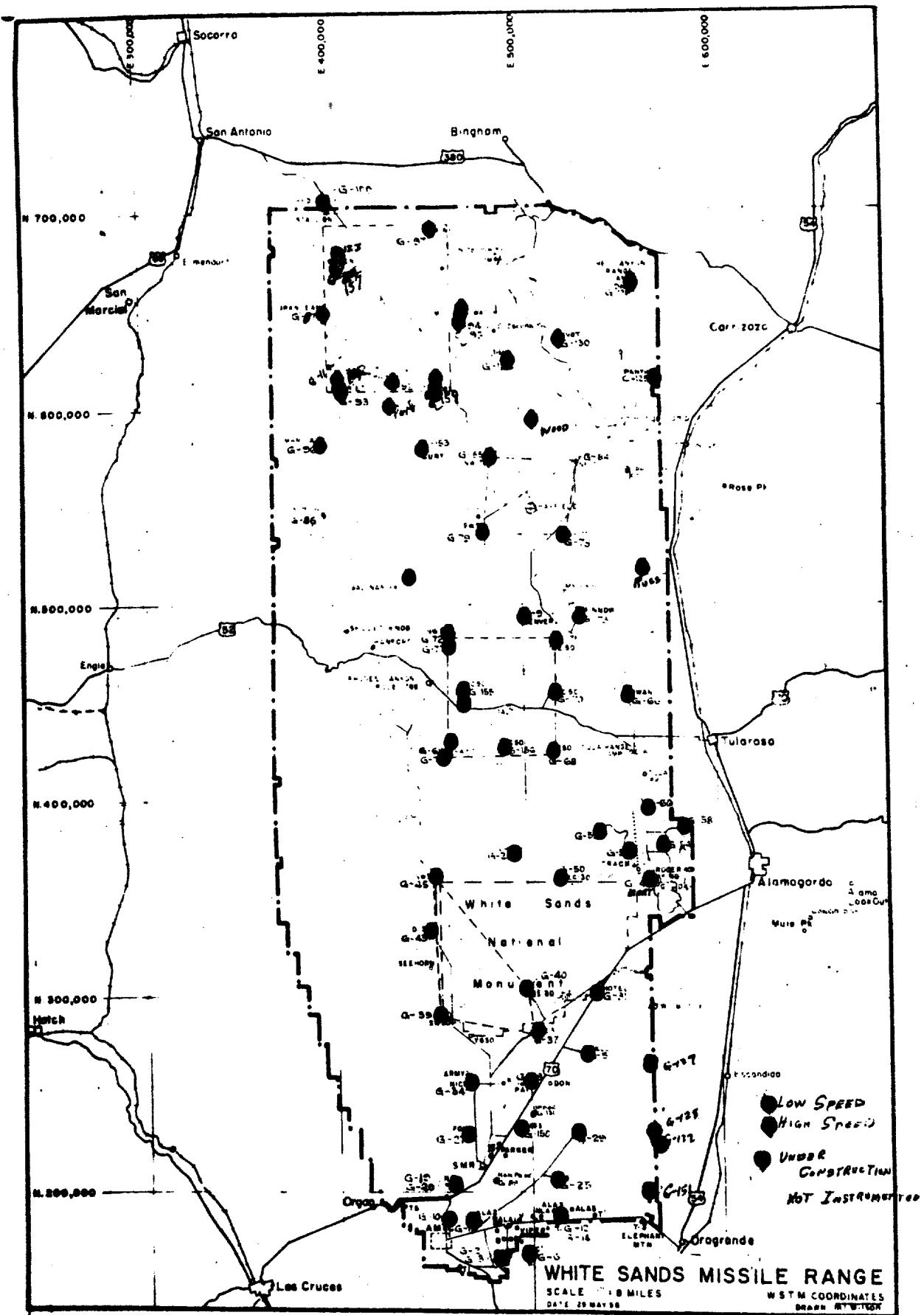


FIGURE 1. CINETHEODOLITE INSTRUMENTATION

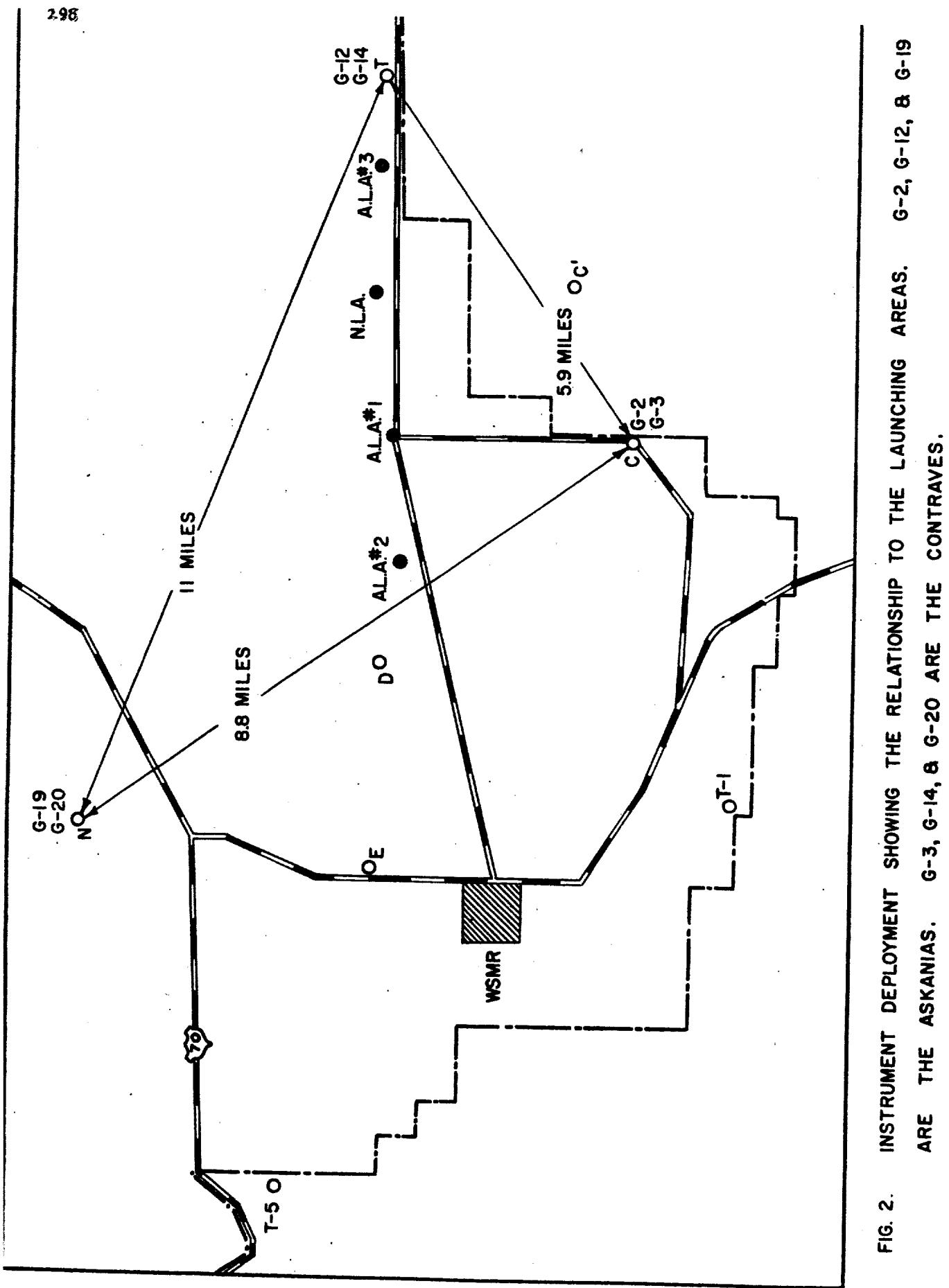
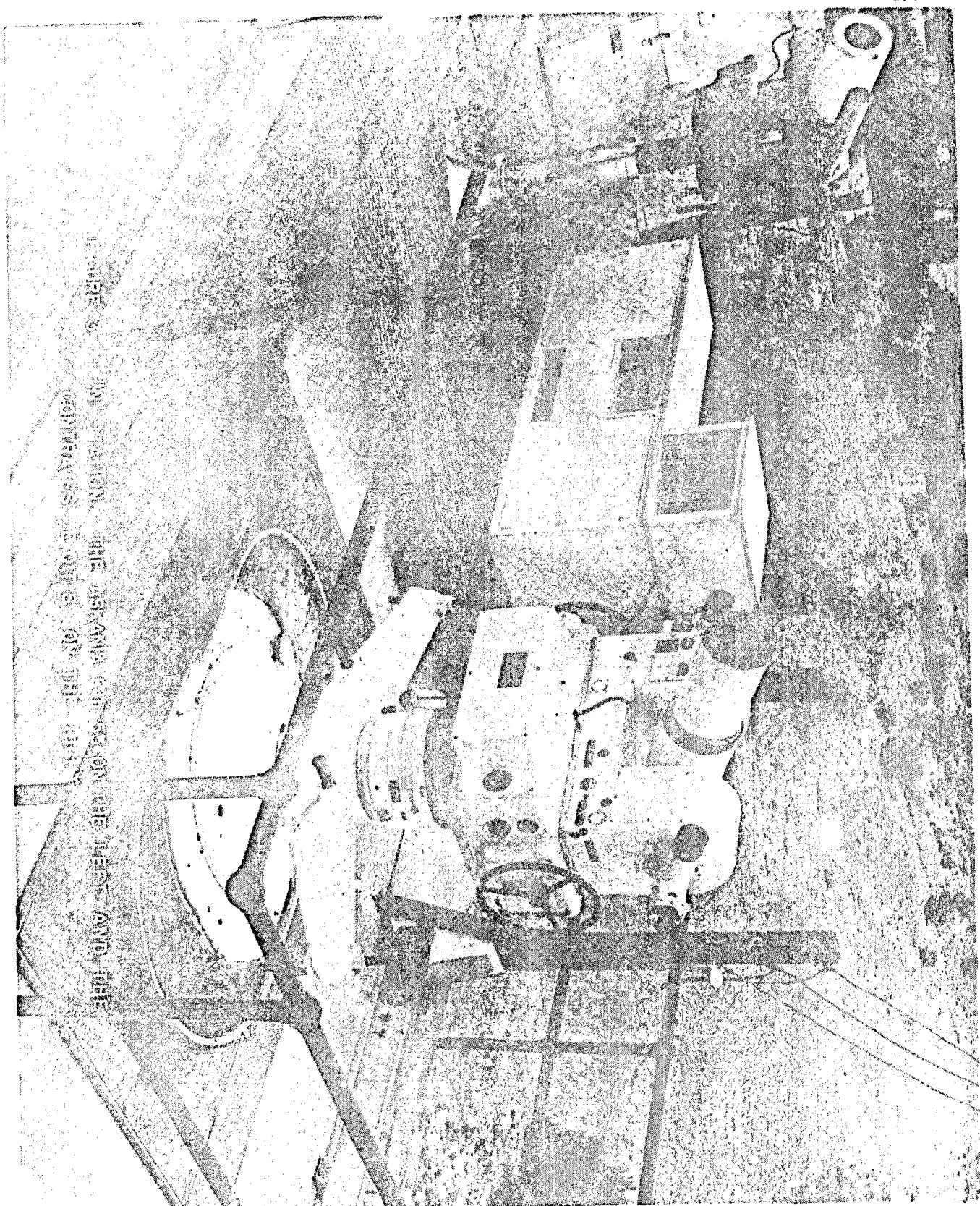


FIG. 2. INSTRUMENT DEPLOYMENT SHOWING THE RELATIONSHIP TO THE LAUNCHING AREAS. G-2, G-12, & G-19 ARE THE ASKANIAS. G-3, G-14, & G-20 ARE THE CONTRAVES.

299



	ASKANIA	CONTRAVES
TRUNION AXIS ERROR	X	X
COLLIMATION ERROR	X	X
MIS-LEVEL	X	X
LENS DISTORTION	X	X
LENS SAG	X	
CIRCLE ECCENTRICITY	X	

FIGURE 4 CALIBRATIONS

VARIABLE	PARAMETER	LEVEL		
		HIGH QE BALLISTIC	LOW QE BALLISTIC	MANOUVRED
ASKANIA	5/SEC B&W	X	X	X
	5/SEC COLOR	X	X	X
	5/SEC B&W	X	X	X
	5/SEC COLOR	X	X	X
CONTRAVES	10/SEC B&W	X	X	X
	10/SEC COLOR			
	20/SEC B&W	X	X	X
	20/SEC COLOR			

FIGURE 5 EXPERIMENT MATRIX

NOTE: EACH POINT IN THE MATRIX REPRESENTS
FROM 1500 TO 6000 DATA POINTS.

FIGURE 6. TABULATION OF VARIATE DIFFERENCES. G-2, G-12, & G-19 ARE THE ASKANIAS. G-3, G-14, & G-20 ARE THE CONTRAVES.

ESTIMATING THE PARAMETERS OF A MODIFIED POISSON DISTRIBUTION*

A. Clifford Cohen, Jr.
University of Georgia

Errors in observing and reporting sample data often complicate the problem of estimating parameters of the distribution being sampled. If neglected, such errors may lead to seriously biased estimates. There exists a large general class of such estimation problems involving numerous different distributions, different types and varying degrees of observational errors. This paper is limited, however, to maximum likelihood estimation in a Poisson distribution which has been modified to the extent that a proportion θ of the ones are reported as being zeros. An inspector who sometimes fails to see or at least fails to report items containing only a single Poisson distributed defect, while correctly observing and reporting results of inspecting items containing two or more defects, produces sample data of the type under consideration. Estimators are derived both for the Poisson parameter and for θ . Asymptotic variances and covariances are derived and an illustrative example is included.

1. INTRODUCTION. In observing a Poisson distributed random variable, it sometimes happens that values of one are erroneously observed or at least reported as being zeros. For example, in determining the number of defects per unit or item examined, an inspector may err by reporting units which actually contain a single defect as being perfect or free of defects. Of course there is also a similar possibility of erroneous observation when the actual number of defects per unit is in excess of one, but here we are concerned only with the case in which some though not necessarily all ones are reported as zeros.

Suppose the number of defects actually present per unit is a Poisson distributed random variable with parameter λ , and that the probability of misclassifying an item containing one defect by reporting it as containing zero defects is θ . The probability function of the random variable x , the observed (reported) number of defects per item, may then be written as

$$(1) \quad p(x; \lambda, \theta) = \begin{cases} e^{-\lambda}(1 + \theta\lambda), & x = 0, \\ (1 - \theta)\lambda e^{-\lambda}, & x = 1, \\ e^{-\lambda}\lambda^x/x!, & x = 2, 3, \dots, \end{cases}$$

where $\lambda > 0$ and $0 \leq \theta \leq 1$.

In an abstract sense, (1) may simply be considered as the probability function of a two parameter modified Poisson distribution, and in this

* This paper was originally published in the Journal of the American Statistical Association (March 1960). Permission to reproduce it here is greatly appreciated by the editors.

paper we are concerned with maximum likelihood estimation of its two parameters λ and θ . The problem under consideration here is a special case of a more general class of estimation problems involving erroneous sample observation which has been encountered, for example, by Neyman and Scott [6] in connection with counting galaxy images on photographic plates and by Toulouse [9] in connection with attribute sampling. It is closely related to the estimation of the Poisson parameter from truncated and censored samples, a problem which received attention from David and Johnson [4], Moore [5], Plackett [7], Rider [8], this writer [2], [3], and various others.

2. DERIVATION OF ESTIMATORS. Consider a sample consisting of N observations of the random variable x with probability function (1) in which n_0 designates the number of zero observations and n_1 the number of ones. The likelihood function for such a sample is

$$P(x_1, \dots, x_N; \lambda, \theta) = [e^{-\lambda}(1 + \theta\lambda)]^{n_0} [(1 - \theta)\lambda e^{-\lambda}]^{n_1} II^* e^{-\lambda \sum_i x_i} / x_i!,$$

where II^* is the product over all x^* 's that are neither 0 nor 1. We write this result in simpler form as

$$(2) \quad P(x_1, \dots, x_N; \lambda, \theta) = e^{-N\lambda} (1 + \theta\lambda)^{n_0} (1 - \theta)^{n_1} \lambda^{\sum_i x_i} [II^* x_i!]^{-1}.$$

Taking logarithms of (2), differentiating with respect to λ and θ in turn, and equating to zero yields the estimating equations

$$(3) \quad \begin{aligned} \partial L / \partial \lambda &= -N + n_0 \theta / (1 + \theta\lambda) + \sum_1^N x_i / \lambda = 0, \\ \partial L / \partial \theta &= n_0 \lambda (1 + \theta\lambda) - n_1 / (1 - \theta) = 0 \end{aligned}$$

where L is written for $\ln P$.

The required M.L. estimators $\hat{\lambda}$ and $\hat{\theta}$, when they exist, will be found by simultaneously solving these two equations. We follow the customary notation of employing $(\hat{\cdot})$ in this paper to distinguish maximum likelihood estimators from the parameters estimated.

To facilitate their solution, the above equations are reduced to

$$(4) \quad \begin{aligned} \lambda^2 - (\bar{x} - 1 + n_0/N) \lambda - (\bar{x} - n_1/N) &= 0, \\ \theta &= [n_0 - n_1/\lambda] / (n_0 + n_1), \end{aligned}$$

where \bar{x} is the sample mean ($\bar{x} = \sum_1^N x_i / N$).

The first equation of (4) results from eliminating θ between the two equations of (3), while the second results from solving the second equation of (3) for θ . A similar pair of equations can be obtained by first eliminating λ between the two equations of (3) and thus obtaining an equation which is quadratic in θ . Estimates are easier to calculate, however, using the results given above in (4).

We note that $(\bar{x} - 1 + n_0/N) > 0$ and $(\bar{x} - n_1/N) > 0$ except when (i) all sample observations are zeros, or (ii) all observations are ones. With these two exceptions, the coefficients of the first equation of (4), which is quadratic of the form $g(\lambda) = 0$, thus exhibit one change of sign, and likewise the coefficients of $g(-\lambda)$ exhibit one change of sign. It then follows from Descartes' well known "rule of signs" that $g(\lambda) = 0$ has exactly one positive and one negative root. The positive root of this equation is the required estimator of λ , and on solving by means of the quadratic formula, we obtain

$$(5) \quad \hat{\lambda} = [(\bar{x} - 1 + n_0/N) + \sqrt{(\bar{x} - 1 + n_0/N)^2 + 4(\bar{x} - n_1/N)}]/2.$$

With $\hat{\lambda}$ thus determined, the second equation of (4) enables us to calculate

$$(6) \quad \hat{\theta} = (n_0 - n_1 \hat{\lambda})/(n_0 + n_1).$$

When $\theta = 0$, (1) becomes the ordinary Poisson probability function without modification, in which case the first equation of (3) yields the familiar estimator $\hat{\lambda} = \bar{x}$. We now turn our attention to three special types of samples, two of which were listed as exceptions in the preceding paragraph. Although samples of these types are unlikely to arise in practical applications envisioned for the results of this paper, they are of theoretical interest and are considered here for that reason.

Special type (i). All observations are zeros; $n_1=0$, $n_0=N$, and $\bar{x}=0$. The likelihood equation (2) for a sample of this type becomes

$$P = e^{-N\lambda} (1 + \theta\lambda)^N.$$

On taking logarithms, differentiating with respect to λ and θ in turn and equating to zero, estimating equations corresponding to (3) become

$$-N + N\theta/(1 + \theta\lambda) = 0,$$

$$N\lambda(1 + \theta\lambda) = 0.$$

Maximum likelihood estimates $\hat{\lambda}$ and $\hat{\theta}$ do not exist in this case, however, since the above estimating equations are simultaneously satisfied only when $\lambda = 0$ and $\theta = 1$, whereas $p(x; \lambda, \theta)$ is defined only for $\lambda > 0$.

Special type (ii). All observations are ones; $n_0=0$, $n_1=N$, and $\bar{x}=1$. Maximum likelihood estimates $\hat{\lambda}$ and $\hat{\theta}$ fail to exist in this case also since estimating equations (3) are not simultaneously satisfied by any pair of values of λ and θ for which $p(x; \lambda, \theta)$ is defined. Although the first equation of (3) with $n_0 = 0$ is satisfied when $\lambda = 1$, the second is only satisfied in the limit as $\theta \rightarrow \infty$, whereas $p(x; \lambda, \theta)$ is defined only for $0 \leq \theta \leq 1$.

Special type (iii). No zeros or ones are observed; $n_0=n_1=0$. In this case the likelihood equation (2) is independent of θ , which therefore cannot be estimated from available sample information. The Poisson parameter, however, is estimated by (5), which for a sample of this type, reduces to $\hat{\lambda} = \bar{x}$.

It is not difficult to construct other samples for which (5) and (6) fail to give acceptable estimates of λ and θ . However, when N is large such samples will be very improbable and their occurrence in practical applications should be interpreted as a suggestion that probability function (1) might not be applicable to the random variable actually observed.

3. SAMPLING ERRORS OF ESTIMATES. The asymptotic variance-covariance matrix of $(\hat{\lambda}, \hat{\theta})$ is obtained by inverting the information matrix whose elements are negatives of expected values of the second order derivatives of logarithms of the likelihood function.

The second partial derivatives of L follow from (3) as

$$(7) \quad \begin{aligned} \partial^2 L / \partial \lambda^2 &= -n_0 \theta^2 / (1 + \theta \lambda)^2 - N \bar{x} / \lambda^2, \\ \partial^2 L / \partial \theta^2 &= -n_0 \lambda^2 / (1 + \theta \lambda)^2 - n_1 / (1 - \theta)^2, \\ \partial^2 L / \partial \lambda \partial \theta &= \partial^2 L / \partial \theta \partial \lambda = n_0 / (1 + \theta \lambda)^2. \end{aligned}$$

Since $E(\bar{x}) = \lambda(1 - \theta e^{-\lambda})$, $E(n_0) = N e^{-\lambda}(1 + \theta \lambda)$, and $E(n_1) = N(1 - \theta)\lambda e^{-\lambda}$, where $E(\cdot)$ denotes expected value, elements of the information matrix follow from (7) as

$$(8) \quad \begin{aligned} E(-\partial^2 L / \partial \lambda^2)/N &= (1 + \theta \lambda - \theta e^{-\lambda})/\lambda(1 + \theta \lambda), \\ E(-\partial^2 L / \partial \theta^2)/N &= \lambda e^{-\lambda}(1 + \lambda)/(1 + \theta \lambda)(1 - \theta), \\ E(-\partial^2 L / \partial \lambda \partial \theta)/N &= E(-\partial^2 L / \partial \theta \partial \lambda)/N = -e^{-\lambda}/(1 + \theta \lambda). \end{aligned}$$

On inverting the information matrix, the asymptotic variances and covariance follow as

$$(9) \quad \begin{aligned} V(\hat{\lambda}) &\sim \lambda(1 + \lambda)/N(1 + \lambda - e^{-\lambda}), \\ V(\hat{\theta}) &\sim (1 + \theta \lambda - \theta e^{-\lambda})(1 - \theta)/N \lambda e^{-\lambda}(1 + \lambda - e^{-\lambda}), \\ \text{Cov}(\hat{\lambda}, \hat{\theta}) &\sim (1 - \theta)/N(1 + \lambda - e^{-\lambda}). \end{aligned}$$

The correlation coefficient between estimates $\hat{\lambda}$ and $\hat{\theta}$ follows as

$$(10) \quad \rho_{\hat{\lambda}, \hat{\theta}} = \text{Cov}(\hat{\lambda}, \hat{\theta}) / \sqrt{V(\hat{\lambda})V(\hat{\theta})} \sim \sqrt{(1 - \theta)e^{-\lambda}/(1 - \lambda)(1 + \theta \lambda - \theta e^{-\lambda})}.$$

The variances and covariance given in (9) and the correlation coefficient given in (10) are applicable in all cases where maximum likelihood estimators $\hat{\lambda}$ and $\hat{\theta}$ exist. Even with samples of special type (iii), $V(\hat{\lambda})$ as given in (9) is applicable. Since N , the total sample size, is fixed n_0 and n_1 are random variables and although they may assume the value zero in particular samples, their expected values as given in the preceding paragraph are in excess of zero. Of course $E(n_0) \rightarrow 0$ and $E(n_1) \rightarrow 0$ as $\lambda \rightarrow \infty$.

Furthermore, when λ is large $V(\hat{\lambda})$ as given by (9) differs but slightly from λ/N which applies when λ is estimated from a sample of size N from an ordinary Poisson distribution without modification.

4. AN ILLUSTRATIVE EXAMPLE. To illustrate the practical application of results of this paper, data from Bortkiewicz's [1] classical example on deaths from the kick of a horse in the Prussian Army have been suitably altered. The original data were collected from records of a certain group of ten Prussian Army Corps over the twenty year period 1875-1894. The study thus included 200 annual reports; that is, 200 observations of the random variable involved. For the purpose of this illustration it has been assumed that twenty of the records which should have shown one death each were in error by reporting no deaths. Both the original and the altered data for this example are given below.

Number Deaths per Army Corps per Year	Number Observations	
	Original Data	Altered Data
0	109	129
1	65	45
2	22	22
3	3	3
4	1	1
5	0	0

Summarizing the altered (misclassified) data, we have: $n_0 = 129$, $n_1 = 45$, $N = 200$, $\bar{x} = 102/200 = 0.51$, $n_0/N = 0.645$, $n_1/N = 0.225$, $(\bar{x} - 1 + n_0/N) = 0.155$, and $(\bar{x} - n_1/N) = 0.285$. On substituting these values into (5), we calculate

$$\hat{\lambda} = \left[0.155 + \sqrt{0.155^2 + 4(0.285)} \right] / 2 = 0.617$$

Subsequent substitution into (6) yields

$$\hat{\theta} = (129 - 45/0.617) / (129 + 45) = 0.322.$$

The estimate $\hat{\lambda} = 0.617$, obtained above is to be compared with 0.610 which follows from the original unaltered data. The estimate $\hat{\theta} = 0.322$ is to be compared with $20/65 = 0.308$, which is the proportion of ones that were misclassified in the process of altering the original data for this illustration.

With λ and 0 replaced by their estimates $\hat{\lambda}$ and $\hat{\theta}$, (9) and (10) enable us to calculate

$$\begin{aligned} V(\hat{\lambda}) &\doteq 0.0046, \\ V(\hat{\theta}) &\doteq 0.0097, \\ \text{Cov}(\hat{\lambda}, \hat{\theta}) &\doteq 0.0031, \end{aligned}$$

$$\rho_{\hat{\lambda}, \hat{\theta}} \doteq 0.47.$$

$V(\lambda) \doteq 0.0046$ as calculated above for $\hat{\lambda}$ based on the altered data is to be compared with $V(\hat{\lambda}) \sim \lambda/N \doteq 0.610/200 = 0.00305$ for $\hat{\lambda}$ based on the complete (unaltered) sample.

REFERENCES

- 1 Bortkiewicz, L. von, Das Gesetz der Kleinen Zahlen, Leipzig: Teubner, 1898.
- 2 Cohen, A. C., Jr., "Estimation of the Poisson parameter from truncated samples and from censored samples," Journal of the American Statistical Association, 49 (1954), 158-68.
- 3 Cohen, A. C., Jr., "Estimating the parameter in a conditional Poisson distribution," In press for publication in Biometrics.
- 4 David, F. N. and Johnson, N. L., "The truncated Poisson," Biometrics, 8 (1952), 275-85.
- 5 Moore, P. G., "The estimation of the Poisson parameter from a truncated distribution," Biometrika, 39 (1952), 247-51.
- 6 Neyman, Jerzy and Scott, Elizabeth L., "Large scale organization of the distribution of galaxies," Handbuch der Physik, 53 (1959), 416-44.
- 7 Plackett, R. L., "The truncated Poisson distribution," Biometrics, 9 (1953), 485-88.
- 8 Rider, Paul R., "Truncated Poisson distributions," Journal of the American Statistical Association, 48 (1953), 826-30.
- 9 Toulouse, Julian H., "Psychological bias in attribute sampling," Industrial Quality Control, 14 (June 1958), 1-8.

THE DETECTION OF GUESS RESPONSES IN THE RATING
OF STATEMENTS BY THE METHOD OF SUCCESSIVE CATEGORIES

Lee E. Paul and Howard W. Hembree
QM Research and Engineering Field Evaluation Agency

This problem arose in the process of constructing equal interval rating scales for the clothing characteristics of fit, comfort, protection, durability and over-all acceptability. Guilford's method of successive categories was used to determine the scale values of a number of descriptive statements covering the full range of the dimensions studied. To accomplish this, the statements were administered to a sample similar to the population which will eventually use the scales and the subjects were asked to place each statement in one of 11 categories, category 1 the least favorable, category 11 the most favorable. A numerical value was computed for each of the 11 categories such that they tended to normalize the frequency distribution of the ratings for each statement. These category values were used to determine the mean and standard deviation of each statement. To construct a scale of any length, one simply selects statements such that the means are equidistant and the standard deviations are low.

Past experience in the administration of questionnaires to enlisted personnel reveals some small proportion who are not highly motivated in pursuing a task, the goal of which seems rather remote. Some of these men might be conservatively described as indifferent. However, since they have been told to rate these statements they must comply one way or another. It then becomes necessary to identify those respondents who did not rate the statements according to the instructions, either because they did not understand the task the English language or because of a lack of motivation. There are two ad hoc methods for detecting these non-conformists. One is simply to look at their responses, (see Table 1 at end of this article), with an eye to detecting patterns, that is, some mechanical scheme for responding that the subject feels will go undetected before he leaves the test session. Subject number 25 chose a rather unimaginative method. To cope with the more complex "guessers" another method has been used which consists of taking pairs of statements, one of each pair obviously favorable, another obviously unfavorable and looking for reversals in the ratings. With a large number of respondents, this can immediately be seen to be rather tedious and time consuming. The most important shortcoming of these methods, however, is that they are not objective in that there is no standardized procedure for their application nor do they provide any information on a cutting point, i.e., some score that indicates the respondent doesn't belong in a normal conscientious population.

The problem, then, is to identify two populations, one which understands the instructions, the English language and is motivated enough to make an honest effort, the other having shortcomings on one or more of these characteristics.

The solution proposed here is quite simple. A number of criterion statements were selected such that 50% or more of the total sample rated each statement in the two most extreme categories: 1 and 2, or 10 and 11,

(see Table 2). This is, of course, somewhat arbitrary. The method further defines a "guess" as a rating of 6 or less on a favorable criterion statement, 6 or more on an unfavorable one.

The criterion statements shown in Table 1 are 1, 5, 7, 10, 12, 15, 16, 22, only 8 meeting the requirements in this case. Now one simply descends the criterion columns and circles any response of 6 or less or 6 or more, depending on whether the majority finds statements are favorable or unfavorable (6 was included in both to catch the guesser whose system, if he has one, includes a lot of 6's). Simply counting the circles in a row provides a guess score for each subject.

It was noted earlier that each respondent rated two sets of statements. The rating of statements was carried out in 5 sessions as shown in Table 3.

Table 3

Session	Characteristics	Criterion Statements	R	N
1	Fit	10	.78	56
	Over-all Acceptability	8		
2	Durability	10	.929	61
	Protection	9		
3	Comfort	10	.668	51
	Protection	9		
4	Fit	10	.865	67
	Acceptability	8		
5	Durability	10	.534	59
	Comfort	10		
				294

Ten of the "fit" statements met the conditions for criterion statements, 8 for "over-all acceptability," 10 for durability," 9 for "protection" and 10 for "comfort." N is the number of subjects in each session, while R is the product-moment correlation of the guess scores obtained from the two sets of criterion statements. It seems apparent that the respondents were fairly consistent throughout the rating and that the criterion statements from the different characteristics were measuring the same thing with the possible exception of those from "comfort."

Figure 1 represents the frequency distribution of guess scores for all five characteristics with each individual represented twice, once for each characteristic rated. This distribution is a bimodal J curve with the primary mode at 0 and a smaller one at 4. Each man rated an average of 9.39 criterion statements and had 6 chances in 11 of getting "caught" at each guess. Thus, the mean for all guessers should be $9.39 \times 6/11$ or 5.11.

In order to get better separation between the guessers at the rest of the population, the guesses of each individual on the two characteristics he rated were combined and the frequency distribution showed in Figure 2. The two populations already suggested now seem evident. One population, the larger, has a J shaped distribution, which brings to mind Allport's J curve of social conformity. One explanation of this distribution is that it is a normal distribution that is quite insensitive at one end. In this case it seems probable that not all those with 0 scores were equally able or conscientious.

The other distribution, which consists of "guessers," is apparently normal. This distribution is based on an average of 18.78 criterion statements for each subject and should have a mean of 10.22 or $18.78 \times 6/11$. A binomial expansion indicates the standard deviation of this distribution should be approximately 2.2. $[\sqrt{NPQ} = (18.78 \times 6/11 \times 5/11)^{1/2}]$ Actually the S.D. will be slightly larger since some subjects were scored on 18 criterion statements, some 19, and some 20.

The dotted line in Figure 2 represents a smoothed extrapolation of the J curve. The inset distribution of guessers is the total curve less the smoothed J curve. While the inset curve doubtlessly includes some misclassified subjects, it is a reasonably close approximation of a symmetrical normal curve with a mean of about 10 and a standard deviation of a little over two.

Using a cutting point of 6 (mean - S.D. = $10.22 - 1.92 \times 2.2$) ought to eliminate about 97% of the guessers. This figure is very close to the intersection of the two curves and so approximates a maximum likelihood ratio.

As a somewhat independent evaluation of the guess scores, some product-moment correlation coefficients were computed between the median rating of the 32 statements for "fit" and the ratings of a number of individuals. These respondents were selected so as to include different guess scores. The correlations were plotted against the guess scores, and the results are shown in Figure 3. While it is obvious that people with high guess scores do not agree with the majority as expressed by median ratings, the number of near 0 and negative correlations reveal that many respondents actually guessed, in the literal sense, throughout the task.

The application of this procedure led to the elimination of the ratings of approximately 15% of the original sample. Since the ratings of most of the rejectees were apparently random with respect to the median value of the statements, it is believed that their inclusion in further computations would have led to a spuriously high estimate of the standard deviations of the statements.

STATEMENTS (O.A.I)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1	9	4	8	7	8	7	5	9	9	10	5	3	6	8	11	2	4	11	5	9	4	3	4	6	4	7	6	3	4	7	7	6
2	4	11	6	4	5	7	5	6	4	7	4	4	5	6	4	4	5	4	4	5	4	5	6	4	7	6	3	4	7	7	4	
3	8	6	5	6	7	5	6	1	6	6	7	5	3	2	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
4	11	4	9	5	1	5	1	11	4	11	5	1	5	1	5	1	5	1	5	1	5	1	5	1	5	1	5	1	5	1	5	
5	6	9	6	7	4	5	6	1	7	6	8	4	1	4	7	10	3	5	1	1	6	5	4	5	4	5	4	5	4	5	4	
6	7	11	6	10	3	1	6	2	1	7	2	1	5	1	6	10	11	4	1	1	6	5	4	5	4	5	4	5	4	5	4	
7	8	11	9	8	2	1	7	2	9	4	10	3	1	6	5	1	1	6	5	1	1	6	5	4	5	4	5	4	5	4	5	
8	9	11	5	9	2	1	5	1	5	9	10	7	11	1	1	6	7	11	1	1	6	5	4	5	4	5	4	5	4	5	4	
9	10	8	7	11	6	5	11	5	11	6	10	7	11	1	1	6	7	11	1	1	6	5	4	5	4	5	4	5	4	5		
10	11	11	4	8	6	1	7	6	1	7	6	10	11	1	1	6	7	11	1	1	6	5	4	5	4	5	4	5	4	5		
11	11	11	4	8	6	1	7	6	1	7	6	10	11	1	1	6	7	11	1	1	6	5	4	5	4	5	4	5	4	5		
12	8	6	7	6	7	6	1	10	4	11	2	11	2	1	1	6	7	11	1	1	6	5	4	5	4	5	4	5	4	5		
13	4	10	9	5	11	7	1	6	1	6	1	5	1	9	5	11	2	1	1	6	5	4	5	4	5	4	5	4	5	4	5	
14	11	2	8	3	1	6	7	1	6	1	5	1	8	1	6	5	10	3	2	1	6	5	4	5	4	5	4	5	4	5		
15	11	11	7	9	5	10	2	3	1	6	1	5	1	6	1	5	1	6	1	5	1	6	5	4	5	4	5	4	5			
16	11	11	4	10	4	2	1	6	1	5	1	9	5	11	2	1	1	6	5	1	1	6	5	4	5	4	5	4	5			
17	8	5	10	2	3	1	6	1	5	1	6	1	5	1	6	1	5	1	6	1	5	1	6	5	4	5	4	5	4			
18	9	5	6	6	7	1	6	1	5	1	6	1	5	1	6	1	5	1	6	1	5	1	6	5	4	5	4	5	4			
19	11	11	6	10	9	3	2	1	6	1	5	1	6	1	5	1	6	1	5	1	6	1	5	1	6	5	4	5	4			
20	21	11	9	10	5	2	1	6	1	5	1	6	1	5	1	6	1	5	1	6	1	5	1	6	5	4	5	4	5			
21	21	11	6	10	9	3	2	1	6	1	5	1	6	1	5	1	6	1	5	1	6	1	5	1	6	5	4	5	4			
22	22	9	5	8	4	1	6	1	5	1	6	1	5	1	6	1	5	1	6	1	5	1	6	5	4	5	4	5	4			
23	23	11	10	9	3	2	1	6	1	5	1	6	1	5	1	6	1	5	1	6	1	5	1	6	5	4	5	4	5			
24	24	10	8	11	5	1	1	6	1	5	1	6	1	5	1	6	1	5	1	6	1	5	1	6	5	4	5	4	5			
25	25	10	1	9	5	1	1	7	2	10	5	11	3	1	7	8	10	3	1	7	8	10	3	1	7	8	10	3	1	7		
26	26	8	11	7	7	4	5	8	4	5	9	7	9	5	4	5	3	4	5	3	4	5	6	3	4	5	6	3	4	5		
27	27	11	7	10	3	2	8	2	10	5	11	3	1	6	5	10	2	3	7	2	3	7	2	3	7	2	3	7	2	3		
28	28	11	3	10	8	1	6	2	10	5	11	3	1	6	5	10	2	3	7	2	3	7	2	3	7	2	3	7	2	3		
29	29	11	4	9	3	1	6	1	10	4	10	3	1	6	5	10	2	3	7	2	3	7	2	3	7	2	3	7	2	3		
30	30	9	4	8	3	1	6	1	10	4	10	3	1	6	5	10	2	3	7	2	3	7	2	3	7	2	3	7	2	3		
31	31	11	4	11	2	2	6	1	7	4	10	5	1	6	2	8	1	2	3	7	2	3	7	2	3	7	2	3	7	2	3	
32	32	11	5	11	2	1	6	1	10	5	11	3	1	6	2	8	1	2	3	7	2	3	7	2	3	7	2	3	7	2	3	
33	33	11	6	11	2	1	6	1	10	5	11	3	1	6	2	8	1	2	3	7	2	3	7	2	3	7	2	3	7	2	3	
34	34	6	7	7	4	5	8	4	5	9	7	9	5	4	5	3	4	5	3	4	5	6	3	4	5	6	3	4	5	6	3	4
35	35	6	5	4	5	3	6	4	5	6	7	10	1	1	7	1	1	3	1	1	6	1	1	6	1	1	6	1	1	6	1	1
36	36	11	9	10	3	1	6	1	10	5	11	3	1	6	2	8	1	2	3	7	2	3	7	2	3	7	2	3	7	2	3	
37	37	1	5	1	3	1	5	1	9	4	8	3	1	6	1	1	3	1	6	1	1	6	1	1	6	1	1	6	1	1	6	
38	38	11	7	9	2	1	5	1	9	4	8	3	1	6	1	1	3	1	6	1	1	6	1	1	6	1	1	6	1	1	6	

MEDIAN

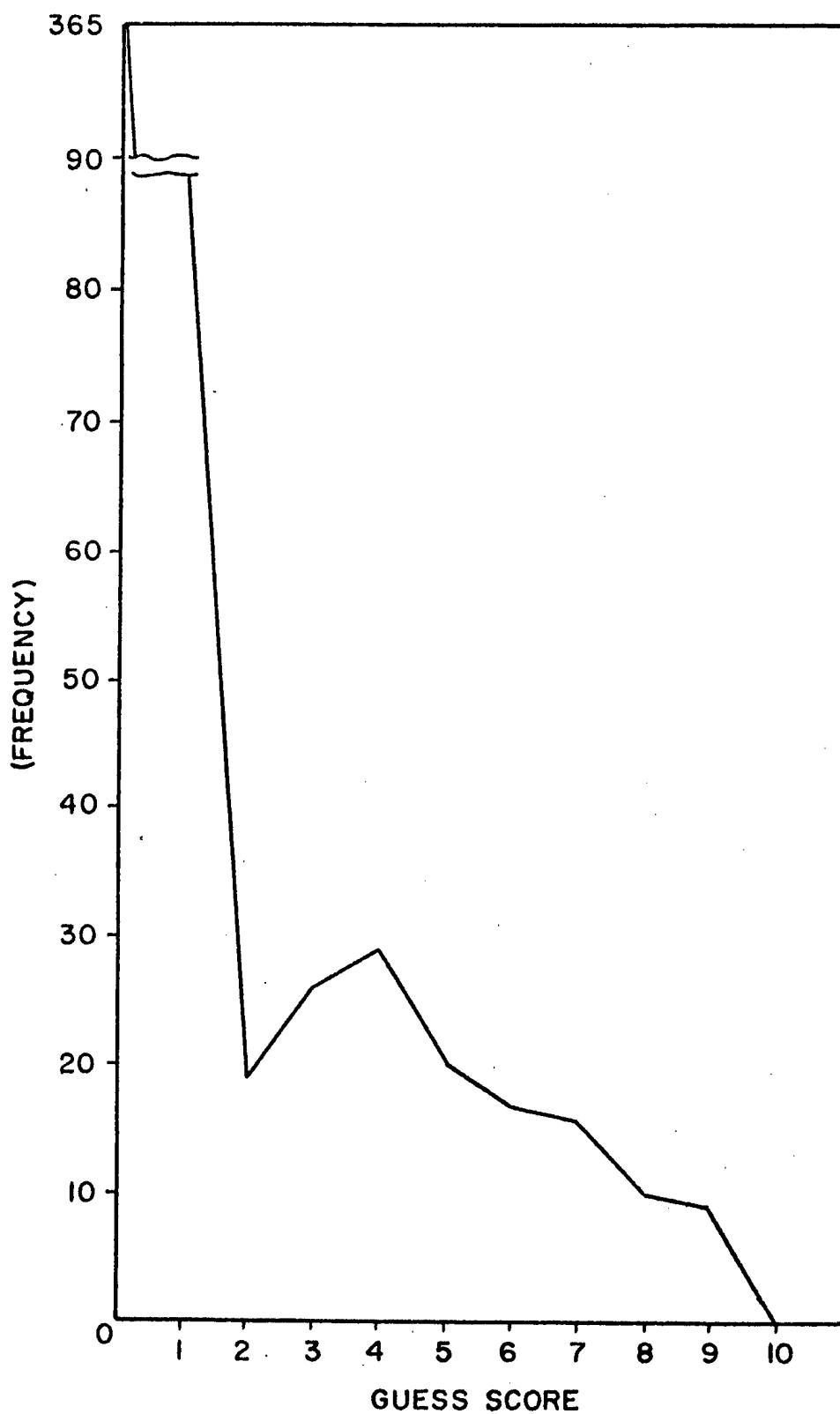
SUBJECTS →

(Table 1)

(Table 2)

<u>Statements</u>	Percent of Ratings in Each Category										
	1	2	3	4	5	6	7	8	9	10	11
*PERFECT IN EVERY RESPECT	0.8	1.6	0.8	2.5	0.8	5.7	1.6	4.9	6.6	7.4	67.2
NOT GOOD ENOUGH FOR EXTREME CONDITIONS	1.6	7.4	7.4	18.0	16.4	15.6	13.1	4.1	8.2	4.1	4.1
VERY GOOD	0.8	0.8	2.5	1.6	3.3	4.1	7.4	20.5	24.6	24.6	9.8
BARELY ADEQUATE	4.1	11.4	27.1	23.8	23.0	5.7	1.6	0.8	1.6	0.0	0.8
*VERY UNSATISFACTORY	54.4	21.1	0.8	0.8	5.7	4.9	2.5	3.2	0.0	3.2	3.2
MODERATELY GOOD	2.5	0.8	0.8	8.2	22.1	30.3	21.3	9.0	3.3	1.6	0.0
*VERY POOR	41.0	28.7	11.5	3.3	4.1	2.5	3.3	1.6	0.0	2.5	1.6
UNUSUALLY GOOD	0.0	2.5	4.1	0.0	5.8	8.3	9.1	12.4	23.1	28.9	5.8
NOT QUITE ADEQUATE	4.1	5.7	11.5	43.4	19.7	7.4	4.9	2.5	0.8	0.0	0.0
*EXTREMELY GOOD	0.8	1.7	0.8	0.0	3.3	4.1	7.4	14.1	9.1	29.8	28.9
NOT VERY SATISFACTORY	3.3	10.7	26.4	34.7	13.2	5.8	2.5	2.5	0.0	0.8	0.0
*EXTREMELY POOR	60.7	18.9	2.5	4.9	1.6	4.1	0.8	2.5	1.6	2.5	0.0
ABOUT AVERAGE	0.8	2.5	3.3	4.9	16.4	36.1	11.5	14.8	4.9	2.5	2.5
VERY GOOD IN MOST RESPECTS	1.7	1.7	0.8	0.8	5.8	11.6	20.7	24.0	19.8	9.9	24.8
*EXCELLENT	0.8	0.8	1.6	3.3	1.6	2.5	1.6	2.5	6.6	21.3	57.4
*BETTER THAN NOTHING	27.1	28.7	14.8	10.7	4.9	5.7	1.6	2.5	2.5	0.0	1.6
NEEDS MAJOR CHANGES	10.7	9.9	22.3	24.8	11.6	5.0	5.8	2.5	3.3	1.7	2.5
MORE THAN ADEQUATE	1.7	0.0	3.3	9.1	11.6	15.7	17.4	19.8	14.9	4.1	2.5
NOT GOOD ENOUGH FOR GENERAL USE	16.4	13.1	18.9	23.0	9.0	4.9	5.7	4.1	2.5	1.6	0.8
ADEQUATE	1.6	0.0	6.5	10.7	28.7	30.3	9.0	4.9	4.9	2.4	0.8
BARELY ACCEPTABLE	5.7	24.6	19.7	22.1	18.0	4.1	1.6	0.8	0.0	1.6	
*POOR	28.7	29.5	20.5	5.7	3.2	2.5	3.3	2.5	0.0	0.8	3.3

*Criterion statements



FREQUENCY DISTRIBUTION OF GUESS SCORES FOR ALL FIVE CHARACTERISTICS WITH EACH INDIVIDUAL REPRESENTED TWICE, ONCE FOR EACH CHARACTERISTIC RATED. (N=588)

FIGURE I.

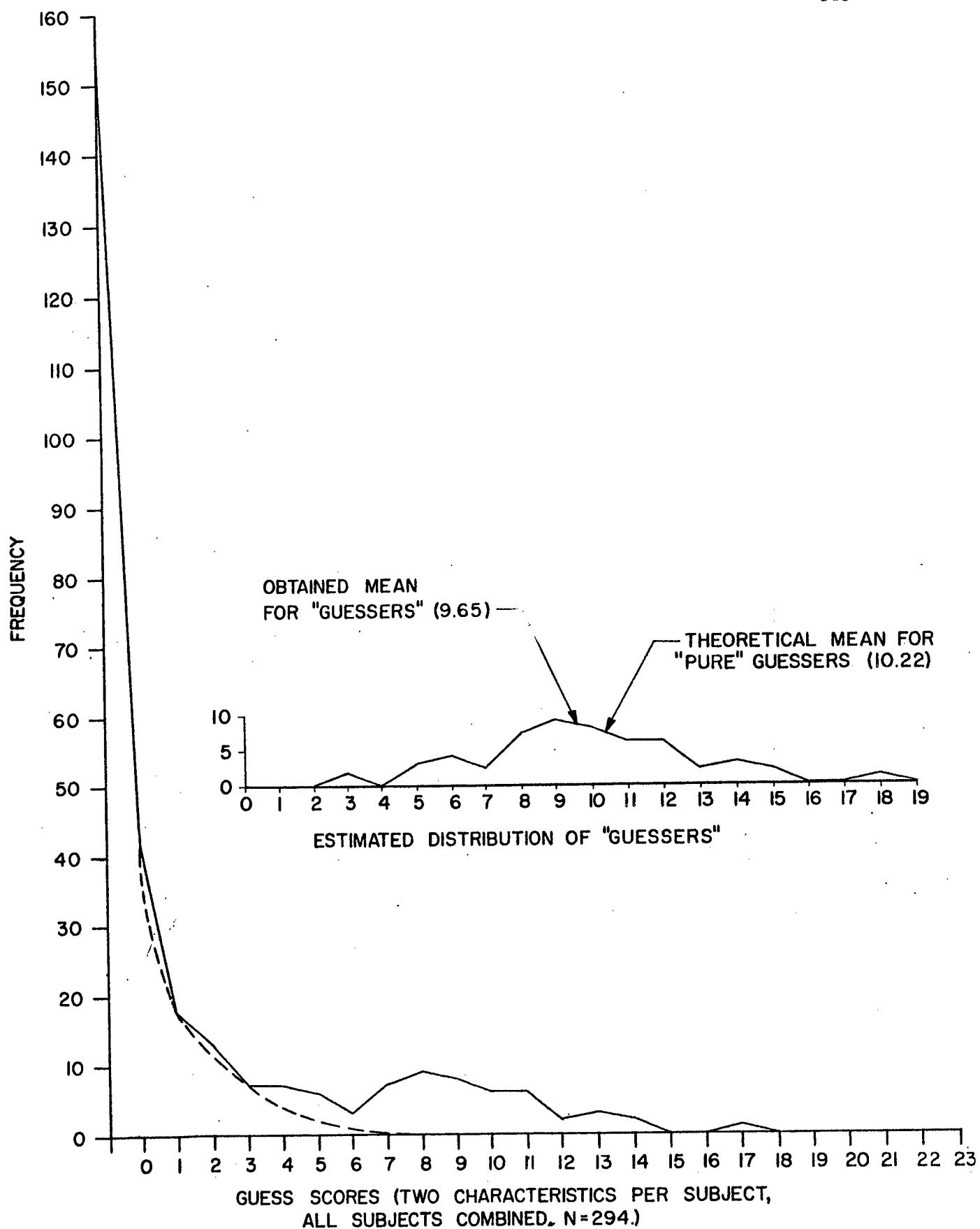
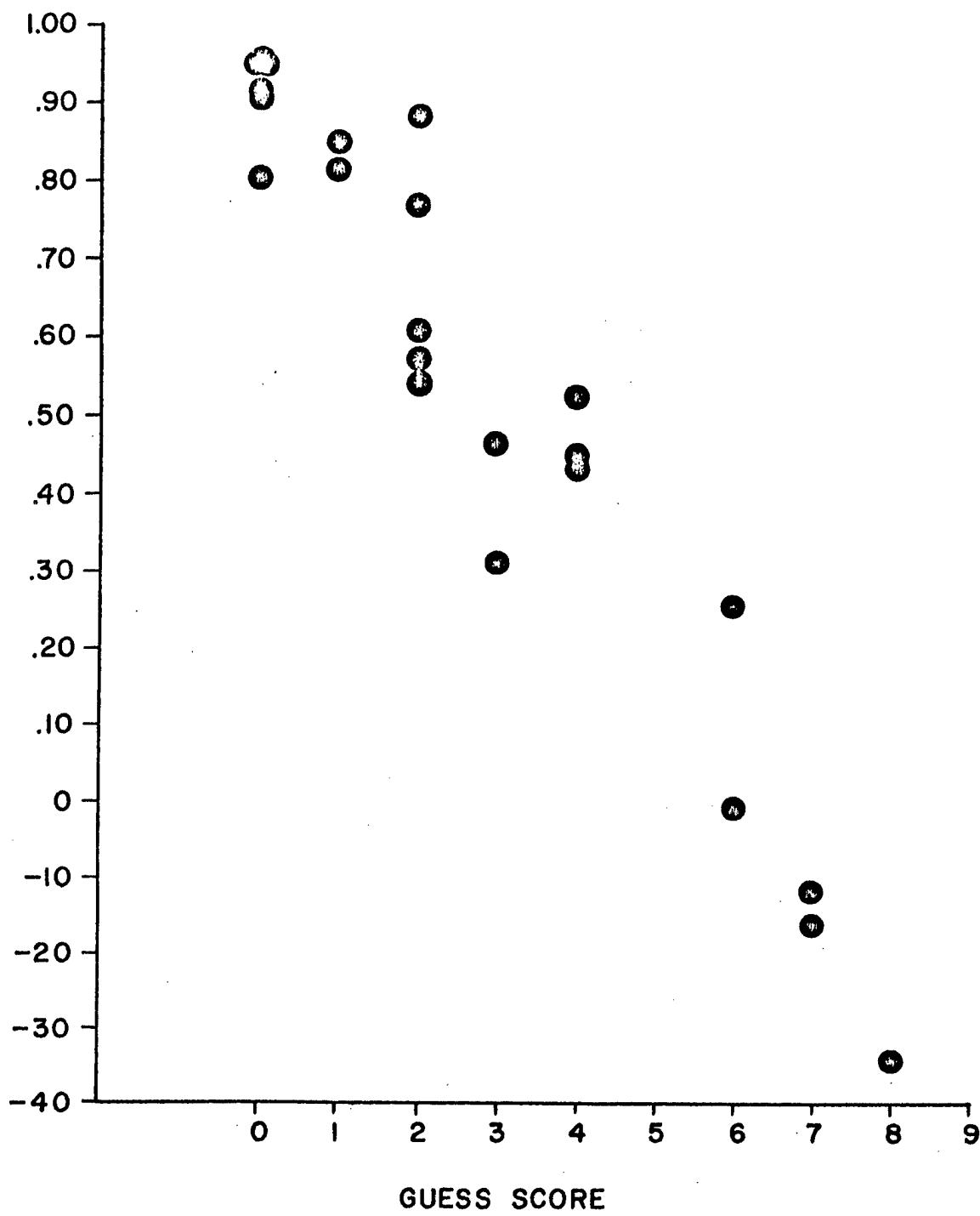


Fig 2. FREQUENCY DISTRIBUTION OF GUESS SCORES

CORRELATION OF INDIVIDUAL RATINGS WITH MEDIAN RATING



(Figure 3)

CORRELATION OF INDIVIDUAL STATEMENT RATINGS WITH MEDIAN RATINGS
PLOTTED AGAINST INDIVIDUAL GUESS SCORES.

DESIGN FOR ESTIMATION BY COVARIANCE TECHNIQUE

M. Rhian

Aerobiology Division, U. S. Army Biological Warfare Laboratories

The problem presented in this paper is similar to those discussed by Cox in the First Conference on Design (1), by Maloney in the Second Conference (2), and by Morrison in Biometrics in 1956 (3). The problem is the estimation of an observation which is purposely not observed. The methods, or mechanics of estimating one or more missing values seems to be of little concern, because there are so many recent descriptions of such procedures. Among these may be mentioned analysis of incomplete data by Wilkinson (4) nature and use of covariance by Cochran (5), and analysis of covariance as a missing plot technique by Coons (6). These and similar articles explore the mathematical bases for calculating missing observations and describe the procedures. These articles also suggested the title of this problem, but your discussion need not be limited to the consideration of covariance analysis.

Specifically, our problem is to estimate the dose of certain micro-organisms required to cause disease in certain animals, when members of the animal species cannot be used in direct experimentation. An approach to a solution of this problem may be obtained from brief consideration of the meaning of estimated, values designs used in other situations to obtain doses for estimating, and examples of observed and calculated values.

A few years ago the estimate of a missing value was regarded as a computational convenience and "was not intended as an estimate of the missing datum" (7). This view was challenged in 1954 by Nelder (8) who said that "whether or not x is intended to be an estimate of the missing datum, it is an estimate of the missing datum, and an unbiased one where the mathematical model used is true." This view was supported by Norton (9) who corrected a typographical error in Nelder's formula for calculating the variance of the missing value. Smith (10) then pointed out "that the error variance of the estimate depends on what is intended to estimate. This must be decided first and then the other aspects fall easily into place." "The variance is, therefore, relevant and Norton's discussion is essentially correct." Smith also pointed out the correct variance test for judging whether an estimated value is preferable to one that was rejected because it seems incompatible with the rest of the data. Rejection of a suspected observation seems to imply that the calculated value is more apt to be valid than the observed value. The view that a calculated value may be "what would have been observed" has not been challenged in Biometrics since 1954. Maybe it will be today.

Cox (1) described two designs for making small samples do double duty, the cross or butterfly design and an x design.

The cross design is illustrated in Slide 1 (at end of this article), taken from Cox's Table 9. He says "assuming no interactions exist it is possible to estimate an expected value for any of the 12 possible combinations of A and B which are given in Table 9 whether or not the combination has data assigned to it." The x design is illustrated in Slide 2, also from Cox (1). In this case three restrictions are necessary to make it

possible to calculate a value for any of the nine combinations. If these designs are valid, they illustrate what may be extreme examples of estimation of about one-half of the total observations.

Morrison describes fractional replications for mixed series and illustrates the estimation of unobserved values (3). One example is taken from Morrison's Table VI. For a $2^4 \times 3$ experiment, all 48 data points had been obtained then it was decided to set up an estimation design and to compare the results of the half-replication. Slide 3 shows only 1/8 of this table. "The standard deviation of the difference between an observation and an estimate is 0.58. For the 24 points estimated, the maximum discrepancy is equal to about 4 standard deviations (of a difference)." "It may be noted that at the 5% level of significance the results of the half- and full-replicate agree."

To illustrate the analysis of incomplete data by covariance Wilkinson (4) used data on blood sugar of rabbits treated with insulin. The original data consisted of observations on 8 rabbits in 4 phases, so there were 32 observations each on percentage fall and initial blood sugar. Three observations on percentage fall in blood sugar were discarded at random, then the corresponding initial values were discarded. The missing values were then estimated by the procedures described by Wilkinson. The values which were observed and discarded, and the values estimated are shown in Slide 4.

These examples from Morrison and Wilkinson indicate that under proper conditions unobserved values can be estimated with acceptable accuracy and precision. We would like to do as well in our situation.

Now what do we have to work with? The basic observations must be dose-response relationships, in which dose is expressed as numbers of deposited organisms and response presumably may be either disease terminated by recovery or disease terminated by death.

In the example chosen for this problem, the doses are numbers of spores deposited and the responses are time to death of the diseased animal. A possible array of data is illustrated in Slide 5. For host species D, doses of organisms are unknown, but the time responses can be obtained from case histories.

As presented in Slide 5 all the missing doses occur in one row, and this seems to complicate the analysis. Can the experiment be designed to change this arrangement?

DeLury has presented procedures for analysis of latin squares when one column or more is missing. Can similar approaches be used in our case?

Slide 1

	B ₁	B ₂	B ₃	B ₄
A_1			9	
			10	
			12	
			13	
			11	
A_2	7	9	11	13
	7	13	12	19
	8	11	14	15
	8	11	13	17
	9	12	16	16

The Cross Design (Cox, Table 9).

Slide 2

	B ₁	B ₂	B ₃
A_1	3		6
	8		8
	4		13
	5		9
	7		12
A_2		5	
		11	
		7	
		6	
		8	
A_3	4		8
	7		9
	10		13
	8		16
	5		17

X Design (Cox, Table 13)

Slide 3

Observation	Dependent Variable	
	Observed	Estimated
11213	20.15	
11223	28.15	26.36
11212	19.80	20.12
11222	26.70	
11221	19.20	
	25.75	24.40

Comparison of Observed and Estimated Values
(Morrison (3))

Slide 4

Variable	Observed and discarded values	Estimated values
y	33.9	33.2
	24.1	25.6
	35.9	39.8
x	96.9	94.0
	73.9	78.0
	79.9	85.0

Observed and Estimated Values (Wilkinson (4))

Slide 5

Host	Situation A			Situation B		
	DA ₁ TA ₁	DA ₂ TA ₂	DA ₃ TA ₃	DA ₁ TA ₄	DA ₂ TA ₅	DA ₃ TA ₆
A	DA ₁ TA ₁	DA ₂ TA ₂	DA ₃ TA ₃	DA ₁ TA ₄	DA ₂ TA ₅	DA ₃ TA ₆
B	DB ₁ TB ₁	DB ₂ TB ₂	DB ₃ TB ₃	DB ₁ TB ₄	DB ₂ TB ₅	DB ₃ TB ₅
C	DC ₁ TC ₁	DC ₂ TC ₂	DC ₃ TC ₃	DC ₁ TC ₄	DC ₂ TC ₅	DC ₃ TC ₅
D#	XD ₁ TA ₁	XD ₂ TA ₂	XD ₃ TA ₃	XD ₁ TD ₄	XD ₂ TD ₅	XD ₃ TD ₆
N	DN ₁ TN ₁	DN ₂ TN ₂	DN ₃ TN ₃	DN ₁ TN ₄	DN ₂ TN ₅	DN ₃ TN ₆

#Unavailable for direct challenge

Type of data that can be obtained
for estimation of Dose to Host "D".

REFERENCES

1. Cox, Paul C. Some Design Techniques Used for Increasing Cell Size with Special Emphasis in the Missile Field. Proceedings of the First Conference on the Design of Experiments in Army Research, Development, and Testing. Oct. 19-21, 1955. Office of Ordnance Research.
2. Maloney, C. J. Methods of Estimating Lethal Dose for Man. Proceedings of the Second Conference on Design of Experiments in Army Research, Development, and Testing. Oct. 17-19, 1956. Office of Ordnance Research.
3. Morrison, Milton. Fractional Replication for Mixed Series. *Biometrics* 12, 1-19, 1956.
4. Wilkinson, G. N. The Analysis of Covariance with Incomplete Data. *Biometrics* 13, 363-72, 1957.
5. Cochran, W. G. Analysis of Covariance: Its Nature and Uses. *Biometrics* 13, 261-81, 1957.
6. Coons, Irma. The Analysis of Covariance as a Missing Plot Technique. *Biometrics* 13, 387-405, 1957.
7. Snedecor, G. W. Answer to Query 96. *Biometrics* 8, 383-4, 1952.
8. Nelder, J. A. A Note on Missing Plot Values. *Biometrics* 10, 400-01, 1954.
9. Norton, H. W. A Further Note on Missing Data. *Biometrics* 11, 110, 1955.
10. Smith, H. F. Missing Plot Estimates. *Biometrics* 13, 115-18, 1957.
11. DeLury, D. B. The Analysis of Latin Squares When Some Observations Are Missing. *J. Am. Stat. Assn.* 41, 370, 1946.

DESIGN OF AN EXPERIMENT TO EVALUATE A BIO-ASSAY
WITH NON-PARALLEL SLOPES

Albert L. Fernelius
Process Research Division, U. S. Army Biological Warfare Laboratories

Graded response virulence estimates of three treatment conditions: ungerminated spores (U), germinated spores (G) and vegetative cells (V) of Bacillus anthracis were made with the mouse as the test animal. For testing virulence the graded response median-time-to-death (MTD) assay takes the form

$$MTD = a + D^b$$

where D is the concentration of organisms administered to the host, a is the intercept, and b is the dose-response slope. For a single intraperitoneal test dose consisting of approximately 10^8 cells the MTD values for the three treatment conditions were: U-15.5 hours, G-11.3 hrs, V-8.0 hrs which led to the conclusion that the decreasing order of virulence for treatment condition was U < G < V. In subsequent trials, four doses spaced at one log intervals were given to the host. The results in Handout 1 at end of this article indicate that the decreasing order of virulence for a 10^5 dose was G < U < V and the MTD values for this dose were: U - 24.2 hours, G - 42.8 hours, V - 19.8 hours. Plots of the dose-response curves for the three treatment conditions are shown in Handout 2. Estimated slopes of these curves are given in column 5 of Handout 1. Germinated spores generated a response slope approximately two times greater than that given by vegetative cells or ungerminated spores, so it seems possible that any comparison of treatment conditions must be based on sensitivity of the host to changes in dose, i.e., the slope, rather than on MTD values alone. When identical slopes are obtained, then MTD values can be directly compared.

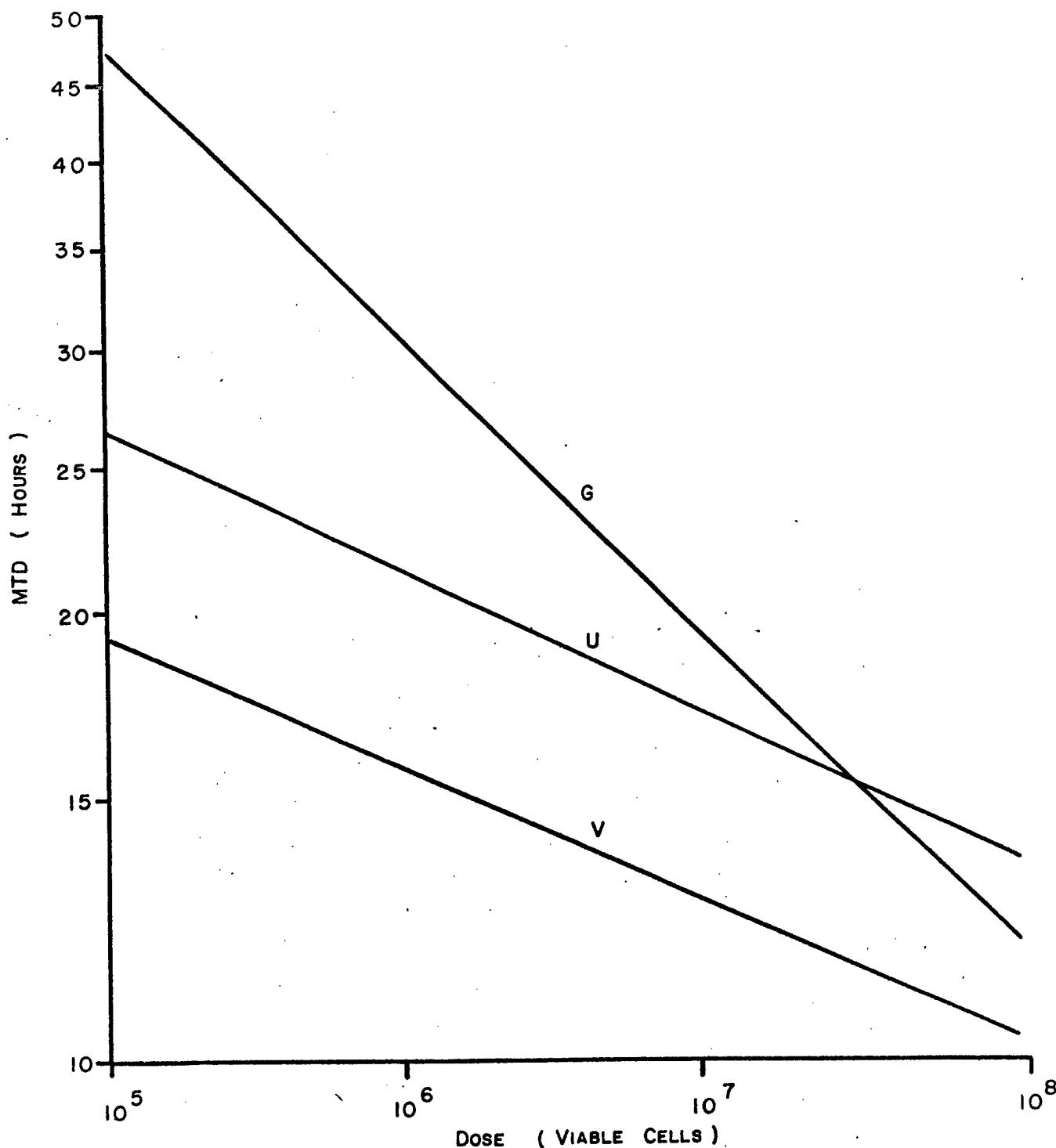
The problem I would like to present to this panel is: How can one design an experiment to compare and evaluate treatment conditions which elicit non-parallel responses in the host? Is there any method of combining the parameters of slope and graded response (MTD) values so that heterogeneous data can be directly compared? Obviously when MTD values are compared, their relative values will be governed by the point selected on the dose-response curve. How do you determine the dose to titrate, or must one always titrate multiple doses for any comparative purpose?

Handout 1

Mouse Median-Time-To-Death and Log Slope Values for Virulence tests
of Bacillus anthracis organisms in three stages of the Spore-Vegetative
Cell Cycle

Cyclic stage	\bar{x} Dose* (cells)	\bar{x} MTD (hours)	\bar{x} MTD for a $10^{6.5}$ dose	\bar{x} log slope
Ungerminated	10^8	13.2		
	10^7	18.3	19	-.090
	10^6	22.5		
Germinated	10^5	24.2		
	10^8	11.4		
	10^7	22.3	24	-.187
Vegetative	10^6	32.6		
	10^5	42.8		
	10^8	9.2		
	10^7	13.3	14	-.090
	10^6	17.1		
	10^5	19.8		

* All values are means of three replications.



HANDOUT 2. DOSE RESPONSE FOR UNGERMINATED, GERMINATED, AND
VEGETATIVE BACILLUS ANTHRACIS ORGANISMS IN MICE

THE ORO AIRCRAFT VULNERABILITY EXPERIMENT*

Charles A. Bruce and Bruce Taylor
Operations Research Office, the Johns Hopkins University

INTRODUCTION. This paper describes some techniques planned for the analysis of results from the ORO part of the Aircraft Vulnerability Experiment performed at The Combat Development Experimentation Center.

A major reason for our interest in these techniques is the large amount of data generated during the experiment. The original data were taken on 200,000 feet of film and were the equivalent of around two million individual numbers (or readings). After the reduction of this data, which is of necessity a computer operation, there will result about 19,000 numbers. These numbers will be inputs to the techniques described in this paper. We are looking for major trends and highlights in the reduced data. The methods used must both considerably reduce the 19,000 input numbers and also provide valid indications of important conclusions. These conclusions are concerned with tactics and design of Army aircraft, and air defense weapons.

The techniques planned are graphs, curvilinear regression, contingency tests, linear correlation, and analysis of variance. It would simplify matters considerably if the data turns out to be predominantly deterministic and lacking in noticeable change fluctuations. In this case graphs will be drawn for special cases and limiting conditions, and will be followed by a curvilinear regression analysis. However, if sizeable statistical fluctuations appear, contingency tables, linear correlations, and analysis of variance will be used to detect completely random effects and to place some bounds on the fluctuating variables. The problem is not so much one of devising new techniques, but of knowing which of the standard techniques are applicable to this particular experiment.

BACKGROUND INFORMATION. The experiment was undertaken to determine the vulnerability of low flying aircraft to forward area ground fire, and in particular, how this vulnerability depends upon the velocity, altitude, and crossing range of the aircraft with respect to the ground weapons, and also the alert status and line-of-sight terrain masking of the ground troops. The weapons used were Redeye, Quad .50 Cal. machine gun, Twin 40mm antiaircraft weapon, .50 Cal. machine gun on armored personnel carrier, the BAR, and M1 Rifle. The raw data were basically gun-camera film to measure aiming errors, radar and phototheodolite data to give aircraft position, and pen records for timing information. Fig. 1 shows precisely what is meant by mask angle.

In regard to work completed to date, the experiment was designed as a complete factorial with two replications, but only one replication was run due to time and equipment limitations. The experiment was run continuously for a period of one month, during which time approximately 500 single aircraft-passes were made over ground troops. The raw data has been transferred to IBM cards and then to Univac magnetic tape. At the present time, a small amount of data from a single aircraft pass is being

* The authors would like to acknowledge the work of Dr. Jack C. Rogers in developing the curvilinear regression technique described in this paper.

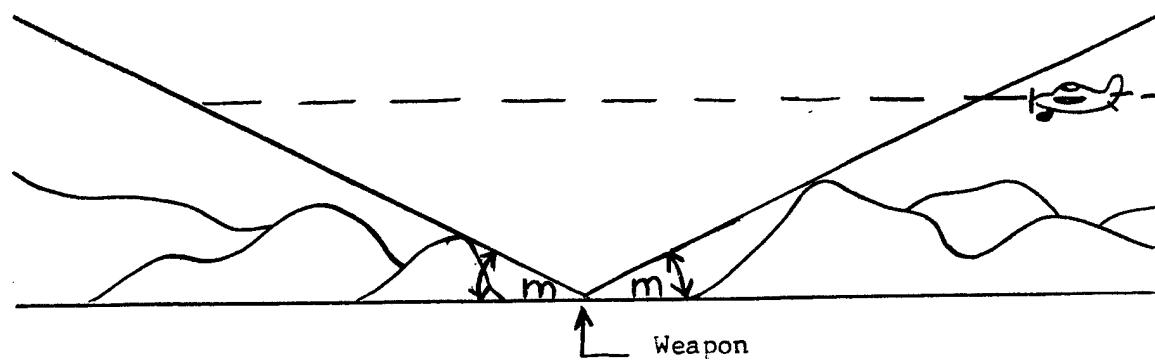


Fig. 1
Terrain Mask Angle

analyzed in detail by hand, while at the same time computer programs are being written which will duplicate the hand method automatically. When the computer programs have been tested, the data reduction will start. The reduced data will then be the basis for analysis of the kind described in this paper.

The aircraft vulnerability of interest here might better be called engagement-vulnerability to distinguish it from the target-vulnerability of the aircraft itself. The engagement-vulnerability is defined by a number of quantities shown in the following figure. These are time under fire (T_1), response time (T_2), rounds fired (simulated) (F), mean miss distance (D), hits (H), and kills (K). Going down the list one probably gets a better definition of engagement vulnerability, but more assumptions must be made in computing them from the experimental data. Of special interest are the cases when T_1 is zero or very small. This means that little or no rounds could be fired at the passing aircraft. The response time T_2 for a weapon crew to respond to a sudden appearance of an aircraft is of interest in itself and probably not a good measure of vulnerability. At any rate the six quantities shown in Fig. 2 are the results we want for every weapon and every set of experimental conditions.

AIRCRAFT ENGAGEMENT-VULNERABILITY	
Per Weapon - Per Ground-Air Engagement	
Dependent Variables	
T_1	Time under fire
T_2	Response time of crews
F	Rounds fired
D	Mean miss distance
H	Hits
K	Kills

Fig. 2

Design of Experiments

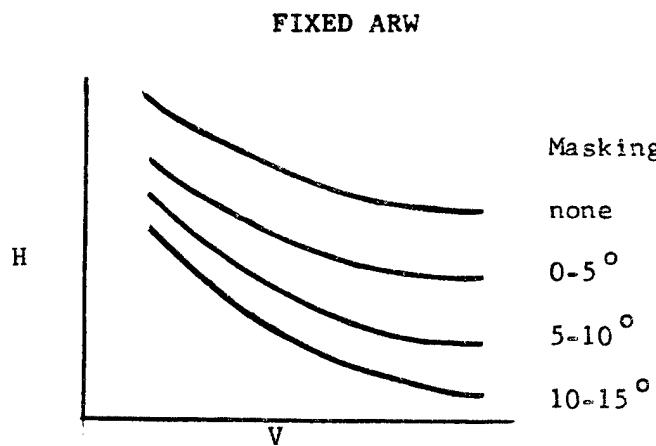
The quantities in Fig. 3 were varied systematically during the course of the experiment. The number of levels for each variable is also shown.

OPERATIONAL VARIABLES	
Aircraft and Ground Troops Operating Conditions	
Independent Variables	
V	Velocity of aircraft (4)
A	Altitude (3)
M	Mask angle of terrain (4)
R	Crossing range (2)
W	Warning (2)

Fig. 3

There were a number of factors in the nature of parameters of fixed conditions, not subject to systematic variation and study during the experiment. These were weapon types and troops, along with aircraft evasive action, identification, and target-vulnerability. There were a fixed set of weapon types, while the ground troops were all given the same training, and were rotated around the various ground positions. The aircraft always flew a straight and level path (from different directions, however) and the troops were not required to identify the aircraft before firing.

SOME TECHNIQUES. For exploratory purposes, the effect of aircraft velocity on vulnerability is graphed as shown in Fig. 4.

Fig. 4
Velocity Graph

The effect of altitude is graphed as shown in Fig. 5

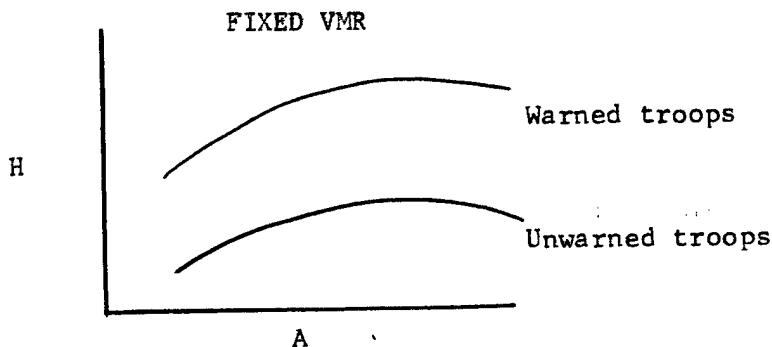


Fig. 5
Altitude Graph

The presence of large fluctuations might already be revealed in the position of the data points on these curves.

Using a regression analysis, it is planned to find all the coefficients F_{ijk} in a polynomial which gives each of the dependent variables (result variables) such as kills (K) or kill probability, as a function of the multilevel variables velocity (V), altitude (A), and mask angle (M). The two-level variables crossing range (R), and alert status (W) will be parameters for the regression. The number of terms in the polynomial increases rapidly with the degree of fit (n), going from 8 to 27 to 64, as n goes from 1 to 2 to 3. The basic method is that of polynomial approximation by orthogonal polynomials and equally spaced points, described by Milne. This requires only a slight adjustment of the experimental data since V, A, and M were designed to have equally spaced points, but were not quite equally spaced due to such difficulties as keeping a light aircraft on a level course. The basic method is repeated a number of times, once for each independent variable until all the coefficients are determined. This process is illustrated in Fig. 6, for n=1. The dependent variable K is expressed as a linear function of V for all possible values of A and M. The resulting (intermediate) coefficients are hence functions of A and M. These coefficients are next expressed as linear functions of A, with the resulting coefficients being functions of M. These latter coefficients are then expressed as linear functions of M. Now by working backwards and substituting the explicit linear form for each coefficient into the previous expressions, the complete polynomial for K is obtained.

CURVILINEAR REGRESSION		
$K = \sum_{ijk=0}^n F_{ijk} V^i A^j M^k$	n	terms
	1	8
	2	27
	3	64
Fixed RW		
$K = F_0(A, M) + F_1(A, M) V$		
$F_0(A, M) = F_{00}(M) + F_{01}(M) A$		
$F_1(A, M) = F_{10}(M) + F_{11}(M) A$	Optimum regions	
$F_{00}(M) = F_{000} + F_{001} M$	Sensitivity analysis	
$F_{01}(M) = F_{010} + F_{011} M$	Error analysis	
$F_{10}(M) = F_{100} + F_{101} M$		
$F_{11}(M) = F_{110} + F_{111} M$		

Fig. 6

In regard to the goodness of fit, a sum of squared deviations is obtained and this will allow an estimate of the amount of randomness involved. The polynomial resulting from this method is adaptable to finding maximum and minimum regions for aircraft vulnerability, a very useful result.

Figure 7 shows a table in which the dependence of vulnerability (in this case hits) upon aircraft crossing range is tested. There are N rounds being considered, and these are classified according to the hits H and non-hits \bar{H} , and also according to the rounds fired R at an aircraft flying a path at 300 meters crossing range and the rounds \bar{R} fired at an aircraft flying a path which leads directly over the weapon position. Crossing range has no influence on hits provided it is found that the proportion of hits and non-hits are the same for paths at a crossing range as it is for all rounds generally. Fluctuations of this proportion due to pure chance will also be considered.

SIGNIFICANCE TESTS (Contingency Tables)				
	R	\bar{R}		
H	(HR)	$(H\bar{R})$	$\rightarrow (H)$	H Independent of R
\bar{H}	$(\bar{H}R)$	$(\bar{H}\bar{R})$	$\rightarrow (\bar{H})$	Provided
	$\downarrow (R)$	$\downarrow (\bar{R})$	$\rightarrow N$	$\frac{(HR)}{(\bar{H}R)} = \frac{(H)}{(\bar{H})}$

Aggregated V-A-M-W

Fig. 7

Figure 8 shows a correlation table containing plus signs for positive correlation coefficients, minus signs for negative ones, and question marks for cases where even the sign of the coefficient is in doubt at our present stage of knowledge.

LINEAR CORRELATIONS						
Independent Variables						
DEPENDENT VARIABLES	V	A	M	R	W	
	T	-	+	-	+	+
	F	-	+	-	+	+
	D	+	-	+	-	-
	H	-	?	-	-	+
	K	-	?	--	-	+

$C = \frac{\sum XY}{\sigma_x \sigma_y}$

Fig. 8

In the application of the analysis of variance to the experimental data, the lack of a second replication requires special consideration. It was generally not the case that the experimental conditions (treatments) were repeated, although a few were run more than once. This means that there is only one result (yield) for every condition, or one vulnerability measure for each combination of the independent variables V, A, M, R, and W. With these variables there are $4 \times 3 \times 4 \times 2 \times 2$, or 192 different conditions. In order to introduce some variation for analysis, a number of approaches suggest themselves. First, if it is known from preliminary analysis that some variable such as crossing range (R) has a negligible effect upon vulnerability, this effect will not be investigated and consequently will be randomized. This gives $4 \times 3 \times 4 \times 2$ or 96 different conditions, each condition now having two results. A complete factorial with two replications is thus obtained. Other approaches are based on a lack of sensitivity of vulnerability to a multilevel variable such as altitude (A), or mask angle (M). In this case either A or M could be reduced to two levels, called high and low altitude, or high and low mask angle. In either case, a complete factorial with two replications would be obtained. With these two replications a completely randomized design is applicable with 95 degrees of freedom for the variation due to error, 1 for replications, and 95 for main effects and interactions.

Another approach is to ignore the highest order interaction, namely V with A, M, R, and W, and treat it as an error term. In this case, there is one replication with 173 degrees of freedom for the main effects and interactions, and 18 for experimental error.

The results of these various approaches is to permit a test of the hypothesis that the mean result for each condition is the same. Since we are rather confident that they are not the same, it will then be desirable to find out which conditions might be equivalent and which of them have the largest effect on vulnerability.

SUMMARY. Some techniques planned for the analysis of results from the ORO part of the Aircraft Vulnerability Experiment have been described, along with some background information on the experiment and a statement of the data analysis accomplished to date.

DESIGN FOR A PROPOSED FIELD EXPERIMENT
WITH LIGHT AT WEAPONS

R. E. Tiller, J. D. Reed, J. P. Young
Operations Research Office, the Johns Hopkins University

OBJECTIVES. The experiment described here was designed to determine accuracies under simulated tactical conditions of a family of current and prototype shoulder-fired AT* weapons, and to develop data which will serve as guide-lines for research and development leading toward optimized weapons.

BACKGROUND. AT warfare places stringent requirements on hand-held infantry weapons. It is essential that these be light in weight, rugged, simple to operate and maintain with minimum training, and most important, they must offer a high probability of hitting and killing the tank with the first round. Tanks very seldom work singly, and even when in sections or platoons usually are accompanied by protecting infantry. Disclosure by fire is therefore a serious problem to the AT gunner; he can fire only one round, and must then move quickly to another position, or be killed. This problem emphasizes the need for a first-round hit.

Hit probability is influenced by a number of factors which may be arbitrarily divided into "ballistic" and "gunner" errors. Even on the training range at known distances, firing at clearly visible targets, the gunner's error greatly exceeds the error of the weapon and ammunition, and in the operational situation, we can in nearly every instance ignore the ballistic factor. If a weapon is acceptable to the Army, the problems associated with ballistic engineering are in most cases minor compared to those introduced by the gunner. Velocity will be acceptably constant, the shot group at a given range will meet the stated specifications, and the weapon will be satisfactorily rugged.

The gunner's errors are another thing, however. His errors in aiming, in canting the weapon, and most important to the shoulder-fired AT weapons, (which have characteristically low velocities), in range estimation, are of primary importance in the system accuracy.

Interaction of velocity with range estimation error therefore determines to a large extent, the probability of the infantryman hitting the enemy tank.

TEST ITEMS. The weapons to be tested are the cuurent 3.5 Rocket Launcher (M20); a new prototype rocket launcher which will be fired in two ways, a new recoilless rifle, this last weapon to be tested with 3 fire control systems and the cal. 30 rifle (M1) which will simulate a hypothetical flat trajectory high velocity weapon.

Theoretical analysis of hit probability as a function of range indicates an elliptical normal distribution, due to the large vertical component of error introduced by range estimation. To check the equation which will be used for determining P_H and determine the constants in that equation, it will be necessary to test the weapons at a number of ranges appropriate to each weapon. A large number of firings will be required, determining the error associated with each firing.

* Antitank

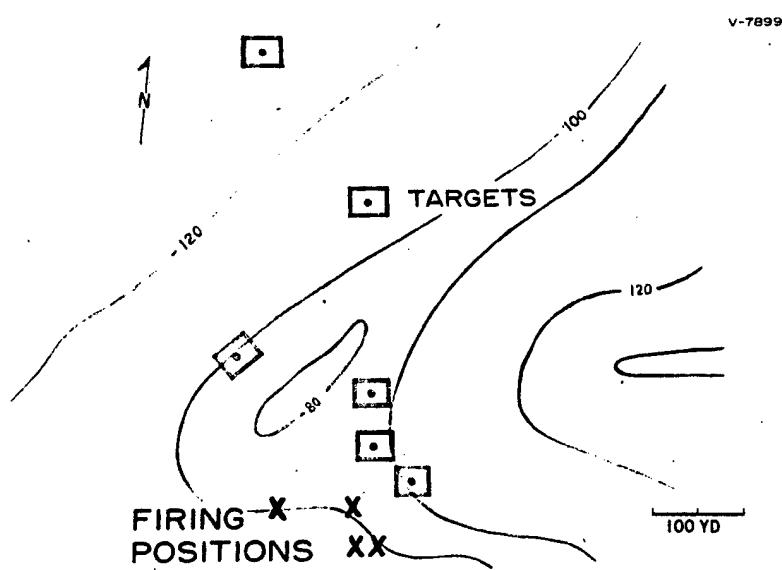


Fig. 1. Hypothetical layout of firing positions and targets for Phase I.

Test Subjects and Training. Fifty-two soldiers, preferably men with no particular M.O.S., who have recently completed Basic Infantry Training, and selected to eliminate critical physical differences, will serve as test subjects. Twenty-six will be trained as gunners and 26 as loaders. Prior to the experiment, after receiving instruction on all test weapons, they will be assigned to four groups of 6 gunner-loader teams (plus spares).

PROCEDURE. It will be necessary to conduct three separate experimental phases to achieve all of the desired objectives.

The first will be a determination of hit probability against static targets, for these we will use 7½ ft square O.D. panels, electro-mechanically controlled to permit exposure in the desired sequence.

The second will employ moving targets ("buttoned-up" tanks on a prescribed course using accepted evasive tactics). The last phase will deal with the determination of second round hit probability as influenced by the first round, and will again employ the panel targets used in Phase I.

TARGET AREAS. Two similar target areas will be required for Phase I of the experiment. A hypothetical target layout is shown in Figure 1.

On these areas the 7½ ft panel targets will be placed at ranges where .25, .50, and .75 hit probabilities are expected on the basis of theoretical calculations; some ranges will overlap and permit the use of the same targets for different weapons.

To simulate more closely actual operating conditions, a series of explosive charges will be detonated near the firing points and near the targets.

EXPERIMENTAL DESIGN. The test schedule is designed to minimize specific learning of ranges. The subject will fire the M-1 rifle and one other weapon in an area each half day. The locations of the firing points and targets will be changed, so that the subjects will not be able to transfer specific information from one situation to another. The order of appearance of the targets will also be varied. As a result, the subjects will gain only a general knowledge of the target area, similar to what might be expected in an operational situation. Furthermore the subjects will not be able to profit by talking to other men who have just completed firing, because each order faces a slightly different situation, and no firers will be permitted to enter the target area at any time.

PHASE I--STATIC TARGETS. The experimental factors to be investigated are:

- 24 men
- 2 firing positions
- 7 weapons or weapon combinations
- 3 ranges for AT weapons; all ranges for the M1 rifle

Each will be systematically varied in a balanced experimental design, utilizing the 24 gunners (with loaders as required) in four groups of six

V-7897

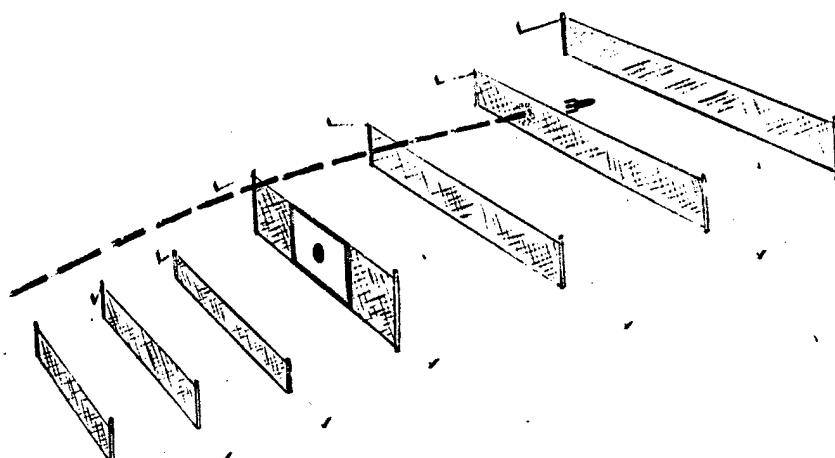


Fig. 2. Diagram of target with associated screens designed to determine "miss distances."

V-7898

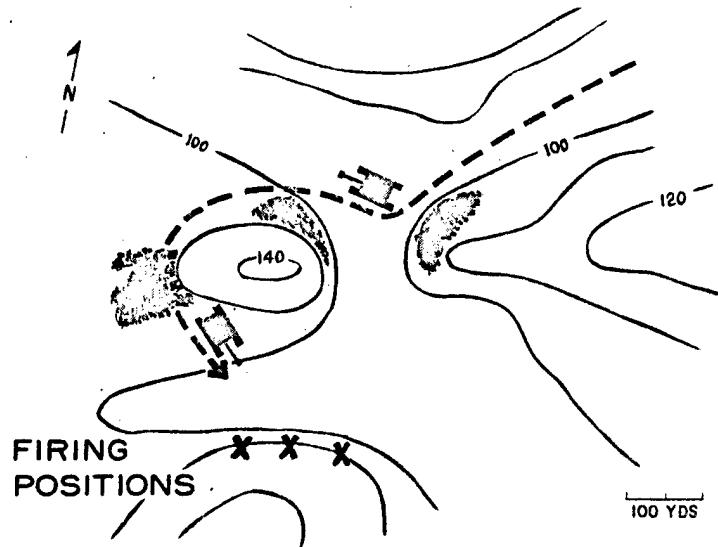


Fig. 3. Hypothetical layout of firing positions and terrain features required for Phase III.

who will fire individually, using one type of weapon in each area in each time period, with the exception of the M-1 which will be fired at all targets.

Weapons fired and their ranges (close, mid-range or distant) will be systematically varied among the firers as well as among firing position, area and time of day.

Twenty-four men, firing in two positions, will yield 48 pieces of information for each range and inasmuch as aiming error expresses itself in angular error, this procedure may offer data for one range which supports the other two ranges.

At a minimum, this design will yield 48 pieces of data at a given range which even if unrelated to the other ranges, will provide satisfactory accuracy for our determination of hit probability.

Adequately complete data for our determinations of P_H will involve not only the distribution of hits on the $7\frac{1}{2}$ ft panels, but will require also measurement of the dispersion of at least 95% of the rounds which fall short or pass over the target. A fully satisfactory and effective technique has not yet been determined, but this problem is currently under study.

The feasibility of installing a series of vertical wire screens appropriately located to the front and rear of the target position is one of the suggested methods. Figure 2 is a diagrammatic representation of a typical target with its associated screens. The height, width and location of the screens is calculated to accomodate the weapon having the sharpest angle of fall at the given range.

Preliminary field testing indicates that although this system does not pose any problem by obscuring the target or by detracting seriously from the tactical realism of the target area (when painted O.D., the fences are invisible at ranges greater than 75 yards), they are highly vulnerable to damage by ricochets.

A test is scheduled in the near future to evaluate the accuracy with which misses can be located by observing the point of impact through a B.C. scope, and determining by standard surveying techniques the horizontal and vertical displacement of the round in relation to the target.

PHASE II. In this exercise a tank will cover a prescribed course, unknown to the gunner, and will use accepted evasive tactics (changes of speed and direction). Three weapons will be used, each at two ranges.

As indicated in Figure 3, terrain will provide a feature allowing the tank to appear for 20 seconds at a 50% P_H range for two of the test weapons. It will then disappear, and reappear at 50% P_H range for the third, shorter range weapon. Each weapon will fire one round at each range.

From these data on weapons of comparatively high and low velocities, we hope to evaluate effectiveness of intermediate velocities.

Design of Experiments

No difficulty for the observers is anticipated in identifying the performance of the weapons, since each has a characteristic velocity and trajectory.

As in Phase I, two areas will be needed. Two groups will fire Phase II while the others are firing Phase I.

PHASE III--SECOND-ROUND HIT PROBABILITY. Determination of improvement in P_H on firing a second shot will require the use of live ammunition for the first round for realistic sensing.

The areas described for Phase I will be used. Two weapon combinations will be employed.

Each man will fire two rounds at each of five targets. The two weapons are so dissimilar that they may be fired concurrently. The target sequence will be varied, but only one firing position will be used. As in Phase I, each group will fire in one area for one period to achieve balance.

It is our hope that this proposed field test will not only yield realistic values for P_H with existing shoulder-fired AT weapons and for several prototype weapons, but will indicate with validity not obtainable from engineering tests, the critical parameters for designing the best weapons in this category.

UNBIASED ESTIMATION BASED ON TRANSFORMED VARIABLES,
WITH PARTICULAR REFERENCE TO CLOUD SEEDING EXPERIMENTS*

Jerzy Neyman and Elizabeth L. Scott
University of California, Berkeley

1. INTRODUCTION. Because of the notorious skewness of the distribution of meteorological observations, their statistical analysis frequently begins with a transformation of variables. Thus, for example, instead of dealing with amounts of precipitation measured in inches or with runoff measured in acre-feet, one works with the square roots of these quantities, or perhaps with their logarithms, etc. With a certain amount of good luck, the distribution of the transformed variables approaches the normal distribution and also satisfies certain other conditions. As a result, a reliable analysis may be performed using the standard statistical techniques, which were developed on certain restrictive assumptions.

Granting that the transformation chosen is satisfactory, certain parts of the analysis can be performed entirely in terms of the transformed data, without any reference to the natural units (inches, etc.) in which the original observations are expressed. However, this is not true with problems of point estimation of the parameters. For example, in order to be intelligible, the estimates of the average increase in precipitation or in runoff ascribable to seeding must be expressed in units appropriate to these quantities, not in square roots or in logarithms. The customary procedure for obtaining estimates in the original units is to calculate estimates in the transformed units and then transform these backwards (by squaring or by taking antilogarithms, etc.). However, it happens that the backward transformation applied to an unbiased estimate leads, generally, to a biased estimate. Occasionally, this bias is trivial and is overshadowed by the uncertainty of the estimate due to random fluctuations. In other cases, the bias is quite important. Furthermore, even if the bias of an estimate is small compared to its random error, the bias may acquire importance when this estimate is combined with other biased estimates derived from other sets of data.

A case in point is the combined evaluation of a number of cloud seeding experiments. Because of the great variability of the observations and the relative scarcity of data, the estimated increase in precipitation obtained in any particular experiment ordinarily is shaky. Thus, it appears reasonable to try to combine the results of a number of experiments so as to obtain an estimate of the average effect of seeding. If the estimates calculated for particular experiments are all unbiased, then the combined unbiased estimate is easy to obtain by calculating the appropriate weighted average. On the other hand, the averaging of biased estimates, particularly if the bias always has the same sign, may lead to serious errors: the random errors attached to the single estimates will tend to average out, but not the bias. In fact, by examining a recently published combined evaluation of a number of commercial cloud seeding operations, the authors found that, due to the bias involved in the

* This paper was prepared with the partial support of the Office of Ordnance Research, U. S. Army under Contract DA-04-200-ORD-171, Task Order 3.

estimates, the average effect ascribed to seeding is likely to have been overestimated by a factor of two.

The purpose of the present paper is to provide formulas for the unbiased estimation of parameters in their natural units when the analysis is based on transformed variables. While we are primarily concerned with the evaluation of cloud seeding experiments, and, more specifically, with unbiased estimation of the effects of seeding, the same formulas are likely to be useful in other cases.

2. NOTATION AND ASSUMPTIONS. In order to estimate the increase in precipitation ascribable to seeding, it is necessary to estimate the precipitation which would have fallen in the target in the absence of seeding. A common method uses the precipitation in one or more, say s , comparison areas presumed to be free of any effect of seeding and so considered as control areas. Then regression analysis provides the estimate needed.

The observations may be the amounts of precipitation, measured in inches, falling in the areas considered during specified intervals of time, or the amounts of runoff. The intervals of time may be variously defined "storms," twelve-hour periods, days, months or years. The exact nature of the observations and the particular intervals of time to which they refer are irrelevant to the discussions which follow. For this reason, and for the sake of simplicity in wording, we shall speak of the amounts of precipitation from a storm.

The observations will consist of a certain number m of seeded and a certain number n of not-seeded storms, and we shall assume that both groups represent random samples from the same well-defined population of storms. We need symbols to denote the amounts of precipitation in the target and in the controls, first, generally, for a storm of the particular category and then for the j -th storm of the available sample. For a seeded storm in general, the amounts of precipitation in the s controls will be denoted by

$$(1) \quad *X_1, *X_2, \dots, *X_s,$$

respectively. Occasionally, it will be convenient to use one symbol to denote these s variables; we shall use the symbol $*X$. for this purpose, so that $*X = (*X_1, *X_2, \dots, *X_s)$. The corresponding seeded precipitation in the target will be denoted by $*Y$. All this applies to a seeded storm "in general." When referring to the j -th seeded storm of the available sample, the corresponding symbols will be

$$(2) \quad *X_{\cdot j} = (*X_{1j}, *X_{2j}, \dots, *X_{sj}) \text{ and } *Y_j,$$

respectively. Here, then, $j = 1, 2, \dots, m$. The notation for the not-seeded storms, either generally or for the j -th member of the available sample, will be the same except that we shall omit the asterisks. Thus, for example, Y_j will denote the target precipitation from the j -th unseeded storm.

Symbols involving the letters X and Y will denote the amounts of precipitation measured in the original units, that is, in inches, etc. We now introduce a corresponding set of symbols to denote the transformed variables, replacing X by U and replacing Y by V. In this way, for example, $*V_j$ will mean the target precipitation in the transformed units from the j-th seeded storm.

In this paper we are not concerned with the choice of the function for transforming from the original variables X or Y to the transformed variables U or V. The literature on this subject is extensive. We are interested in the transform back into the original units. For this reason, it is convenient to denote the function carrying the original variables into the transformed variables by f^{-1} , perhaps with subscript, while the inverse transformation carrying the transformed variables back into the original variables will be denoted by f, with appropriate subscript. It will be seen that since our interest is reversed, the notation is also reversed. For short, we call f the transforming function.

Thus, it will be assumed that each of the original variables X or Y is functionally related to the corresponding transform, and that this relation is the same for seeded and for not-seeded storms. In other words, we postulate the existence of s+1 monotone functions $f_0, f_1, f_2, \dots, f_s$ such that

$$(3) \quad *Y = f_0(*V) \quad \text{and} \quad Y = f_0(V)$$

and, for $i = 1, 2, \dots, s$

$$(4) \quad *X_i = f_i(*U_i) \quad \text{and} \quad X_i = f_i(U_i).$$

Frequently, there s+1 functions all coincide, in which case the identifying subscripts are superfluous.

As stated in the Introduction, the subject of this paper is limited to the estimation of certain parameters. For this reason it will be assumed throughout that the many pitfalls involved in the evaluation of cloud seeding experiments are successfully avoided and, in particular, that the transformed variables U. and V corresponding to not-seeded storms satisfy exactly the following condition (i) and that the variables $*U$. and $*V$ corresponding to seeded storms satisfy either condition (ii) or condition (iii):

(i) For any not-seeded storm with transformed precipitation U. in the control areas equal to $u. = (u_1, u_2, \dots, u_s)$, the transformed target precipitation V is normally distributed with mean

$$(5) \quad E(V|U.=u.) = \alpha_0 + \sum_{i=1}^s \alpha_i u_i = \mu(u.), \quad \text{say,}$$

and with a fixed variance σ^2 , independent of $u.$. The variance σ^2 will be called the residual variance. Here, the α and σ are unknown constants.

We shall use two alternative conditions, say (ii) and (iii), regarding the transformed precipitation from seeded storms. Then each combination, (i) with (ii) on the one hand and (i) with (iii) on the other, will serve to specify a separate problem of estimating the effects of seeding.

(ii) Nothing is assumed regarding $*U.$ and $*V$ except that, for each value $*u.$ of $*U.$, the variable $*V$ has a finite mean $E(*V|*u.)$, which may or may not equal $E(V|*u.)$ in (5).

(iii) For each possible set of precipitation amounts $*u.$ in the control, the transformed precipitation $*V$ in the target is normally distributed about a mean.

$$(6) \quad E(*V|*u.) = \beta_0 + \sum_{i=1}^s \beta_i *u_i = \mu(*u.), \text{ say,}$$

with a fixed variance $*\sigma^2$. Here the β represent constant coefficients which may or may not equal the α in (5). Also, the residual variance $*\sigma^2$ may but need not equal the residual variance σ^2 .

3. SPECIFICATION OF THE PROBLEM. As we have said, the problem of unbiased estimation of the effects of seeding has somewhat different specifications according to whether we assume condition (ii) or the more restrictive condition (iii).

The hypothesis basic for the evaluation is, of course, condition (i). For any preassigned conditions in which the transformed control precipitation has values $u. = (u_1, u_2, \dots, u_s)$ this assumption determines the distribution of the corresponding transformed target precipitation V to be observed without seeding. The expectation of this not-seeded precipitation in the target expressed in the original units is simply, say,

$$(7) \quad \theta(u.) = E\{f_0(V)|u.\} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} f_0(v) e^{-\frac{1}{2\sigma^2} [v-\mu(u.)]^2} dv.$$

If the coefficients α and the variance σ^2 were known, the expectation (7) could be computed easily. However, these constants are unknown and the best that can be done is to use the n not-seeded storms to obtain unbiased estimates of them. The formulas leading to these estimates are given in a number of textbooks, for example, in [1]. For the two most important cases $s=1$ and $s=2$, these formulas are reproduced in Section 6. Using the estimates $\hat{\alpha}_i$ of α_i we obtain the unbiased estimate

$$(8) \quad \hat{\mu}(u.) = \hat{\alpha}_0 + \sum_{i=1}^s \hat{\alpha}_i u_i$$

of (5). The estimate $\hat{\mu}(u.)$ is known to be normally distributed about $\mu(u.)$ and has variance $\sigma^2(u.) = \lambda^2(u.) \sigma^2$, where $\lambda^2(u.)$ is a known function of $u.$. In general, $\lambda^2(u.)$ is a quadratic in $u.$, which attains its minimum value of $1/n$ when $u.$ is equal to the average amounts of transformed control precipitation from not-seeded storms and then increases without limit when $u.$ diverges from these averages.

Also, the same familiar normal theory implies that the sum of squares of residuals

$$(9) \quad S^2 = \sum_{j=1}^n (v_j - \hat{a}_0 - \sum_{i=1}^s \hat{a}_i u_{ij})^2$$

is a statistic independent of $\hat{\mu}(u.)$ and, when divided by σ^2 , is distributed as χ^2 with $v = n-s-1$ degrees of freedom. As a result, the quotient S^2/σ^2 is an unbiased estimate of the residual variance.

As we shall see below, the basic problem of evaluating a cloud seeding experiment consists in using the two statistics $\hat{\mu}(u.)$ and S^2 , computable from data on not-seeded storms, in order to obtain an unbiased estimate of $\theta(u.)$ as defined by (7).

We turn now to the conditions (ii) and (iii) regarding the seeded storms. Condition (ii) does not imply any link between the seeded target precipitation from any two storms. Thus, for example, (ii) admits the possibility that the effect of seeding one type of storm may be positive and that of seeding another negative. Further, under condition (ii) it is possible to draw conclusions as to the seeded target precipitation only in the situations prevailing during the storms that actually were seeded. For this reason, under condition (ii), the evaluation of the experiment must be reduced to comparing the average actual seeded target precipitation and the average expected not-seeded target precipitation in the conditions of the m actual seeded storms. In other words, the quantity to be estimated is

$$(10) \quad \frac{1}{m} \sum_{j=1}^m E(*Y_j) - \frac{1}{m} \sum_{j=1}^m \theta(*u_j).$$

The only unbiased estimate of the first term in (10) is the simple average

$$(11) \quad \frac{1}{m} \sum_{j=1}^m *Y_j = *\bar{Y} .$$

In order to estimate the second term in (10), we need a general formula for estimating (7).

Under condition (iii) the situation is somewhat more flexible. Here the evaluation of the experiment need not be restricted to the amounts

observed in the storms actually seeded, which may happen to be atypical. For example, the evaluator may fix in advance arbitrary amounts of control precipitation u_* , perhaps representing "normal" amounts per storm observed over a number of years, and then estimate from the data the expected effect of seeding. Now the quantity to be estimated is the difference

$$(12) \quad *θ(u_*) - θ(u_*),$$

where $*θ(u_*)$ is defined like $θ(u_*)$ in formula (7), namely,

$$(7a) \quad *θ(u_*) = E\{f_0(*V)|u_*\} = \frac{1}{*\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} f_0(v) e^{-\frac{1}{2*\sigma^2} [v-*μ(u_*)]^2} dv.$$

The problem of estimating $*θ(u_*)$ is the same as that of estimating $θ(u_*)$ except that the data on seeded storms are used.

In conclusion, then, whether we adopt hypothesis (ii) or (iii), the problem of evaluating the experiment requires the formula for an unbiased estimate of a quantity of type (7).

4. TRADITIONAL SOLUTION AND ITS BIAS. The traditional method of estimating $θ(u_*)$ consists in computing $\hat{\mu}(u_*)$, the estimated expected transformed target precipitation without seeding, in conditions such that the transformed control precipitation is u_* , and then applying the backward transformation. In other words, the traditional estimate of $θ(u_*)$ is $f_0[\hat{\mu}(u_*)]$. We now show that, with the usual transforming functions, this estimate is biased. This means that its expectation is not equal to the quantity $θ(u_*)$ to be estimated. There is a systematic error so that even with an infinite number of observations the estimate would not equal the true value. Also, if a number of such estimates are averaged, the bias will not tend to average out.

It is well known that the conditional distribution of the precipitation Y in the target, given a specified amount of precipitation X_i in the i -th control, will generally have a variance that increases with an increase in X_i . One of the main purposes of the transformation of variables is to stabilize this conditional variance. There are many functions f^{-1} which will accomplish this; some are used in one case, some in another. But, in order to be useful, the functions f^{-1} must "shrink" the larger observations. Hence, the transforming function f must be concave. In other words, as illustrated in Figure 1, the graph of the transforming function f is a curve that lies entirely above a tangent straight line, no matter where on the curve this tangent is drawn.

Figure 1 shows three of the transformations used in the evaluation of cloud seeding experiments. The first panel corresponds to the square root transformation, that is, to the case $Y = f_0(V) = V^2$. The second panel corresponds to the logarithmic transformation so that

$Y = f_0(V) = 10^V = e^{kV}$, where k denotes the natural logarithm of 10.

Finally, the third panel corresponds to the equiprobability gamma transformation advocated by Thom [2]. Here $Y = f_0(V)$ is defined by the relation

$$(13) \quad \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^Y t^{\alpha-1} e^{-t/\beta} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^V e^{-t^2/2} dt,$$

where α and β are certain positive constants.

Now we show that the traditional estimate is biased. We begin by examining the relation between the quantities $\theta(u.)$ and $f_0[\mu(u.)]$. As formula (7) indicates, $\theta(u.)$ is the expectation of the variable $Y = f_0(V)$. Thus, $\theta(u.)$ is the weighted average of all possible values of $f_0(V)$, each value weighted by the probability of that value. For each value of V the corresponding value of $Y = f_0(V)$ is equal to the ordinate of a curve similar to those in Figure 1. The difficulty arises because V is not perfectly determined by the observed precipitation $u.$ in the controls; V is a random variable normally distributed with mean $\mu(u.)$.

Let us compare the tangent at the ordinate $f_0[\mu(u.)]$, corresponding to the abscissa $\mu(u.)$, with the curve itself. The equation of this tangent is, say,

$$(14) \quad Z(V) = f_0[\mu(u.)] + [V - \mu(u.)] f'_0[\mu(u.)].$$

The weighted mean of $Z(V)$, subject to the variability of V , is $f_0[\mu(u.)]$ since the mean of $V - \mu(u.)$ is zero. However, since the curve is above the tangent, the weighted mean, with the same weights, of its ordinates $f_0(V)$, namely $\theta(u.)$, must be greater than the mean $f_0[\mu(u.)]$ of the tangent.

Incidentally, the occurrence of the difference between the two means does not depend on the normality of V but persists irrespective of the weights used in averaging. For example, the simple arithmetic mean of the two numbers $x_1 = 2$ and $x_2 = 4$ is $\bar{x} = 3$. The square of this average is $\bar{x}^2 = 9$ while the average of the squares of the same numbers is $(2^2 + 4^2)/2 = 10$ which is larger than \bar{x}^2 . There is another conclusion which can be drawn heuristically from the above discussion: The greater the variability of V , that is, the greater its variance σ^2 , the greater the difference between $\theta(u.) = E[f_0(V)]$ and $f_0[\mu(u.)]$. This results from the simple remark that, with an increase in the variance σ^2 , values of V substantially different from the mean $\mu(u.)$ and, therefore, values for which the difference between $f_0(V)$ and $Z(V)$ is large will have greater weight. This is true in general. For example, the two numbers $x_3 = 1$ and $x_4 = 5$ have the same average $\bar{x} = 3$ as the numbers x_1 and x_2 above. However, x_3 and x_4 have a greater spread. As a result, the mean of their squares $(x_3^2 + x_4^2)/2 = 13$ is greater than the mean 10 of the squares of x_1 and x_2 .

COMMON TRANSFORMATIONS

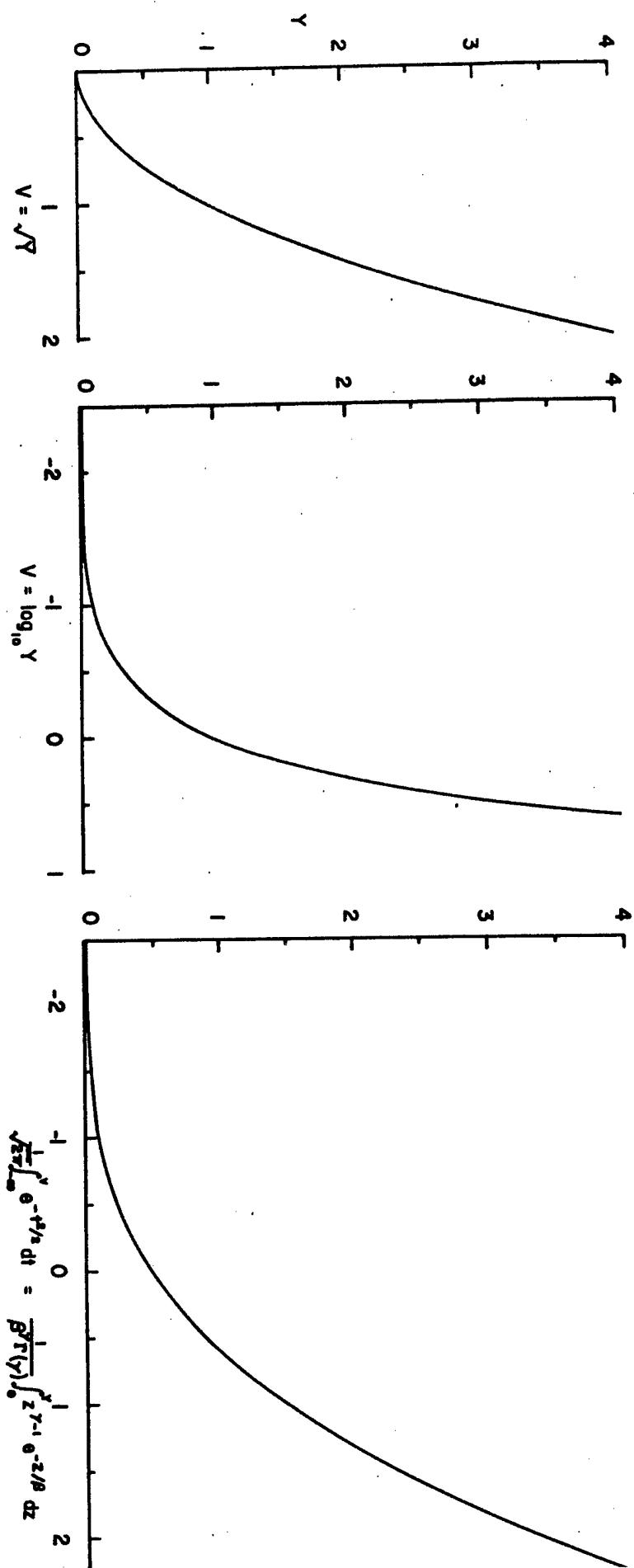


Figure 1

The two conclusions, (a) that the mean of a concave function is greater than the value of this function taken at the mean of its argument, and (b) that the difference between the mean of a concave function and the value of the junction at the mean increases with an increase in the variability of the argument, jointly explain the reason for the bias in the traditional estimate $f_0[\hat{\mu}(u.)]$.

As we have said, $\hat{\mu}(u.)$ is a normal variable with mean $\mu(u.)$ and variance $\lambda^2(u.)\sigma^2$. The mean of V is also $\mu(u.)$ while its variance is σ^2 . Ordinarily $u.$ will not differ very much from the averages of the transformed control precipitation in not-seeded storms. In this case, the value of $\lambda^2(u.)$ is less than unity and so the variance of $\hat{\mu}(u.)$ will be less than that of V . Consequently,

$$(15) \quad f_0[\mu(u.)] < E\{f_0[\hat{\mu}(u.)]\} < E\{f_0(V)|u.\} = \theta(u.),$$

so that $f_0[\mu(u.)]$ will systematically underestimate $\theta(u.)$. On the contrary, if $u.$ is far away from the average transformed amounts of control precipitation of not-seeded storms, then $\lambda(u.) > 1$, and we have

$$(16) \quad f_0[\mu(u.)] < E\{f_0(V)|u.\} = \theta(u.) < E\{f_0[\hat{\mu}(u.)]\},$$

so that the traditional estimate $f_0[\mu(u.)]$ will tend to overestimate $\theta(u.)$ in this case.

5. GENERAL FORMULAS FOR THE UNBIASED ESTIMATE OF $\theta(u.)$. In a recent paper [3] we considered in some detail the problem of a minimum variance unbiased estimate of a quantity of the type of $\theta(u.)$. In the present section we give without proof two general formulas applicable to a broad class of transforming functions f_0 . Before proceeding to details we interpose two remarks.

Remark 1. As we saw in Section 4, the expected value of V , namely $\mu(u.)$, does not by itself determine the value of $\theta(u.)$; the variance σ^2 plays an important role. As a result, it is clear that an unbiased estimate of $\theta(u.)$ cannot depend solely on the estimate $\hat{\mu}(u.)$ of $\mu(u.)$ but must involve the statistic S^2 which serves to estimate σ^2 . Consequently, any unbiased estimate of $\theta(u.)$ will be a function, say $\hat{\theta}[\hat{\mu}(u.), S^2]$ of the two arguments $\hat{\mu}(u.)$ and S^2 .

Remark 2. The two statistics $\hat{\mu}(u.)$ and S^2 form a so-called "sufficient" system for the parameters $\hat{\mu}(u.)$ and σ^2 . Also, this sufficient system is "boundedly complete." According to a theorem of Lehmann and Scheffé [4], if a function of these two statistics is an unbiased estimate of $\theta(u.)$ then it is necessarily the minimum variance unbiased estimate and is the unique estimate of this kind. Consequently, each of the formulae below giving $\hat{\theta}[\hat{\mu}(u.), S^2]$ is the expression for the minimum variance unbiased estimate of the corresponding $\theta(u.)$. Any other formula would either be equivalent to that produced, or give a biased estimate, or give an unbiased estimate with greater variance.

The first formula is of somewhat greater generality than the second. In addition, it is easier to apply for certain types of transforming

functions, although more difficult for other types. In order that the first formula be applicable, it is necessary and sufficient that the transforming function f_0 have derivative of all orders and that the two series

$$(17) \quad \sum \frac{1}{n!} f^{(2n)}(0)z^n \quad \text{and} \quad \sum \frac{1}{n!} f^{(2n+1)}(0)z^n$$

have infinite radii of convergence. If these conditions are satisfied, the minimum variance unbiased estimate of $\theta(u.)$ is given by

$$(18) \quad \hat{\theta}[\hat{\mu}(u.), s^2] = f_0(0) + \sum_{n=1}^{\infty} f_0^{(n)}(0)T_n,$$

where

$$(19) \quad T_{2n} = \sum_{k=0}^n \frac{(2n)!}{(2k)!(n-k)!} \hat{\mu}^{2k}(u.) \left\{ \frac{s^2}{4} [1 - \lambda^2(u.)] \right\}^{n-k} \frac{\Gamma(\frac{3}{2})}{\Gamma(\frac{3}{2} + n - k)}$$

and

$$(20) \quad T_{2n+1} = \sum_{k=0}^n \frac{(2n+1)!}{(2k+1)!(n-k)!} \hat{\mu}^{2k+1}(u.) \left\{ \frac{s^2}{4} [1 - \lambda^2(u.)] \right\}^{n-k} \frac{\Gamma(\frac{3}{2})}{\Gamma(\frac{3}{2} + n - k)}.$$

Although formula (18), combined with (19) and (20), looks complicated, it is easy to apply to some of the transforming functions advocated by Tukey [5]. These functions are of the form

$$(21) \quad f_0(v) = v^p - a,$$

where a is a fixed number and p a positive integer. It will be seen that at $v = 0$ all the derivatives of (21) vanish with the exception of

$$(22) \quad f_0^{(p)}(0) = p! .$$

Consequently, formula (18) reduces to

$$(23) \quad \hat{\theta}[\hat{\mu}(u.), s^2] = -a + T_p,$$

where T_p has the form (19) or (20) according to whether p is even or odd.

The square root transformation is a particular case of (21), with $p = 2$ and $a = 0$. In this case, then, after simplification,

$$(24) \quad \hat{\theta}[\hat{\mu}(u.), s^2] = T_2 = \hat{\mu}^2(u.) + \frac{s^2}{4} [1 - \lambda^2(u.)] .$$

The second general formula for the unbiased estimate $\hat{\theta}[\hat{\mu}(u.), s^2]$ was deduced for a special category of transforming functions, which we call recursive. The formula is easily applicable to these functions, which are rather common. These functions are characterized by the

differential equation of the second order which they must satisfy, namely,

$$(25) \quad f''(v) = A + Bf(v) ,$$

where A and B are two constants. We assume that at least one of these constants must be different from zero; otherwise f would be a linear function in which case it would not stabilize the residual variance.

If the transforming function f is recursive in the above sense, and $B \neq 0$, then the minimum variance unbiased estimate of $\theta(u.)$ is given by

$$(26) \quad \hat{\theta}[\hat{\mu}(u.), s^2] = \Phi\{B[1 - \lambda^2(u.)]s^2, v\}\{f[\hat{f}(u.)] + \frac{A}{B}\} - \frac{A}{B} ,$$

where, generally,

$$(27) \quad \Phi(x, v) = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{1}{2}+n)} \cdot \left(\frac{x}{4}\right)^n$$

$$= 1 + \frac{x/2}{1!v} + \frac{(x/2)^2}{2!v(v+2)} + \frac{(x/2)^3}{3!v(v+2)(v+4)} + \dots .$$

When $B = 0$, which is when f is a quadratic function, formula (26) reduces to

$$(28) \quad \hat{\theta}[\hat{\mu}(u.), s^2] = f[\hat{\mu}(u.)] + \frac{AS^2}{2v} [1 - \lambda^2(u.)] .$$

For the simple square root transformation $A = 2$ and formula (28) coincides with (24).

One might regret that formula (26) involves the infinite series (27). However, in several examples computed by the authors, this series happens to converge very rapidly so that a satisfactory approximation was reached by computing its first two or three terms.

The transformations that are currently most useful in meteorological work are the square root and the logarithmic transformations. They appear to work satisfactorily with precipitation and runoff data, one or the other having a slight advantage, depending upon the circumstances. However, with the growth of experimentation on weather control, certain other transformations are likely to become useful. For example, in the evaluation of experiments with lightning and/or hail prevention the arcsine and the archyperbolic-sine transformations may prove efficient in transforming the frequencies of the relevant events so as to satisfy the conditions of normal tests. For these reasons, we apply the general formulae to deduce the unbiased estimate $\hat{\theta}[\hat{\mu}(u.), s^2]$ appropriate to each of the several cases. For the sake of completeness, we reproduce first the formulae for $\hat{\mu}(u.)$, $\lambda^2(u.)$ and s^2 ,

for the two most important cases of $s = 1$ and $s = 2$. These are known formulae, from the theory of least squares.

6. SOME FORMULAE FROM THE THEORY OF LEAST SQUARES. Although the theory of least squares was developed some 150 years ago, particularly by Gauss [6], and although the relevant formulas are continuously being "developed" anew, we reproduce them here for the sake of completeness of the present paper. Particularly, we consider the two most important cases where the evaluation of the experiment is based on $s = 1$ and on $s = 2$ control areas. We deal with the method of estimating the transformed target precipitation expected to fall from a non-seeded storm when the transformed control precipitation has preassigned values: $u_+ = u_1$ in the case $s = 1$ and $u_+ = (u_1, u_2)$ in the case $s = 2$. Thus, all the operations indicated are to be performed on the data (u_j, v_j) referring to the observed non-seeded storms. If we decide to accept assumption (iii) and it is necessary to calculate a similar estimate for seeded storms, then the calculations indicated by the formulas will have to be performed on data $(*u_j, *v_j)$, referring to the observed seeded storms.

All of the relevant formulae are expressed most conveniently in terms of sample means and of sample variances and covariances. Given some n numbers a_1, a_2, \dots, a_n , the symbol \bar{a} will denote their arithmetic mean and the symbol s_a^2 their sample variance,

$$(29) \quad s_a^2 = \frac{1}{n} \sum (a_j - \bar{a})^2 = \frac{1}{n} \left[n \sum a_j^2 - (\sum a_j)^2 \right] .$$

where all the summations extend over $j = 1, 2, \dots, n$. For a given group of n pairs of numbers (a_j, b_j) , the sample covariance is defined by

$$(30) \quad s_{ab} = \frac{1}{n} \sum (a_j - \bar{a})(b_j - \bar{b}) = \frac{1}{n} \left[n \sum a_j b_j - \sum a_j \sum b_j \right] .$$

The sample variances of the transformed amounts of precipitation in the two controls will be denoted by s_1^2 and s_2^2 , respectively. The sample variance of the transformed precipitation in the target will be denoted by s_v^2 . The three covariances will be denoted s_{12} , s_{1v} and s_{2v} . Further, \bar{u}_1 , \bar{u}_2 and \bar{v} will indicate the observed mean amounts of rain, in transformed units.

We consider first the case $s = 1$ of one control area, say the first control area. If some other control area is the unique area, the formulae for it may be written by analogy. Suppose that the particular preassigned

transformed control precipitation, for which we want to estimate the expected precipitation in the target, is designated simply as u_1 . Then we have

$$(31) \quad \hat{\mu}(u.) = \bar{v} + \frac{s_{1v}}{s_1^2} (u_1 - \bar{u}_1).$$

Further

$$(32) \quad \lambda^2(u.) = \frac{1}{n} \left[1 + \frac{(u_1 - \bar{u}_1)^2}{s_1^2} \right].$$

It is clear that $\lambda^2(u.)$ is quadratic in u_1 , attains its minimum of $1/n$ when $u_1 = \bar{u}_1$, and grows without limit as u_1 diverges from \bar{u}_1 . Finally

$$(33) \quad s^2 = \sum \left[v_j - \bar{v} - \frac{s_{1v}}{s_1^2} (u_{.j} - \bar{u}_{.}) \right]^2 \\ = n \left[s_v^2 - \frac{s_{1v}^2}{s_1^2} \right].$$

The corresponding number of degrees of freedom is $v = n-2$.

In the case of $s = 2$ control areas,

$$(34) \quad \hat{\mu}(u.) = \hat{\mu}(u_1, u_2) = \bar{v} + \alpha_1(u_1 - \bar{u}_1) + \alpha_2(u_2 - \bar{u}_2),$$

where

$$(35) \quad \alpha_1 = \frac{1}{\Delta} (s_2^2 s_{1v} - s_{12} s_{2v})$$

$$(36) \quad \alpha_2 = \frac{1}{\Delta} (s_1^2 s_{2v} - s_{12} s_{1v}),$$

with

$$(37) \quad \Delta = s_1^2 s_2^2 - s_{12}^2.$$

Further,

$$(38) \quad \lambda^2(u.) = \lambda^2(u_1, u_2) = \frac{1}{n} \left\{ 1 + \frac{1}{\Delta} s_2^2 \left[(u_1 - \bar{u}_1)^2 - 2s_{12}(u_1 - \bar{u}_1)(u_2 - \bar{u}_2) + s_1^2 (u_2 - \bar{u}_2)^2 \right] \right\}$$

$$= \frac{1}{n} \left\{ 1 + \frac{(u_1 - \bar{u}_1)^2}{s_1^2} + \frac{s_1^2}{\Delta} \left[(u_2 - \bar{u}_2) - \frac{s_{12}}{s_1} (u_1 - \bar{u}_1) \right]^2 \right\}$$

The second line of formula (38) indicates that, in this case also, the minimum value $1/n$ of $\lambda^2(u.)$ is attained when $u_1 = \bar{u}_1$ and $u_2 = \bar{u}_2$.

Further, as u_1 diverges from \bar{u}_1 and u_2 diverges from \bar{u}_2 , the value of $\lambda^2(u.)$ grows indefinitely. Finally, the sum of squares of residuals is

$$(39) \quad S^2 = n \left\{ s_v^2 - \frac{1}{\Delta} \left[s_2^2 s_{1v}^2 - 2s_{12} s_{1v} s_{2v} + s_1^2 s_{2v}^2 \right] \right\}.$$

Here, the number of degrees of freedom is $v = n-3$.

7. UNBIASED ESTIMATE OF EXPECTED PRECIPITATION IN THE TARGET WHEN THE SQUARE ROOT TRANSFORMATION IS USED. As already indicated, the minimum variance unbiased estimate $\hat{\theta}[\lambda(u.), S^2]$ of expected precipitation in the target without seeding, when the transformed precipitation in the control has values $u. = (u_1, u_2, \dots, u_s)$ is given by formula (24). When the number of controls is either $s = 1$ or $s = 2$, the quantities $\lambda(u.)$, $\lambda^2(u.)$ and S^2 appearing in this formula are calculated from (31), (32) and (33) in case $s = 1$, and from (34) to (39) when $s = 2$. Only one further detail need be added.

If we want to evaluate the experiment under condition (iii) and to estimate the expectation of the extra precipitation ascribable to seeding that may have fallen in the target on the average, during all the m storms that were actually seeded, as indicated in formula (10), then, generally, this involves separate calculation of distinct estimates $\hat{\theta}[\hat{\mu}(*u_j), S^2]$ of $\theta(*u_j)$, where $j = 1, 2, \dots, m$, and then computing their arithmetic mean, say

$$(40) \quad \hat{\theta}_0 = \frac{1}{m} \sum_{j=1}^m \hat{\theta}[\hat{\mu}(*u_j), S^2].$$

However, with the square root transformation, certain shortcuts are possible which might be worthwhile. In fact, using formula (24), we have

$$(41) \quad \hat{\theta}_0 = \frac{1}{m} \sum_{j=1}^m \left[\hat{\mu}(*u_j) \right]^2 + \left[1 - \frac{1}{m} \sum_{j=1}^m \chi^2(*u_j) \right] \frac{s^2}{v} .$$

If there is just one control and, thus $s = 1$, then simple algebra reduces (41) to

$$(42) \quad \hat{\theta}_0 = \left[\bar{v} + \frac{s_{12}}{s_1^2} (*\bar{u}_1 - \bar{u}_1) \right]^2 + \frac{s_{12}^2 * s_1^2}{s_1^4} + \left\{ 1 - \frac{1}{n} \left[1 + \frac{*s_1^2 + (*\bar{u}_1 - \bar{u}_1)^2}{s_1^2} \right] \right\} \frac{s^2}{v} ,$$

where, as formerly, the asterisk on the left of a symbol indicates that the value of this symbol is to be calculated for the seeded storms. Thus, for example $*s^2$ means the sample variance of the m transformed precipitation amounts deposited by the seeded storms in the control area.

If $s = 2$, so that there are two control areas, then, by a similar procedure,

$$(43) \quad \hat{\theta}_0 = \left[\bar{v} + \hat{\alpha}_1 (*\bar{u}_1 - \bar{u}_1) + \hat{\alpha}_2 (*\bar{u}_2 - \bar{u}_2) \right]^2 + \hat{\alpha}_1^2 * s_1^2 + 2\hat{\alpha}_1 \hat{\alpha}_2 * s_{12} \\ + \hat{\alpha}_2 * s_2 + \left\{ 1 - \frac{1}{n} \left[1 + \frac{1}{\Delta} \left[s_2^2 (*s_1^2 + (*\bar{u}_1 - \bar{u})^2) \right. \right. \right. \\ \left. \left. \left. - 2s_{12} (*s_{12} + (*\bar{u}_1 - \bar{u}_1)(*\bar{u}_2 - \bar{u}_2)) + s_1^2 (*s_2^2 + (*\bar{u}_2 - \bar{u}_2)^2) \right] \right] \right\} \frac{s^2}{v} ,$$

where $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are the estimates (35) and (36) of the partial regression coefficients of V on U_1 and U_2 , respectively.

Remark. It is interesting to notice that the formulae given here referring to the square root transformation do not require that the transformed variables be normally distributed. The only conditions of their validity is that the regression of the transformed target rain on the transformed control rain be linear and that the residual variance be constant.

8. UNBIASED ESTIMATE OF EXPECTED PRECIPITATION IN THE TARGET WHEN THE LOGARITHMIC TRANSFORMATION IS USED. If the normalizing transformation is logarithmic, then the transforming function is

$$(44) \quad Y = f_0(V) = 10^V = e^{kV}$$

with k denoting the natural logarithm of 10. It is easy to see that this function belongs to the category we call recursive. In fact, differentiating (44) twice with respect to V we obtain

$$(45) \quad f_0''(V) = k^2 e^{kV} = k^2 f_0(V).$$

With reference to (25) it follows that in this case $A = 0$ and $B = k^2$. Thus, from (26), the minimum variance unbiased estimate of $\theta(u.)$ is given by

$$(46) \quad \hat{\theta}[\hat{\mu}(u.), s^2] = \phi\{[1 - \lambda^2(u.)] k^2 s^2, v\} 10^{\hat{\mu}(u.)}.$$

Unfortunately, in this case no shortcut exists similar to that available for the square root transformation and, if it is desired to estimate the average expected increase in rain from m seeded storms, formula (46) has to be computed for each of these storms separately. Unpleasantly, this will involve the evaluation of the series (27) m times, for m different values of the independent variable.

9. THE ANGULAR TRANSFORMATION OF FREQUENCIES. If an experiment is reduceable to the observation of several sets of units, each set containing the same number n of units and if the observations yield numbers Y of those units which are distinguished by the presence of certain specific characteristics, then a satisfactory "normalization" of the data is occasionally achieved by the so-called angular transformation

$$(47) \quad V = \arcsin \sqrt{\frac{Y+c}{n+2c}}.$$

Anscombe [7] suggests the value of $c = 0.3$ or 0.4 . With reference to weather control experimentation, n may mean the number of seeded storms in a locality to be compared with an equal number of non-seeded storms. Y may mean the number of those storms in either group which are accompanied by hail.

The transforming function corresponding to (47) is

$$(48) \quad f_0(V) = (n+2c) \sin^2 V - c = \frac{1}{2} [n - (n+2c) \cos 2V].$$

Differentiating (48) twice and performing easy transformations, it is found that

$$(49) \quad f_0''(V) = 2(n+2c) \cos 2V = 2n - 4f_0(V)$$

It follows that the transforming function (48) is of the recursive type, with $A = 2n$ and $B = -4$. Consequently, the minimum variance

unbiased estimate of $\theta(u.)$ is given by the second general formula (26) which reduces to

$$(50) \quad \hat{\theta}[\hat{\mu}(u.), s^2] = \phi\{-4[1 - \lambda^2(u.)]s^2, v\}\{f[\hat{\mu}(u.)] - n/2\} + n/2 \\ = \frac{1}{2}[n - (n+2c)\phi\{-4[1 - \lambda^2(u.)]s^2, v\}\cos 2\hat{\mu}(u.)]\dots$$

10. HYPERBOLIC ARCSINE TRANSFORMATION. For integer valued variables X having a rather skew distribution, the hyperbolic arcsine transformation, first used by Beall, might be useful. In particular, this transformation might be applicable to the normalization of such variables as the number X of lightning strikes per storm. We will write it in the form

$$(51) \quad V = \operatorname{arcsinh} \sqrt{X} .$$

The corresponding transforming function is

$$(52) \quad X = f_0(V) = [\sinh V]^2 = \frac{1}{4}[e^{2V} + e^{-2V} - 2] .$$

Differentiating (52) twice and performing easy calculations, we find

$$(53) \quad f_0''(V) = 2 + 4f_0(V) .$$

Hence, (52) is a recursive function with $A = 2$ and $B = 4$ and, according to (26),

$$(54) \quad \hat{\theta}[\hat{\mu}(u.), s^2] = \frac{1}{2}[\phi\{4[1 - \lambda^2(u.)]s^2, v\}\cosh 2\hat{\mu}(u.) - 1] .$$

REFERENCES

- 1 A. M. Mood, Introduction to the Theory of Statistics, McGraw-Hill, New York, 1950.
- 2 H. C. S. Thom, "A statistical method of evaluating augmentation of precipitation by cloud seeding," Final Report of the Advisory Com. on Weather Control, Vol. 2 (1957), pp. 5-25.
- 3 J. Neyman and E. L. Scott, "Correction for bias introduced by a transformation of variables," Annals Math. Stat., Vol. 31 (1960), to appear.
- 4 E. L. Lehmann and H. Scheffé, "Completeness, similar regions and unbiased estimation, Part I," Sankhyā, Vol. 10 (1950), pp. 305-340.
- 5 J. W. Tukey, "On the comparative anatomy of transformations," Annals Math. Stat., Vol. 28 (1957), pp. 602-632.
- 6 C. F. Gauss, Abhandlungen zur Methode der kleinsten Quadrate, Berlin, 1887. Transl. from Latin by A. Borsch and P. Simon.
- 7 F. J. Anscombe, "The transformation of Poisson, binomial and negative-binomial data," Biometrika, Vol. 35 (1948), pp. 246-254.

MATHEMATICAL AND STATISTICAL PRINCIPLES UNDERLYING
CHEMICAL CORPS INSPECTION PROCEDURES FOR PRODUCT VERIFICATION

Henry Ellner
U.S. Army Chemical Center and Chemical Corps
Materiel Command, Directorate for Quality Assurance

1. INTRODUCTION. In accordance with Standardization Manual M205, issued 9 April 1958 by DOD, each military specification is required to include in the beginning of the quality assurance provisions, section 4, the following statement:

"Unless otherwise specified herein the supplier is responsible for the performance of all inspection requirements prior to submission for Government inspection and acceptance. Except as otherwise specified, the supplier may utilize his own facilities or any commercial laboratory acceptable to the Government. Inspection records of the examinations and tests shall be kept complete and available to the Government as specified in the contract or order."

The contractor is thus obliged to assure himself that supplies he offers for acceptance conform with contractual requirements. The Army inspector, under AR 715-20* is enjoined to verify the inspection performed by the contractor and to establish the reliability of the contractor's inspection records prior to acceptance of the submitted supplies. Limited product inspection by the Government inspector and review of the supplier's inspection system are the means prescribed for authenticating the contractor's inspection data.

As stated by Standardization Manual M205, sampling is an important factor in determining compliance with requirements. While details of sampling will vary with the commodities, where applicable, MIL-STD-105 (Sampling Procedures and Tables for Inspection by Attributes) is referenced in specifications. In conjunction with MIL-STD-105, the quality assurance provisions of the specification include one or more classifications of defects. Now using the acceptable quality level (AQL) in the specification as an index to the plans of MIL-STD-105, the contractor is obligated to sample and inspect as prescribed. Sampling inspection is also the modus operandi of the Government inspector for verifying the inspection data recorded by the contractor.

2. VERIFICATION BY SAMPLING. The Government cannot expect its verification data to duplicate the contractor's inspection results exactly since the number of defectives in each sample is a variate dependent upon the true (but unknown) percent defectives in the lot sampled. Chance can, therefore, be responsible for the difference in the proportions of defectives observed in any two samples compared. Wide discrepancies in results may be due to non-random (biased) samples or to failure to recognize a defective as spelled out in the classification of defects. The problem is to set up criteria so that disparities arising by chance alone are differentiated, considering costs and consequences, from disparities arising from improper inspection practices. Furthermore, any general

* Policy background pertaining to product verification inspection is discussed in Appendix 1. This Appendix is under the authorship of Joseph Mandelson.

Design of Experiments

tendency for significant discrepancies to arise must be recognized by frequent testing so that appropriate action can be taken to safeguard Government interests.

When the problem is as stated above the purpose and procedures for accomplishing verification sampling are conceptually simplified. A decision whether to accept or reject a lot, in accordance with acceptance criteria of a sampling plan of the single, double or multiple type selected from MIL-STD-105, is not involved. The decision as to reliability of contractor inspection results is distinct from the decision to accept or reject a lot, although the latter decision may be contingent upon the former. Verification sampling inspection then has as its primary purpose the establishment of the validity of the contractor's inspection data by checking his sampling results against independent sampling inspections. The size of the sample required and the frequency of performing verification sampling depend upon the power of the test to detect significant differences between the paired samplings and the establishment of an objective degree of rational belief in the existence of a state of statistical control over the fluctuating differences observed. This belief can be bolstered by evaluation of the contractor's inspection system and by independent assessment of the quality of product offered to the Government for acceptance.

3. HOMOGENEITY OF ATTRIBUTE SAMPLING DATA. As defined in Supply and Logistics Handbook H105, in inspection by attributes the unit of product is classified simply as defective or nondefective with respect to a given requirement or a set of requirements. The requirement may be an individual checkpoint and the set may be a group of characteristics of equal importance listed under a single AQL in the specification. In the following development, we shall assume that, even when measurement along a continuous numerical scale is possible, such measurement will be classified as conforming or non-conforming with specification limits.

Let us now suppose that the contractor has drawn a random sample in accordance with MIL-STD-105 from an inspection lot and has noted the number of conforming and non-conforming items in the initial sample. The Government inspector has proceeded likewise by selecting an independent sample from the same lot. In the analyses that follow, we shall assume that the lot size is large relative to the total sample size (say, at least 8:1); or, if the lot size is proportionately small, that the samples are drawn without replacement until a comparison has been made. The results of inspection are denoted symbolically in a 2 X 2 table as below:

TABLE 1.
Notation for 2 X 2 Contingency Table

	Defective	Nondefective	Total
Contractor's Sample	d_c	$n_c - d_c$	n_c
Government's Sample	d_g	$n_g - d_g$	n_g
Total	d_t	$n_t - d_t$	n_t

The data are recorded to decide whether the results of inspecting two samples, one the size n_c and the other the size n_g , which are found to contain d_c and d_g defectives, respectively, are significantly different. A common test of significance, for attribute data classified in two ways as shown, is the chi-square test [13] and equivalent alternates. When the expected number of defectives is small, say less than five, Fisher's exact test ([9], Section 21.02) is generally advised. For routine testing these techniques all involve extensive computation, and consequently are not suitable for verification purposes. Short cut procedures [8, 11, 12, 16] devised to meet this problem, including nomograms and extensive tabulations of Fisher's "exact" test, are likewise wanting in that multiple entries are necessary or that tables required are too lengthy and numerous.

A test for homogeneity, applicable when the overall proportion of defectives d_t/n_t is small, say 0.20 or less, is one which compares samples from populations known to give the Poisson type of distribution. Przyborowski & Wilenski [15] considered two observations (in our notation: d_c and d_g) originating from two Poisson-distributed populations with unknown means, and for the symmetrical case $n_c = n_g$ they proposed an "exact" test for the equality of these means. Barnard [1] extended their method to the case $n_c \neq n_g$ reducing the procedure to a simple test for the variance - ratio F. Bross and Kasten [5] derived a related technique for the case $n_c \neq n_g$ and published charts for avoiding or greatly reducing computations for the analysis of fourfold contingency tables. What was apparently a very different test from Barnard's was proposed by Cox [6], but their similarity has been shown by Barton [2]. However, Cox's method has certain advantages over Barnard's which make it preferable for use in product verification.

4. "EXACT" TESTS OF SAMPLES FROM TWO POISSON SERIES. Before the advantages of Cox's method can be discussed it will be necessary to derive the "exact" test for comparing two Poisson-distributed observations. Suppose d_c and d_g of Table 1 approximately follow independent Poisson distributions so that:

$$(1) \quad P(d_c, d_g | p'_c, p'_g) = P(d_c) \cdot P(d_g) = \frac{e^{-p'_c n_c} (p'_c)^{d_c}}{d_c!} \cdot \frac{e^{-p'_g n_g} (p'_g)^{d_g}}{d_g!},$$

where:

p'_c = the expected fraction defective in the contractor's sample n_c

p'_g = the expected fraction defective in the Government inspector's sample n_g .

Under the null hypothesis $p'_c = p'_g = p'_o$ so that Equation (1) reduces to:

$$(2) \quad P(d_c, d_g \mid p'_o) = \frac{e^{-p'_o(n_c+n_g)} (p'_o)^{d_t} \frac{d_t}{n_c} \frac{d_c}{n_c} \frac{d_g}{n_g}}{d_c! d_g!},$$

which can be rewritten as:

$$(3) \quad P(d_c, d_g \mid p'_o) = P(d_g \mid d_t) P(d_t \mid p'_o)$$

$$= \frac{d_t! \frac{d_c}{n_c} \frac{d_g}{n_g}}{d_c! d_g! (n_c+n_g)^{d_c} (n_c+n_g)^{d_g}} \cdot \frac{e^{-p'_o(n_c+n_g)} (p'_o n_c + p'_o n_g)^{d_t}}{d_t!}$$

But we need the probability of getting some pair of results having the same total $d_c + d_g = d_t$; and so the relative probability, on the null hypothesis, of getting the pair (d_c, d_g) out of all results with the same total d_t is:

$$(4) \quad P(d_g \mid d_t) = \frac{P(d_g \mid d_t) P(d_t \mid p'_o)}{P(d_t \mid p'_o)}$$

$$= \frac{d_t!}{d_c! d_g!} \left(\frac{n_g}{n_c + n_g} \right)^{d_g} \left(\frac{n_c}{n_c + n_g} \right)^{d_c}$$

If we let $r = \frac{n_c}{n_g}$ then:

$$(5) \quad P(d_g \mid d_t) = \frac{d_t!}{d_c! d_g!} \left(\frac{1}{1+r} \right)^{d_g} \left(\frac{r}{1+r} \right)^{d_c}$$

We note that conditionally on d_t , d_g is binomially distributed with parameters $\frac{1}{1+r}$ and d_t , which can be used as the basis for a significance test. Accordingly:

$$(6) \quad F(y) = \sum_{y=d_g}^{d_t} \binom{d_t}{y} \left(\frac{1}{1+r} \right)^y \left(\frac{r}{1+r} \right)^{d_t-y} = I_{\frac{1}{1+r}}(d_g, d_c + 1),$$

where $I_x(p, q)$ is the incomplete β -function representation of a sum of binomial probabilities.

If the only admissible alternative to the null hypothesis $p'_c = p'_g = p'_0$ is $p'_g > p'_c$ then the appropriate critical region, in the Neyman-Pearson sense, for rejection of the null hypothesis is defined by $d_g \leq k_1 (d_t, \alpha)$ or $d_g \geq k_2 (d_t, \alpha)$,

where α is the risk of the first kind of error and where

$$(7) \quad P \left\{ d_g \geq k_2 (d_t, \alpha) \mid d_t, p'_c = p'_g \right\} \leq \alpha .$$

For the "exact" test this may be expressed by:

$$(8) \quad I_{\frac{1}{1+r}}^{(d_g, d_c + 1)} \leq \alpha .$$

This inequality may be written in terms of the probability distribution function $P_{f_1, f_2}(F)$ of the F distribution with (f_1, f_2) degrees of freedom since:

$$P_{f_1, f_2}(F) = I_x(p, q)$$

where $f_1 = 2q$, $f_2 = 2p$ and $F = \frac{p}{q} \frac{1-x}{x}$ with the result that

$$(9) \quad P_{2d_c+2, 2d_g} \left(\frac{r d_g}{d_c + 1} \right) \leq \alpha .$$

Inequalities (8) and (9) establish a level of significance which does not exceed α . The true level of significance depends upon the unknown p'_0 and may in some cases for small (d_c, d_g) be considerably less than α .

5. COX'S "APPROXIMATE" TESTS FOR POISSON VARIATES. In inverse Poisson sampling, with d fixed, the number of sample items n drawn in sequence up to the d th event is distributed as $(2p')^{-1} \chi^2_{2d}$, where χ^2_{2d} denotes a chi-square variate with $2d$ degrees of freedom and p' represents the true rate. For direct Poisson sampling in which the number of events d occurring in a fixed n is observed, we have

$$(10) \quad P(x \geq d) = \sum_{x=d}^{\infty} \frac{e^{-p' n} (p' n)^x}{x!} = P\left(\frac{1}{2p'} \chi^2_{2d} \leq n\right), \text{ and}$$

$$(11) \quad P(x \geq d+1) = P\left(\frac{1}{2p'} \chi^2_{2d+2} \leq n\right) .$$

Cox suggested an approximation to $P(x > d)$ in which d is treated as a continuous variate by taking a quantity intermediate between (10) and (11):

$$(12) \quad P(x > d) \approx P\left(\frac{1}{2p} \chi^2_{2d+1} \leq n\right),$$

which implies that probabilities are calculated as if

$$(13) \quad 2p'n \text{ is distributed as } \chi^2_{2d+1}$$

When two populations with fraction defectives p'_c, p'_g are compared by means of samples n_c, n_g which exhibit d_c, d_g defectives, then, from (12) we compute the ratio:

$$(14) \quad \frac{2p'_c n_c}{2d_c + 1} \div \frac{2p'_g n_g}{2d_g + 1}$$

which is distributed approximately as F with $(2d_c + 1, 2d_g + 1)$ degrees of freedom. Thus, we may test the hypothesis that $p'_c = p'_g = p_0'$ against the alternate hypothesis that $p'_g > p'_c$ by referring

$$(15) \quad F = r \frac{(d_c + 0.5)}{\frac{d_g + 0.5}{r}}$$

to the F tables with $(2d_c + 1, 2d_g + 1)$ degrees of freedom for the appropriate α percent point.

This may be represented by

$$(16) \quad P_{2d_c + 1, 2d_g + 1} \left(r \frac{d_g + 0.5}{d_c + 0.5} \right) \leq \alpha$$

or

$$(17) \quad I \frac{1}{1+r} (d_g + 0.5, d_c + 0.5) \leq \alpha$$

It is now clear that the "exact" tests given by (8) and (9) have been modified slightly to yield the approximate tests of (16) and (17). The modification has the effect of making the true level of significance less dependent upon the unknown p_0' and to approximate the nominal value of α when averaged over d_t .

6. POWER FUNCTION OF TESTS FOR POISSON VARIATES. The Neyman-Pearson theory of tests considers all tests of the same size and lays down objective standards for selecting the best test. The theory introduces the term, "power of a test," relative to the alternate hypothesis, to denote the probability of correctly rejecting the null hypothesis when an alternative is true. Of all tests at a given significance level, the most preferred is the one which has the maximum power relative to all the alternate hypothesis considered. The probability of rejecting the null hypothesis H_0 , regarded as a function of H' , where H' is any of the admissible alternates to H_0 , is called the power function of the test. If we commence with the determination of the critical region subject to (7) we can calculate the power function of a given test of significance. Thus, for the "exact" test all points satisfying (8) or (9) are entered in (1) and the absolute probabilities are summed. Similarly, for the "approximate" test all points satisfying (16) or (17) are entered in (1) for addition of the absolute probabilities. Tables 2 and 3 provide the actual probabilities associated with the respective tests for a one-sided test of the null hypothesis $p_g^r = p_c^r$ against the alternatives $p_g^r = 3p_c^r$ and $p_g^r = 4.5 p_c^r$ for $r = 1, 2, 3, 5$ and 8 , respectively, over a range of nuisance parameters, $p_c^r n_c^r$, which may be encountered in practice.

The arrangement of Tables 2 and 3 clearly reveals that the significance level α is a function of the expected number of defectives in the contractor's sample and the ratio of the contractor's sample size to the size of the Government's verification sample. For the "exact" test, under the null hypothesis, $p_g^r = p_c^r$, the quantity α increases about tenfold on the average as $p_c^r n_c^r$ increases from 0.75 to 9.00 . In contrast, for the "approximate" test, α increases only 1.5 times on the average over the same range of $p_c^r n_c^r$. Furthermore, the average level of significance of the thirty entries summed over the five tabular values of r for the "exact" and "approximate" tests are 0.015 and 0.052 , respectively. The conclusion is that the "approximate" test more effectively controls the size of the test at the significance level of 0.05 than the "exact" test.

Since we can generally estimate $p_c^r n_c^r$ from the contractor's record of inspection results and the AQL under which he is operating, we can select the power of test by adjusting the sample size ratio r commensurate with relative fraction defective, p_g^r / p_c^r , which should be detected if it exists.

This power can be further augmented by simple pooling of inspection results for a given r until the expected number of defectives for the contractor's samples exceeds the desired value of $p_c^r n_c^r$. Birnbaum [4] has considered various methods of comparing two Poisson processes in terms of the ratio of their parameters, and suggests for fixed samples an accumulation of observations until the total number of defectives d_t is sufficient to yield the power of test desired.

7. COMBINATION OF TESTS OF POISSON VARIATES. When the sample size ratio r is varied or the class of defects considered is not maintained constant so that pooling of inspection results from a sequence of lots is inappropriate for the methods represented by (8) or (9) and (16) or

TABLE 2

381

Power of Extended P-W* 'Exact' Test at Nominal Significance Level 0.05 for Hypothesis $p_g^! = p_c^!$ Against Alternatives $p_g^! > p_c^!$

$p_g^! n_c$	r = 1			r = 2			r = 3			r = 5			r = 8		
	$p_g^! / p_c^!$			$p_g^! / p_c^!$			$p_g^! / p_c^!$			$p_g^! / p_c^!$			$p_g^! / p_c^!$		
	1	3	4.5	1	3	4.5	1	3	4.5	1	3	4.5	1	3	4.5
.75	.001	.040	.139	.003	.062	.144	.001	.023	.065	.005	.042	.082	.003	.026	.059
1.50	.004	.173	.433	.011	.177	.372	.004	.084	.217	.011	.093	.180	.009	.070	.136
2.25	.010			.020			.008			.014			.013		
3.00	.015	.428	.790	.024	.351	.650	.013	.244	.513	.017	.190	.372	.015	.135	.280
4.50	.022			.029			.020			.020			.020		
9.00	.032			.031			.027			.026			.028		

*Przyborowski, J. & Wilenski, H. [15]

TABLE 3

Power of Cox's* 'Approximate' Test at Nominal Significance Level 0.05 for Hypothesis $p_g^! = p_c^!$ Against Alternatives $p_g^! > p_c^!$

$p_g^! n_c$	r = 1			r = 2			r = 3			r = 5			r = 8		
	$p_g^! / p_c^!$			$p_g^! / p_c^!$			$p_g^! / p_c^!$			$p_g^! / p_c^!$			$p_g^! / p_c^!$		
	1	3	4.5	1	3	4.5	1	3	4.5	1	3	4.5	1	3	4.5
.75	.020	.214	.407	.028	.188	.352	.014	.102	.202	.070	.200	.284	.044	.144	.198
1.50	.049	.389	.652	.054	.336	.554	.029	.220	.409	.072	.241	.356	.049	.184	.275
2.25	.062	.488	.774	.066	.429	.678	.040	.318	.550	.062	.264	.432	.050	.213	.354
3.00	.065	.564	.862	.070	.495	.761	.048	.390	.652	.056	.315	.515	.051	.255	.424
4.50	.060	.702	.955	.068	.591	.869	.055	.500	.790	.053	.394	.653	.053	.315	.535
9.00	.049	.928	.999	.053	.810	.986	.056	.743	.959	.049	.593	.865	.053	.466	.754

* Cox, D. R. [6]

(17) an omnibus type of test is required. This test can serve to combine all of the evidence obtained by means of verification sampling to provide a single measure of confidence in the contractor's inspection results.

From the α risks associated with the "exact" and "approximate" tests under the null hypothesis we can expect a certain frequency of significant differences. Further, from the β risks associated with these tests we can expect a certain frequency of erroneous acceptances of false hypotheses. Accordingly, it is not correct to reject or accept the general hypothesis that the contractor's inspection data are as a whole unreliable as a consequence of the individual lot comparisons, which taken separately appear to yield either significant or non-significant results. The over-all test calls, therefore, for the combination of a number of independent tests of significance. Fisher ([9], Section 21.1) has given a general method for combining the probabilities of several mutually independent tests. A number of other writers have discussed and illustrated this problem, but Birnbaum [3] has shown that Fisher's method is to be preferred for its somewhat more uniform sensitivity to the alternatives of interest.

The over-all test developed by Fisher deals with continuous variables. It will yield biased results if applied directly to probabilities derived from the "exact" test for Poisson variates. Lancaster [10], David and Johnson [7], Tocher [17] and Pearson [14] have considered the difficulties encountered by the combination of tests based on discontinuous variates. Since Cox's "approximate" test treats the number of events, d_c, d_g as continuous variates the probabilities obtained can be handled on a practical basis by application of Fisher's probability integral, which may be defined generally as follows:

Let $p(\chi)$ be the probability density function of a continuous random variable χ in the interval $a \leq \chi \leq b$, where $p(\chi) = 0$ for $\chi < a$ or $\chi > b$. Then if

$$(18) \quad P = \int_a^b p(\chi) d\chi,$$

P is uniformly distributed in the interval $(0,1)$ and $x = -2 \log_e P$ is distributed as χ^2 with 2 degrees of freedom.

If now we combine k independent probabilities, the combined probability is the product of the k separate probabilities, or

$$(19) \quad \sum_i (z_i) = -2 \log_e (P_1 P_2 \dots P_k) \\ = -2 \sum_{i=1}^k \log_e P_i,$$

and so has the χ^2 distribution with $2k$ degrees of freedom. Thus, by means of the probability integral transformation, any number of probabilities P_1, P_2, \dots, P_k may be converted to a χ^2 value and, using the additive

properties of the χ^2 distribution, may be summed together with the degrees of freedom to yield from published tables an over-all probability. The application of these results to continuous populations is straightforward.

For discrete populations, such as the binomial represented by (5), the over-all probability is biased when the null hypothesis is true. The expectation of χ^2 for discontinuous variates is always below the theoretical value of 2. Thus, for the case $d_g + d_c = 4$ and $r = 1$ we obtain, under the null hypothesis, the binomial $(1/2 + 1/2)^4$ and find from Table 4 below for a one-sided comparison that the expectation of $-2 \log_e P_i$ is 1.241 and the variance of the distribution is 3.527.

TABLE 4

Distribution of Probability Integral Transformation Applied to "Exact" Test for Case of Binomial $(1/2 + 1/2)^4$

No. of Events d_c	d_g	Relative Frequency of d_c, d_g	Cumulative Probability P_i	Probability Integral Transform $-2 \log_e P_i$
4	0	0.0625	1.0000	0
3	1	0.2500	0.9375	0.1291
2	2	0.3750	0.6875	0.7494
1	3	0.2500	0.3125	2.3263
0	4	0.0625	0.0625	5.5452

	Expectation	Variance
χ^2 with 2 D.F. (theoretical)	2.000	4.000
$-2 \log_e P_i$	1.241	3.527

Similarly, for the case of the binomial $(1/3 + 2/3)^5$ which can be derived from (5) the expectation of χ^2 is 1.314 and the variance of the distribution is 2.482. In contrast, Cox's "approximate" method for the same distribution as shown in Table 5 below yields a χ^2 expectation of 2.042 and a variance of 4.393.

TABLE 5

Distribution of Probability Integral Transformation Applied
to "Exact" and "Approximate" Tests for Case of Binomial (1/3+2/3)⁵

No. of Events	Relative Frequency of d_c, d_g	Probability Integral Transforms for Probabilities Derived from "Exact" Test "Approximate" Test
d_c	d_g	
5	0	0.131687 0.0000 0.0796
4	1	0.329218 0.2824 0.6570
3	2	0.329218 1.2357 2.0488
2	3	0.164609 3.1225 4.4886
1	4	0.041153 6.1903 8.3082
0	5	0.004115 10.9862 14.6404
χ^2 with 2 D.F. (Theoretical)		Expectation Variance
"Exact" Test Probability Integral Transformation		2.000 4.000
"Approximate" Test Probability Integral Transformation		1.314 2.482
		2.042 4.393

There is clearly considerable bias when the probability integral transformation is applied to the probabilities derived from the "exact" test. In contrast, Table 6 below indicates comparative lack of bias in the behavior of the "approximate" test when we wish to combine its results for a series of independent determinations to verify a common hypothesis, i.e., that the contractor's inspection records are reliable. The numerical results of Table 6 show that even for an extremely small number of observed defectives the continuity correction of the "approximate" test is very effective.

TABLE 6

Expectances and Variances of Binomially-Distributed
Probability Integral Transformations Derived from
"Approximate" Tests of Poisson Variates
(one-sided comparison)

$n \setminus p$	$\frac{1}{1+r} = 1/2$	$\frac{1}{1+r} = 1/3$	$\frac{1}{1+r} = 1/4$	$\frac{1}{1+r} = 1/6$	$\frac{1}{1+r} = 1/9$					
d_t	E(z)	Var(z)	E(z)	Var(z)	E(z)	Var(z)	E(z)	Var(z)		
5	2.045	4.364	2.042	4.393	2.044	4.392	2.056	4.409	2.084	4.391
4	2.050	4.316	2.051	4.253	2.056	4.485	2.072	4.485	2.111	4.453
3	2.045	4.108	2.067	4.540	2.074	4.604	2.101	4.585	2.159	4.431
2	2.024	3.540	2.086	4.463	2.106	4.686	2.158	4.678	2.257	4.458
1	1.905	2.259	2.084	3.630	2.170	4.165	2.302	4.377	2.474	4.146

NOTES: (contd. from Table 6)

$$(a) z = -2 \log_e I \frac{1}{1+r} (d_g + 0.5, d_c + 0.5)$$

(b) z is distributed as $\left(\frac{1}{1+r} + \frac{r}{1+r}\right)^{d_t}$ where $d_t = d_g + d_c$;
viz. $(p+q)^n$

$$(c) E(\chi^2)_{D.F.=2} = 2.000 ,$$

$$\text{Var } (\chi^2)_{D.F.=2} = 4.000 .$$

8. TABLES FOR ACCOMPLISHING VERIFICATION INSPECTION. Data recorded as shown in Table 1 can be conveniently tested for statistical significance by means of a table providing critical limits.

For a given number of total defectives, d_t , observed in both the contractor's and Government inspector's samples, limits can be set for either d_g or d_c as indicated by (17) for a specified α . This arrangement enumerates the boundary points of the critical region of the test of significance. However, the Government inspector is more concerned with comparing his sample results, for a given d_g recorded by the contractor, against an "allowable number." Accordingly, the critical value for d_g , designated as $d_g(A)$ can be obtained from (17) for a specified r . When the critical number $d_g(A)$ is reached or exceeded, the Government inspector adopts a course of action on the premise that a discrepancy actually exists in the contractor's inspection system.

Critical limits for indicating a discrepancy in paired attribute sampling inspections are presented in Tables IA through IE* of Section III of the Chemical Corps Verification Handbook [19]. The five sections, A through E, correspond to sample size ratios of 1, 2, 3, 5 and 8, respectively. Two standards of significance were set: $\alpha \leq 0.05$ for $d_g(A)$ and $\alpha \approx 0.10$ for a "warning" limit $d_g(W)$. When the "warning" limit is reached, the Government inspector is alerted to look for a possible discrepancy in the contractor's inspection system.

The probability integral transformation,

$$(20) z = -2 \log_e I \frac{1}{1+r} (d_g + 0.5, d_c + 0.5)$$

for a given d_g , d_c at a specified r can be readily derived from Tables of the Incomplete Beta- Function [18] and natural logarithm tables, and tabulated for comparison against critical values of χ^2 in accordance with (19). To simplify the procedure for the Government inspector, a Table II,** "Check Ratings for Paired Attribute Sampling Inspection" and a Table III,**

* Table IE is illustrated in Appendix 2.

** Portions of Table IE and Table III are illustrated in App. 2.

"Upper Critical Limits for Cumulative Check Ratings" have been included in Section III of the Chemical Corps Verification Handbook, Table II, which is subdivided into five sections corresponding to $r = 1, 2, 3, 5$, and 8, yields directly for a pair of values, d_c, d_g , the quantity $1/2 z$.

Table III is an extended table of the percentage points of the χ^2 distribution for even-numbered degrees of freedom. As Table III is used in conjunction with Table II, the critical values tabulated are $1/2 \chi_{2k}^2$ for

$2k$ degrees of freedom, where k is the number of probabilities to be combined, i.e., number of lots verified. The warning and action limits in Table III have been set at the 0.10 and 0.01 significance levels, respectively, and the median value at the 0.50 level.

The accumulation of check ratings serves to summarize all available information concerning the reliability of the contractor's inspection results. Furthermore, the ratings establish an objective degree of confidence, in the existence of statistical control over the contractor's inspection practice compared with the Government's standards. Visual representation of the check ratings on semi-logarithmic graph paper, with $1/2 z$ plotted on the log scale and critical limits of $1/2 \chi_{D.F.=2}^2$ imposed will be found useful for recording serially a common set of tests of significance.

9. ESTIMATING PRODUCT QUALITY. Product verification sampling has as its primary purpose the checking of the supplier's inspection records. However, the verification sampling results are also useful in providing an independent estimate of the contractor's "process average" and in furnishing an unbiased estimate of the quality of the conforming lots offered by the contractor for Government acceptance. Since the contractor's process average determines whether reduced, normal, or tightened inspection should be used, its validity should be established. The tolerance limits in MIL-STD-105 for a specified AQL can also be applied to the results of verification sampling. Normally, these limits should be applied to the process average derived from the results of the non-conforming lots as well as the conforming lots, since the process average reflects the average quality of product on which the supplier performs inspection.

The supplier's inspection results serve to segregate his inspection lots into conforming and non-conforming segments, and to determine the average percentage of defective items in the product represented by the samples inspected. Only when all lots are in conformance with acceptance criteria or when the product is manufactured under statistically controlled conditions can the process average computed by the contractor be used to furnish an unbiased estimate of the quality of product offered for Government acceptance.

Consider the O-C curve of a single sampling plan:

$$(21) \quad Lp^c = \sum_0^c \binom{n}{c} q^c n-c p^c$$

where L_p' denotes the probability of acceptance of lots binomially controlled at quality p' . If the plan n, c is designed or selected so that $0 < L_p' < 1$, then a portion of all lots from the controlled process will

yield samples in conformance with the acceptance criterion, c , and the rest of the lots will yield samples which are not in conformance. For the conforming fraction, the number of defectives, d_c , in each accepted sample will vary from 0 to c , and for the non-conforming fraction d_{c^*} will vary from $c + 1$ to n . Since n is fixed for all lots submitted for inspection the mean number of defectives in the samples from conforming lots will be less than the mean number of defectives in samples from the non-conforming lots. The apparent difference in the estimated quality between the two fractions of lots submitted for inspection against the acceptance plan (n, c) contradicts the original premise that the production of all lots was binomially controlled at a fixed p' . Accordingly, it is evident that sample results used to segregate lots cannot furnish an unbiased estimate of the respective fractions. This argument can be extended to the common case by using inverse probability for lots productd from different binomially controlled processes to demonstrate that acceptance sampling results cannot furnish unbiased estimates of the quality of the conforming segment offered for Government acceptance.

For an unbiased quality estimate, the following generalization can be used to obtain the best linear estimate of the percent defective p' of any lotted portion of product sampled independently by the Government inspector:

$$(22) \quad \hat{p} = \frac{N_1 \hat{p}_1 + N_2 \hat{p}_2 + \dots + N_k \hat{p}_k}{N_1 + N_2 + \dots + N_k},$$

where $\hat{p}_1, \hat{p}_2, \dots, \hat{p}_k$ are the respective estimates of lot quality derived from Government sampling results of inspection lots with lot sizes N_1, N_2, \dots, N_k .

If lot sizes are approximately equal the following estimate is an unbiased estimate of p' :

$$(23) \quad \hat{p} = \frac{\sum_{i=1}^k (d_g)_i}{\sum_{i=1}^k (n_g)_i},$$

where $(d_g)_i$ denotes the number of defectives found in a sample of size $(n_g)_i$ drawn from the i th lot checked by the Government.

When skip lot sampling is practiced by the Government inspector the lots selected for verification should be randomly selected to assure an unbiased estimate of p' . Usually skip lot sampling will not be as efficient as proportional sampling from each lot for estimating p' .

REFERENCES

- [1] Barnard, G. A., "A test for homogeneity of Poisson series," Advisory Service on Statistical Method and Quality Control, Technical Report No. Q.C./R/18, Ministry of Supply, London (1944).
- [2] Barton, D. E., "On the equivalence of two tests of equality of rate of occurrence in two series of events occurring randomly in time," Biometrika, Vol. 45 (1958), pp. 267-268.
- [3] Birnbaum, Allan, "Combining independence tests of significance," Journal of the American Statistical Association, Vol. 49 (1954), pp. 559-574.
- [4] Birnbaum, Allan, "Statistical methods for Poisson processes and exponential populations," Journal of the American Statistical Association, Vol. 49 (1954), pp. 254-266.
- [5] Bross, Irwin D. J., and Kasten, E. L., "Rapid analysis of 2x2 tables," Journal of the American Statistical Association, Vol. 52 (1957), pp. 18-28.
- [6] Cox, D. R., "Some simple approximate tests for Poisson Variates," Biometrika, Vol. 40 (1953), pp. 354-360.
- [7] David, F. N., and Johnson, N. L., "The probability integral when the variable is discontinuous," Biometrika, Vol. 37 (1950), pp. 42-49.
- [8] Finney, D. J., "The Fisher-Yates test of significance in 2x2 contingency tables," Biometrika, Vol. 35 (1948), pp. 145-156.
- [9] Fisher, R. A. Statistical Methods for Research Workers, 8th Ed., Oliver & Boyd, Ltd., Edinburgh and London (1941).
- [10] Lancaster, H. O., "The combination of probabilities arising from data in discrete distributions," Biometrika, Vol. 36 (1949), pp. 370-382.
- [11] Latscha, R., "Tests of significance in a 2x2 contingency table: extension of Finney's table," Biometrika, Vol. 40 (1953), pp. 74-86.
- [12] Mainland, Donald and Murray, I. M., "Tables for use in fourfold contingency tests," Science, Vol. 116 (1952), pp. 591-594.
- [13] Paulson, Edward and Wallis, W. A., "Planning and analyzing experiments for comparing two percentages," Techniques of Statistical Analysis, Statistical Research Group Columbia University, McGraw-Hill Book Co., Inc., New York (1947).

- [14] Pearson, E. S., "On questions raised by the combination of tests based on discontinuous distributions," *Biometrika*, Vol. 37 (1950), pp. 383-398.
- [15] Przyborowski, J. and Wilenski, H., "Homogeneity of results in testing samples from Poisson series," *Biometrika*, Vol. 31 (1940), pp. 313-323.
- [16] Swaroop, Satya, "Exact significance of difference in responses under two treatments," *Indian Medical Research Memoirs*, Memoir No. 35.
- [17] Tocher, K. D., "Extension of the Neyman-Pearson theory of tests to discontinuous variates," *Biometrika*, Vol. 37 (1950), pp. 130-144.
- [18] Tables of the Incomplete Beta-Function, edited by Karl Pearson, Biometrika Office, University College, London (1934).
- [19] Tables for Accomplishing Product Verification Sampling Inspection, U. S. Army Chemical Corps Verification Inspection Handbook, Section III, prepared by Headquarters, U. S. Army Chemical Center and Materiel Command, Directorate for Quality Assurance.

APPENDIX 1.

Basis for Preparation of Section III, "Statistical Sampling and Assessment"
of the Chemical Corps Verification Inspection Handbook

Joseph Mandelson

U. S. Army Chemical Center and Chemical Corps
Material Command, Directorate for Quality Assurance

1. References:

- a. AR 715-20 dated 3 September 1957
- b. Change 1, AR 715-20, dated 2 January 1958

2. Section III of the Chemical Corps Verification Inspection Handbook is intended to implement one important phase of Army inspection policy as established by References 1a and 1b. At this time it provides statistical tools for accomplishing the objectives of product verification inspection by attributes.

3. Par. 3e, Reference 1a, defines "...verification inspection to include:

"(1) Army evaluation of contractor's inspection systems to determine compliance with clause 5.e. Standard Form 32 (General Provisions-Supply Contract), or a similar inspection clause contained in the contract.

"(2) Army product inspection performed to measure quality of product offered for acceptance."

4. Par. 4a, Reference 1b, makes contractors "...responsible for controlling product quality and for offering to the Army for acceptance only those items...considered by them to conform to contractual requirements." Clause 5.e., cited in Par. 3e above, makes it a contract requirement that the contractor establish and maintain a system of inspection acceptable to the Government. Par. 3f, Reference 1a, refers to "...records of results..." as integral parts of the contractor's inspection system. Pars. 4b and 5b(1), Reference 1b, place an upper limit on Army verification inspection which "...will not exceed...the total of that inspection set forth in the Quality Assurance Provisions of the specification or contract." Furthermore, the References contain several general and specific allusions to the objective of reducing Army verification inspection when the contractor's quality inspection system is found to be reliable. In particular, Par. 4b, Reference 1b, states: "The extent of Army verification inspection to determine compliance with the Quality Assurance Provisions and other requirements of the contract will be adjusted to reflect the following factors:

"(1) The pertinency, completeness, and reliability of the supplier's inspection records.

"(2) The previous quality history of the supplier's product.

"(3) The unit cost of the item."

5. Par. 5b, Reference 1b, states: "When definitive specifications are the basis for procurement, the inspection system of the contractor... will be considered acceptable when quality of produced supplies or services is consistently acceptable, and it includes, as a minimum, the performance of those Quality Assurance Provisions stated in the specification and not reserved for sole performance by the Government." In prescribing the categories of verification inspection, Par. 5d, Reference 1b, states: "The amount of verification inspection will be adjusted to make maximum utilization of the contractor's quality control system and the quality history of the product..."

6. The cited provisions of References 1a and 1b gave rise to the following deductions:

a. Contractors are responsible for controlling product quality and must offer for acceptance by the Army only those items which the contractor considers to conform to contractual requirements.

b. To insure that this responsibility is fully discharged by the contractor, the Government makes it a contractual requirement that the contractors establish and maintain a system of inspection acceptable to the Government; records of inspection results are considered integral parts of the contractor's inspection system.

c. The Government considers the contractor's inspection system acceptable when it includes, as a minimum, performance of all quality assurance provisions of the specification not reserved for sole performance by the Government, and when the quality of material or services produced is consistently acceptable.

d. Army verification inspection, on the other hand, must not exceed the inspection prescribed "as a minimum" for performance by the contractor. In addition, several references are made to the (downward) adjustment of Army verification inspection, depending upon the reliability of the contractor's quality inspection system.

7. From the above, it is clear that contract provisions written in accordance with the references require the contractor to furnish both supplies and a quality inspection service related to the supplies. This inspection service is intended to cover all elements of product inspection which, prior to September 1957, were required of the Government inspector. It is obvious that if the contractor performs this service diligently and with a validity equal to that of a Government inspector then, with assurance that such is and remains the case, the Government inspector could confidently accept contractor inspection results as though he, himself, had performed the inspection.

8. CmIC has always stressed the importance of "feedback" of inspection data as an essential element in controlling quality, and indeed inspection. Product verification insures, among other things,

the existence of a valid body of independent data which, through proper "feedback," can be used as a self-governor to limit government inspection to that required to protect the Government's interests. CmIC considers contractor quality history to be an essential element in quality assurance, useful in preaward surveys and as a check on contract performance. In its implementation of AR 715-20, CmIC uses product verification as an essential, independent estimate of contractor quality history which can also be used to optimize the economics of government inspection consistent with quality requirements. It is noted that this viewpoint dates back to 1944 when CmIC first introduced into its specifications a "quality control" paragraph which stated in effect that if a contractor operates under a system of quality control acceptable to the Government and consistently produces high quality material, the Government might "...modify the whole or in part..." the sampling and testing requirements of the specification. The object, of course, was to reduce Government inspection.

9. The unabridged dictionary defines "verify" to mean "to prove to be true;...to confirm, as by comparison with facts;...to check or test the accuracy or exactness of; to confirm or establish the authenticity of;..." Thus, the objective of the "verification inspection" described in par. 3 above, is to prove, confirm or authenticate (as the dictionary puts it) the validity of the contractor's inspection system by comparing his inspection results with the independent factual findings of the Government inspector. The problem lies in assuring the reliability of the contractor's inspection system. To do this completely required careful check of the administrative and technical phases of the contractor's quality inspection activity, followed by independent product verification inspection by the Government inspector. The validity of the contractor's inspection data is established when no discrepancy is noted in the administrative and technical phases of his inspection work and when no statistically significant difference is found in measuring his inspection data against those generated by Government verification inspection of the same material.

10. The viewpoint in par. 9 conforms with the definition of verification inspection contained in par. 3e, Reference 1a. It combines Army evaluation of the contractor's inspection system to determine compliance with Clause 5.e. of the Supply Contract, with measurement of quality of the product offered for acceptance. Since product inspection is normally the most laborious single activity of the Government inspector, reduction of Government inspection, with concomitant savings of man hours and dollars, could most appropriately be made here. Therefore, the objective is to reduce product verification inspection to the extent possible through dependence on contractor inspection results provided the validity of these results has previously been assured by thorough-going verification inspection and product verification by the Government.

MEASURING A COMPLEX FIELD OPERATION

K. L. Yudowitch
Operations Research Office, The Johns Hopkins University

We who profess to practice operations research for the Army are confronted with a dual problem: the inaccessibility of the operation and the complexity of the operation. I have chosen to discuss what appears to be a relatively simple military operation: the combat firing of a rifle, or rifle-like weapon. The inaccessibility is inherent in the word "combat" -- there is no available combat operation to provide the proper context. Second, the word "firing" implies the activity of a human being, unfortunately frivolously complex. In order to get meaningful measurement therefore, two procedures are essential. First we must simulate the operation, second we must isolate or randomize the complex parameters. This means that we are obliged to find out how rifles are fired in combat, and then to imitate this repetitively under an appropriate variety of conditions. This will provide I hope a good illustration of the considerations, both quantitative and qualitative affecting the design of such an experiment.

Before proceeding to the design, it is necessary to know just what is being measured. We are interested in the combat effectiveness of rifle fire. The effectiveness component of interest has been designated as the combat accuracy. It has already been determined that combat rifle accuracy is critically dependent upon the error of aim, and negligibly affected by other identifiable errors, generally categorized as interior or exterior ballistics. Our interest in this particular study is further arbitrarily limited to what has been called "snap shooting" -- that is, firing the rifle with a very brief aiming or pointing time.

The study was motivated by a desire to pin down the effects of certain selected rifle characteristics on snap shooting accuracy. These characteristics are four: configuration, recoil, sights, and weight. Much experience and more folk lore gave rise to heated debate on the effects of one or another of these parameters on rifle accuracy. In addition, a fifth effect on accuracy is inherent in the apparent difference between first round and succeeding round accuracy on a single target. Sixth, the skill of the rifleman doubtlessly affects accuracy.

It is clearly a complex matter when one considers that we have undertaken to examine the effects of six possibly interacting parameters. The number of conditions for experiment is inherently quite high. Even if our experiment is extremely rough and examines only two values for each of these six parameters, there are 2^6 or 64 possible sets of conditions. Recall also that we must effect repetitions because of the "random" variations.

5 QUESTIONS

1. Which Conditions Most Difficult?
2. Amount and Confidence of Differences?
3. Which Interactions Negligible?

4. Adequate Parameter Increments?

5. Relative Value of Effects?

How do we now determine a design? Five questions are basic: First, which set of conditions are too difficult or expensive? Second, how much random or statistical variation is permissible? Third, which interactions may be ignored? Fourth, how many values of each parameter are sufficient? Fifth, which effects are of more interest, and which of less interest?

Let's look at the example. For configuration we select the M14 as military standard and (on advice of experts) a popular hunting rifle like the Winchester Model 70. Recoil is M14 standard and reduced (say half) load. Sights for examination are military aperture and hunting type open sights. There is no problem supplying either sight or recoil to both these rifle configurations. Weight is less simple. The Winchester 70 is available in both light and heavy versions -- and the heavy Winchester just about matches the M14 weight. However there is no lightweight version of the M14. This immediately eliminates 1/4 of our 64 sets of conditions, unless we supply special rifles. Our first qualitative question is asked; and in this case special rifles are not deemed economically feasible. Succeeding fire after first round is only of interest for semi-automatic fire. The Winchester is manual only, eliminating another 1/4 of our 64 sets of conditions, unless special rifles are supplied. Again the first question is asked, and special rifles ruled out. Our design has already been reduced to half, 32 sets of conditions.

SETS OF CONDITIONS

4 Sets

8 Sets

Config.	Weight	Mode	Recoil	Sights	Skills
(M14	Std.	Sing.)	High	Open	Exp.
M14	Std.	Mult.	Low	Aper.	Mark.
W70	Std.	Sing.			
W70	Light	Sing.			

The eight combinations of recoil, sights and skill values apply to all remaining sets of configuration, weight and mode. As the M14 single mode data are included in the multiple mode data, the M14 single mode set may be deleted without loss, leaving only 3 times 8, or 24 sets of conditions. It is only necessary to identify the first bullets fired at each target -- easily done by painting those bullets. The paired comparisons provided by these imperfectly balanced 24 sets of conditions are listed:

COMPARISONS

CONFIGURATION 8 (Recoil x Sights x Skill)
 (M14 vs. W70)

WEIGHT 8 (Recoil x Sights x Skill)
 (Std. vs. Light)

RECOIL 16 (Config. x Wt. x Sights x Skill)
 (High vs. Low)

SIGHTS 16 (Config. x Wt. x Recoil x Skill)
 (Open vs. Aper.)

We see that without repetitions, these conditions provide 48 comparisons: 8 configuration, 8 weight, 16 recoil and 16 sight. Having answered the first of the 4 questions, and thrown out half of the possible conditions as impractical, we next ask about statistical reliability.

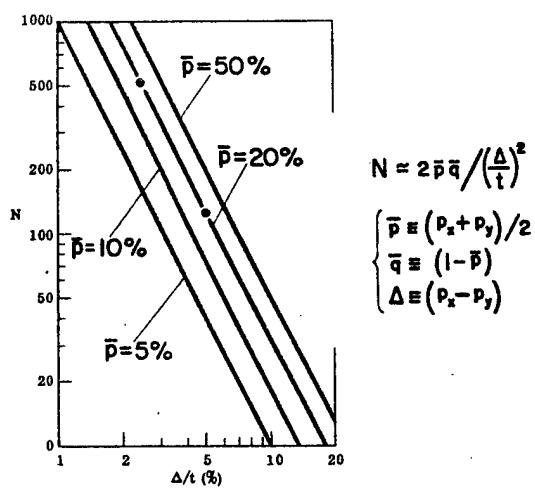
How many repetitions are needed to provide sufficient confidence in our results? This is clearly dependent on how large a difference we are seeking to identify. If the rifle accuracy is characterized by a hit probability p_x , and the changed condition results in p_y , we may note the change in p as $p_y - p_x$. It is then a simple exercise to relate this change to the number of measurements N and the familiar student's t , measuring confidence level (See graph.).

Estimating an average rifle hit probability of 20%, and an objective of identifying a 5% change in this probability with 95% confidence ($t = 2$), we get a requirement for 500 measures. If we are satisfied with identifying a 10% change in hit probability, only 125 measures are needed.

If we use 7 firers of each skill level, and each man fires at 7 targets, we should need to run through only 2 or 3 times for a 10% difference $[125/(7 \times 7) \approx 2\frac{1}{2}]$. A 5% difference required 10 replications $[500/(7 \times 7) \approx 10]$.

The decision on number of replications calls for consideration of our third question. The third question is the educated guess about which interactions are negligible. Surely the 8 comparisons of the configuration are not all independently different. Similarly for the 8 comparisons of weight and the 16 comparisons each of sights and recoil. If there were no interactions at all, we could expect to have adequate statistical reliability to identify a 5% difference from a single run-through. Clearly we want somewhere between 1 and 10 replications. Value judgments finally determine the number of replications.

Experience predicts experimental running time. If it is desired to complete the experiment in one week in the field, one comes up with 3 to 4 replications. Of course the multiple round modes of fire supply more



data per run (by a factor of the number of rounds fired per target). Thus a reasonable design permits 4 single mode replications (64 runs), and perhaps 2 multiple mode replications (16 runs).

As value judgment indicates this experiment is worth about 1 week, and statistical reliability obtained in that time is just about adequate, the fourth question regarding the number of values per parameter is answered; No. Any refinement of the effect of more than 2 increments of the four basic input parameters is best postponed to a succeeding effort.

Our fifth question evaluates the several comparisons which the experiment makes. If there is especially great interest in the effect of recoil for example on the current military standard rifle, we might incorporate extra runs for both recoils with Marksmen firing the M14 with aperture sight in multiple mode. Similarly, reduced interest in one of the parameters might dictate deletion of some conditions. However deletions must be made with care, for each condition is used in an average of two comparisons, and this advantage of the semblance of balance that remains in our design is quickly lost by deletions.

A further incidental advantage of the minimal experimental design is its relatively lower susceptibility to biases of learning, fatigue, weather changes, etc.

My purpose has been to illustrate how practical considerations of time, value and cost impose over statistical considerations to define an experimental design. The result is generally unbalanced in a statistical sense, but balanced in an operational and value sense.

THE CONDUCT OF MILITARY FIELD RESEARCH ON A SHOESTRING

Andrew J. Eckles, III
Operations Research Office, The Johns Hopkins University

There is, I believe, a certain poetic justice in the fact that this paper is the last one to be given at this conference. For after all, our subject matter deals not with what we would really like to have, if we "had our druthers;" but rather with a last resort, stop-gap method of doing our necessary field research.

In recent years, our methods of wargaming and other analytical techniques have increased in accuracy and reliability to such an extent that their need for reliable input data have by far outstripped all available facilities. To some extent, this urgent need for valid performance data under actual field conditions has been recognized; and as a result, CDEC, an agency admirably suited for the conduct of sophisticated, controlled field experiments in the area of tactical operations was established. But, as most of you know, the conduct of operational field experiments is not only time consuming, but extremely expensive in manpower and equipment as well as money. Therefore, all of our field research facilities together, with the means available to them, cannot hope to keep pace with the increasing appetites of our model makers for the basic data upon which to perform their manipulations.

In most cases, our military analysts have been forced to resort to guesstimate, or even worse, manufacturer's specifications, for the vast majority of their basic input parameters--which, of course, regardless of how sound the model, reduces most results to little more than science fiction.

As a result of this shortage of reliable input data obtained under adequate field conditions, we at ORO have been investigating some short-cut methods of conducting small scale operational field studies. In this paper, then, we will present one possible solution to the cost factors involved in the conduct of complex field experiments. We will discuss, literally, the conduct of such studies on a shoestring.

Perhaps I should emphasize that what we are discussing here is nothing new--in fact, the basic ideas have even been presented at one of these conferences a few years back. But we have made some efforts to codify our techniques for greater efficiency, and have even coined an expression "SYMBION," to connote the underlying principles. Essentially, the concept of Symbion is to superimpose experimental designs and data collection techniques upon carefully selected, and at times modified, phases of the Army training Programs. This provides us with a means whereby we can tap the vast reservoirs of manpower, equipment and ammunition expended in the normal training cycle as a potential source of valuable weapons systems performance data.

We have recently completed several studies utilizing this idea of Symbion, and presently have another under way. So before discussing the principle itself, with its good and bad points, I would like to describe briefly two of these studies.

Good - if possible

Bear in mind, now, that this work which I shall describe is not exactly what we would like to have done. But we found ourselves in the condition of the poor soldier with mighty ambitions. For example, his ideal weapons system might be something like this (See Slide 1), complete with air conditioning and fully stocked bar. But there can be certain economic and other obstacles in his way. And yet, he has a job to do--with whatever equipment he has available or can scrounge. So, like this hypothetical soldier, in order to get some place to do something, we had to be satisfied with a little more primitive situation. (Slide 2)

The first study which we will discuss was designed to investigate the effectiveness of tank high explosive fire against hastily prepared antitank gun positions. The data collection phase of this study lasted from September 1958 through March 1959; though the actual time involved in data collection was less than three weeks. To some extent this was an unusual situation in that we wanted to test our concepts of Symbion as well as obtain the data we sought.. To do this, we threw the entire burden for the conduct of the study on the Fort Stewart, Ga., compliment. Our only participation consisted of advice and criticism, and in the development of the initial equipment to be used, which was then duplicated and used by post personnel.

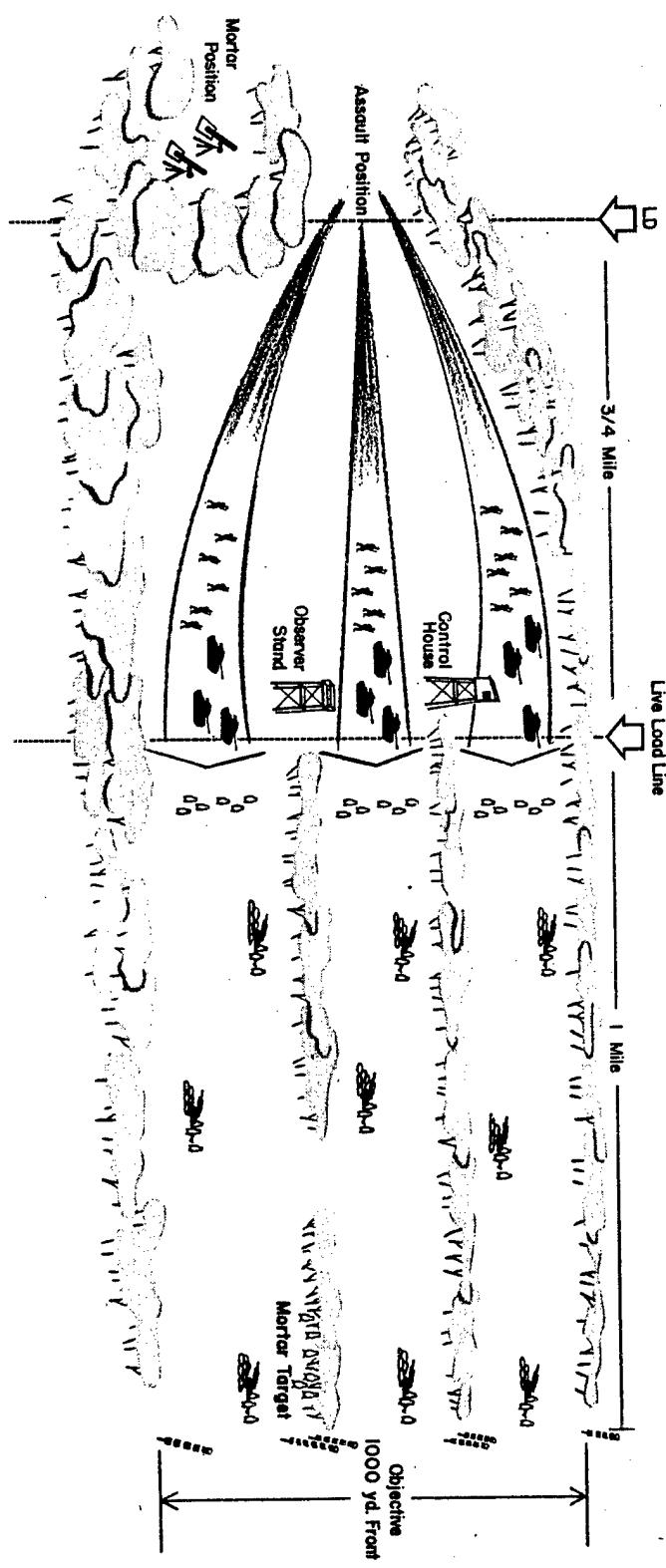
Our first step, in this case, was the development of a realistic antitank gun target which would respond appropriately to H. E. ammunition, and yet which would be realistic enough, cheap enough, and easy enough to use that the Army Post itself would provide all of the equipment which was needed for the conduct of the experiment. Part of this problem was easily solved by modifying a previously used gunfire simulator so that it could be built by Post Signal personnel utilizing only surplus drone target parts which had been salvaged. Our Electronics Laboratory built the first such system, and the remaining ones were then built by Post Signal. To obtain a realistically killable target to serve as gun crewmen, we developed one which consisted simply of a toy balloon inserted in a canvas bag. Such a target possesses amazingly realistic characteristics. After our first run, Ft. Stewart constructed their own target bags, but until this killable target was adopted by the Army as a training aid, our Office still had to purchase the balloons--which, oddly enough, were not standard army equipment at that time and difficult to justify to the comptroller.

After we had developed the prototype target system, an experimental course was set up by the Post, following our proposed plans very closely. The field setup looked something like this. (Chart of experimental setup--Slide 3.) Essentially, this was a Company live fire problem. However, for convenience of data collection, the problems were conducted as three simultaneous platoon courses. As shown here, there were three separate lanes, one for each platoon. Each lane consisted primarily of three antitank gun targets located in the following range brackets: 300-400 yards, 450-550 yards, and 1200-1500 yards.

Each antitank gun target consisted of the following items: a mock antitank gun; a gun fire simulator which would detonate, on command, up to three charges of TNT (representing the AT gun firing); and a crew consisting of our newly developed "killable" balloon targets.

Often
we must be
more primitive





TABULATED RESULTS

V7951

Crew Number	Gun positions		Crew members		Rds Fired	Kill probabilities per round	
	Available	Killed	Available	Killed		Gun position	Crew member
3	9	4	27	5	54	.07	.09
4	24	6	96	20	144	.04	.14
5	12	5	60	5	72	.07	.07
6	12	9	72	20	120	.07	.17
Totals	57	24	255	50	390	.06	.13

(Unfortunately, we did have one serious fly in the ointment. This was the availability of balloons. We started with four small targets and two large targets making up each AT gun crew. However, the number of crewmen serving each gun varied from phase to phase of the problem as we ran out of balloons, and as our supply of funds varied for the purchase of additional ones. Indeed this study was conducted on a shoestring.)

In addition, to enhance the realism of the test situation, overhead mortar fire was delivered on the objective, and infantry supported the tanks in the attack. Additional killable targets were utilized for mortar targets and for infantry targets. These gave us, purely as a by-product, some information on the effectiveness of mortar and .30 caliber firing.

We have, to date, obtained fairly reliable information on twenty-one different tank platoons that have gone through this test course. Our primary information, concerned with the kill probability of a round of H.E. fired against an unarmored gun crew, resulted from the expenditure of a total of 390 rounds of 90-mm H.E. ammunition. The total cost of this information to our Office, not counting the time of one analyst, and development of prototype equipment, amounted to somewhere between \$1.50 and \$2.00 per data point.

Had we conducted this very same study as a standard field experiment, again not counting the time of one analyst and development of prototype equipment, the cost per data point would have been somewhere around \$300.00.

A rough summary of the primary results in this study are shown on this chart. (Slide 4) Since this is more a report of methodology than results of a specific experiment, I won't spend too much time giving you our results, but briefly this table is interpreted as follows:

The first column gives the number of killable targets in each gun position, for a given phase. The second column gives the number of gun positions available in each phase (always in multiples of three, since there were three gun positions per lane). The third column gives the number of gun positions killed—that is, a gun position was considered killed if one or more of the crew members serving that gun was killed by an HE fragment. The fourth column gives the number of crew members available to be killed in each phase (found by multiplying the number of crew members per gun by the number of gun positions). The fifth column gives the number of crew members killed by HE fire. The sixth column gives the total number of rounds fired during a given phase. The seventh column gives the probability of a given round inflicting a kill on a gun position. And finally, the last column gives the probability of a given round inflicting a kill on a given crew member.

A quick glance at the kill probabilities would indicate that an AT gun position has a very good chance against an individual tank firing one round of HE. However, as indicated in the first two columns, when an objective of this type is attacked by three tanks, a given gun position has only a little better than a 50% chance of survival from the tanks 90-mm HE fire. The use of the tanks sub-caliber weapons would reduce these chances of survival even further. Also, observation of the actual

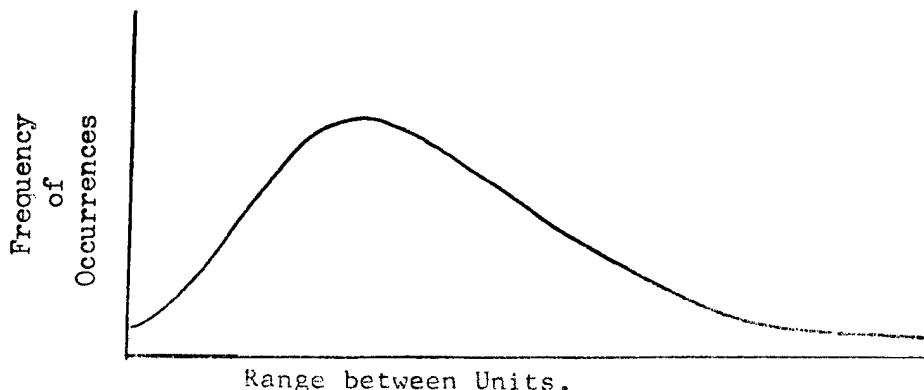
situations would lead one to think that rigid fire discipline would decrease the survival rate of AT guns even more.

I would now like to discuss another project which demonstrates how these shoestring techniques can be utilized to collect usable basic data. This study is presently being conducted by the Communications Group at ORO, and the data collection phase will continue, at periodic intervals, through next Spring.

It is well known that adequate communications have always been a problem in the wars of history, and will no doubt continue to be even more so in wars of the future. In an effort to cope with these problems, it has been usual for the soldier to ask constantly for radios with longer range, greater reliability, etc., but communications threaten to become the tail that wags the dog. However, there is a group at ORO presently working on the anticipated requirements for communications in the next time frame--and perhaps for the first time a realistic attempt is being made to try to determine just what communications facilities will be needed, and what can be discarded as dead weight or excess clutter on the air-waves.

Considerable information concerning the future need in this area has been obtained by means of wargaming. However, there is as yet no data available which could show a correspondence between the results obtained from our analytical models and the results which can be expected in the real world. The present study, then, is an effort to provide data which would help define the relationships being studied.

I have time to show only one facet of this study, as an example of its objectives. One measure obtained from analytical techniques which can help provide guidelines for communications requirements is a frequency distribution showing range between units which need to communicate. Such a hypothetical distribution might look like this:



But one very serious question is, just how valid is such a distribution which is obtained only from mathematical models; is the shape even similar to what might be found in a real life situation? Is the total message count realistic?

It is our hope that this "ongoing" study will provide data which could help answer these questions.

Our subjects, in this case, will be individual tank battalions with supporting infantry. They will, to the greatest extent presently possible, be placed in a quasi-combat situation for 72 hours, in which they will engage an aggressor force under a variety of situations. During this period of test, the ORO group will monitor the battalion net and record each radio transmission together with the distance between the units concerned. From this information, a frequency distribution will be made as from the war games. The correspondence between these curves should provide at least one base point for estimating the correspondence between the results of our war games and what might conceivably happen in the real world. Of course much additional information will be obtained--but the above is a prime example. By this spring, we hope to have data from at least four battalions.

At the present time we do not know just how many data points will be obtained from each battalion. However, if we assume that we obtain only about 500 data points from each battalion, by using the techniques of SYMBION, the approximate cost from research funds will be somewhere in the order of \$2.00 per data point (again not counting the analysts' salaries, which would be constant with either method). On the other hand, to conduct an identical situation as a standard field experiment (and, by the way, the situation which we would have set up is not very different from the one now being conducted as a training program) the cost of this information would be somewhere in the neighborhood of \$500.00 per data point.

And besides this, we could never hope to obtain the number of different subject battalions which are easily available to us in the normal training cycle by merely extending our data collection phase.

Now that you have seen some of the capabilities of this "cheap-skate" method of data collection, I would like to discuss some of the basic principles behind the technique.

Essentially, the principle objective of the concept of Symbion is to increase the efficiency of our available research facilities--that is, to get the very maximum out of every research man-hour and research dollar (and it is very difficult to say which is more scarce at present).

The real value of this technique is shown whenever we attempt to obtain performance data under quasi-combat conditions as opposed to manufacturer's performance data or proving ground type data. For it is here that the cost of operational field studies mounts to prohibitively high figures--and yet it is also here where the soldier is most willing to work closely with the scientist in order to increase as much as possible the realism of his training programs.

The procedures for conducting a Symbion type project are relatively simple. It involves first the selection of an appropriate phase of the Army Training Program which will provide the basic situations required for

Design of Experiments

our experiments. Every effort is then made to develop equipment which will provide the necessary realism to qualify those situations as quasi-combat. If the equipment developed is "practical" enough, from the over-all military viewpoint, then the cost of this equipment can even be borne by the training program, and not out of research funds; however, in some cases it will be necessary to utilize research funds to furnish this equipment. All of the data collection equipment will normally have to be provided by and operated by the research agency.

Let us now examine some of the differences between a specifically conducted experiment and one conducted by means of these "shoestring" techniques. The primary differences can (but not necessarily must) exist in the following areas: experimental control; precision of measurement; time to obtain the required information; sample sizes and representation; and problems involving the general area of experimental design. I will touch briefly on each of these five areas.

Experimental control is a major problem in any complex operational field problem--and an area in which many concessions must be made due to such necessary factors as safety, limited terrain, etc. But these problems must be faced no matter how we conduct our research. It is true, however, that there will be times when, under a Symbion type program, we will not have the control which we would like to have over all of the factors which may affect the results. Often, once we have initiated the desired quasi-combat courses and installed our data collection equipment, we are then reduced to the position of the astronomer--that is, we may be limited in the actual manipulations which we are allowed to perform, but must obtain most of our data by simply observing and recording events as they occur. Such, for example, is the case in the communications study which was described earlier. (But there, to a large extent, we would not exert any more control even in a specifically set up experiment--largely due to the fact that as yet we do not know exactly what factors we wanted to control in this particular type case.)

This, at times limited control, then leads us to question the precision of the data obtained by these methods.

As for precision, data collected by means of a Symbion-type program can run the gamut of quality just as those obtained by any other means. In the great majority of cases, if the basic program is properly organized and set up, the resulting data will be every bit as accurate as that obtained by any other means. But, again, there will be times when, using our quick and dirty techniques, we will have to accept data which are not as precise as those which could be obtained from a specific experiment. It then behooves us to weigh carefully the additional cost (in manpower, equipment, and time as well as dollars) of each increase in precision required.

When evaluating this additional cost, it should be remembered that, unlike most scientific research, where the results of our efforts often form lasting foundations for pyramiding results, military research to a very great extent is dated and all too quickly becomes obsolete as weapons systems and the characteristics of war change. Thus much of what the

military researcher does has only temporary value. Knowing this, we must consider the cost of our work in the light of its often time-limited usefulness. It is therefore usually much wiser to be satisfied with "broad-brush" type answers in order to have timely solutions to our problems, and then spend our remaining efforts on other situations that are also crying for solution. We just cannot afford to waste precious time in obtaining data that are somewhat more accurate, but are useful for only a short period. We must of necessity develop short-cut methods of obtaining as rapidly as possible the data we require.

We have now somewhat sneakily entered the area of time required to obtain our data. And it is here that our "quick and dirty" techniques can prove to be very quick indeed. A developed Symbion-type program is perhaps the only hope we have in the foreseeable future of providing the vast quantities of performance data required in any realistically acceptable time frame. A completed series of "quasi-combat" training programs of the type which we have described would provide the military scientist with full-blown, completely equipped "laboratories in the field" which would be readily available for his needs as they develop. These "laboratories" would be staffed and run by competent and experienced soldiers already acquainted with the equipment required for a quasi-combat course, since it would have become part of their everyday performances. In addition, they would have become even somewhat inured to the often rigid demands of the scientist. With such a program, the scientist would be prepared to examine rather complex problems with a minimum of delay, and in most cases obtain some of his required data within months rather than years of an expressed need.

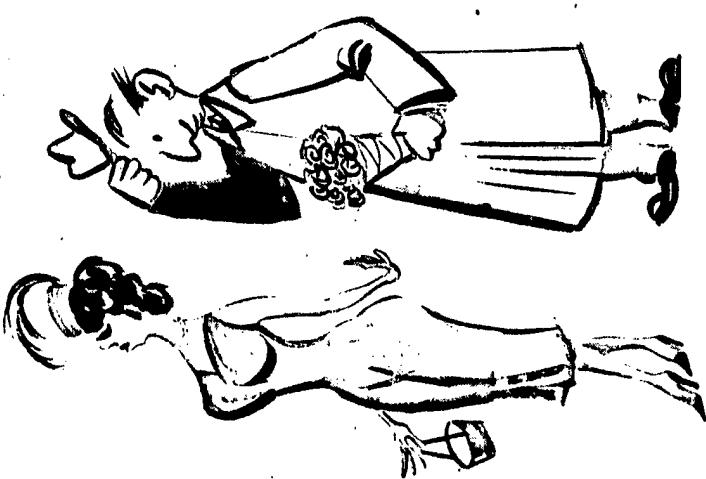
We now come to the area of sample size and sample representativeness. In most large scale field experiments, we are forced to limit severely our sample sizes, and make repeated use of "test-wise" troops. This can, to a very large degree, limit the extrapolations which could be made from the results of our studies to "troops in general." In many cases, then, it would seem to be very advantageous for us to utilize the Army Training Program in order that our samples be large, and that they more closely resemble the army-wide population in such important factors as degree of training, motivation, intelligence, etc. For example, in a specific experiment, a sample of four or five test-wise experimental companies might be considered a large sample; but in a Symbion type program, we could consider a sample of ten naive companies relatively small.

(As an aside, it should be mentioned that we can to some degree compensate for lack of control in this type study through increases in sample size.)

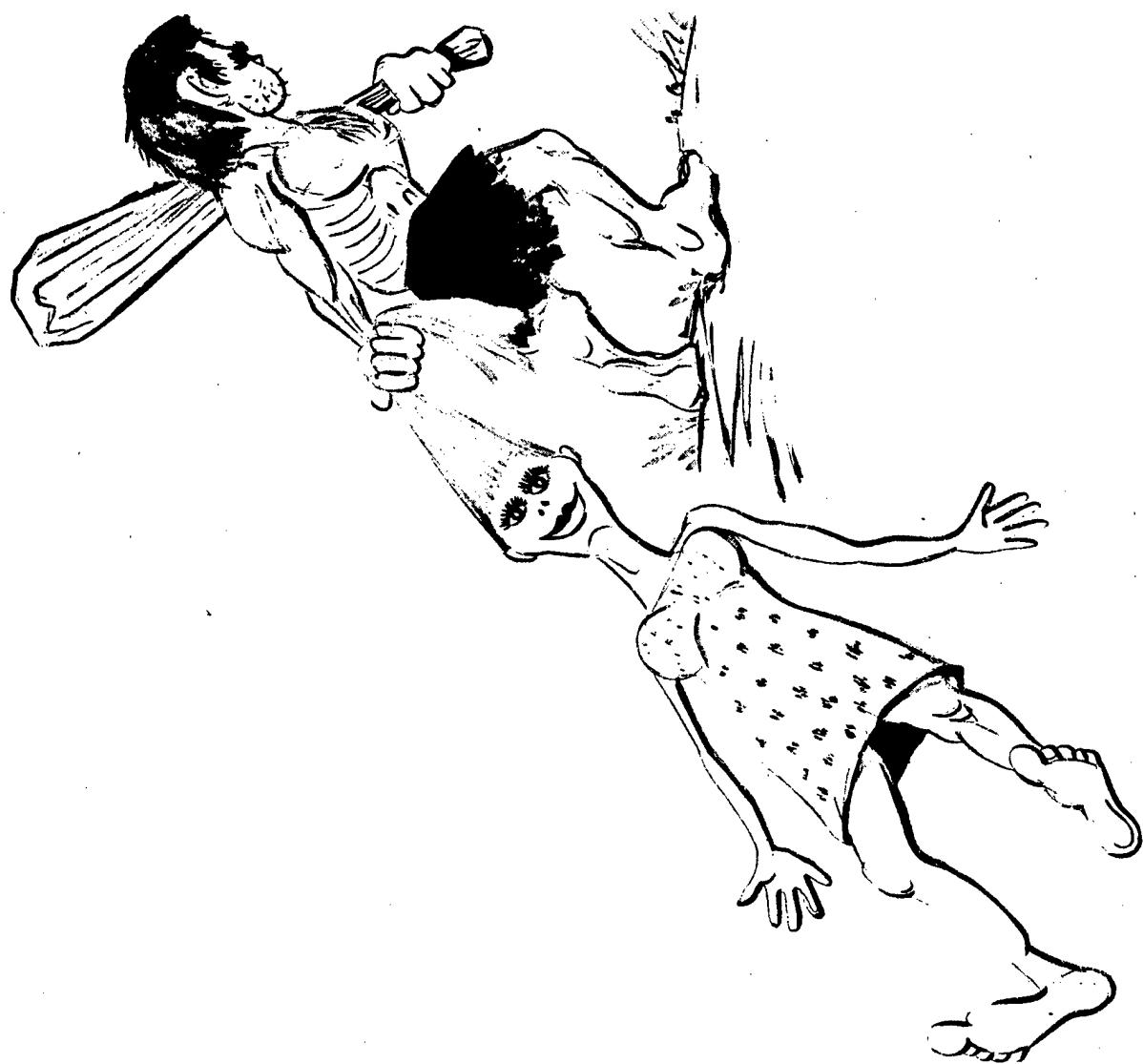
The final problem area of major differences is that of experimental design. The design problem is often, it is true, made much more difficult in our shoestring type situation than in specifically conducted experiments. More often than not we will have to be satisfied with the more crude, less sensitive designs and their companion analytical techniques. We simply have to utilize designs which can be fit into an already complex and crowded training schedule. But when the final results are compared, we will find in most cases that we have lost nothing of practical significance (as

414

V7851-10 / 59



Slide 5



distinguished from statistical significance). And this is especially true when we are faced with the choice of either no data at all (other than guesstimates) or data obtained under less than ideal conditions.

The completed program of this type will always be conducted by the training officer, and the scientist is reduced to the role of advisor and observer. But it is surprising how much common ground can be found between the soldier and the scientist especially when, after a long day in the field, they stand side-by-side at the bar. We have found that the participating soldiers are not only willing but eager to assist us at every opportunity--and almost every suggestion made to implement scientific validity has been followed as far as safety requirements and available resources have allowed.

And I should add that we have also found that our most vociferous, and often most helpful critic has been some frustrated soldier trying to persuade a piece of balky data-collection equipment to operate in the rain, and do his own job at the same time. But the soldier, too, is interested in objective measures of his performance under combat-type conditions, and will consider his extra efforts well worth while.

In summary, then, we feel that these economic methods of data collection do enable us to obtain usable weapons systems performance data under quasi-combat conditions with a minimum expenditure of research time and money--both of which are extremely limited. In addition, the training officer is quite eager to participate in any program which he feels will add realism and motivation to his programs.

A judicious utilization of appropriate phases of the Army Training Program can, then, enable us to conduct much of our necessary field research on a shoestring. And, in addition, the participating soldier gains as much of immediate value as does the scientist when they join forces.

Before closing, I would like to add one word of warning. We cannot overemphasize the fact that the research methods which we have described here are not used merely by preference. They are "make-shift," "stop-gap" techniques--which, nevertheless, provide us with a source of usable data which would otherwise be unobtainable. Perhaps I could demonstrate this best by means of a cartoon. (Slide 5)

Like this gentleman, we often have our goals clearly in mind, and are well aware of the more sophisticated and refined techniques which we would like to use. But these techniques do cost money, and they do take time--and sometimes we find ourselves short of both. In such cases, then, we might find that the more direct approach used by our ancestors, even though somewhat crude, could work wonders. (Slide 6).

PROPOSAL FOR FIELD CALIBRATION OF A TRACKING RADAR

Victor B. Kovac
RCA Missile Test Project, Patrick Air Force Base

PURPOSE.

The purpose of this experiment is twofold: to determine the feasibility and adequacy of automatic tracking of sun data to evaluate several types of angular misalignments in the radar mechanical system; and to investigate the feasibility of obtaining ship's position accurate enough to serve as origin for data reduction purposes.

INTRODUCTION.

The coordinates of a missile's position recorded during radar track at 0.100 second intervals are azimuth, elevation, and range. These coordinates and time are channeled directly into a digital computer for "real time" data presentation, and to a tape recorder for subsequent data reduction. The angular resolution of the AN/FPS-16 Radar is 0.05 mil and the tracking precision is 0.1 mil. Absolute errors are not large, but are believed to exceed 0.1 mil. Hence, it is desired to remove systematic errors so that ultimate, smoothed data is not only precise but accurate.

Owing to the use of several radar stations to track long range missiles, it is essential that they are accurately aligned with the common geodetic reference, so that overlapping spans in time also agree in position. During the data reduction process, small misalignments can be corrected if they are known. Evaluation of these misalignments in the mechanical system as a function of the radar's azimuth and elevation orientation is one approach to the solution.

During tracking of a missile, there must exist small deviations of the target which the radar attempts to correct automatically, or else it could not track. This dynamic characteristic of tracking tends to produce angular deviations in data which are difficult to separate from the small alignment biases. Furthermore, refraction (bending) of the radar beam in the atmosphere is a complex function that has to be accounted for. Partial correction of refraction errors in data has been effected, but the oscillatory nature of the dynamic error requires further study.

Figure 1 (at the end of this article) is an example of seven seconds of typical raw elevation data during an early portion of a missile track. It shows:

1. Time (abscissa) at 0.100 second intervals
2. Raw elevation points (radar)
3. Refraction error (from separate sources)
4. Reference standard (position data from cine-theodolite cameras near the target translated into elevation at the distant radar origin.)
5. Elevation error (bias between a radar data point corrected for refraction, and the theodolite point).

The elevation scale has been purposely exaggerated in order to illustrate: The magnitude of refraction error, the oscillatory nature of the dynamic track, and the difficulty of evaluating angular biases in which we are interested.

Figure 2 illustrates the three orthogonal axes of the radar's mechanical system. A deviation between the encoder bearing and the true bearing results in a constant azimuth index error. A small rotation about the OX or OY axis (or both) corresponds to turntable mislevel or tilt. An elevation axis tilt with respect to the horizontal is called a standards error.

Electrical (or optical) axis deviation from normal to the elevation axis, constitutes a departure from a truly orthogonal system. In this instance, the deviation of the electrical axis from the plane of the elevation circle causes an error in azimuth. Dial eccentricity denotes eccentric deviation of the dial shaft from the true center.

MISALIGNMENTS and THEIR EFFECTS (Figure 3)

Mislevel.--Misalignment of the radar's vertical axis with respect to the local vertical results in a turntable mislevel from the horizontal. The resulting elevation error is sinusoidal, having a period equal to one complete rotation in azimuth, an amplitude equal to the angle between turntable and horizontal planes, and a phase (θ) defined by the direction of the intersecting planes and North. (Equation 1). Two components of error denoted by a and B at 0° and 90° azimuth, respectively, also describe a particular sine wave (Equation 2). The effect of a tilted turntable on azimuth error is a tangent function of elevation as shown in Equation 3. Thus a 1-mil tilt results in a sinusoidal azimuth error with 1-mil amplitude at an elevation of 45° . In Figure 4, the increasing magnitude of the amplitude is indicated by the dash-dot line.

Standards Error.--This error is caused by a misalignment of the elevation axis with respect to the horizontal, and is analogous to mislevel, except that it is independent of azimuth. This is so because the standards, which support the elevation axis, rotate with the turntable and thus always present the same aspect with respect to the target. Both standards error and mislevel may be evaluated by direct measurement, using the boresight telescope and leveling bubbles in each case. Both direct methods have a resolution of 0.05 mil. Evaluation of component errors by independent means simplifies the resolution of the total error, (i.e., reduces the number of unknowns).

Electrical (Optical) Axis Deviation.--A line which deviates from the elevation circle by an acute angle (ℓ) generates a flat cone during a complete revolution about the elevation axis. Thus, the radar's electrical axis may deviate from the elevation circle. The net effect is an error in azimuth. An axis deviation of 1 mil results in an azimuth error that is a secant function of elevation, as shown by the dashed line in Figure 4.

Bending Error (Antenna Unbalance).--The bending effect due to antenna unbalance is nearly compensated by a pair of spring equilibrators mounted on the radar standards. The cosine function of elevation listed in Equation 5 (Fig. 3) is intended to match the true error that may arise at various elevations. However, the true condition is unknown, hence the form of the mathematical expression as well as the unknown coefficients of bending error require careful analysis.

Dial Eccentricity.--There exists a gear train between the respective azimuth and elevation shafts and the data encoders, so that the successive contributions of error due to eccentricity in each gear result in a complex wave. According to Barton, (1) the measured values in azimuth and elevation are 0.03 mil rms and 0.005 mil rms respectively. Equations 9 and 10 are not applicable directly because they are correct only for cases of one dial. (2). Their form may be of use in the analysis of the residuals.

Index Errors.--Index errors are constants. They denote index deviations from North and from the horizontal.

METHOD

Optical and electronic instruments are subject to error with regard to absolute position of a target in space. Unless the geometry of the other instrument and target is favorable, so that the bias as seen from the radar is negligible, then such instrument cannot be used as a standard. Thus, for a long-range radar, there are few instruments or conditions which offer sound comparison. (Actually, an effort is now being made to track aircraft simultaneously with radar and still cameras which use star background as reference points). Radar, being a point source, requires that calibration be performed over a hemisphere centered at the radar. Furthermore, the coordinates should be in absolute agreement with the geodetic system. Finally, in order to reduce angular error to a minimum, the distance to the reference points should be very long, and the elevation angles should be high enough to avoid multipath effects ($>5^\circ$).

In view of the considerations above, it appears that direct measurement of celestial bodies may be the best means to calibrate radars. Star calibration using telescope (or camera) mounted on the elevation axis has proven successful (3). Recently, automatic tracking of electro-magnetic radiation from the sun has been performed, so that direct comparison of computed azimuth and elevation with raw radar data is feasible.

Investigation of the range of azimuth and elevation coverage available from sun track indicates that this source should furnish most, if not all information needed. Some auxiliary points, such as a surveyed reference (Boresight Tower), and the star, Polaris, should be included to complete azimuth coverage and to check electrical-optical axis collimation. Observations of two celestial points at azimuths 60° apart may be reduced to geodetic coordinates sufficiently accurate to serve as a fix for the radar origin in data reduction. Thus, our purpose is to determine whether a number of bursts of sun track, combined with a few auxiliary points, furnish sufficient information to provide angular calibration of the radar.

Design of Experiments

To test this, the proposed experiment requires an additional, (optical) star calibration to serve as standard. In any case, it is essential to find a dependable means to determine the magnitude of misalignments present in a radar.

TEST DESIGN, REDUCTION, AND ANALYSIS.

A discussion of the reference frame, computation of coefficients and evaluation of the residual errors is in order. First, the range of azimuth is 360° , or a normal sector of 180° plus an equal "extension." Then, the elevation ranges from 0° to 90° , and the "dumped" position from 90° to 180° (Figure 4). Thus we can obtain two observations, or two coverages of the same hemisphere. The sun track will furnish points in the two southern azimuth quadrants with peak elevations ranging from about 60° to 80° , depending on the station's latitude and the season. Low elevations, say below 20° , are undesirable, being subject to considerable refraction error.

Reference to the equations of component errors indicates that they are well-behaved functions of radar orientation. Bending error, being one that cannot be readily isolated, requires that the instrument be dumped. As a consequence, it is necessary that sun tracks be performed in the normal and dumped positions. In order to provide valid standard, it is evident that at least one optical calibration pass should be made completely overhead, (e.g., North-South), and that two azimuth passes at constant elevation (say, $E_1 = 45^\circ$, $E_2 = 71.6^\circ$) should be made.

The tentative scheme for evaluating the coefficient and making an analysis of the residuals is as follows: The array of azimuth and elevation errors will be treated as two separate problems. The elevation errors will be treated first in order to establish the mislevel coefficients. These can be checked against actual measured values using level bubbles. There should be sufficient points to fill four 1×4 matrices ($m = 16$). This means that there should be at least four sun observations before noon (repeated with dumped observations) and another set after noon. Each observation consists of three 10-second bursts of data spaced one minute apart. The reason for this is to permit computing three Solar positions at one minute intervals so that their slope may be known. Also raw data trend and bias can be evaluated more readily from three bursts than from one, particularly since the fitting of a trend is needed to narrow down the radar's wander about the Sun's center.

The azimuth errors will be entered in matrix form and one known coefficient removed (either b from measured standards error, or one of the mislevel coefficients determined by the elevation data). This will leave a 4×4 matrix to be solved. Actually, there will be four such sets.

Now a check on agreements among the four separate sets of solutions will be made. First, a least-squares solution will be attempted. The entire array of data will then be treated to removal of the systematic errors. The residuals will then be inspected to determine whether cyclic effects are still evident. Here we can effect a review of bending error

and possibly eccentricity. The remaining residuals, in factorial form, should reveal whether further adjustments in the index errors need be made. Analysis of variance should yield a clue to trends as a function of azimuth and elevation.

As an alternative measure, the factorial format of the four solutions (four quadrants) should reveal complimentary reversals. These can be interpreted as adjustments to be made in the coefficients. Once the coefficients show reasonable agreement, a final review of the residuals can be made with regard to oscillations due to bending effects and eccentricity.

The final comparison of coefficient magnitudes and experimental error is between sun track and optical star calibration results. Here the effect of collimation error between electrical and optical axes may have to be taken into account.

Selecting sun track data at approximately 150° and 210° azimuth, a solution of the radar's geodetic position will be obtained for comparison with surveyed values. If the agreement is good, this means of position indicating may be of use for ship-borne radar stations. This procedure may furnish a check on the magnitude of refraction errors.

CONCLUSIONS.

1. Samples of automatic sun track taken during a day should furnish sufficient information to permit evaluation of the angular errors present in the radar's mechanical system.
2. Star calibration using the boresight camera can be used as standard for determining the accuracy and adequacy of the automatic sun track data to evaluate misalignments.
3. An error surface developed in terms of the radar's azimuth and elevation orientations can be reproduced by analytical expressions suitable for correcting data by machine process.
4. As a possible by-product, two day-time sun orientations may be used to establish a ship's geodetic position.

REFERENCES

1. D. E. Barton, "Instrumentation Radar AN/FPS-16 (XN-1), Evaluation and Analysis of Radar Performance," R.C.A. Missile and Surface Radar Dept, Moorestown, N. J., Jan., 1959.
2. H. Schmid, BRL Report No. 764, "Systematic Errors of Cine-Theodolites," Aug., 1951.
3. K. E. Pearson, "Evaluation of the AN/FPS-16 (System Nr 1) At White Sands Missile Range," Tech Memo 606, U. S. Army Signal Missile Support Agency, White Sands Missile Range, New Mexico, Feb 1959.

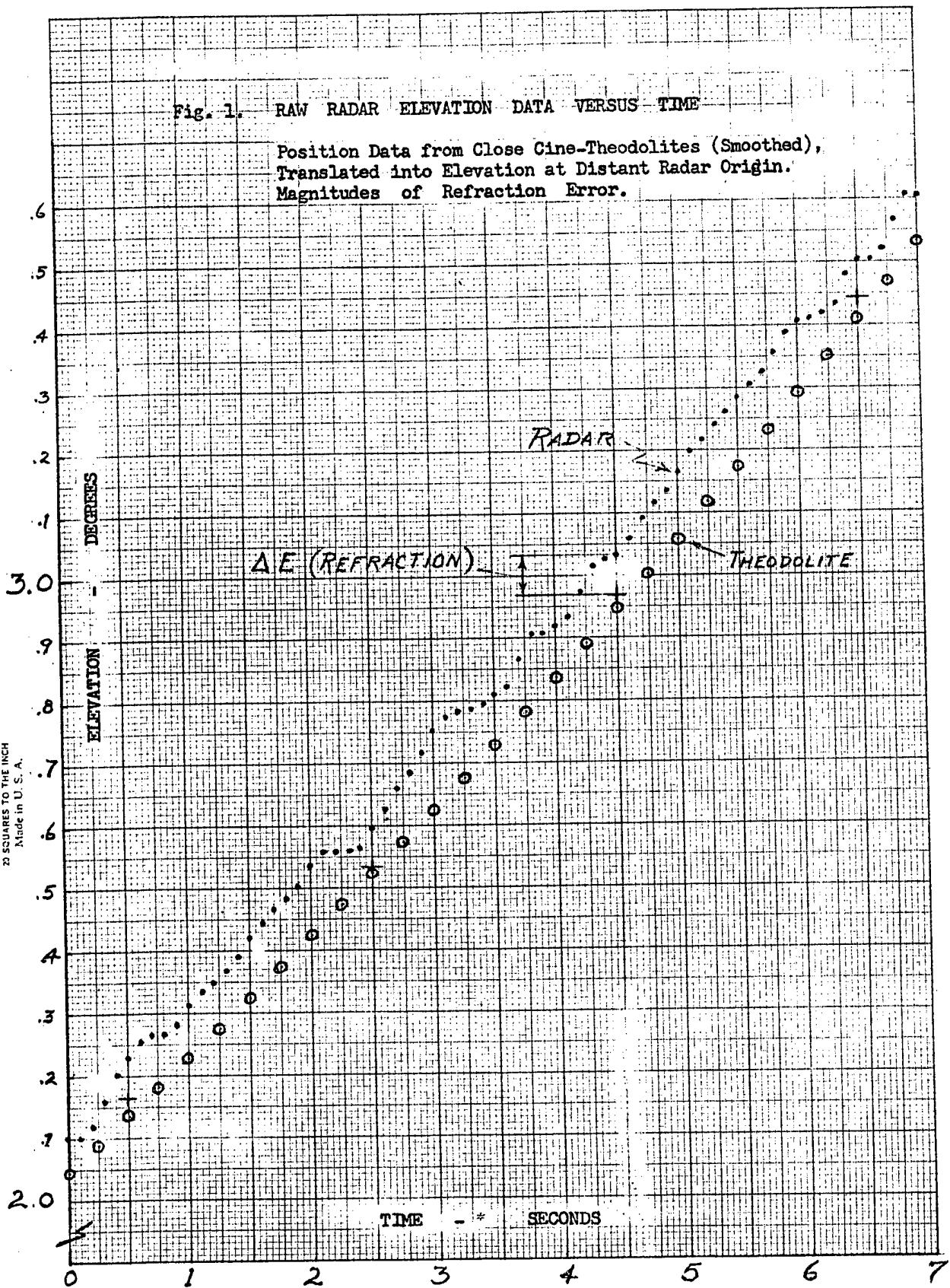


Fig. 2. SCHEMATIC DIAGRAM of a RADAR'S SYSTEM of AXES

Showing Correct Alignment With Geodetic North And Local Vertical.

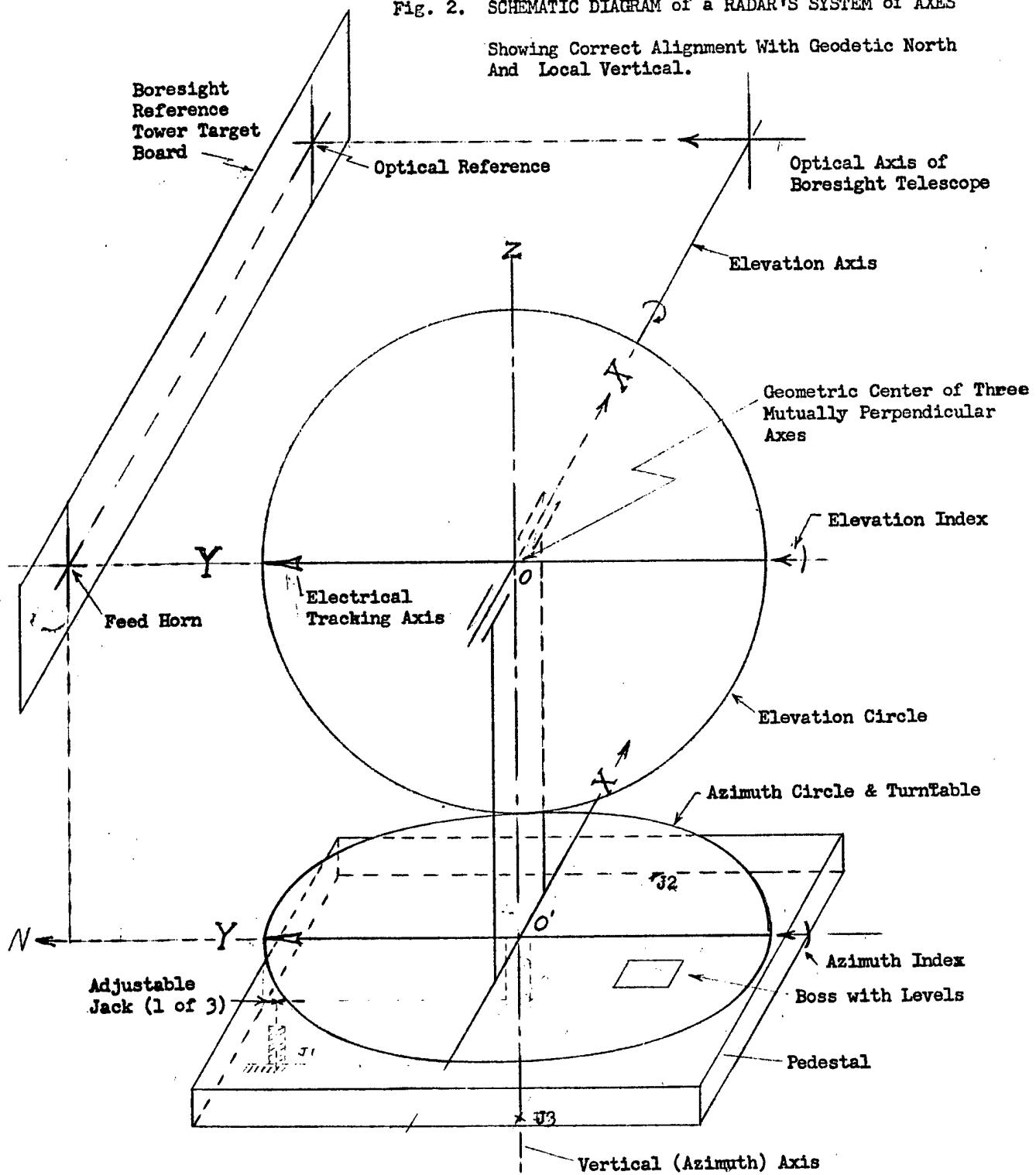


Fig. 3. ALIGNMENT ERRORS and their EFFECTS ON ΔA , ΔE in
TERMS of RADAR ORIENTATION (A , E).

MISLEVEL

$$(1) \quad \Delta E = e_m \sin (A + \theta)$$

$$(2) \quad \Delta E = \alpha \cos A + \beta \sin A$$

$$(3) \quad \Delta A = \alpha \tan E \sin A - \beta \tan E \cos A$$

STANDARDS ERROR (ϕ)

$$(4) \quad \Delta A = b \tan E \quad (b = \tan \phi)$$

BENDING ERROR

$$(5) \quad \Delta E = d \cos E \quad (\text{or: } d \cos (E - E_m))$$

ELECTRICAL AXIS DEVIATION (ℓ)

$$(6) \quad \Delta A = c \sec E \quad (c = \ell \text{ in radians})$$

INDEX ERROR

$$(7) \quad \Delta A = g$$

$$(8) \quad \Delta E = h$$

ECCENTRICITY OF DIALS

$$(9) \quad \Delta A = (e/r) \sin (A - A_o)$$

$$(10) \quad \Delta E = (e/r) \sin (E - E_o)$$

FIG. 4. EFFECTS ON ΔA OF 1-MIL ERRORS IN MISLEVEL, ELECTRICAL AXIS DEVIATION, & STANDARDS ERROR.

