REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE June 1957	3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE
Proceedings of the First Conference on the Design of Experiments in Army Research, Development and Testing

5. FUNDING NUMBERS

6. AUTHOR(S)
Not Available

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Army Mathematics Advisory Panel

8. PERFORMING ORGANIZATION
REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U. S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING
AGENCY REPORT NUMBER

ARO-OORR 57-1

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12 a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This is a Technical report resulting from the Proceedings of the First Conference on the Design of Experiments in Army Research, Development and Testing.

14. SUBJECT TERMS

15. NUMBER OF PAGES
278

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT **UNCLASSIFIED**	18. SECURITY CLASSIFICATION ON THIS PAGE **UNCLASSIFIED**	19. SECURITY CLASSIFICATION OF ABSTRACT **UNCLASSIFIED**	20. LIMITATION OF ABSTRACT **UL**

OORR 57-1

Office of Ordnance Research

PROCEEDINGS OF THE FIRST CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH, DEVELOPMENT AND TESTING

OFFICE OF ORDNANCE RESEARCH, U.S. ARMY
BOX CM, DUKE STATION
DURHAM, NORTH CAROLINA

OFFICE OF ORDNANCE RESEARCH
Report No. 57-1
June 1957

PROCEEDINGS OF THE FIRST CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH DEVELOPMENT AND TESTING

held at
Diamond Ordnance Fuze Laboratories
and
National Bureau of Standards
19 - 21 October 1955

Office of Ordnance Research
Ordnance Corps, U. S. Army
Box CM, Duke Station
Durham, North Carolina

Initial Distribution

The initial distribution list of the Proceedings of the First Conference on the Design of Experiments in Army Research, Development and Testing includes those who attended the meeting and/or the government installations with which they are associated. For economy, only a limited number of copies have been sent to each. Additional copies will be transmitted upon request.

TABLE OF CONTENTS

Table of Contents (Continued) Page

* This paper was presented at the Conference. It is not published in these Proceedings.

** This paper can be found in a classified security information (Confidential appendix of this Report.

Table of Contents (Continued) Page

* This paper can be found in a classified security information (Confidential, appendix of this Report.

** This paper was presented at the Conference. It is not published in these Proceedings.

FOREWORD

In 1954 the Army Mathematics Advisory Panel (AMAP) was established by the Office of Ordnance Research to provide advice on the mathematical needs of the Army to the Chief, Research and Development, Office, Deputy Chief of Staff for Plans and Research, Department of the Army. In carrying out its duties the AMAP made an extensive survey of the mathematical activities and requirements of more than 30 Army research, development and testing facilities. One of the most frequently mentioned needs expressed by the scientists and engineers at these establishments was greater knowledge and use of the modern statistical theory of the design and analysis of experiments made in the course of their work.

On the basis of this expression of interest the AMAP considered the possibility of organizing an Army-wide conference on the design of experiments. Upon making further inquiries it was found that a number of research workers at various facilities expressed an interest in contributing papers to such a conference. Others had unsolved or partially solved problems which they wished to present for discussion.

The AMAP decided to organize a three-day conference on the design of experiments with three kinds of sessions. The first group of sessions would consist of invited papers by well-known authorities on the philosophy and general principles of the design of experiments. The second group would consist of technical papers contributed by research workers from various Army research, development and testing facilities. The third group would be clinical sessions consisting of presentations and discussions of partially solved and unsolved problems which had arisen in these establishments. The program of the 3-day conference is included in the first part of these Proceedings.

The conference was held on October 19 - 21, 1955 at the Diamond Ordnance Fuze Laboratories and the National Bureau of Standards in Washington, D. C. It was attended by over 230 registrants and participants representing some 50 organizations. Speakers and other participants in the conference came from the Bell Telephone Laboratories, Johns Hopkins University, Princeton University, Virginia Polytechnic Institute, Bureau of Ships, National Bureau of Standards, and 18 Army facilities.

The present volume of the Proceedings contains 28 papers and an appendix which contains 2 classified papers, all of which were presented at the conference. The papers are being made available in this form as a contribution toward a wider use of modern statistical principles of the design of experiments in the research, development and testing work of the Army.

The members of the AMAP take this opportunity to express their thanks to those research workers in the various Army research, development and testing facilities who participated in the Conference; to Lt. Colonel J. A. Ulrich, the Commanding Officer of the Diamond Ordnance Fuze Laboratories, and Dr. A. V. Astin, the Director of the National Bureau of Standards, for making available the excellent facilities of their two organizations for the Conference; to Mr. John A. Wheeler who handled the details of the local arrangements for the Conference at both installations; and to Dr. F. G. Dressel of the Office of Ordnance Research who carried through the details, including all correspondence involved in organizing the Conference.

S. S. Wilks
Professor of Mathematics
Princeton University

PROGRAM

CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH DEVELOPMENT AND TESTING

19-21 October, 1955
Diamond Ordnance Fuze Laboratories
and
National Bureau of Standards

19 October 1955

On Wednesday all sessions of the Conference will take place in the East Building Conference Room of the National Bureau of Standards.

REGISTRATION: 0900 - 0930 (Eastern Standard Time)

MORNING SESSION: 0930 - 1145

Chairman: Professor S. S. Wilks
Princeton University

Introductory Remarks: Lt. Colonel J. A. Ulrich, Ordnance Corps, Commanding Officer of the Diamond Ordnance Fuze Laboratories

The Philosophy Underlying the Design of Experiments
Professor W. G. Cochran, The Johns Hopkins University

Design of Experiments in Industrial Research and Development
Dr. W. J. Youden, National Bureau of Standards

LUNCH: 1145 - 1315

AFTERNOON SESSION: 1315 - 1600

Chairman: Colonel P. N. Gillon, Ordnance Corps, Commanding Officer of the Office of Ordnance Research

The Principle of Randomization in the Design of Experiments
Dr. Churchill Eisenhart, National Bureau of Standards

Finding Optimum Condition by Experimentation
Dr. M. E. Terry, Bell Telephone Laboratories

20 October 1955

A total of four Technical Sessions will be conducted Thursday, 20 October 1955. The security classification of morning Session II is Confidential, and that of afternoon Session IV is Secret. No clearances will be required for Session I (morning) or Session III (afternoon).

TECHNICAL SESSION I: 0900 - 1145: Chemistry Bldg. Lecture Room, NBS

Chairman: George Glockler, Office of Ordnance Research

Experimental Design in Personnel Test Research
F. K. Thomson, The Office of the Adjutant General

Operational Experiments
Paul Michelsen, Operations Research Office

Operational Gaming
W. E. Cushen, Operations Research Office

TECHNICAL SESSION II: 0900 - 1145: Connecticut Ave., Annex Conference Room - DOFL

Security Classification - CONFIDENTIAL

Chairman: Nicholas Smith, Operations Research Office

Some Design Techniques for Increasing Cell Size with Special Emphasis in the Guided Missile Field
P. C. Cox, White Sands Proving Ground

The Application of Statistics to Electronics (U)
L. M. Court, Diamond Ordnance Fuze Laboratories

Progress and New Techniques in Surveillance
Miles Hardenburgh and David Howes, Chemical Corps

LUNCH: 1145 - 1315

TECHNICAL SESSION III: 1315 - 1600; Chemistry Bldg. Lecture Room, NBS

Chairman: Joseph Weinstein, Signal Corps Engineering Laboratories

Analysis of the Results of Designed Experiments on Standard Punch Card Machinery
C. J. Maloney, Chemical Corps

Some Examples of the Use of High Speed Computers in Statistics
J. M. Cameron, National Bureau of Standards

Comparative Characteristics of Medium Fully Automatic Computers for Statistical Applications
Sam Alexander and Mary Stevens, Chemical Corps

TECHNICAL SESSION IV: 1315 - 1600; Bldg. 83 Conference Room - DOFL

Security Classification - SECRET

Chairman: Charles Bicking, Office of the Chief of Ordnance

Some Problems of Experimental Design Dealing with Military Units of Company Size and Greater
D. M. Meals, Combat Operations Research Group

The Use of Experimental Design in Ammunition Surveillance (C)
J. R. Johnson, Ballistic Research Laboratories

The Design of Experiments in Stability Testing
J. W. Mitchell, Frankford Arsenal

21 October 1955

Of the four Clinical Sessions that will be conducted Friday morning, 21 October 1955, Sessions A and B require no clearances. The security classifications of Sessions C and D is Secret. The Panel Discussion on Friday afternoon will be open to the public.

CLINICAL SESSION A: 0830 - 1130; Chemistry Bldg. Lecture Room, NBS

Chairman: C. J. Maloney, Chemical Corps

Panel Members: J. M. Cameron, National Bureau of Standards
W. J. Youden, National Bureau of Standards

Sensitivity Testing of Explosives
A. Bulfinch, Picatinny Arsenal

Statistical Approach to the Evaluation of Electric Initiators
F. Lawrence, Picatinny Arsenal

CLINICAL SESSION A (Cont'd)

Monte Carlo Approach for Developing a Method for Sensitivity Testing
Pvt. S. Ehrenfeld, Picatinny Arsenal

Analysis of Variance Applied to Evaluating the Reproducability of a New Piece of Apparatus
K. R. Fisch, Frankford Arsenal

CLINICAL SESSION B: 0830 - 1130; Manse Conference Room, NBS

Chairman: F. E. Grubbs, Ballistic Research Laboratorie

Panel Members: Churchill Eisenhart, National Bureau of Standards
John Tukey, Princeton University and Bell Telephone Laboratories

A Plan for Testing a Missile Unit in the Laboratory
P. C. Cox, White Sands Proving Ground

The Problem of Grouped Firings
P. C. Cox, White Sands Proving Ground

Application of Sequential Analysis to Quality Inspection of Small Lots of an Ordnance Item
L. Stout, Frankford Arsenal

Estimating an Average or Standard Trajectory
P. C. Cox, White Sands Proving Ground

CLINICAL SESSION C: 0830 - 1130; Conn. Ave. Conference, DOFL

Security Classification - SECRET

Chairman: Ira Cisin, Human Resources Research Office

Panel Members: M. E. Terry, Bell Telephone Labs.
S. S. Wilks, Princeton University

Long Term Exposure Tests of Various Ordnance Materials
S. L. Eisler, Rock Island Arsenal

The Design of an Experiment to Determine the Cause(s) of Projectile Break Ups and/or Prematures (U)
Benjamin Shratter, Lake City Arsenal

Determining the Effectiveness of Cutting Oils in Reducing Machine Tool Wear
S. L. Eisler, Rock Island Arsenal

CLINICAL SESSION C (Cont'd)

Controlled Operational Field Experimentation (U)
K. L. Yudowitch, Operations Research Office

CLINICAL SESSION D: 0830 - 1130; Bldg. 83 Conference Room, DOFL

Security classification - SECRET

Chairman: Paul Meier, The Johns Hopkins University

Panel Members: Boyd Harshbarger, Representative of Redstone Arsenal
Paul Meier, The Johns Hopkins University

Sequential Bioassay?
G. S. Woodson, Chemical Corps Medical Labs.

Corrections for Multiple Restrictions in Range
W. A. Klieger, The Office of the Adjutant General
(Presented by K. F. Thomson)

Establishing Hypotheses When the Experimental Factors are not at Present Under the Control of the Experimenter
K. R. Wood, Quartermaster Food and Container Institute

Another Type of Experiment is Needed
M. M. Algor, Diamond Ordnance Fuze Laboratories

LUNCH: 1130 - 1300

AFTERNOON SESSION: 1300 - 1500; East Building Conference Room,
National Bureau of Standards

Chairman: John Tukey, Princeton University and Bell Telephone Laboratories

Panel Members: Cuthbert Daniel, Private Consultant
Besse Day, Bureau of Ships
Churchill Eisenhart, National Bureau of Standards
M. E. Terry, Bell Telephone Labs.
S. S. Wilks, Princeton University

Panel Discussion on How and Where Do Statisticians Fit In.

THE PHILOSOPHY UNDERLYING THE DESIGN OF EXPERIMENTS

William G. Cochran
Professor of Biostatistics
The Johns Hopkins University

Introduction. In ordinary speech, the word "experiment" has a broad meaning. It denotes trying out anything new. In our conferences here we shall be using the word in a narrower sense. The essence of an experiment is that we deliberately introduce one or more changes into some process, and take measurements in order to find out the effects of these changes. The changes whose effects are being compared are often called the experimental treatments or more simply the treatments.

The ability to do experiments is one of the most powerful weapons that man has for making advances in his understanding of the world. When conflicting claims or conflicting theories can be put to a crucial test, the workers in a branch of knowledge cannot long remain in error. In fields like economics, sociology and history, on the other hand, where experiments are rarely if ever feasible, it is difficult to get at the true causes behind the events that are observed. Having read that all the economic experts forecast a prolonged rise in the stock market, you may take your meager savings from under the bed and purchase a few good-looking stocks. When the market promptly falls, you may become slightly mad at the experts. Instead, you should be sympathetic and understanding: these men cannot do experiments , and it is hard for them to unravel the complex forces behind the market.

The origin of an experiment usually lies in some information that we would like to have, or in some questions to which we want answers. After carefully phrasing the questions, we select a set of treatments such that comparisons of the effects produced by these treatments will answer the questions. We must then consider the environmental conditions in which these treatments should be applied, and the most suitable kinds of measurement to take. At the end, if the experiment is successful, we find that the results do enable us to answer the questions.

Failures in experimentation. There is much to be learned by considering the causes of failures in experimentation - that is, experiments which do not produce the desired information. Although such failures can be classified in various ways, the following rough grouping will serve my purpose.

1. We may have asked the wrong questions in the beginning. This is perhaps of more frequent occurrence in fundamental research, where the ability to ask the significant questions distinguishes the outstanding from the second-rate scientist. In both fundamental and applied research one can find experiments on questions that have already been answered, since the increasing volume of research makes it harder to keep up with what has been done. And in applied research there are examples where the questions asked were the unimportant rather than the important ones that would have to be faced in putting the results into practice. Before starting to plan an experiment, it is always advisable to ask: "Is this the most informative question to try to answer right now?"

2. We may have asked the right questions, but the treatments selected are incapable of providing answers to some of the questions. This should not happen in simple experiments with only two or three treatments, but the danger is present in a complex experiment designed to throw light on a number of different questions. It is a hazard particularly associated with experiments that are planned by committees, especially if they contain several strong-minded persons who don't agree with one another. The principal safeguard is to sit down, after the treatments have been selected, and verify for each question the treatment comparisons that will be made in order to answer the questions.

3. In applied research, the conditions under which the experiment is conducted are often strikingly different from those in which results will be applied. Treatment effects that are found in the experiment may not hold up under the conditions of application. There is no entirely safe way out of this difficulty, because much experimentation has to be done with small-scale equipment in a specialized environment, in order to keep down costs and to obtain precise results. In many types of work, the practice is to use the small-scale experiment primarily for screening. Promising candidates from this screening are tested again under conditions that more closely approximate those that will prevail in applications.

I mention this difficulty because it is always well to realize how the conditions of experimentation differ from those of application. Even in pilot experiments it may be possible to include some comparison that help to bridge this gap. Suppose that three different models of some piece of equipment are used in practice and that the properties of these models have already been worked out, so that there is no need to experiment further in comparing them. It may still be advisable to include all three models in experiments designed to test other factors, rather than take the simpler path of confining the experiments to one model. In this way we obtain some check as to whether the other factors perform the same way on every model. Similarly, we may sometimes reject a refinement in experimental technique that otherwise seems attractive, on the grounds that the refinement makes the conditions of the experiment too remote from those of application.

4. We may obtain erroneous results for the effects of the treatments. It has happened several times in medical research that a dramatic cure is found in a first experiment, and perhaps confirmed in a second, causing much excitement in the press. But later and more careful experiments repeatedly fail to find any effect, and after a time medical science reluctantly concludes that the cure doesn't exist. Erroneous results of this type usually happen because some unsuspected bias has crept into the results. Such biases are one of the most frequent causes of failures in experimentation.

5. Finally, the results may be so indecisive as to be useless. This happens when all that we can conclude at the end of the experiment is that the difference in effect between treatment 1 and treatment 2 lies somewhere between a large positive value and a large negative value. In other words, we haven't learned which treatment is superior, nor can we even assume that the two are approximately equal in their effects. Those of you who are new to the design and analysis of experiments may protest that surely we can get

more definite information than this out of an experiment. Unfortunately, even with well-conducted experiments, vague conclusions of this kind are often all that can be drawn.

This type of failure is perhaps less serious than the preceding types. It arises because the experiment was not precise enough for its purpose, and it can be remedied by repeating the experiment sufficiently often with the same treatments. This, however, is no consolation if decisions have to be taken before there is time for more experimentation.

In this catalogue of failures in experimentation I have not, of course, mentioned all types of failures. I have been told of an agronomist who laid out an experiment in the semi-arid backlands of Australia, and then as harvest approached could not remember where he had put it. There are even cases where the professor forgot that he had started an experiment until long after the time when the results should have been recorded.

In the remainder of my remarks, I shall concentrate on failures that arise from biased results and from indecisive results. These are the types of failure in which statistical ideas appear to have been able to help most.

Variability and experimental errors. One of the most pervasive features of experimentation is the presence of variability in the results. The easiest way to find out how much variability you face in your own results is to apply the same treatment several times and see how well the results agree in these several repetitions. A repetition does not mean just repeating the final readings that are made, but running through from the beginning the whole process of applying the treatment and taking the measurements. In some lines of work, these repetitions agree within one or two percent. In this event you may count yourself lucky, in that the experimental errors are small. Often, however, the variation in results from one application of the same treatment to another is much larger: sometimes it is enormous. In certain experiments in immunology, for example, the amount of protective serum that produces a given colorimetric response is measured. When a treatment is applied a second time, about all that we can be sure of is that the dose of serum producing the same response will lie somewhere between one quarter and four times the dose that was needed at the first trial.

What causes these variations? They can enter at any stage in the conduct of the experiment. They may be due to lack of uniformity in the raw material to which treatments are applied. In experiments in which the raw material is living, as with animal or human subjects, this is one of the most important sources of variability. Uncontrolled changes in the environment or in the equipment or machinery used, variability in the human operators and errors in the measuring devices are all contributory causes.

What do experimental errors do to us? If we are not careful, we may finish the experiment with results that are biased and misleading. If an experiment with two treatments takes two days to carry through, one way of doing it that often seems natural to the experimenter is to perform all the work and take all the measurements on treatment 1 on the first day.

Treatment 2 is handled on the second day. This prevents any danger of mixing up the two treatments, but it is the surest way to invite bias. If the raw material used is somewhat different on the two days, or the equipment is more worn on the second day, or the observers are more careless, all the results obtained for treatment 2 will be subject to a bias. Moreover, the standard methods of statistical analysis give no warning of the presence of bias. These techniques assume, in fact, that no bias is present - a point that is not sufficiently emphasized in introductory courses in statistics.

Some tests were carried out during the last war of three preventatives for sea sickness. Available for the tests were four small ships used for carrying troops. It was proposed to have the ships follow a course a short distance behind one another and to give a specific pill to all the men on a single ship. Administratively, this is the most convenient way to conduct the test. The objection was raised that this procedure might result in a bias if one of the ships proved to be less subject to rolling and pitching than the others. This was not thought likely, because the ships had been built to the same specifications in the same shipyard. However, it was finally agreed to carry out the administratively more difficult plan of giving each pill to one-third of the men on each ship. When the trials were completed it was found that there were adequate amounts of sea sickness on two of the ships. But on the third ship hardly anyone was sick no matter what pill he received, even though one of the pills contained just sugar. Further investigation showed that this ship had given trouble during its seaworthiness trials and that the shipyard had dumped in an extra load of ballast which made it unusually stable.

Even if we are able to avoid biases, experimental errors may result in the kind of indecisive conclusions to which I have already referred. In several repetitions of a test we may find that sometimes treatment 1 wins and sometimes treatment 2. Although treatment 1 wins often enough so that we are convinced of its superiority, it may not be clear how closely we can trust the estimate of the amount of superiority - which is often the important quantity for the practical use of the results.

Naturally, experimental errors are important mainly in relation to the size of difference that the treatments produce, or to the size that we are interested in detecting and studying. The experimenter who faces experimental errors of the order of one or two percent and is dealing with treatments that produce differences of the order of twenty or thirty percent has no problem of this kind. But sooner or later, in most lines of work, there comes a time when we have skimmed off the cream and are no longer working with treatments that produce large differences. When the treatments are producing differences that are of the same order of magnitude as the experimental errors, we have to find some way of coping with these errors.

What can we do about experimental errors? A three-point program might run as follows:

1. Try to find out the main causes of the experimental errors to which you are subject. Do they lie in the raw material, in equipment that gives erratic performance, in wear or fatigue, in the environment or in errors of the measuring devices? This task may sometimes require an extensive investigation.

2. Having discovered the principal contributors to experimental error, consider for each one what feasible steps, if any, can be taken to reduce or remove its effects. There are many possibilities. I shall discuss a few of them later.

3. After surveying all these proposed steps, select those that will produce the needed amount of reduction in experimental errors most economically and conveniently.

Improvements in technique. One class of methods for cutting down the effects of the principal contributors to experimental errors may be called improvements in technique. If the principal difficulty lies in the variability of the raw material, can we procure more uniform raw material? At one time I was engaged in experiments on the nutrition of pigs in England. We found that our experimental errors were of a size that made precise results difficult to obtain. On the other hand, the rival establishment, Cambridge University, which was doing the same kind of experimentation, had experimental errors low enough so that they had satisfactory precision. A careful comparison of methods revealed only two relevant differences between the two places. We had better statisticians, but Cambridge had better pigs. In Cambridge the pigs had been carefully bred so as to be uniform in their weight gains, while our pigs appeared to have been purchased in a bargain basement and showed a regrettably high degree of variability in their weight gains. Since the only pigs that we could afford were bargain basement pigs, and since an offer to trade a statistician for 20 pigs would probably have been refused by Cambridge, we abandoned this line of experimentation until better resources could be obtained.

Under the same heading come the purchase of better equipment and measuring devices, the standardization of the environment through temperature and humidity controls and so on. Naturally, these facilities cost money and may delay a program of experimentation.

There are three methods of dealing with experimental errors that have been extensively worked upon by the statisticians. These are local control (sometimes called grouping or balancing); randomization; and replication.

Local control. Local control may be illustrated by an experiment with only two treatments, each of which we intend to apply six times in order to get some replication of the results. The general principle is to divide the experiment into six separate little experiments. In each of these we take all precautions that are feasible to ensure that the comparison of treatment 1 and treatment 2 will be an accurate one.

To illustrate, an experiment was conducted in order to find out whether a dose of x-rays might enable a rat to withstand better the effects of a poison gas. There was some reason to believe that this would be so. The experiment contained two groups of rats, one receiving a preliminary dose of x-rays, the other no preliminary treatment. To receive the poison gas the rat was placed under a bell jar into which a steady stream of gas was fed. The time taken for the rat to die was measured.

What are the principal sources of error variation in this experiment? One is, of course, the rat. Rats vary in their toughness in remaining alive under doses of the gas. Hence it is important that the two groups contain equally resistant rats. The resistance of a rat presumably varies with its sex, its age, its weight and with other factors. The flow of poison gas into the bell jar might be a second source of variability, since this flow could not be kept quite uniform from one test to another.

In order to apply local control, therefore, the experimenters selected for a single trial two rats that were of the same sex and came from the same litter. This made their genetic backgrounds somewhat similar, which might affect their ability to resist, and also ensured that they were of the same age. The two rats in any one trial were both put into the bell jar together, so that variations in the amount of flow of the gas from one occasion to the other did not affect the accuracy of the comparison in any single trial. This is a good example of the use of local control to make sure that a number of potential sources of experimental error affect each of the treatments equally.

Although the experimenters had evened out the variables that I have mentioned, they were not able to control weight. The two rats in a pair differed more or less in weight. The experimenters decided always to give the x-rays to the lighter rat of the two. They argued that if x-rays showed a beneficial effect when given to the supposedly weaker rat of the pair, this would make the final results still more convincing in favor of x-rays.

Notice the logical confusion in this decision. A series of steps designed to make the comparison fair and precise is followed by a step that is designed to make the comparison unfair. The experimenters soon learned the error of their ways. In each of the first 3 trials, the smaller rat, the one receiving the x-rays, died first. What conclusion could they draw? It was time to stop and think.

There are several ways in which the experimenters could have dealt with the problem presented by variation in weights. What they did was to toss a coin at each subsequent trial to decide whether the lighter or heavier rat should receive the x-rays. This is the method of randomization. It doesn't attempt a complete equilization of the disturbing variable, but merely ensures that the trial shall be a fair game with respect to this variable. Randomization is not the best way of handling major sources of experimental errors, because a careful balancing will take care of them more adequately. It is very useful, however, for dealing with sources of variation that remain after we have exhausted the resources of balancing. We hope these sources are minor, but if we are wrong, randomization gives each treatment the same chance of benefitting from them.

Another method that they could have used was to make the experiment up in pairs of trials, giving the x-rays to the lighter rat in the first trial and to the heavier rat in the second trial. This method, based on 2 x 2 latin squares, gives a better balancing out of the weight effect than randomization. Alternatively they could have recorded the weights and then at the end adjusted the results so as to equalize weights by an objective statistical technique known as analysis of covariance.

A friend of mine, after making several attempts to read a book on experimental designs written by Miss Cox and myself, remarked that the subject seemed to be a very complicated one. It is true that the subject abounds with strange names for particular types of designs such as latin squares and graeco-latin squares, and recently with more formidable creatures like partially balanced incomplete blocks, doubly balanced incomplete blocks and so on. Although these designs are unavoidably somewhat complex in detail, their purpose is simple. They are all devices for enabling the experimenter to balance out the effects of the major disturbing variables in a great variety of different situations. It is worthwhile to have many ways for applying the notion of local control, because local control often costs practically nothing to apply, involving merely careful advance thinking about the way in which the experiment should be done.

Replication. Finally, increased precision can always be gained by repeating the experiment enough times, making sure that in each replication the test is independent of the previous replications so that the experimental errors have a chance to average out. In this way good experiments can be done with crude equipment and variable material if we replicate enough times. Of course, replication is not the answer to all our problems because it too costs money and materials.

In this connection there are methods available by which one can make rough estimates, before starting an experiment, of the number of replications needed to detect treatment differences of a given size. More frequent use of these advance estimates would avoid much wastage in experimental work. During the war I had to recommend courses of action on the basis of a summary of the results of the experiments that had been conducted on some scientific question. In a number of these situations the experimental data were practically worthless. Variability was so high and replications so few that the results were too erratic to be relied upon. The point to be emphasized, however, is that in many of these cases it could have been predicted in advance that experiments of the size and type that were done would be almost certain to give indecisive results.

Statistical analysis. By careful technique, local control plus randomization, and use of enough replications we can hope to reduce the effects of experimental errors on the average results for the different treatments to a tolerable amount. In writing our conclusions we must, however, take proper account of the experimental errors that do remain in the estimated treatment effects. The calculations by which this is done may seem mystifying to the beginner, since they derive from the theory of probability. In the standard methods of analysis, each experiment furnishes its own estimate of the magnitude of the experimental errors, making the appropriate allowance for any local control that was employed and for the number of times that the experiment was replicated. The calculations do not allow for biases that have crept into the comparisons, and there seems no way in which this can be done. Constant vigilance against bias should therefore be the watchword of the experimenter.

Summary

This paper discusses some general principles that should govern controlled experimentation. By way of introduction, some of the main reasons why experiments may fail to provide useful information are outlined, as follows.

1. The wrong questions were asked in planning the experiment.
2. The experimental treatments that were selected were incapable of furnishing answers to some of the questions.
3. The conditions under which the experiment was conducted were too remote from those in which the results were to be applied.
4. The results obtained were biased.
5. Although unbiased, the results were so erratic and indecisive as to be useless.

Although the points above are of equal importance, the remainder of the paper concentrates on the last two, on which the statistical viewpoint has the most to contribute.

Since biased and imprecise results arise from uncontrolled variability that affects the results of the experiment, the experimenter should make it his business to find out how large his experimental errors are and what sources of variation are the principal contributors to them. Various methods for reducing the effects of experimental errors and avoiding bias are discussed. These include improvements in technique, local control, randomization and replication. In any given situation, the experimenter is advised to utilize the method that seems to promise the greatest returns.

DESIGN OF EXPERIMENTS IN INDUSTRIAL RESEARCH AND DEVELOPMENT

W. J. Youden
National Bureau of Standards

At the very outset I think it appropriate to remark that the phrase "research and development" has been in use much longer than the words "design of experiment." Research and development got along without any formal design of experiments for a long time. Even today the majority of operations that are grouped under the heading of researdh and development are conducted without specific recognition of recent advances in the theory of experimental design. We can set 1925 as the earliest date experimental problems were viewed in a systematic general manner in contrast with the consideration of each research as an individual and isolated problem for the experimenter.

It seems to me that the Conference, in this first of several sessions, is concerned with the nature and function of design of experiments, just as you might first explain to a visitor from another planet the nature and purpose of automobiles, leaving to a later time the matter of explaining how to drive one. It seems important to spend a little time on the relationship of experimental design to the kind of learning process that we mean by research and development because even the textbooks on experimental design are almost completely devoted to the "how to drive" aspect of the subject. Naturally enough, these texts assume that we've bought the idea of experimental design and go on from there.

It is necessary, I think, to consider for a little this whole problem of learning. Certainly the first formal learning, undertaken in the elementary and secondary schools, deals with material that, in theory, is to be learned *in toto*. No choice is involved. At least if the student takes algebra the subject matter is set forth without any opportunity for selection by the student. Of course I don't mean the student learns all that is presented - he may learn just enough to get by but even then selection hardly enters the picture. Note, too, that the material is considered as authoritative, without doubt or uncertainty in any of the facts or evaluation.

Even in the undergraduate college years these authoritative and nonselective characteristics predominate. True, the student chooses a field of study, selects electives, but apart from this there is very little selective element in what is learned. Well, sometimes there is a kind of selection process that comes into play. Fraternities collect examinations from previous years and these may be a guide to the selection of a rather small fraction of the material in the course which will be enough to pass the next examination. But I cannot see that training in this kind of selection is going to be helpful in acquiring the kind of skill in selection required in research and development.

What I have been leading up to is the statement that people engaged in research and development are people who have achieved the ability, in one way or another, to select with good discrimination what they will try to learn with their available facilities and resources. I want to try to show that design of experiments is a discipline that serves to accelerate the acquisition of good discriminatory skill in *what* we learn and *how* we learn.

It is worth dwelling a little longer on this learning procedure. Sometimes the college senior is given a modest problem. Either that or he finishes and gets a job and is confronted with the problem of looking up something. He goes to handbooks and reference works, monographs, or the literature or an abstract journal. If he is lucky what he wants to know is there somewhere, along with an immense amount of unwanted material. There is selection here certainly, but rather easy kind because the searcher knows precisely what he is looking for. Even this kind of selection is not extensively taught in undergraduate work.

Sometimes search reveals that apparently the desired information just doesn't exist. There is a hole in the fabric of learning and somebody, if its important enough to him, will try to weave in the missing material. I suppose this process comes under the heading of research, but it seems to me a fairly elementary level of investigation.

I would like to illustrate this type of problem with an example from my own experience. The question had to do with the vapor pressure of sulfur - in particular, the vapor pressure at temperatures such as occur outside on hot days. It was important because sulfur, widely used as an agricultural dust, might give damaging concentrations of sulfur vapor on hot days. Now lots of measurements on the vapor pressure had been made - all of them above 100°C. It seemed, at first, a simple matter of extending the vapor pressure curve down to lower temperatures. We knew exactly what we wanted but nevertheless it turned out to be a challenging research problem. It was soon clear that existing procedures for determining the vapor pressure of sulfur would not cope with the extremely small amounts of sulfur given of at 30°C. The amounts of sulfur, if I recall correctly, were rather less than that of gold in sea water. In order to carry conviction a new method, when developed, ought to extend to temperatures already in the record, so that the new piece of the curve could be added with confidence to the established curve. It may seem unnecessary to comment on this reasonable requirement to give confidence in the new data. But it is, as shown by its general acceptance, something which is inherent in good experimental design in research.

I'll mention two other examples because they bring aspects of design that are not given much attention in textbooks the subject. The first concerns a visit by a pair of young men who were testing a small airtight metal container for an electronic item vital in some military hardware. The weighed containers, without the electronic component, but containing some phosphorus pentoxide, were to be cycled from heat and steam to cold and vacuum for several months. The visitors were concerned with getting an estimate of the number of containers they should test. The idea was that leaks would be revealed by gains in weight for the containers. I remarked that even if 1,000,000 containers showed no gain in weight a real sceptic might demand proof that leaks would give detectable weight increases. It would be so easy to spike this critic's guns by putting in a few containers known to have microscopic holes and collect data to show that these containers did in fact gain weight. Otherwise there might be another six months of cycling needed to cope with the sceptic. Well, I maintain that this, too, is design of experiments because we must be prepared to defend the interpretation given the data.

The other example was even more simple - so simple that the young man told his chief that there was no need to see a statistician. There were several varieties of an electrical gadget and a search was under way to find one that should not lose more than some specified percent of its performance when exposed continuously for a month to a rather high temperature. By "pure accident" I was in the vicinity shortly after the experiment started. Periodically the gadgets were taken out and brought to a standard room temperature for performance measurements. The lad showed me the first three or four points for each variety lying smoothly along various curves. "You see" he said "no statistics is needed here." "Quite right." I said, "But can you tell me if the fall off in performance is due to the added hours at high temperature or to these periodic temperature shocks when you take them out for measurement? In use," I said, "They are continuously at the high temperature, isn't that so?" Once the question was stated the lad was perfectly capable of modifying his procedure to protect himself against the question raised. Here again I include this aspect of an investigation under the heading of good experimental design.

I think it is clear by now that what I call design of experiments is inextricably interwoven with the research study itself. It is a very limited concept of experimental design that accepts without question an experimenter's program and just shuffles the work schedule into a Latin square or some other standard statistical design.

The relationship between the statistician and the investigator is indeed a very nice (I mean sensitive) balance. The experimenter has, without question, the full power to decision of what it is he wants to find out. On the other hand the statistician has responsibility to test by skillful questioning whether the experimenter's program will in fact meet the experimenter's needs.

The statistical consultant often starts with seemingly irrelevant questions about what it is that is being measured and how the measurements are made. Somehow the statistician must acquire an understanding of what is going on. Eventually, however casually introduced, questions like these are put to the experimenter: "What is the purpose of getting these data?" or "How did you come to undertake this work?" These are softer versions of the question the statistician is quite tense about - i.e., Why are you trying to learn these particular things? What will you do with the information?

I know these questions sound very much as though the statistician is impinging on territory not his. But the fact is that the statistician is has seen the results of a lot of experiments and most particularly where data were brought to him together with questions that the data wouldn't answer. Go back to the examples I mentioned. Taking the results, as first planned, to a statistician will not answer the questions raised about these projects. The answers have to be built into the project. In the modern era of complicated experimentation it is well to see a statistician first. The finest materials and workmanship may be put into a house, but calling in an architect after the house is built is not the way to use the special skills of an architect.

We are, at this stage, considering the larger aspects of research and development. Consider a proving ground with various types of terrain for testing vehicles. Suppose a number of competitive vehicles are put through their paces over a selected course. Clearly the make up of the course has been determined by the experimenter. There may be muddy, sandy, hilly and other sorts of terrain. Should the course have equal stretches of the various kinds of terrain? If the vehicles are at all selective in their ability to withstand the different kinds of rough going the make up of the course may largely determine the ranking of the vehicles. It is no answer to say that a vehicle, to be acceptable, must not fail regardless of the course. This may bring up matters such as speed and cost as prohibitive limitations. Well, what should the course be? We try to guess what the actual demands on the vehicle may be. We try to discover what are the strong and weak features of the various vehicles.

Perhaps the easiest way to emphasize the magnitude of the problem is to imagine that two independent test organizations are each given a supply of test vehicles and resources to construct whatever proving facilities they deem desirable. If the two organizations differ markedly in their findings the danger attendant on accepting the verdict of one orgnaization, if only one had made tests, should be clear to all. In a broad sense, design of experiments has the task of helping to insure that independantly conducted tests will concur in their conclusions.

Perhaps the more familiar aspect of design of experiments is its involvement in the small details of the research program. It is this aspect that I usually elect to talk about. Time will permit only a small excursion in this realm of experimental design.

I have a favorite example involving a realistic test of the merits of leather and a synthetic as a material for soling shoes. One way to conduct the test is to prepare a pair of shoes with leather soles and present them to one man and prepare a second pair with synthetic soles and give these to another man. Obviously if a difference in wear is found after a month this may be a result of a difference in the walking habits of the two men rather than a difference between the materials. But if many pairs of each kind are available and many men are included in the test then the walking habits would tend to average out to about the same. A difference, if found, could fairly be ascribed to the materials. But how much easier it would be to achieve this equality in exposure to wear by making pairs, one pairs, one shoe soled with leather, the other shoe soled with synthetic. There is a marked tendency of one foot to go along with the other. This simple paired comparison can greatly reduce the number of tests required.

Suppose there are several test materials but only two shoes to the pair. Then again it is the role of the statistician to indicate various combinations of materials that could be used for the various pairs of shoes and also to insure that there is adequate repetition to provide a basis for judging any observed differences. As a statistician I am inevitably committed to the idea of a reasonable amount of repetition. Yet when one is personally involved one may be less insistent. A recent clot in a vein of my right leg set up an elegant paired comparison with the control left leg. I am, however, most unwilling to consider a repetition of this experiment.

I have another favorite problem, the gun problem, that I use to illustrate the way design of experiments is tied up in the detailed conduct of a research and development project. It is easy to spend an hour on this example But I'll try to get my main point across in a paragraph or two. The gun problem concerns a test on five different ammunitions for 16 inch guns. There are just four rounds available of each ammunition - i.e., 20 rounds altogether I further postulate that there is considerable gun barrel wear - far too much to be neglected in the firing of 20 rounds. Now one way to schedule the firing sequence is to fire the four rounds of one ammunition, then the four rounds of another and so on like this:

Firing Order 123...............20

Ammunition AAAABBBBCCCCDDDDEEEE

The objection to this firing sequence, of course, is that ammunition A is favored by being tested on the gun when new and E gets rated on the gun after considerable wear on the barrel. A false rating of the ammunitions may result partly because the check rounds of any ammunition are fired in succession and will show pretty good agreement.

If this is not a good sequence in which to fire the rounds what would be a good one? I expect you to be surprised, at least I was, when I report that there are 305,540,235,000 different distinguishable sequences in which these 20 rounds may be fired. It would take a long time to examine all these.

The contribution that design of experiments has to make is very considable. This immense number of sequences can be classified into about ten important classes - each class having particular properties from the viewpoint of experimental design. The problem of choosing a firing sequence is now immensely simplified. One can select the class of design with the desired characteristics and then pick any sequence that is a member of that class. Design of experiments brings organization and direction into the selection of the firing sequence.

I think some time should also be given to the present limitations of design of experiments. Consider the advanced research problem posed by the development of a new material with certain desired characteristics and properties. To be a bit more specific consider this material is in the general area of plastics. The experimenter is confronted with a considerable choice of raw materials, their proportions, and the conditions for their interaction. Every experimenter is caught between the two extremes of expending his appropriation in an intensive search in a small area or in a superficial search over a broad range. Statisticians have made little more than a beginning on this difficult problem.

Perhaps an analogy will throw some light on the nature of this problem. Consider an oil painting several feet in each dimension hanging in a dark room. The experimenter has at his disposal a limited number of small spotlights. Each spotlight illuminates a small area, say two inches in diamenter. Furthermore, once a spotlight is aimed at a point within the frame of the picture it cannot be moved. These spotlights are the experiments. The two dimensional canvas is a simplified version of the usual research and development problem. The task of the leader of this project is to direct his available spotlights so as to form the best idea he can of the hidden picture.

Now various schools of thought exist about this problem. One school advocates running a traverse of lights horizontally across the picture at some arbitrary height. Then, picking the most provocative spot in this traverse, to expend the remaining lights on a vertical string through this point. The most interesting part of the picture may easily be missed using this approach.

Another school proceeds on a more flexible basis. A couple of shots are taken at random (i.e., on hunches!) and an attempt made to follow up anything that looks interesting. I mention that these data are difficult to examine statistically.

Statisiticians are just beginning to think about this problem. At first they proposed a rigid set of coordinates to be followed without regard to the evidence obtained as each shot was taken. Now it seems more profitable to try and design a thin systematic coverage using some of the spotlights, then, after appraisal of these, to concentrate the remaining lights in the most promising area. There is, I think, a real appeal in a preliminary systematic coverage, which might, in the analogy I've chosen, indicate the presence of an attractive maiden in the picture. All will agree that considerable care should be taken to locate strategically the remainging spotlights.

I look back on these remarks and seek for the major idea that I've been trying to establish. It is, I think, just this: Design of experiments is part and parcel of research and development. It has always been so. Senior experimenters have always come to use in their planning the same basic concepts that are the fundamental concepts in design of experiments. The growth of design of experiments as a separate discipline means that we are trying to get down and extend the hard bought experience of senior investigators. More than that, we want to make this experience more easily available to our junior investigators. Statisticians and experimenters will need to work in close cooperation to develop new techniques. The overriding consideration, as demonstrated by this Conference, is to make every research and development program more effective, and to get results that will stand up whatever the future holds in the way of tests with all the chips down..

THE PRINCIPLE OF RANDOMIZATION IN THE DESIGN OF EXPERIMENTS

Churchill Eisenhart
National Bureau of Standards

Synopsis*

I

ADMINISTRATIVE ADVANTAGES OF RANDOMIZATION

1. Avoids Personal Responsibility for Selections and Allocations Employed.
2. Is Widely Accepted as Fair, Just, and Objective.
3. Can Eliminate All Possibility of Personal Bias, Conscious, Subconscious or Unconscious.

   a. Bias from Conscious Acts

   Choice of "controls" such as to insure success of "treatment"
   Leaning over backwards so far to avoid favoritism that serious bias in opposite direction results
   The fallacy and pitfalls of selecting the "poorest" for "treatment," leaving the "better" for controls: Card trick.

   b. Bias from Subconscious or Unconscious Acts

   "Blindfolding" a necessary adjunct to randomization.

4. Can be a Useful Strategy in Coping with Both Men and Nature.

II

FUNDAMENTAL ROLE IN EXPERIMENTAL SCIENCE

1. Provides an Opportunity for Effects of Individual Idiosyncracies and Uncontrolled Factors to Balance Out.

   Random Positioning of a Scale, to Minimize Effects of Ready Errors and Imperfections of the Scale.
   Reduces Systematic Error by Transferring Some Constant Errors into Random Errors Which Tend to Balance Out as Replications.

* Present plans call for publication in full of this paper in the Proceedings of the Second Conference on the Design of Experiments in Army Research, Development and Testing.

2. Is Essential for Validity of Measures of Precision and Methods of Statistical Inference Based on the Mathematical Theory of Probability.

   Serves to Separate Constant (or Systematic) Errors from Components of Imprecision by Requiring Consideration of Exactly What Would Constitute a "Repetition" of the Experiment.

3. Should Always Be Done Formally, and the Resulting Allocations Strictly Adhered To.

   An Exception: When Resulting Pattern Points up to the Fact that a Better Design is Required.

EXPERIMENTAL METHODS OF DETERMINING OPTIMUM CONDITIONS*

J. S. Hunter **
American Cyanamid Company

One Dimensional Experimental Designs

Oftentimes an experimenter is interested in exploring the relationship between some response y as a continuous function of some single quantitative variable x , the variable x being under the control of the experimenter. Such studies, requiring the control of but one single variable, give rise to what are termed 'one dimensional' experimental designs.

Initially nothing may be known about the nature of the relationship; the variable y being simply an unknown function of x , i.e.,

$$y = f(x) . \qquad (1)$$

As a first step in exploring this association between y and x it is frequently assumed that the response y is a simple straight line function of x . The mathematical model expressing this linear relationship can be written as

$$y = B_o + Bx \qquad (2)$$

where B_o is the intercept of the fitted line on the y axis, and B is the slope of the line. The choice of this linear model may be based on the knowledge of analogous experimental studies, or may simply be the knowledge that quite complicated functions can be illustrated by straight lines over limited ranges. Estimates of the coefficients B_o and B can be quickly obtained using least squares. The actual least squares formulas for estimating these coefficients are given in many texts (1,2,3) and are quite easy to use.

After the straight line model has been fitted to the data it may be apparent that it inadequately represents the relationship between y and x . A statistical measure of this lack of fit of the derived equation is possible provided a valid estimate of the natural variability of the response y is available. Also, it frequently happens that the experimenter knows beforehand that a straight line is inadequate, and he may from the very start wish to fit a quadratic curve to the data, i.e. fit the second order mathematical model

$$y = B_o + B_1x + B_{11}x^2 \qquad (3)$$

The least squares estimates of the coefficients in this model is straightforward, but can become very cumbersome numerically. The calculations are particularly difficult if the response y has been recorded at some haphazard array of settings of the controlled variable x . To reduce this computational load it is usually requested that the response be recorded at equally spaced intervals of the controlled variable. This simpliest of experimental designs then permits the ready estimation of the coefficients in the second order model by using the tables of the Orthogonal Polynomials (4,5). In fact, fitting a cubic or quartic model is similarly very simple provided the suggestion of observing y at equally spaced intervals of x is followed. Since only one variable is controlled, x , this simple array of settings for x is called a one

* This paper was originally published in the Proceedings of the All-Day Conference on Quality Control at Rutgers University, September 1955. Permission to reproduce it here is greatly appreciated by the editors.

** Since Dr. M. E. Terry based his talk on this paper by Dr. J. S. Hunter, he has requested that it be printed in place of his own address.

dimensional experimental design. The principle reason for using this design is the resultant ease of calculations, oftentimes a very considerable item.

Two Dimensional Experimental Designs

Frequently an experimenter may be interested in studying some response variable as a function of two independent controlled variables, i.e.,

$$y = f(x_1, x_2) \tag{4}$$

As in the one dimensional case, the first step in exploring this function may be to fit the first order model

$$y = B_o + B_1x_1 + B_2x_2 \tag{5}$$

This mathematical expression is the equation of a plane.

If the response y is recorded at a haphazard array of settings of x_1 and x_2, the estimates of the coefficients in the model can become quite awkward. However, if the experimenter will pre-select the settings of x_1 and x_2, that is, use an experimental design, not only will the calculations be greatly reduced, but the experimenter's confidence can be made uniform over all the estimated coefficients.

The most popular experimental design for fitting this planar model is the two level factorial design. This experimental design requires that each controlled variable be fixed throughout the experimental program at a high level and at a low level, and that all possible combinations of high and low levels of the variables be run. For example, suppose an experimenter were planning to study the relationship between the expected yield of a chemical process as a function of time and temperature. A two level factorial design would take the form:

Controlled Variable Levels		Experimental Design Levels		Response	Predicted Response
Time	Temp.	x_1	x_2	y	$\hat{y}$
1 hr	240°	-1	-1	43	42
5	240	1	-1	53	54
1	280	-1	1	59	60
5	280	1	1	73	72

The mathematical model is most conveniently fitted to the experimental design levels rather than the actual levels of time and temperature. The coding mechanism associating the design variables x_1 and x_2 with time and temperature are

$${}_1 = \frac{\text{Time in hours} - 3}{2}, \quad x_2 = \frac{\text{Temp } ^\circ\text{C} - 260}{20}$$

The estimates of the coefficients in the planer model are:

$$b_o = \frac{43+53+59+73}{4}, \quad b_1 = \frac{-43+53-59+73}{4}, \quad b_2 = \frac{-43-53+59+73}{4}$$

The fitted equation is then

$$\hat{y} = 57 + 6x_1 + 9x_2$$

The method for computing these coefficients is quickly recognized. The four predicted values of the responses are recorded in the table. The regression analysis table (or analysis of variance table if you insist) is

		df
Total Corrected Sum of Squares	472	3
A Effect	144	1
B Effect	324	1
Residual	4	1

As a check we note that $\sum_{i=1}^{4} (\hat{y}_i - y_i)^2 =$ Residual Sum of Squares

If the settings of the design variables are plotted out on graph paper they will form the vertices of a square. Another experimental design, called a first order experimental design, that will also quickly provide estimates of the coefficients in the model is formed from the vertices of an equilateral triangle.

Estimating a Path of Steepest Ascent

Frequently the question is asked, "What combination of levels of the controlled variables will give the highest response?" In attempting to answer this question several alternative experimental approaches are possible. The experimenter may randomly select different settings of the controlled variables, try them in his laboratory or pilot plant, and with luck gravitate to the maximum point. Or he may run a sequence of experiments over a grid covering the entire region of interest and thus literally map the response. Both these approaches can quickly require a great many experimental runs and are usually avoided.

A favorite attack is the method of one factor at a time. This method requires that the experimenter hold all the controlled variables save one at some constant level, and then vary the remaining single variable until a maximum response is observed. Then holding this variable at its optimum value, a second variable is varied, and so on. The method is illustrated in Figures 1 and 2.

Suppose a response y (% of theoritical yield) is a function of time (measured along the x_1 axis), and temperature (measured along the x_2 axis). Suppose further that the response, when viewed geometrically, has the appearance of a mound with a single maximum point. This response is illustrated by means of the contour diagram in Figure 1.

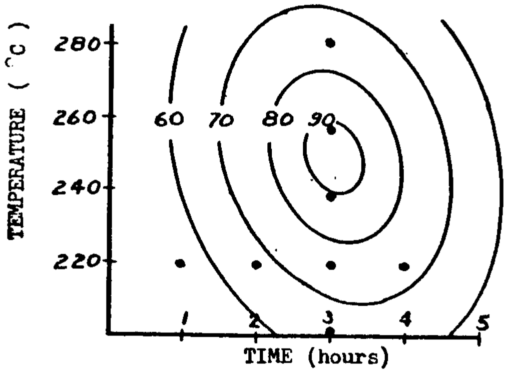


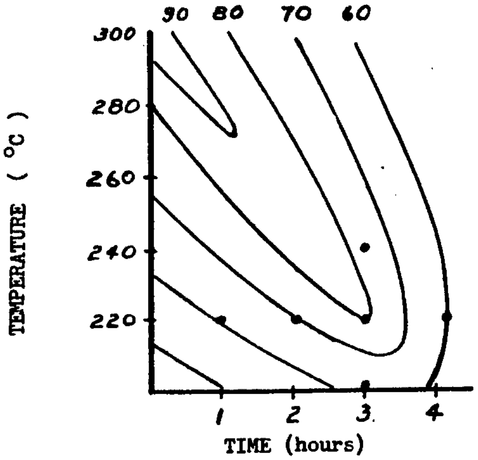
Figure 1

Contours of Equal Response for y % of Theoritical Yield

Illustration of Method of One Factor at a Time

Using the method of one factor at a time, the experimenter might hold hold the temperature constant at 220^o and vary the time in one hour increments. The resultant experimental trials are shown by the horizontal line of heavy dots in Figure 1. Deciding that three hours was the best time, he would then hold the time constant at this value and vary the temperature in, say, increments of 20^o. This gives the series of experiments illustrated by the vertical line of dots. Thus the procedure leads the experimenter to the point of maximum response.

However, imagine that instead of a mound, the response were only slightly more complicated and had tne appearance of a rising ridge as shown in Figure 2.



'Figure 2

Contours of equal Response for y = % of Theoritical Yield

Illustration of Method of One Factor at a Time

The identical procedure illustrated by the dots in Figure 2 leaves the experimenter with the conviction that the setting of three hours and 220^o is optimum for steps higher or lower in time or temperature from this point will produce a decrease in the response. Obviously higher yields are possible. In this instance the experimenter is stuck on a ridge and can only discover the higher yields by varying time and temperature simultaneously.

A method which guarantees that the experimenter will tend towards the maximum point regardless of the form of the response surface (save that it is continuous) is the method of steepest ascents (6,7,3). This technique makes use of the first order mathematical model and either the two level factorial design or the first order experimental design. The basic concept behind the idea of predicting a path of steepest ascent is simply that within a small region a plane will do a good job approximating a curved surface. This is analogous to the old argument that straight lines do a good job of approximating curved lines over small distances. The plan is then to predict the best fitting plane in some small sub-space of the experimental region. Then noting the tilt of the fitted plane, a path of steepest ascent can be predicted. Experiments are then performed along this path until a decline in response is noted. Additional observations taken in and around this point can confirm whether a maximum has been reached, whether a new path of steepest ascent should be predicted, or whether the response surface should be mapped in this important region.

For example, imagine that a two level factorial design had been run as illustrated by the small circles in Figure 3.

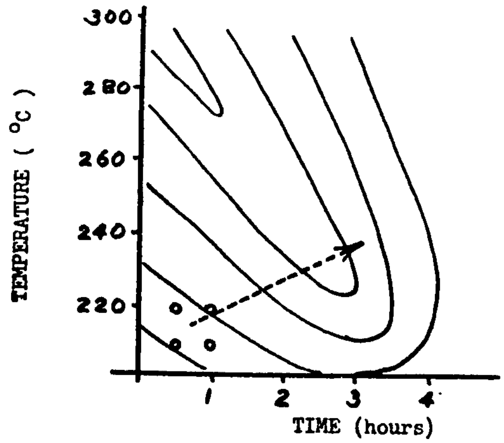


Figure . 3

Contours of Equal Response for y = % of Theoritical Yield

Illustration of Path of Steepest Ascent

The following table of results would be observed:

Controlled Variable Levels		Experimental Design Levels		Response
Time	Temp.	x_1	x_2	y
0.5 hr	210°	-1	-1	51 %
1.0	210	1	-1	57
0.5	220	-1	1	55
1.0	220	1	1	61

Fitting the first order model given in equation (5) one obtains for the best fitting plane in this region

$$\hat{y} = 56 + 3x_1 + 2x_2$$

The path of steepest ascent is now determined by the coefficients of x_1 and x_2 . In this example therefore, we are advised that for every three units x_1 is changed, x_2 should be simultaneously changed two units. The changes requested are porportional to the size and signs of the coefficients. The units that are considered here are the units of the design variables, not the levels of the controlled variables. Thus, starting from the center of the design array, the estimated path of steepest ascent is plotted as shown. Experiments along this path would lead to the creat of the ridge where a second series of experiments would direct the experimenter up the ridge. Should the response be a simple mound as in Figure 1, the very first predicted path should lead the experimenter very close to the maximum point.

Surface Fitting Designs

Although locating the maximum point of a response variable can be valuable, experimenters are often asked to describe a response quite generally over an entire region. This requirement is ideally met if the experimenter can actually construct the contour lines describing the form of the response surface. On other occasions the experimenter may be asked to find a region which is optimum, not with respect to a single response, but with respect to two or more responses considered simultaneously. One interesting means for finding points or regions that are optimum with respect to several responses is to superimpose the contour diagrams of the responses. For example, imagine the response indicated by the contours in Figure 1 as indicating the yield of product A, and the response illustrated in Figure 2 as the yield of the simultaneously produced product B. By superimposing these two contour diagrams, as shown in Figure 4, one can determine the settings of time and temperature which will simultaneously give, say, a yield of 90 for A and a yield of 80 for B.

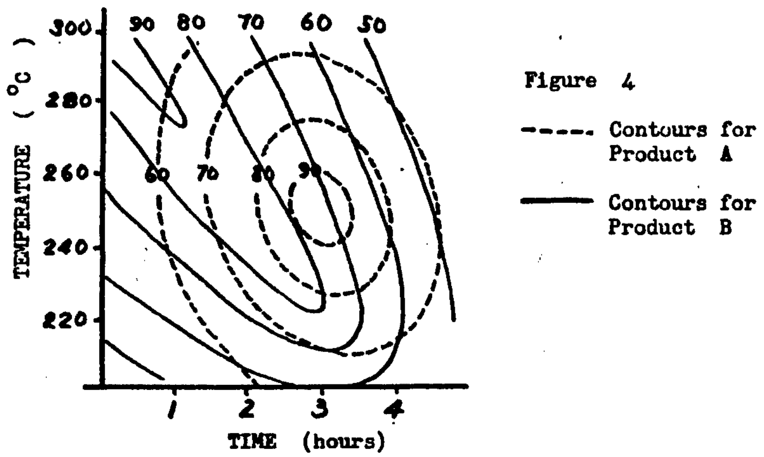


Figure 4

If a response surface is planar then the contour lines become parallel straight lines. Methods of fitting first order models have already been discussed. However, if the surface is thought to be curved, or if the controlled variables interact in affecting the response, then a planar representation is no longer adequate. To estimate a non-planar surface, a second order mathematical model can often be profitably used. The second order model in two dimensions is:

$$y = B_o + B_1x_1 + B_2x_2 + B_{11}x_1^2 + B_{22}x_2^2 + B_{12}x_1x_2 \qquad (6)$$

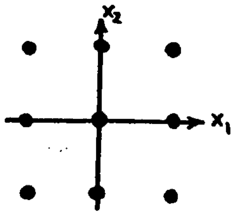
This is a very versatile mathematical model. Imagine that the coefficients in the model are known (or have been estimated). Next, set y equal to some constant, say $y = 80$. The resultant equation will be the equation of a circle, an ellipse, an hyperbola, a parabola, or even a straight line. The actual geometric shape depends, of course, on the signs and magnitudes of the various coefficients. Furthermore, everywhere on the line, regardless of its particular form, y would equal eighty.

Thus, this mathematical model can be used to estimate the contour lines of a response surface (3,7). For example, imagine the 'unknown' response surface given in Figure 1: a simple mound. Imagine further that the response has been recorded at enough settings of time and temperature, x_1 and x_2, so as to permit the estimation of all the coefficients in the second order model given in equation (6). Let us also assume that the error is recording the response y is small. Then, by setting $\hat{y}=90$, 80, 70, etc. the fitted equation would undoubtedly generate three perfectly concentric ellipses, the ellipse for $\hat{y}=90$ being innermost, the ellipse for $\hat{y}=70$ being outermost. These ellipses become in fact the predicted contour lines. Viewing such a set of contours, the form and nature of the response surface should be immediately apparent to the experimenter.

Rotatable Designs

Estimating the six coefficients in the second order model given in equation (6) can become very awkward if care is not taken beforehand in selecting the levels of the controlled variables. One experimental design (not a rotatable design) that can be used to reduce this labor of computation is the three level factorial design. This design requires that each of the controlled variables be held at a high (-1), middle (0), and low (-1) level, and that all combinations of levels and variables be run. In general, the number of experiments required is $N=3^k$, where k equals the number of controlled variables. Thus, the co-ordinates of a two dimensional, three factor factorial design are:

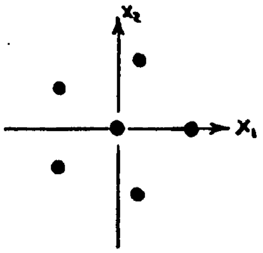
	Experimental Design	Levels
	x_1	x_2
Run # 1	-1	-1
2	0	-1
3	1	-1
4	-1	0
5	0	0
6	1	0
7	-1	1
8	0	1
9	1	1



Plot of the Three Level Factorial Design

However, although this array of points will greatly simplify the calculations required for estimating the coefficients in the model, it provides an unfortunate balance in the variances of the estimates of the linear, crossproduct, and quadratic coefficients. As a matter of fact, the variances of the estimated coefficients will literally change as one moves about in the space of x_1 and x_2., even though one should move in a perfect circle about the center of the design.

A class of experimental designs called 'Rotatable Designs' (8,9) has recently been develpoed which not only possess the quality of easy calculations, but also provide that the variances of the estimated coefficients remain constant as one moves in a circle abbout the center of the design. This can be shown to be a very desirable property for any experimental design. Furthermore, the configuration of points in the space of x_1 and x_2 are easily remembered, for they form the vertices of the regular figures, starting with the pentagon. Thus the simpliest second order rotatable design is illustrated on the next page.



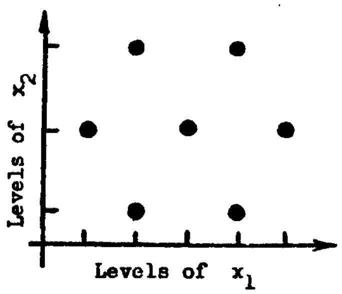
The Pentagonal Design

Design Co-ordinates

x_1	x_2
0.000	0.000
1.000	0.000
0.309	0.951
-0.809	0.588
-0.809	-0.588
0.309	-0.951

A required additional point is one (or more) at the center of the design array. Replication of this single point will not only provide a valid estimate of the experimental error, but is also valuable in building up the predictive power of the resultant fitted model in the interior of the design.

The hexagon, septagon, octagon, etc., all with one or more center points, all provide second order rotatable designs. The choice of one rotatable design over another is usually determined by the number of trials the experimenter wishes to run, and by the number of levels he must maintain for each variable. For instance, a hexagon design requires that x_1 be controlled at five different levels, but x_2 only at three different levels, i.e.,



The Hexagonal Design

Experimental Designs in Three or more Dimensions

Situations in which some response may be the function of three or more variables are not at all uncommon, i.e.,

$$y = f(x_1, x_2, x_3, \ldots, x_k) \qquad (7)$$

The first order model in three dimensions is a simple extension of equations (2) and (5), i.e.,

$$y = B_o + B_1x_1 + B_2x_2 + B_3x_3 \qquad (8)$$

and the general first order model in k dimensions becomes

$$y = B_o + \sum_{i=1}^{k} B_i x_i \qquad (9)$$

These mathematical models can be used to estimate a 'planar' response surface, or to predict a path of steepest ascent, regardless of the number of dimensions (the number of controlled variables) involved.

As is the case of two dimensions, the two level factorial design may be used to estimate the coefficients of a first order model. However since the number of experimental points, $N = 2^k$, quickly becomes large as k increases, fractional factorial designs are used. These designs provide that a $\frac{1}{2}$ replicate, or even a $\frac{1}{4}$ or still smaller fraction, of the total number of required points be run. The fraction permitted depends, or course, on the number of dimensions. For example, in three dimensions the design points for a full two level factorial design, and a $\frac{1}{2}$ replicate ($\frac{1}{2}$ fraction) of a two level factorial are as follows:

x_1	x_2	x_3
-1	-1	-1
1	-1	-1
-1	1	-1
1	1	-1
-1	-1	1
1	-1	1
-1	1	1
1	1	1

Three dimensional
Two Level Factorial.

x_1	x_2	x_3
-1	-1	-1
1	1	-1
1	-1	1
-1	1	1

$\frac{1}{2}$ Replicate of a
Three Dimensional
Two Level Factorial

also a

Three Dimensional
1st Order Design.

If these points are plotted in space those of the full factorial design will form the vertices of a cube, while those of the $\frac{1}{2}$ replicate will form the vertices of a tetrahedron.

The first order design in three dimensions is identical to the $\frac{1}{2}$ replicate of the two level three dimensional factorial, i.e., its co-odinates are the vertices of a tetrahedron.

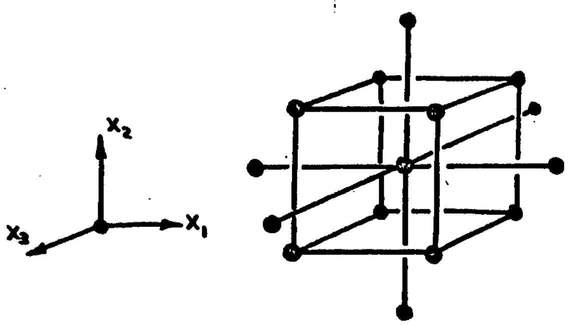
Two level factorial, and first order designs, can be constructed regardless of the number of dimensions involved (10). They become in higher dimensionality the vertices of a hyper-cube (or a particular subset of a hyper-cube if a fractional factorial is used), and the vertices of a hyper-tetrahedron respectively. On those occasions when the number of dimensions is one of the arithmetic series 3, 7, 11, ... a first order design will be found to coincide identically with some fraction of a two level factorial. For dimensions of a number other than these the vertices of a tetrahedron cannot everywhere take on the values of plus or minus unity. For example, the vertices of a tetrahedron in four dimensions, i.e., the first order experimental design in four dimensions are

x_1	x_2	x_3	x_4
-1.581	-0.913	-0.645	-0.500
1.581	-0.913	-0.645	-0.500
0	1.826	-0.645	-0.500
0	0	1.935	-0.500
0	0	0	2.000

It is of course possible to extend the second order mathematical model to higher dimensions, i.e., to consider a curved response as a function of three or more controlled variables. In these instances the experimenter will predict contour surfaces instead of contour lines, and although the interpretation, and visualization, of second order responses in three or more dimensions is not easy, it has been successfully used in several instances (6,7,11,12). The second order model in three dimensions is

$$y = B_o + B_1x_1 + B_2x_2 + B_3x_3 + B_{11}x_1^2 + B_{22}x_2^2 + B_{33}x_3^2 + B_{12}x_1x_2 + B_{13}x_1x_3 + B_{23}x_2x_3 \quad . \qquad (10)$$

Several three dimensional experimental designs are available for estimating the coefficients in this model. The three level factorial will recuire $N = 3^3 = 27$ points, usually too large a number. Three rotatable designs exist: the icosahedron design which is formed by the 12 points at the vertices of an icosahedron plus one or more points at the center, the dodecahedron design with 20 points plus one or more at the center, and finally the cube plus octahedron design -- better known as the 'central composite design' (7) -- formed from the 8 points of a cube, the 6 points of the octahedron, and one or more points at the center of the array. This latter design is illustrated below.



Three Dimensional Central Composite Design

To form a rotatable design from the central composite design, the radius arm of the octahedron must equal 1.68. The co-ordinates of the vertices of the cube all equal plus or minus one. Rotatable designs, in the form of the central composite design, are available in all dimensions.

Conclusion

It has only been possible to paint with a very broad brush the concepts of steepest ascent and surface fitting, and to briefly describe the experimental designs associated with these methods. For a more detailed description of steepest ascent and surface fitting one should read reference (7). Another excellent description, coupled with careful explanations of the numerical details appears in (3). The portion of this book describing these topics is written by Dr. G. E. P. Box, the originator of these methods.

Bibliography

1: Statistical Methods for Chemists: W. J. Youden, J. Wiley & Sons New York, 1951.

2: Introduction to Statistical Analysis: W. J. Dixon & F. J. Massey, McGraw-Hill, New York, 1951.

3: The Design & Analysis of Industrial Experiments: O. L. Davies (Editor), Hafner, New York, 1954.

4: Statistical Tables for Biological Agricultural and Medical Research: R. A. Fisher & F. Yates, Hafner, New York, 1953.

5: Statistical Methods: G. W. Snedecor, Iowa State College Press, Ames, Iowa. L950.

6: Box, G. E. P. & K. B. Wilson: "On the Experimental Attainment of Optimum Conditions" J. Royal Stat. Soc. Ser B, 13, 1951.

*7: Box, G. E. P. "The Exploration and Exploitation of Response Surfaces" Biometrics 10, 1954.

8: Box, G. E. P. & J. S. Hunter: "Multifactor Designs", Institute of Statistics Mimeo Series, Raleigh, N. C., 1954.

9: Hunter, J. S. "Searching for Optimum Conditions", Trans. N.Y. Acad of Science, Ser 11, 17, 1954.

10: Cragle, R. G. et al. "The Effects of Various Levels of Sodium Citrate, Glycerol, and Equilibration Time on Survival of Bovine Spermatozoa After Storage at -79°C" J. Dairy Science, 38, 1955.

11: Pike, F. P. et al. "Application of Statistical Procedures to a Study of the Flooding Capacity of a Pulse Column", Submitted to I & E Chem.

*7. "The Exploration and Exploitation of
Response Surfaces"
Ordnance Project No. TB2-0001 (832)
Dept. of Army Project No. 599-01-004

OPERATIONAL EXPERIMENTS

Paul F. Michelsen
Operations Research Office
Johns Hopkins University

Most of the participants of this conference are concerned with providing assistance in the continuing problem of increasing the Army's effectiveness. Often this assistance is in the form of constructive base for decision for the improvement of the Army's tactics, doctrine, organization, or materiel.

The sources of data to support such reports and recommendations are many and varied. "Good" or reliable data, in general, is difficult to obtain. Historical data such as combat records from World War II, and from the Korean conflict, even if judged reliable as study sources in the past, are now increasingly remote bases for extrapolation as time goes on.

SOURCES OF DATA

COMBAT RECORDS
ENGINEERING TESTS
USER TESTS
MAP EXERCISES
MANEUVERS

Fig. No. 1

The most significant sources of military operational data to supplement combat records are: engineering tests, user tests, (as at the CONARC boards and schools), information derived from map and field exercises, and maneuver records and observations.

Engineering data concerning a weapon, such as its maximum rate of fire, its ballistic dispersion and its reliability under set environmental conditions, is usually readily available, and is in general, quite accurate. It often suffers, however, from being too restricted in its application. Data acquired from the other sources is often judged less usable due to real or assumed inaccuracies in the basic measurements, or due to limited scope, and insufficient number of observations, or the pressure of too many uncontrolled variables. Perhaps the largest problem in engineering test data, however, is the lack of generalization of the specific data into the desired particular military operational context.

Weapon system development is continuous and new tactics are required to utilize (and to combat) such development. There is an urgent need for decisive data for developing appropriate new tactics. A requirement exists, therefore, for methods by which we can develop such decisive data---data that deals with the interaction of men and machines. Since the final criterion for the Army is combat performance, some sort of combat-like evaluation is mandatory. The Operational Experiment offers such an evaluative method. It is intended to employ troops, weapons, and material, in the closest approximation to the tactical situations of actual combat.

The Operational Experiment is a field evaluation technique, designed to test critical aspects of an operation within the framework of the operation itself. When used in conjunction with an Operational Game, (another operational simulation method), the combination forms a new and powerful research tool. This combination can be utilized to develop required data; to evaluate present tactics, doctrine and organization, as well as interactions or combinations; but even more important, the Experiment-Game combination should become an integral element in the development of innovations of those same basic areas of effort. (The specific innovations or tactical inventions are still, primarily the function of the military.

The Operational Experiment as a field evaluation technique possesses most of the same strengths and weaknesses of any other field experiment--it deals with the real world, actual terrain, typical military situations, and military personnel; it suffers from uncontrolled variables due to the same features, plus restraints imposed upon the interaction zone, such as loss of life or destruction of weapons and equipment. These restraints, while presently necessary, lower much of the motivation of the personnel involved, as well as rendering "good" experimental data unlikely.

The Operational Experiment emphasizes the reduction of these undesired restraints, and places the accent on "realism". In addition to "realism", the accumulation of data becomes more automatic.

The Operational Game is the mathematical and probabilistic model of an operation. It is the quantitative formulation of what occurs during a military operation; where the Experiment provides the values of the parameters. The Game, in turn, is of extreme value in determining the selection of variables, their probable range, any precision requirements and some of the other specific considerations that have received the attention of the designer. The game, either map type for the quicker, lower accuracy, efforts, or the computer type for the more involved models, can be performed many times to obtain a distribution of outcomes. Sensitivity of the outcome distribution to the parameters is inherent in the solution of the model. This further reduces the scale of effort for a selected level of experiment.

Let's take a look at how the Operational Experiment--Operational Game combination functions. Before much more is said, the comment should be made that Operational Experiment or the Operational Experiment/Operational Gaming technique is not proposed as an ultimate device. The combination does promise, however, to add many more quantitive factors to what has hitherto been largely an area of qualitative judgments.

The size of action adaptable to this technique varies from a single interaction (placed in proper context) up through company size operations, with the expectation that battalion combat team problems will be handled in the near future as the methods are more fully developed and implemented.

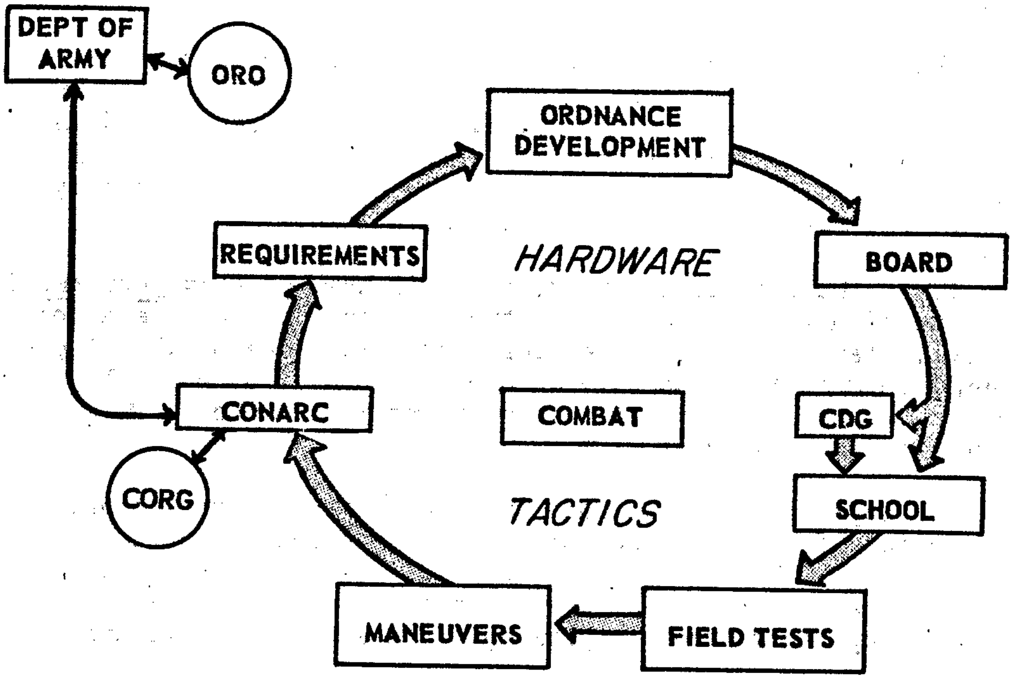


Fig. 2—Development Cycle for Ordnance Units

Figure 2 shows in an elementary way the military development cycle for equipment and, to some extent, doctrine. The Department of the Army has the overall responsibility and defines the mission, and assigns it to Continental Army Command (CONARC). For convenience (at this point) we shall break into the development loop. CONARC determines the general requirements and specifications of the equipment in order to attain the desired mission. The equipment is engineered by the ordnance development groups, prototyped, tested, modified, retested, and when completed the prototype and/or pilot units are delivered to the cognizant board for initial user tests. The Board tests, recommends changes, and when its tests indicate satisfactory performance, the equipment is turned over to the pertinent school for further user tests and maneuvers of wider scope. The School cooperates in developing the tactics, and a good portion of the doctrine for Army use of the equipment. The box marked CDG represents the military Combat Development Groups which assist the schools in the missions under CONARC, and who, as part of their duties, assist the schools in the planning of experiments and operational tests. This School-Combat Development Group effort is closest in its intended purpose to the Operational Experiment concept.

The next step in this development cycle comes when the School has readied the equipment and the associated School-developed tactics for field tests which involve larger combinations of Army units. Such tests usually involve the equipment in maneuvers where the greatest degree of realism in employment is intended.

The box in the center marked "combat" is a reminder that in time of war when actual combat testing exists, the information thereby acquired is of use in all portions of the cycle. It must be remembered, however, that the utilization of even such valid information is severely limited by the practical difficulties that restrict the acquisition, accuracy, and pertinacy of any data in the midst of actual battle.

Since the important consideration is how the Army performs in combat, the type of data most desired is the kind that combat operations alone develop. Next most desirable, for a number of reasons, is the conduct of planned interactions that are as close to combat as possible or necessary for the purposes of acquiring significant information. This is the intent of the Operational Experiment.

The experiment is designed to develop the desired information; to use military forces, and equipment, in a tactical situation, with extensive measurements. The amount of control of the operation is to be minimized, in fact, the conduct of the operation to be as unrestrained as possible with the hope that "free" experiments are in the near future.

Figure 3 shows the way that simulated combat, or Operational Experiments feed evaluations of equipment, tactics, doctrine, or organization, into the development cycle. The greater accuracy over presently used methods is due not so much because to its scope, but primarily to the increased accent on realism, and its objective of eventually obtaining basic measures for truer evaluations. The derived information will affect the development cycle at all the points shown on the chart. It should permit a speed-up of the development rate and a corresponding decrease of lost time, with the net result that the weapon and its accompanying tactics will be in organizational being in much shorter time than the present short range step-by-step, mode of development.

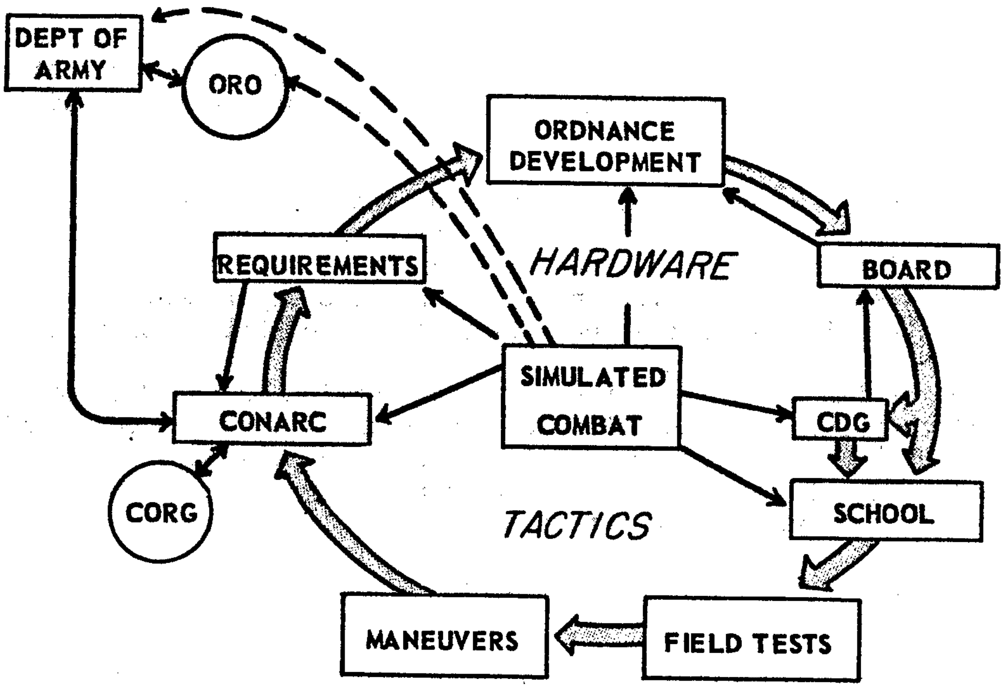


Fig. 3—Proposed Development Cycle

To accomplish the previously defined aims of the Operational Experiment we require:

a) Realism heightening devices.
b) Data acquisition methods and devices.
c) Data handling and storage.
d) Data reduction.

Realism heightening devices have two important qualities: They should duplicate the operational decisions of the real weapon; and they should include as many of the weapons operational characteristics, sound, flash, etc. as consistent with safety requirements.

Data acquisition devices to indicate position, provide interaction information, and record other physical data are being considered. In some cases, the simulators will simultaneously record some of this desired data.

Another possibility for data acquisitions and handling may be found in the adaptation to this measuring use of the many recent automation devices that already exist with coded output or which can be equipped with coding devices.

The concept of coded output measuring instruments now appears to be firmly established and, even if only a few devices are as yet usable from our point of view, our needs will surely be met as the instrumentation effort proceeds. An important point is that the experiment designer should be acquainted with the needs and potentialities of this approach.

Another important element of data acquisition is the determination of appropriate measures to be obtained from an experiment. A good number of these are as yet unknown; some will not be measurable directly; even the presently possible measures need to be carefully evaluated and their interrelationship and required accuracy must be established.

Codifying measures such as "intelligence" will require considerable attention.

Once coded, the handling and storage problem is well within the capabilities of modern systems. The reduction of these huge amounts of data can similarly be handled by automatic computers.

All the preceding points up the strong requirement for a cooperative effort between the analyst, the instrumentation, and the data reduction groups, to evolve systems suitable to productive Operational Experiments.

The Operations Research Office has a developed interest in increasing the realism of maneuvers. It is pushing the substitution of operational simulation devices for certain aspects of the umpiring of maneuvers. These devices can also be readily incorporated into Operational Experiments as can a number of more conventional training aids which are often suitable as decision assisting units.

Simulators are necessary and desirable substitutes where the real effect is too dangerous to human life. The compromises made with the effect they are simulating need to be evaluated within the context of the experiment they are being used in. These compromises are in terms of the functioning of the simulator in relation to the functioning of the real thing.

A simulator now being developed by ORO to establish a good measure of realism to the interaction of tank vs. tank engagements in both maneuvers and experiments is known as the Aimed-gun-fire simulation device. It will instantaneously handle the entire sequence of decisions required of an umpire in a gun duel between tanks, as well as larger tank vs. tank engagements. For example, where presently Tank A informs him that he has fired at Tank B, an umpire must decide can Tank A see Tank B; is Tank A's gun aimed and ranged properly; would Tank A probably hit Tank B. As a basis for his decision the umpire depends upon his knowledge of hit probability tables for the ammunition simulated, the range in question, and the vulnerability of the target tank.

Since his information is in the form of probabilities a further requirement upon the umpire involves a subjective interpretation of the tables in order to make the most "realistic" decision.

In the meantime, or perhaps at the same time, Tank B indicates that he has "fired" at Tank A. The umpire, or umpires, must reach the additional decision as to which fired first. It is easily seen that considerable time is consumed if results are to be "accurate" or if snap decisions are given by the umpire, large errors can accumulate in regard to actual interaction results. The umpires decisions then become primarily a function of his personal experience and opinions. They tend to appear arbitrary, if not actually so, which causes the participating troops to lose the motivation or more direct effects. The information obtained from the interactions then becomes of little use.

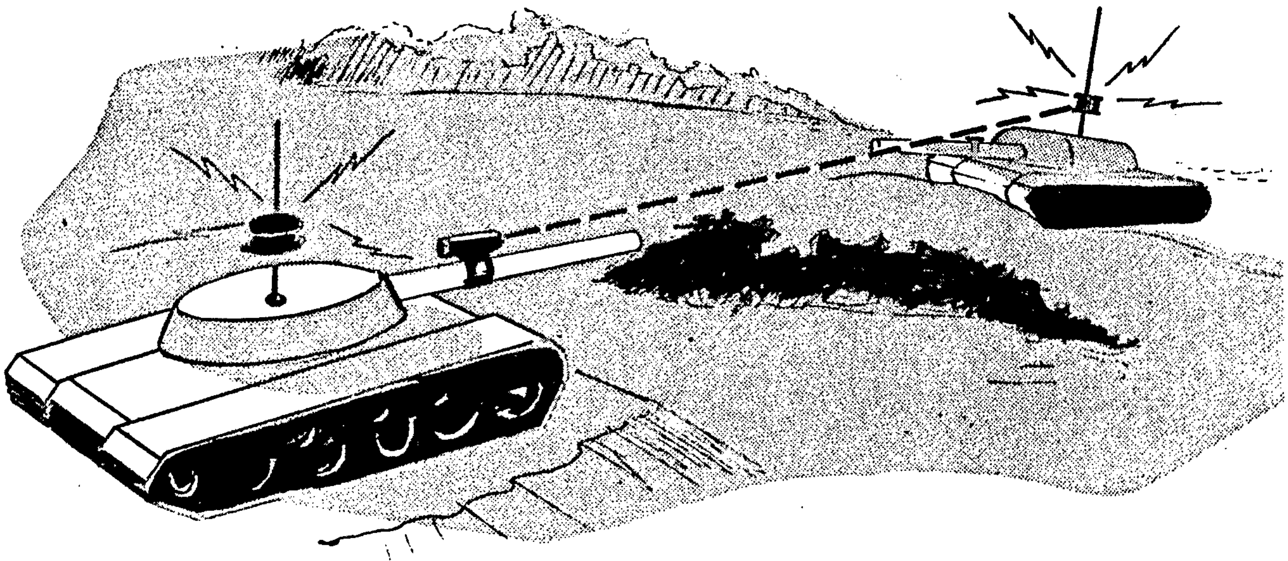


Fig. 4—Aimed-Gun-Fire Simulator Device

Fig. 4 shows a possible form of the Aimed-gun-fire simulation device. It will decide instantaneously the whole series of interactions noted above. Because these decisions are instantaneous, simulated combat of this new type can progress realistically. At the same time the simulated weapon requires the same attention from the operators in leading and training, for example, and thereby injects that further portion of the realism requirement.

In this interaction simulation shown, the tanks have a narrow beam light detector mounted on, and aligned with, the gun barrel. The nearest tank is "sighted" and aimed by means of the light source mounted above the turret of the target tank. Since the visual angle of the simulator device is very small, the gunner of the near tank is required to aim his weapon at least as accurately as he would for an actual firing. The radio antenna above the light source transmits the "hit" information to the target tank, where the appropriate signal operates a stopping mechanism and sets off a "flash-and-bang" unit, thereby simulating the effect of a "hit."

Another simulator conceived at ORO, but which is being brought to fruition by the Combat Operations Research Group of CONARC for their field experiments, is an Anti-tank Mine simulator. Like the Aimed-gun-fire device the decision aspects and effects of the simulator are as rapid as the Anti-tank Mine itself.

These are two operational simulators; the first still early in its development stage, the second nearly complete, which will enormously improve the significance of the results of an experiment which involves tanks, other aimed-gun-fire weapons (such as anti-tank guns) and Anti-tank Mines. Other weapons and weapon effects will subsequently be simulated for use in the Operational Experiments. As a not inconsiderable bonus, they will continue to raise the motivation and training benefits to the using personnel as well.

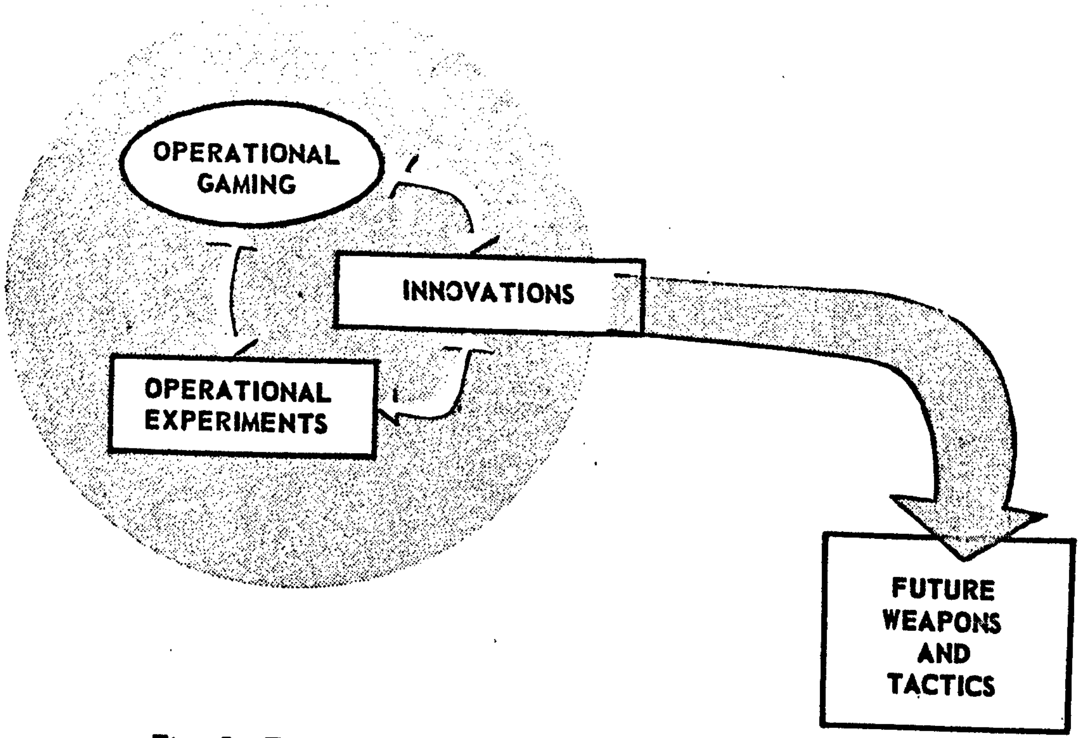


Fig. 5—Evaluation Method for Innovations

Operational Experiments or the Operational Experiment-Operational Game combination can potentially develop a firm basis for the evaluation of present weapons, tactics, doctrine, and organization, however their most unique and valuable role may well prove to be their use for the evaluation of innovations in those same areas of Army interest. (Fig. 5). For example, if a new weapon is indicated from basic tactical or engineering considerations, the new characteristics can be simulated by an operational device. Operational Experiments could verify the desirability and refine the initial specifications of the new weapon, while still in the concept stage. Then while the weapon is being developed and prior to its actual availability, further simulation in other experiments could develop or modify appropriate tactics and doctrine. By the time that the weapon is issued to the troops, tactics and doctrine will be in existence in far more tested form than ever possible before.

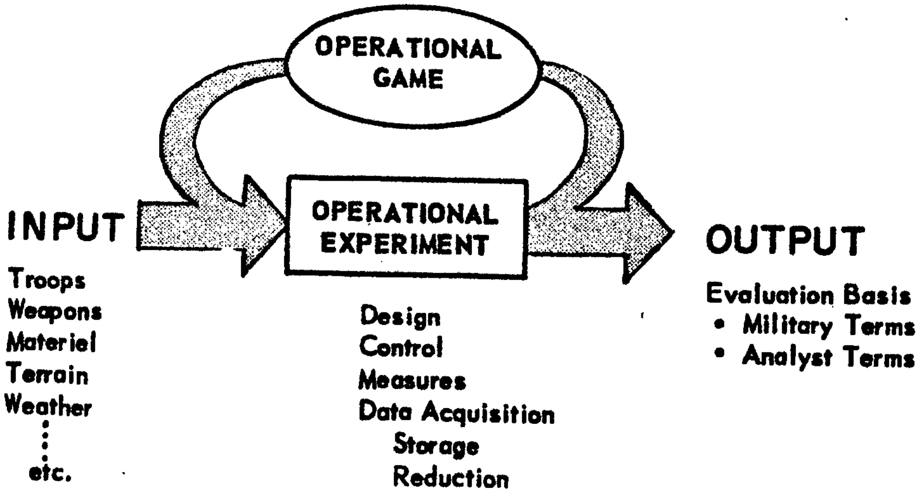


Fig. 6—Operational Experiment Structure

It is of interest to collect the several described parts of the Operational Experiment together. (See Fig. 6). The input of the experiment consists of the military units plus the terrain, weather and similar other preselected elements of the experiment. The analysts function is to design the experiment, specify the levels, and as well as assisting in performing the necessary controls of the equipment. He determines the measures and evaluates the information derived by the data acquisition devices and techniques.

Finally the analyst group reduces the data into two forms of output; in military terms so that the military units can use the results to appraise the effectiveness of the operation and thereby guide them in the development of the operation; and in analysts terms for use both in the accompanying operational game and for other analysts studies. The diagram also shows the feedback arrangement of the Operational Game or games.

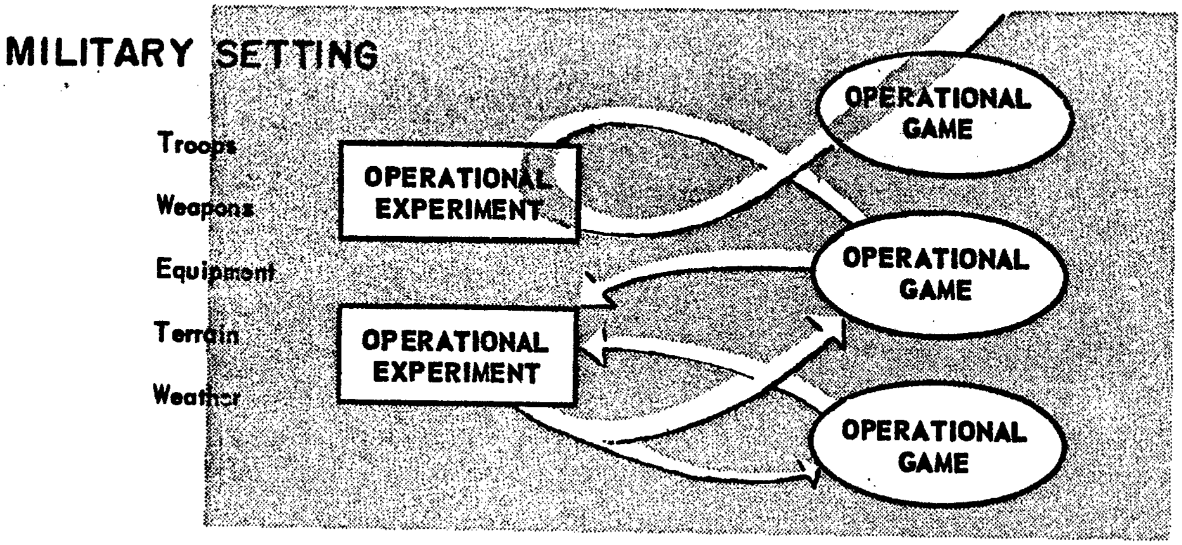


Fig. 7—Operational Experiment/Operational Game Interaction Levels Interaction Levels

This is again shown in the last illustration (Fig. 7) where the Operational Experiment can profit from a lower level, the same level and from higher level, games. The operational game, in turn, is dependent on experimental information obtained from various levels of experiments. The diagram also helps to point out that there is a practical limit to the size of effective unrestrained experiments and that from this stage on, the mathematical game, founded on well verified data from the lower level games and experiments can handle investigations of larger military, and possibly non-military interactions.

TO CONCLUDE

An Operational Experiment is conducted to analyze the interactions of men and machines in order to form a sounder basis for decision. The amount of control, the design of experiment, the amount of data collected, and the number of tests run, is dependent merely on the scope and degree of interest and the possible or desired accuracy.

Recommendations for improvements of tactics, organization and materiel can be substantially strengthened, should be more readily acceptable to the cognizant Army group, and should increase the quality and tempo of the improvement cycle.

OPERATIONAL GAMING AND THE DESIGN OF EXPERIMENTS

W. Edward Cushen
Operations Research Office

Summary. In recent years the traditional military war game has been developed into a research tool capable of examining a large number of important features of a complex system in the context of their interacting effects. The thesis of this paper is that there is a two-fold potential connection between operational gaming and the design of experiments, in the conventional sense of the word, which must be developed to permit useful inferences to be drawn from operational gaming. Reciprocally, the operational game can be expected to indicate the required format of an experimental design for maximum resultant information.

The Importance of the Context. Let us return to the title of the conference: "The Design of Experiments in Army Research, Development, and Testing." Imagine, for a minute, that the ambitious task of preparing the Army R&D program is incumbent upon the reader. Within the budgetary limits of an R&D program, and with a view to the later budgetary restrictions on the procurement of the items developed, it is necessary to invent, improve, modify, and combine the materiel of war in such a way that, when used by the men available, there is maximum expectation of success in a potential armed conflict.

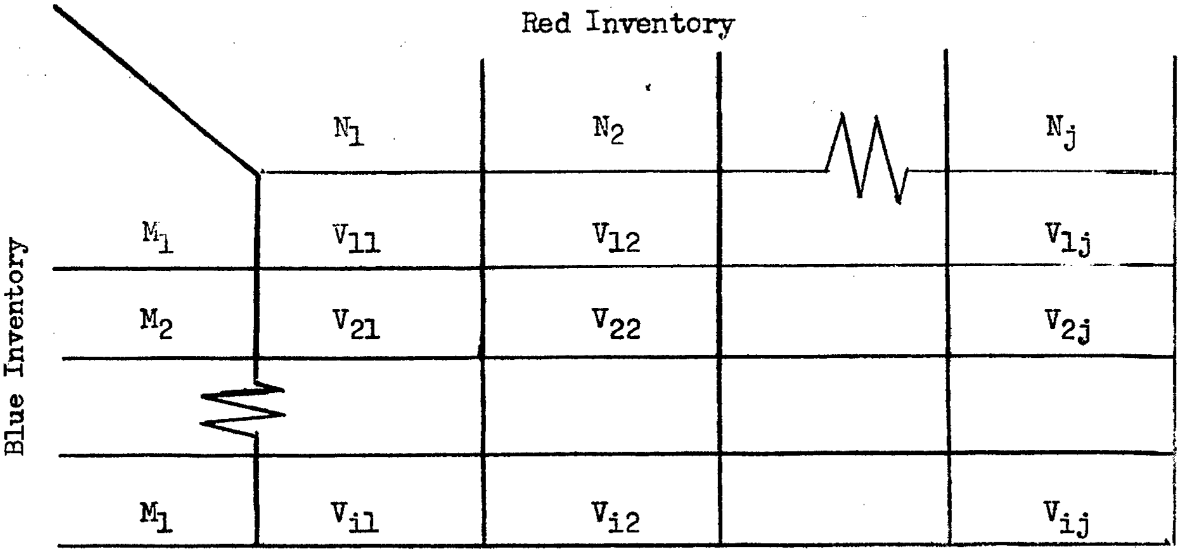
The hardware which must be the subject of research, development, and testing, must therefore be ordered in a priority fashion, in such a way that those items which more importantly affect the expectation of success, receive proportionately greater attention. Although marginal improvements to a weapons system, such as accuracy of aimed fire, are desirable, it is necessary to be assured that such marginal improvements are really worth the investment in time and resources. Each proposed item of research and development must therefore be tested in the crucible of potential value; and the cruicible is characterized by the notion of a calculated risk.

Some means must therefore be used to isolate "important" developments from the "less important." For purposes of this paper, it will be assumed that the choice must be made between inventories of equipment as the sole criterion, although this is clearly an approximation. Other variables bearing on the selection are the strategies and tactics of the two sides, the locale of the combat action, and the morale of the nations involved. "Importance" will be measured against the yardstick of the national objectives: in addition to winning a war, it includes the overtones of "deterring" the incidence of war, the reconstruction after the war, the cultural traditions of the combatants, etc.

It is necessary to observe, in this connection, that the generation of a scalar theory of value to serve as this yardstick is a matter of pressing need. One example of this kind of value calculus is that under development by N. M. Smith of the Operations Research Office.[1]

Assuming that an index of value can be employed to determine which of several choices of weapons inventories is best, there is an additional degree of freedom which is needed before the selection can be made. This degree of freedom is dictated because of the ability of the potential enemy to exercise a similar option. Thus, my selection of weapons mix M_1 may be best only if the opponent selects weapons mix N_1; it may fall far short of being adequate if the enemy selects N_2. The conventional means of illustrating this competition of strategies is through the use of a "strategy matrix". In the case at hand, the matrix is really one of competing weapons mixes, since the strategic values to be achieved are reflected in the evaluation yardstick, and the strategies and tactics of using the weapons mixes are assumed to be dictated by the selection of a given weapons mix.

Following the procedure of the mathematical theory of games, the weapons mix matrix can be constructed, as in Figure 1. The Blue team may elect to develop an inventory of weapons, M_1, M_2, etc.; the Red team may develop N_1, N_2, etc.



	N_1	N_2		N_j
M_1	V_{11}	V_{12}		V_{1j}
M_2	V_{21}	V_{22}		V_{2j}
M_1	V_{i1}	V_{i2}		V_{ij}

Figure 1 The Weapons Mix Matrix

Once this matrix has been completed, the mathematical means for the selection of an "optimum strategy" is available. The values, V_{11}, V_{12}, etc., have been proposed as scalar indices of expected success from the point of view of one of the sides. Lest we lose sight of the generation of the V's, recall that they are the values of the end products of (in this case) the potential war. The items for which the values have been summed are men, tanks, guns, economic capacity, etc.

It is at this point of the argument that the operational game becomes useful. The determination of the results of conflict between the various weapons mixes on each side depends upon an ability to calculate the effects of all the variables which significantly affect the course of the war. The proposition repeated here is that the war game is the most suitable vehicle for determining this "expected result" of a war opposing one Blue weapons system and one Red weapons system.[2] The war game, in the sense of its use in this paper, is the means for filling the boxes in the strategy matrix.

The war game parallels in intent several well-developed mathematical methods, but diverges significantly in method. Among the mathematical techniques which have been applied to the problem of calculating the expected result of an engagement, perhaps the Lanchester equations are the most widely celebrated. These equations generally follow the assumption that the number of kills against one side is proportional to the number of enemy and the enemy rate of kill:

$$\frac{dM}{dt} = k_N N$$

The equations have been generalized to include the effects of different kinds of weapons against different kinds of targets. The k's have been permitted to become variables, thus representing kill potential as a function of remaining enemy and friendly troops. Indeed, comceptually, it is possible to imagine a large number of simultaneous differential equations which define the problem. But the solution of such sets of equations has proved to be beyond the limits of present computing capacity. And for a complete set of equations, it becomes a nearly impossible task to reach agreement on "realistic" coefficients for the various terms of the equations.

A second approach has been widely celebrated in the literature of operations research, and this deals with what has been called "suboptimization." The intent of this approach is the isolation of those parts of a problem which are relatively unaffected or uniformly affected by the remainder of the problem, and reaching an "optimum" solution to each of the subproblems. The solution of the large problem is then reduced to the synthesis of the solutions to the various sub-problems. The difficulty with the suboptimization approach when applied to the selection of preferred weapons inventories is that it is by no means clear that the large problem can be dissected in the way necessary to make suboptimization valid. Try to divide the problem as one may, there are either very sensitive or complicated interactions between the various conceptual divisions. The quest for certainty is therefore thwarted.

What is needed is a method of analysis which examines combat as a whole. For want of a more inspired and elegant scientific approach, the war game may prove to be the heuristic vehicle for such an analysis.

A war game in a research context--an "operational game," as Ellis Johnson has named it, is essentially a very simple thing. It is a simulation of the various portions of combat, or of economics, or of business strategy, or of some other conflict situation. The components of a war game are, like a parlor game, three in number: the board, the pieces, and the rules.

For a war game, the board may be the traditional map. It may be a schematic diagram of the flow of operations in a system,[3] or it may be a conceptualized terrain model.[4] The pieces are those items which are moved about on the board. They may be models of individual tanks, or blocks representing armies, or counters representing supplies. The rules of the game are the conventions according to which the pieces may be moved from ohe place to another, casualties may be inflicted, information given to the opposing side, etc. The war game is, then, simply a scale model of combat.

It is open-ended, in that pieces may be added as desired, the board may be expanded, and the rules of the game changed as needed. It is this open-ended characteristic which permits the assertion that a war game may simulate the entire combat operation in all its important features. Because of the potential comprehensiveness of the game, and because of its nature as a model, it is possible to make the proposition that the play of a game can be associated with the actual history of a real combat situation.

The difficulties with a war game are immediately apparent from the above description. As an approximation to the real system, its results must be interpreted with due regard for the degree of approximation involved. In general, it would be expected that, as the rules of play are expressed in more and more realistic terms, the results of the play would more nearly represent the true expectation. Furthermore, the interactions between variables must be expressed directly. The advantage with the game approach is that it is generally easier to express the interactions in terms of the pieces of the game than in the terms required by the other methods of analysis. Finally, the development of a comprehensive game is still in the category of a fairly long or intermediate range project, and the capital investment in time is fairly significant.

The overriding advantage to the approach is that war gaming may be the methodological "breakthrough" required to facilitate the kind of analysis posed as the problem of the paper.

The remainder of the sketch of the path of analysis is straightforward. The game is played repeatedly, each time introducing a different set of initial conditions--different pieces, different board, different rules. In the comparison of weapons inventories, the game is simply repeated for each of the proposed weapons systems. The resultant values can then be compared on a relative basis--weapons mix M_1 is better than weapons mix M_2, etc.

The value of a given development in a weapon should then be capable of direct measurement. The game is played with and without the given improvement. The value of the weapon in the system is therefore indexed by the difference in the values of the plays.

The Design of Experiments. It has been observed that the degree of realism in a game may well determine its potential usefulness. In this regard, experience with war games has shown that a number of significant gaps wxist in our knowledge of the interaction effects between the weapons being simulated. To the end of improving the game structure, therefore, a recurring feedback from experiments is necessary. The date required to support a scientific gaming enterprise appears to be a natural and direct consequence of an appropriately designed field experiment. This proposal requires some reorientation in experimental design as customarily employed, although the change may well be small.

The reciprocal relation between gaming and the design of experiments is also important. It has been noted that the development of a sufficiently comprehensive game can be made by the annexation of further peices or rules. The end product of this annexation process is a game with a large number of variables. A complete solution to the game, therefore, requires an extremely large number of plays. This requirement is multiplied by a large factor when the play of chance is permitted to be a feature of the rules. It therefore appears that the development of the gaming technique has provided another fruitful field of application for the experimental design technique, in that the isolation of the variance due to the independent variables is of prime concern, and the repetition of the game (now the experiment) is restricted by reasons of the economics of time for game solutions.

Finally, the game can be used as a "test run" of a proposed field experiment. The effect of this application of war gaming should be to indicate the nature and frequency of the observations to be made in the actual experiment, the variables whose quantities are to be recorded, and some guidance as to the unnecessary portions of the experiment.

The conclusion is therefore inescapable. "War Gaming" and "Design of Experiments" form a sort of reciprocal system, each half enriching and giving guidance to the other.

References

1. N. M. Smith, Jr., *et al.*, "The Theory of Value and the Science of Decision --A Summary," *Journal of the Operations Research Society of America*, *1*, 103-113, 1953. "Comments," *Jour. Op. Res. Soc. America*, *2*, 181-187, 1954.

2. W. E. Cushen "War Games and Operations Research," *Philosophy of Science*, *22*, 309-320, 1955.

3. W. E. Cushen, "Operational Gaming in Industry," *Operations Research for Management*, 2, 358-375, 1956.

4. R. E. Zimmerman, "A Monte Carlo Model for Military Analysis," *Operations Research for Management*, *2*, 376-400, 1956.

Some Design Techniques Used for Increasing Cell Size With Special Emphasis in the Missile Field.

Paul C. Cox
White Sands Proving Ground

1.0 Introduction:

The designs which are to be discussed are some which have been considered in the guided missile or rocket fields. More particularly, they have been considered in the field of evaluating complete missiles.

There are three conditions which play a critical role when designing test plans for guided missiles and rockets: (1) Missiles are usually very expensive, and this means that it is economically impractical to have anything but a small sample; (2) it is usually necessary to study the effects of a number of environmental treatments and several levels for each treatment; (3) frequently the dependent variable to be analyzed is expressed as a variance or an attribute.

The three designs to be considered will illustrate the possibility of "stealing" data from certain cells in order that the remaining cells may be built up to include at least five or six items each. The purpose of this build up is to have enough items in each cell that a variance may be computed or the ratio of successes determined for every cell. Furthermore, an effort will be made to indicate the cost of such rearrangements and some of the necessary precautions to take when using these techniques.

2.0 The Replicated Latin Square.

The first plan to be discussed is the replicated or repeated Latin Square. It is the usual purpose of a Latin Square to test only one type of treatments, and then to remove the variation both horizontally and vertically. Our purpose is slightly different. We are trying to measure the effects of these types of treatments in a way that will conserve on the number of cells and at the same time permit the estimation of all possible treatment combinations. Those rows and columns, whose variation we only wanted to remove now represent treatment types which we now want to evaluate. This plan will take a Latin Square and repeat the same design N times in order that a variance or ratio of success may be computed from N observations for each cell.

Consider the example given by table one below. This design permits three types of treatments, three levels for each treatment, and 6 rounds assigned to each cell. For a conventional factorial design, this would require 3x3x3x6 = 162 rounds. It is hoped the replicated Latin Square will give the desired information with only 3x3x6 = 54 rounds.

	Slant Range (Launcher • Target)						Altitude Totals
Target Altitude		SR_1		SR_2		SR_3	
A_1	T_1	-2 yds.	T_2	13 yds.	T_3	-3 yds.	(11)
		3		0		-9	
		2		2		-6	
		-3		5		12	
		-4		-2		0	
		-5		-7		15	
A_2	T_3	15	T_1	-1	T_2	-12	(11)
		13		5		-1	
		-3		-3		-3	
		-9		-5		15	
		-5		-7		19	
		3		-1		-9	
A_3	T_2	0	T_3	-7	T_1	1	(-2)
		-5		-2		3	
		-3		12		-8	
		-2		3		0	
		-3		-3		8	
		8		1		-5	
Slant Range Totals:		0		3		17	20
Temp. Totals:	T_1= -22		T_2 = 15		T_3=27		

Table 1. p component of miss distance [1] for slant ranges (SR_1, SR_2, SR_3,) Propellent Temp. (T_1, T_2, T_3) and Target Altitudes (A_1, A_2, A_3)

[1] The p component is measured parallel to the trajectory.

2.1 Analysis of Variance of Variances:

Probably the information of primary concern when dealing with such data is the effect of the various levels of treatments upon missile dispersion: Therefore, an analysis of variance of the variances will be performed. Table 2 lists the natural logarithms of the variances which were obtained from the corresponding groups of six in table 1.

	SR_1	SR_2	SR_3	Totals
A_1	T_1 2.3096	T_2 3.8330	T_3 4.5675	10.7101
A_2	T_3 4.5757	T_1 2.8449	T_2 5.0845	12.5051
A_3	T_2 3.0634	T_3 3.7542	T_1 3.4843	10.3019
Totals	9.9487	10.4321	13.1363	33.5171
Totals for Temp. - T_1: 8.6388; T_2: 11.9809; T_3: 12.8974				

Table 2. Natural Log of the variances for the p component of miss distance: 3 levels of slant range, Pressure Altitude, and Propellent Temperature.

The analysis of variance of variances is given by table three.

Sources of Var.	d/f	S.S.	M.S.	F.
Between Alt.	2	.9168	.4579	
Between S. R.	2	1.9674	.9837	2.46
Between Temp.	2	3.3494	1.6747	4.18*
Th. Error	∞		.4	
Computed Error	2	.3373	.1687	
Total	8	6.5678		

Table 3. Analysis of Variance for data of Table 2. (* Indicates significance at the 5% level).

The theoretical error which is equal to $2/N - 1$ (N is the number of items in each cell) was used in this analysis, and this is recommended when the cell size is small.[1] However, if there is a significant difference between the computed and theoretical errors one should carefully reexamine the assumptions of the analysis of variance.

From the analysis in table 3 it appears that an increase in propellent temperature will cause an increase in missile despersion, and the effect is significant. An increase in slant range appears to have the same effect, and this is significant at the 10% but not at the 5% level.

One may wish to estimate the variance for some combination of treatments, and this can be done for the 108 treatment combinations for which no data was obtained as well as from the 54 treatment combinationa for which data was collected. Using the formula $y_{ijk} = m + a_i + b_j + c_k$, where $m = Y_{...}/N^2$; $a_i = Y_{i.}/N - m$; $b_j = Y_{.j}/N - m$; $c_k = \Sigma y_k/N - m$; and N is the no. of levels for each type of treatment, let us estimate the variance corresponding to (SR_2, A_3, and T_1). This is a point, incidentally, for which data was never collected. One obtains the value 3.72 - .24 - .29 - .84 = 2.35 for the estimate of the natural log of the variance. The estimated variance would then be 10.5 $(\text{yds.})^2$ for this combination of treatments.

2.2 Analysis of the Basic Data.

It would be possible to analyze the basic data of table one at this time, although it should be kept in mind that the analysis in table 3 indicates that one important assumption, namely homogeneity of variances, does not hold. If it is desired to proceed with the analysis, regardless, the method is demonstrated by table 4. It is clear from the analysis that nothing is significant.

Sources of Variation	d/f	S.S.	M.S.
Between S. R.	2	9.148	4.074
Between Altitude	2	6.259	3.129
Between Temp.	2	72.482	36.249
Between Blocks (2)	5	40.370	8.074
Error	42	2616.334	62.294
Totals.	53	2744.593	

Table 4. Analysis of the data of Table one.

[1] Many prefer to use the computed error at all times, but it is clear that there are not sufficient degrees of freedom for the computed error to be feasible when the Latin Square is less than 4 x 4.

[2] It will be assumed the first items in each cell were fired first in a random order, followed by all second items, third items, etc. These groups will be considered the six blocks.

2.3 Analysis of Variance of Attributes.

The Replicated Latin Square might prove very useful for an Analysis of Variance when the only value measured is simply success or failure. This will be demonstrated by table 5 in which a sample of 108 electronic assemblies were given certain combinations of Vibration (V_k), Shock (S_i) and Temperature Conditioning (T_i). Three levels of each environment were considered, making 27 possible combinations. Nine combinations were used, and 12 assemblies were assigned to each combination. The values in table 5 are the number of successes out of the 12 tested in each cell. The quantity in paretheses is the arc sine of the square root of the ratio of successes to 12. It is this arc sine which is analyzed by methods which are very nearly conventional.

	Temperature			Total Shock
Shock	T_1	T_2	T_3	
S_1	V_3 6 (45.000)	V_2 9 (60.000)	V_1 11 (73.333)	26
S_2	V_1 10 (65.917)	V_3 4 (35.250)	V_2 7 (49.835)	21
S_3	V_2 5 (40.233)	V_1 9 (60.000)	V_3 2 (24.150)	16
Total Temp.	21	22	20	63
Total Vibration	$V_1 = 30$	$V_2 = 21$	$V_3 = 12$	

Table 5. The Number of Successes from 12 Units in Each Cell. The Value in Parentheses is the Arc Sine of the square root of the Ratio of the Number of Successes to 12.

The analysis of the data of table 5 is given by table 6. From this analysis it appears that the extent of vibration has a highly significant effect upon the unit, the effect of shock is significant, but the effect of temperature is clearly not significant. From this data it would be quite simple to go ahead and make estimates of the ratio of successes expected for all 27 possible combinations of (T, S, and V). It is unfortunate that there is so much difference between the computed and theoretical Error. This may lead one to question whether the basic assumptions are actually valid.

Sources of Variation	d/f	S.S.	M.S.	F.†
Between Temp.	2	10.50	5.25	
Between Shock	2	485.13	242.57	3.55*
Between Vibration	2	1500.11	750.06	10.97*
Theoretical Error	∞		68.4 = 821/N	
Computed Error	2	6.72	3.36	
Total	8	2002.46		

Table 6. Analysis of the Data of Table 5.
(* Denotes Significance at the 5% Level)

2.4 Discussion.

The Replicated Latin Square appears to do the desired job with a considerable saving of test items. It should be made perfectly clear, however, that this saving in test items is brought about at a definite cost, for example: (1) It must be assumed there is no interaction for the analysis to be valid. There not only is no way to test for interaction; it definitely must not exist. (2) To estimate the effects of 2/3 of the treatments extrapolation (under the assumption no interactions exist) is necessary, and extrapolation is always dangerous; (3) There is one school of thought which believes that it is not legitimate to test the effects of rows and columns in a Latin Square because complete randomization can be accomplished for only one of the three types of treatments.[1] While this comment is not to be taken lightly, it does suggest a luxury which the guided missile field cannot afford.

Finally, it frequently occurs that we must resort to the Latin Square for economy when attributes or variances are not a matter of concern. For example, suppose it is necessary to test 3 types of treatments each at 4 levels, and 16 test units are all that could possibly be procured. The Latin Square might well be the appropriate solution to the problem.

[1] See Ostle, Bernard, Statistics in Research, P 322. 1954, Iowa State College Press, Ames, Iowa.

3.0 The Cross or Butterfly Design.

The second design, the cross design, is valuable for the same reason as the replicated latin square, i.e. to increase cell size and at the same time keep the same number of comparisons without increasing the overall sample size. One reason why this particular design is valuable is because rounds are sometimes fired according to this pattern to obtain data for constructing firing tables, and it is desirable to take all data available, regardless of why it was obtained, and secure as much engineering information as possible. It is therefore valuable to know how to analyze data when in this form. Actually, it is not difficult to analyze any 2 way design(if the assumption is made that no interactions exist). The techniques is demonstrated on Pages 79 - 87 of Kempthorne [1] and the cross design is but one example covered by this general class.

3.1 The Design and Technique of Analysis.

The design is given by table 7, the formulas for estimating the parameters and the formulas needed for the Analysis of Variance follow, and the Analysis of Variance is given by table 8.

				$x_{1,n}$				$X_{1 \cdot\cdot}$
				$\vdots$				
				$x_{m-1,n}$				$X_{m-1 \cdot\cdot}$
$x_{m,1}$	$x_{m,2}$	$\cdots$	$x_{m,n-1}$	$x_{m,n}$	$x_{m,n+1}$	$\cdots$	$x_{m,s}$	$X_{m \cdot\cdot}$
				$x_{m+1,n}$				$X_{m+1 \cdot\cdot}$
				$\vdots$				
				$x_{r,n}$				$X_{r \cdot\cdot}$
$X_{\cdot 1 \cdot}$	$X_{\cdot 2 \cdot}$		$X_{\cdot n-1 \cdot}$	$X_{\cdot n \cdot}$	$X_{\cdot n\pm 1 \cdot}$		$X_{\cdot s \cdot}$	$X_{\cdot\cdot\cdot}$

Table 7. The Cross Design

[1] See Kempthorne, Oscar, Design and Analysis of Experiments, 1952, John Wiley & Sons, New York.

3.11 Estimates of the Parameters.

Using the model: $x_{ij} = u + a_i + b_j + e_{ij} \begin{cases} i = 1, 2, \ldots r \\ j = 1, 2, \ldots s \\ \underline{k} \text{ replications} \end{cases}$

the estimates are:

$$\hat{u} = \frac{1}{k} \left(X_{\cdot\cdot\cdot} - \frac{s-1}{s} \cdot X_{m\cdot\cdot} - \frac{r-1}{r} \cdot X_{\cdot n \cdot} \right)$$

$$\hat{a}_1 = \frac{1}{k} \left(X_{1\cdot\cdot} - \frac{1}{r} X_{\cdot n \cdot} \right)$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\hat{a}_{m-1} = \frac{1}{k} \left(X_{(m-1)\cdot\cdot} - \frac{1}{n} X_{\cdot n \cdot} \right)$$

$$\hat{a}_m = \frac{1}{k} \left(X_{m\cdot\cdot} + \frac{r-1}{r} X_{\cdot n \cdot} - X_{\cdot\cdot\cdot} \right)$$

$$\hat{a}_m + 1 = \frac{1}{k} \left(X_{(m+1)\cdot\cdot} - \frac{1}{r} X_{\cdot n \cdot} \right)$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\hat{a}_r = \frac{1}{k} \left(X_{r\cdot\cdot} - \frac{1}{r} X_{\cdot n \cdot} \right)$$

$$\hat{b}_1 = \frac{1}{k} \left(X_{\cdot 1 \cdot} - \frac{1}{s} X_{m\cdot\cdot} \right)$$

$$\hat{b}_{n-1} = \frac{1}{k} \left(X_{\cdot (n-1) \cdot} - \frac{1}{s} X_{m\cdot\cdot} \right)$$

$$\hat{b}_n = \frac{1}{k} \left(X_{\cdot n \cdot} + \frac{s-1}{s} X_{m\cdot\cdot} - X_{\cdot\cdot\cdot} \right)$$

$$\hat{b}_{n+1} = \frac{1}{k} \left(X_{\cdot (n+1) \cdot} - \frac{1}{s} X_{m\cdot\cdot} \right)$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\hat{b}_s = \frac{1}{k} \left(X_{\cdot s \cdot} - \frac{1}{s} X_{m\cdot\cdot} \right)$$

3.12 Formulas for the analysis of Variance.

$$T = \sum\sum\sum X^{2}_{ijk}$$

$$R_o = \hat{u}\, X \ldots + \hat{a}_1 X_{1..} + \ldots + \hat{a}_r X_{r..} + \hat{b}_1 X_{.1.} + \ldots + \hat{b}_s X_{.s.}$$

$$= \frac{1}{k}\Big\{\sum X_{i..}^{2} + \sum X_{.j.}^{2} - 2(X_{.n.})(X_{...}) - 2(X_{m..})(X_{...})$$

$$+ X_{...}^{2} + 2(X_{m..})(X_{.n.})\Big\}$$

$$R_b = \frac{1}{k}\Big\{\sum X_{i..}^{2} - \frac{s-1}{s}\; X_{m..}^{2}\Big\}$$

$$R_a = \frac{1}{k}\Big\{\sum X_{.j.}^{2} - \frac{r-1}{r}\; X_{.n.}^{2}\Big\}$$

Where R_o is the reduction in sum of squares due to fitting (u,a, and b); R_b is the reduction due to fitting (u and a); R_a is the reduction due to fitting (u and b).

Sources of Var.	d/f	S.S.
Between Rows	r-1	R_o-R_a
Between Columns	s-1	R_o-R_b
Error	(k-1) (n+s-1)	$T-R_o$

Table 8. Analysis of Variance for the Cross Design

3.2 Example:

In table 9 we have an example in which the cross design is used. There are three A treatments (Rows), 4 B treatments (Columns), 5 replications, and the intersection occurs at $(x_{2,3})$. That is to say: (r=3, m=2, k=5, s=4, and n=3).

	B_1	B_2	B_3	B_4	Totals
A_1			9 10 12 13 11		$X_{1..}= 55$
A_2	7 7 8 8 9	9 13 11 11 12	11 12 14 13 16	13 19 15 17 16	$X_{2..}= 241$
A_3			13 14 15 16 19		$X_{3..}= 77$
	$X_{.1.} =39$	$X_{.2.} =56$	$X_{.3.}=198$	$X_{.4.} =80$	$X_{...}=373$

Table 9. An Example Using the Cross Design.

3.21 Substituting in the formulas of sections 3.11 will give the following estimates.

$\hat{u} = 12.05$	$\hat{b}_1 = - 4.25$
$\hat{a}_1 = - 2.20$	$\hat{b}_2 = - .85$
$\hat{a}_2 = 0$	$\hat{b}_3 = 1.15$
$\hat{a}_3 = + 2.20$	$\hat{b}_4 = 3.95$

Using these estimates and the formula $y_{ij} = u + a_i + b_j$ and assuming no interactions exist it is possible to estimate an expected value for any of the 12 possible combinations of A_i and B_j which are given in table 9 whether or not the combination has data assigned to it. For example the estimate for (A_1,B_4) is 12.05 - 2.20 + 3.95 = 13.80.

3.22 The data for the analysis of variance is given below and the results of the analysis are given by table 10. ($T = 4951$; $R_o = 4873$; $R_a = 4825$; $R_b = 4695$).

Sources of Var.	d/f	S.S.	M.S.	F.
Between A's	2	51	25.5	7.8*
Between B's	3	178	59.3	18.2*
Error	24	78	3.25	

Table 10. Analysis of Variance of the data from table 9.
(*indicates significance)

3.23 The Analysis of Variance of Variance.

The natural log of the variances, computed from the cells in table 9 is given by table 11.

	B_1	B_2	B_3	B_4	
A_1			.9163		$X_{1..} = .9163$
A_2	-.3567	.7885	1.3083	1.6094	$X_{2..} = 3.3495$
A_3			1.6677		$X_{3..} = 1.6677$
	$X_{.1.} = -.3567$	.7885	3.8923	1.6094	$X_{...} = 5.9335$

Table 11. Natural log of the variance for data in table 9.

Estimates of the Parameters.

$\hat{u} = .8265$

$\hat{a}_1 = -.3811$

$\hat{a}_2 = .0108$

$\hat{a}_3 = .3703$

$\hat{b}_1 = -1.1941$

$\hat{b}_2 = -.0489$

$\hat{b}_3 = .4710$

$\hat{b}_4 = .7720$

Data for the analysis of Variance.

$T = 8.6716$

$R_o = 8.6715$

$R_a = 8.3891$

$R_b = 6.4257$

The analysis of Variance is given by table 12.

Sources of Var.	d/f	S. S.	M. S.
Between A's	2	.2825	.1412
Between B's	3	2.2458	.748
Error	∞		.50 = 2/(5-1)

Table 12. Analysis of Variance of the data in table 11. (no significant comparisons).

3.24 An analysis of Variance of Attributes could be conducted, using this design, without any difficulty.

3.3 Discussion.

The cross design, like the Latin Square is useful only if there is good evidence that no interaction effects exist. It is much more versatile than the Latin Square because the point of intersection can be chosen anywhere and the design can be extended to any number of dimensions. Intuitively, the Latin Square appears to make a more equitable distribution of rounds to the various treatment combinations than the cross design does.

Some one might ask, why use the cross design given by table 9 when a simple 2 x3 factorial will require no more cells. The answer is that the 2 x 3 factorial will usually be preferred, but the 3 x 4 cross may be more desirable when there exists satisfactory evidence that there are no interactions and when it is really necessary to test 4 levels of one type of treatment and 3 levels of another type.

Finally, it is interesting that a 3 x 4 cross with 2 replications requires the same number of rounds as a 3 x 4 factorial with one replication. To analyze the data from either design it is necessary to assume no interaction, and in some instances it might be preferable to use the cross design in order that there will be two replications. Incidentally, the cross design cannot be used with only one replication because such a plan would allow for no error term.

4.0 The X Design.

The final design to be considered also omits a few cells in order that the number of elements in the remaining cells may be imcreased. Table 13 presents the X design in its simplest form and gives an example with five replications. Demand for this or some similar design has arisen because engineers are frequently interested in knowing how well a test item proforms at the extreme limits of the design specifications and also whether performance in the center is consistent with performance at the extremes.

	B_1	B_2	B_3	Sum
A_1	3 (X_{11}) 8 4 5 7		6 8 (X_{13}) 13 9 12	$X_{1\cdot\cdot} = 75$
A_2		5 11 (X_{22}) 7 6 8		$X_{2\cdot\cdot} = 37$
A_3	4 7 10 (X_{31}) 8 5		8 9 13 (X_{33}) 16 17	$X_{3\cdot\cdot} = 97$
	$X_{\cdot 1 \cdot} = 61$	$X_{\cdot 2 \cdot} = 37$	$X_{\cdot 3 \cdot} = 111$	$X_{\cdot\cdot\cdot} = 209$

Table 13. The X design in its simplest form and an example with five replications.

4.1 Solution of the Normal Equations.

To solve the normal equations it is necessary in this case to place three restrictions instead of the usual two. Assuming the model $X_{ij} = u + a_i + b_j + e_{ij}$ and placing the restrictions that $2a_1 + a_2 + 2a_3 = 2b_1 + b_2 + 2b_3 = 0$ and $a_2 = b_2$ the following solutions are obtained.

$$\hat{u} = \frac{1}{5\,K}\,(X...) = 8.36$$

$$\hat{a}_2 = \hat{b}_2 = \frac{1}{10K}\left\{5X_2.. - X...\right\} = -.28$$

$$\hat{a}_1 = \frac{1}{40K}\left\{20X_1.. = 9X... + 5X_2..\right\} = -.93$$

$$\hat{a}_3 = \frac{1}{40K}\left\{20X_3.. - 9X... + 5X_2..\right\} = +1.27$$

$$\hat{b}_1 = \frac{1}{40K}\left\{20X._1. - 9X... + 5X_2..\right\} = -2.33$$

$$\hat{b}_3 = \frac{1}{40K}\left\{20X._3. - 9X... + 5X_2..\right\} = +2.67$$

From these solutions it is possible to obtain an expected value for any of the nine combinations of treatments for table 13. For example, if one should extrapolate for the point (A_1, B_2) the expected value would be $X_{12} = 8.36 - .93 - .28 = 7.14$.

4.2 Analysis Omitting the X22 Term.

The analysis presents certain problems. If the mormal equations are used along with the technique described on pages 77-81 of Kempthorne[1] it will become evident immediately that the sum of squares for the main effects is exactly the same as if the X_{22} terms had never been used. Consequently, the analysis will begin by omitting the X_{22} cell and analyzing the data as if the design were a simple 2 x 2 factorial. This analysis is given by table 14.

Sources of Var.	d/f	S.S.	M.S.	F
Between A's	1	24.2	24.2	2.8
Between B's	1	125	125	14.5*
Interaction	1	3.2	3.2	
Error	16	138.4	8.65	
Total	19	290.8		

Table 14. An Analysis of Variance of the Data of Table 13. (Omitting items in cell x_{22})

[1] Kempthorne, Oscar, Design and Analysis of Experiments, 1952, John Wiley & Sons, New York.

4.3 Analysis Including the X_{22} Term.

The next step in the analysis is to determine the influence of the X_{22} Term. The first thing will be to find $\{\Sigma X^2 - (\Sigma X)^2/N\}$ for the X_{22} term. This is found to be 21.2 with 4 degrees of freedom. This can be added directly to the sum of squares for error in table 14 giving a new error term with 20 degrees of freedom and 159.6 for the sum of squares.

It is now necessary to set up some hypothesis to test the effect of X_{22} in terms of the other cells. It would be reasonable to test the hypothesis that $4\ X_{22} = X_{13} + X_{31} + X_{11} + X_{33}$.[1] The sum of squares for such a comparison could be obtained from the formula: $\frac{1}{20K}\{\Sigma X_{11} + \Sigma X_{33} + \Sigma X_{13} + \Sigma X_{31} = 4\Sigma X_{22}\}^2 = 5.76$ with one degree of freedom. The final analysis of variance is given by table 15.

Sources of Var.	d/f	S.S.	M.S.	F.
Between A's	1	24.2	24.2	3.0
Between B's	1	125.0	125.0	15.4*
Interaction	1	3.2	3.2	
$[\Sigma X_{11}+\Sigma X_{13}+\Sigma X_{31}+\Sigma X_{33}$ vs. $4\Sigma X_{22}]$	1	5.8	5.8	
Error	19	153.8	8.1	

Table 15. Complete Analysis of the data of table 13.
(* Indicates significance at the 5% level)

The analysis of variances of variances or an analysis of variance of attributes could follow a similar pattern and would present no difficulty. A Theoretical error would have to be used because no degrees of freedom would remain for a computed error unless the Mean Square for interaction were to be used for the error term.

[1] An alternative set of hypothesis to test would be that the effect along the principal and minor diagonals are both linear.

5.0 Conclusions:

Three designs have been discussed which have the purpose of eliminating certain cells to increase the size of the remaining cells. It is recognized that there is no substitute for an adequate sample size, but it is also recognized that frequently a job must be done and it is either impossible or at least economially impractical to obtain as large a sample as desired. It is then necessary to modify the analysis in such a way that as much as possible of the desired information can be obtained, although it is admitted that a design requiring a smaller sample size will either sacrifice some information or will require additional assumptions.

The first design is a well known design which has been replicated. This idea could easily be extended to other well known designs. For example, if an analysis of Variance of Variances is desired it might be reasonable to use a Graeco - Latin Square and repeat it N times. Also, it might be desirable to use a fractional replication design, repeat it several times and then run an analysis of Variance on the fractional replication of variances.

The second design is simply a design with two types of treatments and N replications. This is not a well known rectangular design, but nevertheless, the analysis is not difficult nor would it be difficult for most other odd shaped designs which have two types of treatments.

The third design required a special step to compare the effect of the X_{22} term, but this was not difficult to determine as would be true of most designs which are some variation of the X design.

The speaker is very appreciative of the many constructive comments made at the close of the presentation. In particular he would like to acknowledge the comments by Dr. John Tukey who, among other things, gave a very forceful argument in favor of using the computed error rather than the theoretical error. Dr. Tukey also made suggestions which led to a satisfactory solution of the "X" design. The speaker would also like to acknowledge the comments by Dr. Boyd Harshbarger who made many useful suggestions both before and after the presentation.

THREE APPLICATIONS OF STATISTICS TO ELECTRONICS

L. M. Court
Diamond Ordnance Fuze Laboratories

It is approximately one hundred and fifty years since Gauss availed himself of his newly invented method of least squares to compute the orbit of Ceres from what would have hitherto been considered an inadequate number of observations, thus rediscovering the new asteroid that had been followed in the skies for a few weeks and then lost sight of. That may have been the first application of what now would be termed mathematical statistics. Since then physicists and astronomers in large numbers have employed other ideas of Gauss in the realm of probability - - to wit, the normal law, on which the method of least squares itself rests - - to calculate means of sets of observation and attach probable errors to them.

Roughly a century after Gauss, Karl Pearson and R. A. Fisher, turning away from the physical sciences to problems in genetics and agriculture,, were busy expanding the methods armory of what now could justly be called mathematical statistics. Fisher particularly devoted himself to the theory of small samples, a powerful tool that we who work in industrial statistics frequently resort too. But it was left to Shewhart to conceive the idea of applying statistics to an industrial process on a wholesale basis, not just computing a few means and probable errors, and to get a continuous authentic picture of what was taking place. Shewhart's approach, concentrating originally on product quality, if only because of the completeness of the statistical record it gives rise to, endows the engineer with an enormous control over the industrial process he is supervising. Because it pinpoints the trouble almost as fast as it arises, the causes are often located in a matter of hours where previously they might have taken days or weeks to uncover.

The statistics predominating in industrial applications is not just the straight forward kind introduced by Gauss, although that too is used, but the more sophisticated sort that has been elaborated since, particularly the theory of small samples. With respect to the species of statistics it relies on, I think that a technological laboratory, even when engaged in research, resembles an industrial plant more than a pure physics research laboratory. I think this applies with particular emphasis to the organizations represented at this Conference. For example, I shall take up this morning three applications of statistics to problems that arose in the Diamond Ordnance Fuze Laboratories, and in none of these was the pure, unvarnished normal law used. As a matter of fact, we fell back on distribution-free methods in the second application and such things as the χ^2 and t distributions in the third; in the first, the normal law is used, not so much in its own right, but to show that another distribution approximates closely to it.

All three of our applications have one thing in common -- they deal with the quality of electrical items, either hand made in the Laboratory or produced by mass methods in a plant. It is true they do not deal with it in the usual quality control sense. Now a production process is characterized by a great many variables besides quality, viz., time, cost, profit yielded by the process, etc., and a change in the quality variable is almost always bound to affect the values of one or more of these other variables. Let us consider a situation of this sort. Suppose that one or more of the components that go to make up a product are meeting finer tolerances than are needed to maintain the product's over-all quality or reliability as specified by the designing engineer. There is then immediately a "waste" of quality. Ultimately this can be translated into a waste of money or cost. Generally there are many ways to eliminate this waste, the best one depending on the nature of the manufacturing process. Suppose, for simplicity, that there is only one component of this kind, and that it is available in the market in a number of grades, several of which are inferior to the grade being used. If one of these inferior grades will still keep the product's over-all quality within the limits specified by the design engineer and can be bought at a reduced price, the solution to our waste problem is to switch to this inferior grade; actually, to switch to the cheapest inferior grade for which this is true.

Suppose, on the other hand, that only one grade of this component is to be had in the market; and that its tolerances are below the one's used by the manufacturer. In all probability, then, he has been upgrading the quality of the purchased component by testing and selection. Since we have assumed that the tolerances of the component as it goes into the product are higher than they need be, the component is being upgraded too much. In the extreme case, no upgrading at all is required to maintain the product's over-all quality. In this case, the waste is eliminated by abolishing the upgrading. The reduction in the component's quality is then immediately translated into a time saving. A manufacturing process can differ from either of the two situations described so that a relaxation of tolerances is converted immediately into a change in the value of some production variable other than cost or time.

In both of the situations described the proper course of action is transparently clear. Often things are not so simple - - it is not self-evident that the tolerances used are too fine. Then some analysis is required, usually of a mathematical nature, to establish this fact. The problem taken up in our first application is of this sort.

I. Altering an Amplifier's Assembly Procedure. In many of our fuzes an amplifier is present as part of the circuit. The one that was the subject of our present experiment multiplies the incoming signal by a factor of 141,000, more or less. I say "more or less" because obviously electrical components vary in their performance, and so long as the final amplification is within three decibels of this figure, the amplifier will serve the purpose for which it was designed. The task of amplification is divided unevenly between the amplifier's transformer and its tubes: the three tubes, this being the number in the amplifier circuit, provide a gain of 10,850, the transformers a mere 13. Since the three tubes are nominally alike, the gain attributable to each is the cube root of 10,850 or 22.1.

Prior to this analysis the practice was to test each tube separately and thus make sure that its gain was very close to 22.1. Afterwards it was possible to take them at random from a stockpile and insert them without additional examination into the amplifier's circuit. Thus three steps in the assembly procedure are eliminated, i.e., there is a considerable time saving. The purport of our analysis is that the upgrading of the tubes by Laboratory personnel was unnecessary, and that, as supplied by the manufacturer, their tolerances were already sufficiently good to build a serviceable amplifier.

Let us see how the analysis proceeded. These tubes were pentodes so that the gain depended on their transconductance, G_m. The manufacturer assured us that the G_m's of his pentodes were normally distributed with a mean μG of 5,000 mhos and a standard deviation σ_G of 300 mhos.

The other factor on which the gain of a tube depends is the load R_L. In our case the load was always constant -- 4,420 ohms, i.e., it could be regarded either as a non-random quantity or a variate with a one-valued distribution. The gain y is given by $y = R_L G_M = 4420\ G_m$, with the result that y is a normally distributed variate with a mean $\mu_y = 22.1$ (the figure mentioned earlier) and a standard deviation $\sigma_y = 1.33$, both calculable from the distribution for G_m.

An elementary theorem on probability states that the sum of any finite number of normally distributed random variates is itself normally distributed. The gain due to the three tubes, however, is the product of the individual gains, thus precluding us from applying this theorem. If we insist on using this theorem to deduce that the total gain is essentially normally distributed, we must somehow convert this product into a sum. The obvious way out is to take the logarithms of the gains, i.e., to measure gain in decibels instead of natural numbers, as the engineers do. If z is the gain of a tube in decibels, then by definition $z = 20 \log_{10} y$.

If the total gain in decibels is to be normally distributed, the individual gains in decibels must be too, and the trouble is that we are ignorant of the form of z's distribution. The fact is that z can*not* be normally distributed since y *is*, and the relationship between the two variates is a logarithmic one. I.e., in the strict theoretical sense. But in the present instance, z's departure from normality is slight enough to be neglected, as a little calculation will show.

Let us develop $z = 20 \log_{10} y$ as a Taylor series, taking as the value y_o about which the development is centered the mean $\mu y = 22.1$ of the independent argument variate. This is a natural choice since in the case of most reasonably well-behaved distributions the arithmetic mean is the core or central value, this being even truer of a normal distribution. Thru the second degree term in $y - \mu_y$ the expansion is given by:

$$Z = 20 \log 22.1 + \frac{20}{22.1} (\log e) (y-22.1) - \frac{20 (\log e)}{2 (22.1)^2} (y-22.1)^2 + ..$$

*

$$Z = 26.88 + 0.393 (y-22.1) - 0.009 (y-22.1)^2 + ..$$

Since y is normally distributed, an interval of $2\sigma_y = 2.66$ about μ_y on both sides, will include 95 percent of all y values. For this overwhelming portion of the y - population, the error introduced is small if the quadratic and higher degree terms in the expansion for z are dropped. In fact, the ratio of the quadratic to the linear term in *) is $\frac{9}{393}$ (y - 22.1). It is a maximum numerically for the interval in question when y is taken at either extreme end of the interval, i.e. when y - 22.1 = 2.66. Its value then is only .061. The relative error in truncating the expansion at the linear term is even less since the series in question is an alternating one.

To within the indicated degree of approximation, z is a linear function of y. To within this degree, its distribution is normal with $\mu_z = 26.85$ and $\sigma_z = 0.523$. Since it is σ_z, or rather a multiple of it, that will finally convince us that the distribution of the total gain is not so widely dispersed as to produce an unreliable amplifier if the tubes are taken at random, we can make things safer for ourselves by overestimating this quantity. This can be done by replacing the 22.1 in the denominator (there only) of equation (*) by the lower limit of the $2\sigma_y$ interval, or better yet the $3\sigma_y$ interval, about μ_y. We then know that $\sigma_z \leqq .638$ with a very high degree of probability.

The total gain in decibels due to tubes, T, is equal to $z_1 + z_2 + z_3$, where each z has the approximately normal distribution just developed. T's distribution is therefore approximately normal with $\mu_T = 80.64$ and $\sigma_T \leqq 1.11$. The amplifier's total gain g is given by g = T + t where t is the transformer gain. t, we mentioned, was constant; in decibels $t = 20 \log_{10} 13 = 22.28$. g's distribution is therefore approximately normal with $\mu_g = 102.92$ decibels and $\sigma_g \leqq 1.11$ decibels.

Like a living organism, every instrument has a margin of error within which its function must be considered normal or satisfactory. The amplifier in question easily meets the requirements of the fuze circuit, of which it is part, even if its amplification strays from the nominal value of 103 decibels by as much as 3 decibels. Since in a normal distribution an interval of 2-1/2 times the standard deviation contains 99 percent of the population, and the analysis has shown that 2-1/2 $\sigma_g \leqq 2.75$ decibels, where the g - distribution was derived on the assumption that the tubes are taken at random, it is plain that picking them in this fashion will hardly affect the amplifier's performance.

Credit should be given to Mrs. M. Hamill, formerly of these Laboratories, for this practical piece of analysis. It was the speaker who observed that this was an example of a situation in which tolerances imposed on a component were too fine, and that they could easily be relaxed with a concomitant time saving.

II. A Power Supply Development Program. We deal here with a program that ran for five years, whose objective was to develop a packaged power supply unit that would function under the most diverse weather conditions, from the arctic to the equator. To place the program in the right perspective, it should be mentioned that the principal business of our Laboratory is the design of fuzes for missiles, the majority of these devices being electronic. An electronic fuze must have a power supply, and customarily it is fed from a source in the missile. There is an advantage in giving the fuze a power supply of its own, making it independent of the missile. One of the contrivances recently proposed for this role is a battery. Since the battery must always be on tap, ready to power the fuze at a moment's notice, it is often referred to as a reserve power supply.

An ordinary battery, such as an automobile's, is totally unsuited for this purpose. For one thing, it is too bulky. For another, it cannot withstand indefinite storage, the battery drawing a minimal current even when not in active operation. Besides these objectives, there is another, even more important, which is the main subject of our present discussion. For technical reasons that have no bearing on the statistical aspects, our reserve battery must have a minimum life of 300 seconds once its terminals are connected.

The program was initiated in 1951, gaining momentum in 1953. A private firm, that had had considerable experience with electrochemistry and batteries, was made responsible for the experimentation, the Laboratory exercising supervision thru certain of its personnel. Enough progress had been made by early 1954 to warrant setting up a serious testing program that would decide whether the objectives of the development program were being attained. This was done largely on the initiative of the Laboratory's supervisory personnel.

Repeated samples were taken from the populations of battery lives throughout 1954 and the early months of 1955. (Batteries of several different voltages were being developed, and there were as many basic populations as voltage types.) The μ's or means were fairly stable, for the most part above the minimum allowable life of 300 seconds. The standard deviations (σ's), however, were large and, what is worse, quite variable, altho they did tend to diminish somewhat as the program continued. I was never provided with the actual figures, but it was clear from the various accounts that the data was statistically unhomogeneous. In the language of quality control, the development program had failed to attain a state of statistical control.

Because the developing firm felt that the obstacles responsible for these inconclusive results were gradually being ironed out, a decision to gather fresh data and reassess the program's progress was made in the second quarter of 1955. This was the data which the speaker analyzed. Since the earlier material was statistically unhomogeneous, it was felt that any assumption concerning the forms of the populations, viz. that they were normal or had some other distribution, was unwarranted. I decided to fall back on distribution-free methods that avoid any reference to this form in testing whether the program was nearing its objective. It was necessary, of course, to state the objective in a precise form before subjecting it to a test. As it finally emerged in a conference with the supervisory Laboratory personnel, it was that with a high confidence coefficient (95 percent), all but a minor fraction (1 percent) of any particular battery population had a life of 300 seconds or over.

Table I (below) gives the figures by classes for an illustrative sample from one of the battery populations (Population III). The raw material on which the table is based was unavailable to the speaker, the data reaching him only in this highly processed form. Fortunately the distribution-free technique adopted did not require him to refer back to the raw data.

TABLE I

ILLUSTRATIVE SAMPLE FROM BATTERY POPULATION III

LIFE-SECONDS	NO. IN CLASS
0 - 100	0
101 - 200	0
201 - 300	2
301 - 400	3
401 - 500	14
501 - 600	15
601 - 700	13
701 - 800	4
801 - 900	0
TOTAL	51 = n

The distribution-free methods we are about to employ rest on a binomial distribution that for its specification does not require an actual knowledge of the underlying distribution but only of the particular percentile point in this last distribution which is being tested. It is for this reason that assumptions about the form of the underlying distribution can be dispensed with.

A study of Table I reveals that, although a good deal of the information contained in the original sample has been lost because of the processing into classes, the earliest item to have a life of 300 seconds or more is the third. (The item count is from the bottom, beginning with the item that had the least life.) If any percentile point of the original underlying population falls above this third item, we can be sure that the corresponding fraction (1 minus the percentile point in question) of this population has a life of over 300 seconds. The probability that this will happen does not require a knowledge of this actual population for its computation but can be deduced from the binomial distribution previously referred to. The thing to do, therefore, is to calculate the probability that the 1 percentile point exceeds the third item, using the binomial distribution, and if this turns out to be 0.95 or more, we can be sure that the development program is fulfilling its objective. More generally, we can carry out this kind of computation for the 3rd, 5th, etc. percentile points. In statistical language, we are computing the confidence coefficients to be attached to confidence intervals for the p-th percentile point, p = 1, 3, 5, etc., the left endpoint of these intervals being the third observed item and the right end point + ∞. Table II is the result of such computations.

TABLE II

CALCULATION OF CONFIDENCE COEFFICIENTS FOR POPULATION III FOR VARIOUS PERCENTILE POINTS

Notation: x_r = r-th item, counting from the least in the sample

λ_p = p-th percentile point

P( ) = Probability (confidence coefficient) of enclosed statement

Basic Formula: $P(x_r \leq \lambda_p) = 1 - \sum_{k=0}^{r-1} C_k^n \, (p)^k \, (1-p)^{n-k}$

$$P(x_3 \leq \lambda_{.01}) = 1 - \sum_{k=0}^{2} C_k^{51} \, (.01)^k (.99)^{51-k} = .014$$

$P(x_3 < \lambda_{.03})$ . . . . = .197

$P(x_3 < \lambda_{.05})$ . . . . = .472

$P(x_3 < \lambda_{.07})$ . . . . = .701

$P(x_3 \leq \lambda_{.10})$ . . . . = .896

$P(x_3 \leq \lambda_{.12})$ . . . . = .953

Basically Table II is a list of statements, relating various percentile points to the third sample item, together with their probabilities of occurrence. Glancing through it, we see that the earliest statement to which a confidence coefficient of 95 percent can be attached is the one referring to the 12 percentile point. To our original objective, i.e. the statement involving the 1 percentile point, a tiny confidence coefficient, namely 1.4 percent, is attached. If the development program is to be judged by our confidence in its ability to preponderately turn out batteries with adequate life spans (over 300 seconds), the inevitable conclusion, based on the samples behind the Tables, is that it has fallen short of the mark. Even if we agree to drop back to the 5 percentile point, i.e. demand that no more than 95 percent of the thermal batteries have life spans of over 300 seconds, a confidence coefficient of only 47 percent can be attached to our statement.

III. A Statistical Study of Thirty Type-5840 Tubes for Transconductance. Type-5840 is a heater type pentode which is to be found in many of the circuits developed by the Guided Missile Fuze Laboratory. It occurs, as has already been mentioned, in several of our amplifiers. Since the amplification which a pentode produces depends upon the transconductance G_m, rather than the amplification factor μ, the former quantity is obviously of prime importance. The G_m of this particular tube has been assigned a nominal value of 5000 micromhos, its upper and lower acceptance limits being 5800 and 4200 micromhos.

The pentodes are produced for the Laboratory in batches or lots by the manufacturer. Lot JBN was manufactured early in November 1952 (Nov. 3 thru 7), but due to the burning-in process and other manufacturing procedures was not ready for delivery until some months later. Something like 3800 tubes were produced in this lot, and most, if not all, of these were included in an order of 5000 tubes of the type in question delivered in August 1953.

As the Laboratory's Reliability Program got underway, it became advisable to round out our picture of the tube by a visit to the manufacturer's headquarters for discussions with some of his key men. A visit of this kind was made on September 15, 1953. The speaker was a member of the visiting group.

Because of certain discrepancies, it was decided at these meetings to select a fixed sample of 30 tubes from Lot JBN and have our Tube Laboratory and the manufacturer independently check their transconductances. In this way the two apparatuses used to make these measurements (the Tube Laboratory's and the manufacturer's vacuum tube bridges) could be correlated. The practice is fairly general with the manufacturer whenever he is supplying premium tubes to a vendee who wishes to check on their electrical characteristics. A member of our Tube Laboratory made the selection, and this is how the sample of 30 tubes referred to in the talk's title arose.

Reasoning correctly, our associate in the Tube Laboratory decided that the idea was to provide something very constant for the apparatuses to measure so that discrepancies in the apparatuses rather than in the things they were operating on would show up. Since tubes with low G_m's are much more likely to deteriorate with time or in passage, he culled them from the correlation sample. To balance this element of arbitrariness, he also excluded tubes with high G_m's. If the difference between the two measuring apparatuses is (theoretically) constant, i.e. does not vary with the G_m of the tube being determined, excluding extremes from the sample will not affect the accuracy of the correlation, on the contrary, for the reason already alluded to, should improve it. Unfortunately, what is good for one purpose may vitiate another. To arrive at definite statistical conclusions, it is essential in most cases to have a random sample, or if the randomness has been tampered with, to know exactly what this tampering is and allow for it. The non-randomness of the correlation sample means that we cannot draw the inferences indicated by the second of the two statistical tests in this talk, at least that we must draw them with tongue in cheek.

It was suggested that the population of the G_m's for the correlation sample tubes be tested for normality. The measurements on the tubes were made in our Laboratory and are recorded for reference in Table I, (page 80). A population of this sort, consisting of a finite number (30) of items, cannot strictly speaking be called normal or Gaussian; the question, properly interpreted, is whether the empirical population it represents is normal. The arithmetic mean and standard deviation were computed for the data in Table I, and with them in hand, this data was broken up into five classes, so chosen that the theoretical frequency of every class was 5 or more. In carrying out the test for normality a reasonable number of classes is desirable, the theoretical frequency in each class being no less than 5, and with an empirical frequency of 30, this was rather hard to do. χ^2 was then computed for these classes and frequencies and compared with the theoretical value of χ^2, taken from tables, for the 95 percent level at 2 degrees of freedom. Since the computed value (3.84) was less than the theoretical value (5.99), it was concluded at the 5 percent level that the correlation-sample population was not significantly different from normal. (The computations, etc. are given in Table II (page 81). As in all such tests, the conclusion, when the null hypothesis is not rejected, is weak; <u>we do not so much accept the hypothesis that the population is normal as reject the hypothesis that it is not normal.</u>

When the visit to the manufacturer's headquarters was made, we received a list of G_m measurements on 18 tubes, made at the time the JBN lot was produced or shortly thereafter. (See Table III, page 82). The arithmetic mean of the manufacturer's sample was 5278 whereas that of the correlation-sample (30 tubes) was 4991. It occurred to the speaker that this difference could be used to test whether the two samples represent the same population, as superficially one would assume, or whether, despite the fact that the tubes bear the same lot designation, in reality they come from different populations. This could come about if the original population (it consisted of 3800 tubes, the approximate number in the JBN lot) were split into two subpopulations, the two samples coming from distinct subpopulations. It is not necessary for the split to be "clean", e.g. the two subpopulations could overlap and have elements in common. Nor is it necessary for there to be a formal splitting; the same situation can arise as a result of a number of other practices such as taking a sample that is non-random with respect to the original population. (A non-random sample with regard to a given population can be random with regard to one of its subpopulations.) Another possibility is that the population changed, for better or worse, with time, so that the population the Component and Test section was measuring in September 1953 was not the same as that measured by the manufacturer in November 1952. In that case the difference between the means of the two samples is one rough index of this change.

The two populations can be referred to as the earlier and later populations. What we can do is test to see whether their means are significantly different. The null hypothesis then is that the means are the same; what we do is either reject it or much less forcefully, we might say "passively", accept it. It is assumed, of course, that the two populations are normal or Gaussian. If they had the same variance (they can still differ in their means), but this common variance is unknown, the test that would apply is the familiar Student's t-test. A simplifying but somewhat unnatural assumption of this kind will not be made here; the variances of the earlier and later populations will not only be treated as unknown but as, in general, different. Then the t-test does not strictly apply, but a modification of it, in which the number of degrees of freedom is computed by a rather complicated formula and need not be integral, can be used as an approximation.

The data for the manufacturer's (earlier) sample is given in Table III; it is seen that the mean $\bar{X}_2$ of the different G_m's = 5278 μ-mhos and the (estimated) standard deviation $s_2 = 241.5 \mu$-mhos. The data for the Laboratory's sample of 30 has already been given in Table I; the mean there was $\bar{X}_1 = 4991 \mu$-mhos and the standard deviation $s_1 = 326.7 \mu$-mhos. The number of degrees of freedom for the approximate t-test that applies when it is not assumed that the unknown variances of the two populations are equal, is computed in Table IV. The conclusion there is that the hypothesis that the two populations have the same mean must be rejected on the basis of the observed samples at a significance level of .05 (equal to 1 - .95).

There is of course the possibility of a systematic error, e.g. that there was a systematic difference between the Laboratory's and the supplier's measuring apparatuses, and that this accounts for the wide difference (287 μ-ohms) between the sample means. This would still mean that the populations from which the samples were drawn were different, not the real populations but the populations of measurements. We might assume that a systematic difference, if it exists, does not exceed 3 percent of the nominal G_m value (a reasonable fraction unless the apparatuses were altogether out of kilter), and see whether if even allowing for such an error, the populations were distinct. The constant error can be attributed to one or the other of the two samples or divided arbitratily between them, but the final conclusion will remain the same. The estimated standard deviation of either sample (s_1 or s_2) remains unaltered, since if each measurement in the manufacturer's sample is (algebraically) increased by a certain amount, the mean of the sample will be increased by this amount, and the difference between the corrected measurement and corrected mean will be the same as before (before the corrections were made). It is only the difference between the sample means $\bar{X}_2 - \bar{X}_1$ that is changed, and this we may assume is reduced by 3 percent of the nominal value, 5000 μ-ohms, e.g. reduced by 150 μ-ohms, in order to favor the hypothesis that the populations are not different, in other words, make it more difficult to prove that they are different. The t statistic as calculated using the altered $\bar{X}_2 - \bar{X}_1$ is reduced to 1.66. The number of degrees of freedom f remains the same (46.09) as in the calculation making no allowance for systematic errors since it depends only on the estimated standard deviations (s_1 and s_2) and the numbers in the sample. The observed t (= 1.66) and $t_{.95}(46.09)$ (=1.68) are about the same; so that at the .05 = 1-.95 level of significance, we must reject the hypothesis that the two populations are the same after allowing for a systematic error of up to 3 percent. The computations are indicated at the bottom of Table IV. (Page 83).

The populations from which the manufacturer's sample of 18 and the Laboratory's sample of 30 were drawn are, therefore, not identical. A number of explanations is possible. Those that come most readily to mind and that have already been mentioned will now be enumerated for convenience:

a) The Laboratory's sample is non-random or non-representative,
b) The manufacturer's sample is non-random or non-representative,
c) There was a systematic difference between the Laboratory's and the manufacturer's measuring apparatuses,
d) The passage of time altered the original tube population so that the Laboratory's and the manufacturer's samples were drawn, without prejudice to the question of randomness, from non-identical populations.

If a) and c) could be excluded (we know however that the Laboratory's sample was definitely not random), the trouble would be in either b) or d). At this stage it is only fair to assume that b) is false, i.e. that the manufacturer's was a random sample of the earlier populations. The spirit in which the discussions were carried on at the manufacturer's headquarters leads us to repose a great deal of confidence in the Company's integrity; if the sample was in any way unrepresentative, it was so by accident, not by intention. If a), c), and b) could all be excluded, then the responsibility would rest with d). Thus we are brought back again to the fundamental problem of the fuze, that of reliability in time: Even assuming that the fuze is perfect in design and workmanship at the moment it leaves the manufactory, how will it and its components stand up during the interval it is shipped, stored, and tested?

TABLE I

THE CORRELATION SAMPLE:
30 TUBES SELECTED FROM LOT JBN AND MEASURED FOR G_m
BY THE LABOTATORY'S COMPONENT AND TEST SECTION IN SEPTEMBER 1953

Tube No.	G_m in μ-mhos	Tube No.	G_m in μ-mhos
1	4720	16	4740
2	5440	17	5540
3	5090	18	5430
4	4510	19	5040
5	5120	20	4300
6	4860	21	4850
7	5160	22	5110
8	5350	23	5230
9	4710	24	4810
10	5150	25	5600
11	4860	26	5210
12	5100	27	4250
13	5140	28	4900
14	4760	29	5000
15	4960	30	4790

N_1 (number in sample) = 30

$\overline{X}_1$ (sample mean) = 4991μ-mhos

s_1 (sample standard deviation) = 326.7μ-mhos

TABLE II

COMPUTATION OF X^2 FOR THE DATA IN THE CORRELATION SAMPLE; COMPARISON AT THE 95% LEVEL TO ESTABLISH ITS ESSENTIAL NORMALITY

Class Interval G_m measured in μ -mhos	$\frac{u_{x_i} - \bar{X}_1}{s_1}$	Theoretical Frequency F_i	Observed Frequency f_i	$\frac{(F_i - f_i)^2}{F_i}$
$-\infty$ - 4700	- .89	5.5	3	1.14
4700 - 4900	- .28	6.2	9	1.26
4900 - 5100	+ .33	7.2	5	.67
5100 - 5300	+ .95	5.9	8	.75
5300 - $+\infty$	∞	5.1	5	.02
				$3.84 = \chi^2_{cs}$

(χ^2 for correlation sample)

From published tables $\chi^2_{.95}(2) = 5.99$

Since $\chi^2_{cs} < \chi^2_{.95}(2)$, the empirical population represented by the correlation sample is <u>not</u> at the 1.00 - .95 = .05 level significantly nonnormal.

LEGEND:

$\bar{X}_1$ = sample mean

s_1 = sample standard deviation

u_{x_i} = upper limit of the i-th class interval

l_{x_i} = lower limit of the i-th class interval

F_i = theoretical frequency of the i-th class interval

$$= 30\,G\left(\frac{u_{x_i} - l_{x_i}}{s}\right)$$ where G = normal probability of indicated interval (Gaussian probability)

f_i = observed frequency of the i-th class interval

TABLE III

THE MANUFACTURER'S SAMPLE:
18 TUBES TAKEN FROM LOT JBN AND MEASURED FOR G_m
BY THE MANUFACTURER IN NOVEMBER 1952

Tube No.	G_m in μ-mhos	Tube No.	G_m in μ-mhos
1	4950	10	5430
2	4960	11	4840
3	5270	12	5610
4	5570	13	5170
5	5480	14	5320
6	5620	15	5230
7	5200	16	4920
8	5300	17	5310
9	5290	18	5530

N_2 (number in sample) = 18

$\bar{X}_2$ (sample mean) = 5278μ-mhos

s_2 (sample standard deviation) = 241.5μ-mhos

TABLE IV.

COMPARISON OF THE MEANS OF THE TWO SAMPLES USING AN APPROXIMATE t-TEST (STANDARD DEVIATIONS OF THE TWO POPULATIONS ASSUMED TO BE UNKNOWN AND UNEQUAL)

$$t = \frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{s_2^2}{N_2} + \frac{s_1^2}{N_1}}}$$

μ_1 = theoretical mean of 1st populatior

μ_2 = theoretical mean of 2nd populatior

$$= \text{(by null hypothesis)} \; \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{s_2^2}{N_2} + \frac{s_1^2}{N_1}}}$$

$$= \text{(from Tables I, III)} \; \frac{5278 - 4991}{\sqrt{\frac{(241.5)^2}{18} + \frac{(326.7)^2}{30}}} = 3.48$$

$$f \text{ (degrees of freedom)} = \frac{\left(\frac{s_2^2}{N_2} + \frac{s_1^2}{N_1}\right)^2}{\frac{\left(\frac{s_2^2}{N_2}\right)^2}{N_2+1} + \frac{\left(\frac{s_1^2}{N_1}\right)^2}{N_1+1}} - 2$$

$$= 46.09$$

From published tables $t_{.95}(46.09) = 1.68$

Since $t > t_{.95}(46.09)$, the means of the earlier and later populations are significantly different at a level of $1.00 - .95 = .05$.

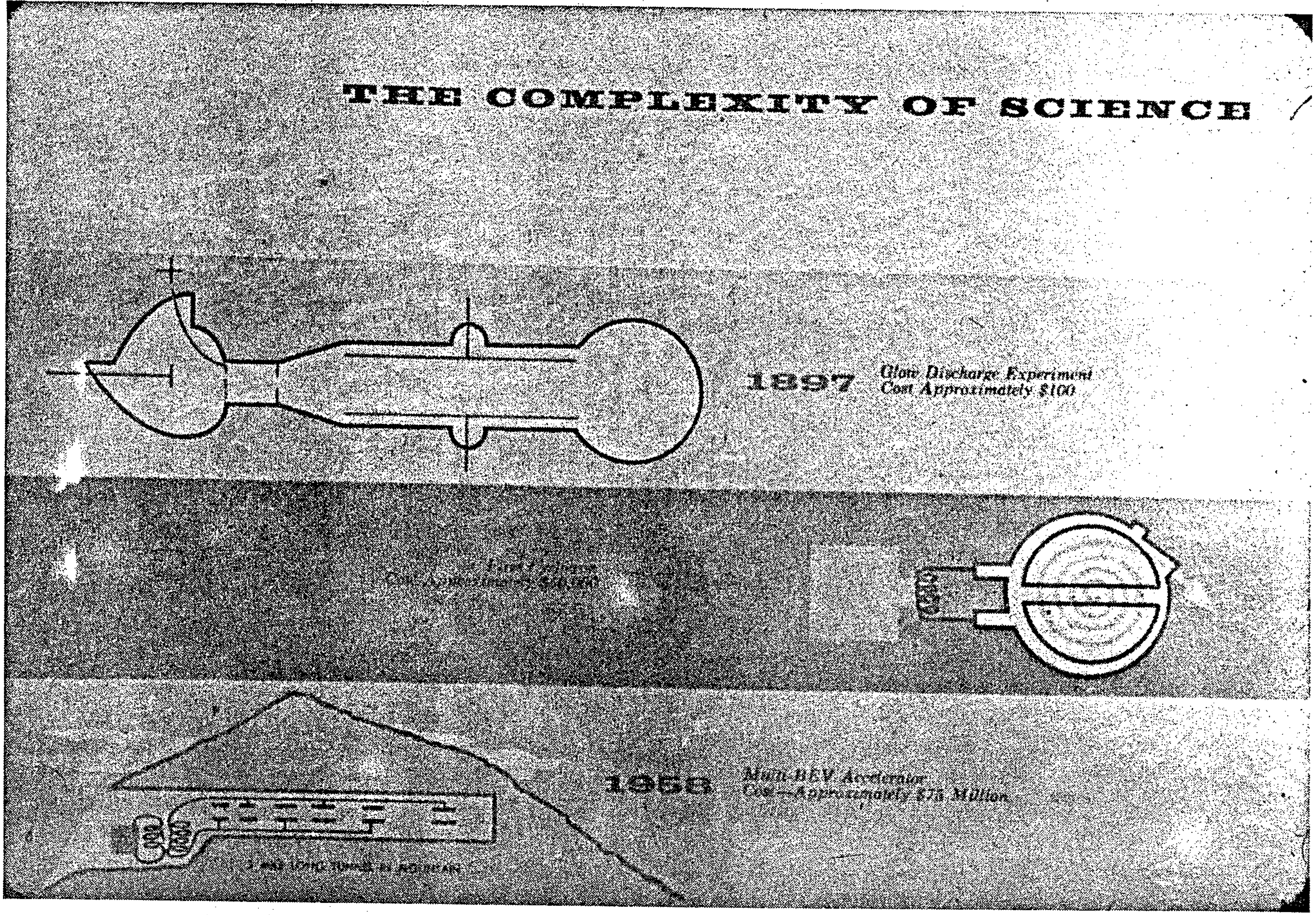
$t_e = t$ (allowing for a 3 percent systematic error = 150μ-ohms)

$$= \frac{(5278 - 4991) - (150)}{\sqrt{\frac{(241.5)^2}{18} + \frac{(326.7)^2}{30}}} = 1.66$$

Since t_e = (approximately) $t_{.95}(46.09)$, the means of the earlier and later populations are significantly different at a level of .05 even after allowing for systematic errors of up to 3 percent.

Slide 32

Complexity of Science (50 billion electron volt accelerator)



THE COMPLEXITY OF SCIENCE

1897 Glow Discharge Experiment
Cost Approximately $100

First Cyclotron
Cost Approximately $10,000

1958 Multi-BEV Accelerator
Cost—Approximately $75 Million

NEW TECHNIQUES IN MACHINE ANALYSIS OF SURVEILLANCE DATA*

Miles R. Hardenburgh and David Howes
Chemical Corps

The Chemical Corps is responsible for the storage of materiel all over the world in quantities valued at many millions of dollars. One of the facets of this responsibility is maintaining continued assurance that the stored materiel is at all times serviceable and ready for immediate issue. Serviceability requirements of Chemical Corps materiel and military supplies in general, are very exacting in that this materiel, when issued, must function within very precise quality limits. The program of maintaining a continuous knowledge of the quality of stored materiel is surveillance.

The Chemical Corps Engineering Agency is responsible for establishing all technical criteria used in the Corps' program of surveillance. Chemical Corps Materiel Command is responsible for the implementation of these criteria and the actual conduct of the inspections and tests that are required. In establishing surveillance criteria, the Agency considers four distinct elements: the basis, the interval, the technical requirements, and the serviceability level. The first of these elements - the basis - is a grouping of homogeneous material from which samples are taken that will be indicative of the quality of the material represented. The basis may be the original manufacturer's lot or a combination of manufacturer's lots, depending upon the characteristics of the material under consideration. The second element, the interval of surveillance inspection, specifies the frequency with which inspections are conducted on any given items. Intervals are usually annual but may be shortened or lengthened, depending upon knowledge of the storage characteristics and rates of deterioration of the material under study. The third element, technical requirements, consists of check points for visual inspection, classification of these check points with respect to quality requirements, functioning capabilities and, wherever necessary, tests to determine compliance with these functioning capabilities. The fourth element, the serviceability level, defines the minimum quality required to insure that the unit of issue will perform within the limits prescribed by the relevant military characteristics. Once established, these criteria are consolidated into a surveillance inspection procedure and published as a military regulation. These regulations are distributed to all commanders charged with storage responsibility and constitute a mandatory phase of the storage mission.

Surveillance, as now applied in the Chemical Corps, is relatively new, having been developed to its present capabilities since the end of World War II. Prior to World War II, only a token program of surveillance was conducted. Essentially, this program consisted of the separate storage of surveillance samples selected from a production lot at the time of production. This method proved ineffective, in that the preselected surveillance samples soon lost identity with the material which they supposedly represented. During World War II, there was no surveillance; nor was there any need. When materials were manufactured, they were issued immediately. Subsequent to World War II, large stores of materiel were returned from overseas installations to domestic depots. Some of this materiel had been subjected to severe storage conditions

*This paper was prepared within the Chemical Corps, Engineering Agency. It was not coordinated with other Activities of the Corps. Therefore, the ideas expressed reflect the opinions of the authors but not necessarily those of the Chemical Corps.

and showed the effects of marked deterioration. On the other hand, much of this materiel was relatively new and was of a quality that could be stored for later reissue. Because the domestic depots were becoming filled with stores of heterogeneous materiel, it was evident that some system must be devised whereby a quality assessment of this materiel could be conducted. As a result of this need, the first surveillance standards of the form that we employ today were devised.

In the Agency, it has always been appreciated that the primary function of a surveillance program is to maintain assurance that depot-stored materiel is serviceable. The Agency has further recognized that the data gained through these recurring surveillance inspections could be invaluable to research and development activities if properly collected, evaluated, and applied. Surveillance procedures require that an inspector, during the conduct of a surveillance inspection, prepare a comprehensive report indicating all observed attribute and variable defects. These reports are submitted to Chemical Corps Materiel Command and to the Engineering Agency. When the Engineering Agency was first organized, these reports were being submitted at the rate of approximately 1,500 a month. It was known that the data contained in these reports could be applied to many engineering problems, however, by sheer bulk alone, only a very minor fraction of the true value of these data was utilized. (See Plate 1 at the end of this paper).

Different methods were employed in an effort to reduce these data to a usable form. The first method consisted of an extraction of information from reports and a compilation on summary sheets. After a few months experience it soon became evident that the amount of work required to make these extractions and compilations was over and above the Agency's man-power capability. Next, a study was conducted to determine the feasibility of handling these data by microfilming. A saving of storage space was the only advantage gained by this method. Actually, it was more difficult to make data analysis from the microfilmed reports than it was to use the reports themselves. The next attempt to solve this problem and arrange the data in a usable manner consisted of an attempt to transfer the data to a McBee Key Punch card system. This system is similar to that used on Army personnel service records. The studies that were conducted indicated that essentially, all of the attribute data could be coded and used from these cards. However, no means was apparent wherein variable data could be used. Even though unsuccessful, this study did lead to the conclusion that all of these data could be reduced to an IBM card deck from which any type of statistical study could be conducted without losing any of the value of the raw data. The three unsuccessful attempts of data analysis, as described, covered a period of approximately 2 years.

When the idea was conceived that IBM methods could be employed, the major obstacle that stood in the way was lack of a suitable code. Fortunately, at that time, the Chemical Corps Engineering Agency had on its staff men with both statistical training and IBM experience. These men initiated work toward the development of a suitable code. This work continued for approximately 1 year, at the end of which time a code had been developed that was considered adequate.

To test the code that had been developed, one item, the M4A2 smoke pot was selected and all the data from approximately 1,000 surveillance reports was transferred to an IBM card deck. All conceivable types of statistical analysis were made from this deck. These analyses provided the information that was needed to make coding modifications consistent with Corps needs. This work was completed approximately 1 year ago. Since that time, the entire backlog of Chemical Corps surveillance reports has been transferred to an IBM card deck, and this work is kept current through the daily transference of incoming surveillance reports to this deck. At this writing, the deck consists of approximately 150,000 cards and is growing at a rate of approximately 3,000 cards a month. The deck will be maintained at a usable size by retiring cards when they reach an age of 7 years.

At the present time, the card deck is being used for statistical studies for the Agency and other activities of the Corps. Further, a monthly report, Monthly Compilation of Field Surveillance Data, is compiled which contains the surveillance data collected during the preceding month. These monthly reports are distributed on a recurring basis to all major Chemical Corps headquarters, and upon request, the Chemical Corps Engineering Agency will perform any special study of existing data that may be required. It is anticipated that through use, many new applications of this monthly report will become apparent. The report is now finding application in procurement planning, supply management, research and development, and item engineering.

This monthly report is printed by IBM equipment and is of the standard IBM report form. The data from each surveillance report is contained in two IBM cards and consequently, the monthly report contains two lines of printed information for each surveillance report covered. The first line of information gives the results of the visual inspection and includes the lot number of the material inspected, the lot size of the material inspected, the storage installation, the sample size, the date inspected, the date manufactured, and, if applicable, the date renovated. In addition, all visually perceptible defects are recorded, such as corrosion, missing components, dents, and abrasions. All packaging defects are similarly recorded and consist of such information as deteriorated or inferior packing materials, broken boards, torn barrier material, inadequate padding, and inadequate preservatives. The second line of the monthly report for any given lot consists essentially of the test data and it includes all defects that have occurred, such as duds, fuze failures, and first fire failure. Also, in the second card, all variable data are included, such as the burning time, fuze delay time, results of chemical analysis, and moisture content. (See Plates 2 and 3)

In establishing the code, devising a method of coding variable data proved to be the most difficult phase of the task. For our purposes, it was determined that the best method of coding variables was the use of a frequency distribution. The class intervals used in the frequency distribution were determined by analysis of existing data obtained from tests of a representative group of lots. The average values of the data were computed and the variability observed. The average observed value was made the center of six classification intervals, with the upper burning limit of the first and lower limit of the sixth group being made to correspond with the upper and lower limits of variability in the lots studied. When data obtained from newly tested lots infringe upon these limits, they are to some extent, atypical of the parent population. The significance of the observed infringement may be determined by common statistical test.

In the succeeding paragraphs, several proposed applications of this report are described. These applications should serve to effect a tremendous economy within the Corps and also to eliminate many of the engineering problems with which we are constantly confronted.

The Chemical Corps Engineering Agency is responsible for applications engineering, or that phase of engineering that is the bridge between research and development and regular production. To fulfill this responsibility, the Chemical Corps Engineering Agency has assigned project engineers to the overall item responsibility for individual items. This report will afford the project engineers a yard-stick of the effectiveness of the items for which they have responsibility. This report will further indicate to them the advisability of specification changes leading to better material at reduced cost. Through the analysis of this report, the project engineers are afforded the opportunity of detecting abnormal trends of deterioration before a major problem has arisen and, as a result, will be in position to take corrective action before the fact, rather than after.

In research and development, emphasis is placed on the design of new and better material; however, a significant part of the development is directed toward either the modification of existing material or toward the addition of an item to an existing family of items. Prior to the start of such development work, analysis of the data contained in these reports will indicate inherent weaknesses in design, raw materials, and fabrication of existing or related items. Through the attainment of this knowledge, the development engineer is then in a position to eliminate these weaknesses from either the modified or the newly designed item. This type of survey is related to the axiom about the chain and its weakest link. These studies afford an opportunity to eliminate the weak links.

In procurement planning, in order to maintain a predetermined stock level, one must be aware of turnovers resulting from issues, deterioration, and surveillance tests. This report contains the latter two of these planning essentials. This information, coupled with the anticipated issues, form a basis for predicting the amount of material that must be manufactured or procured during any given period of time.

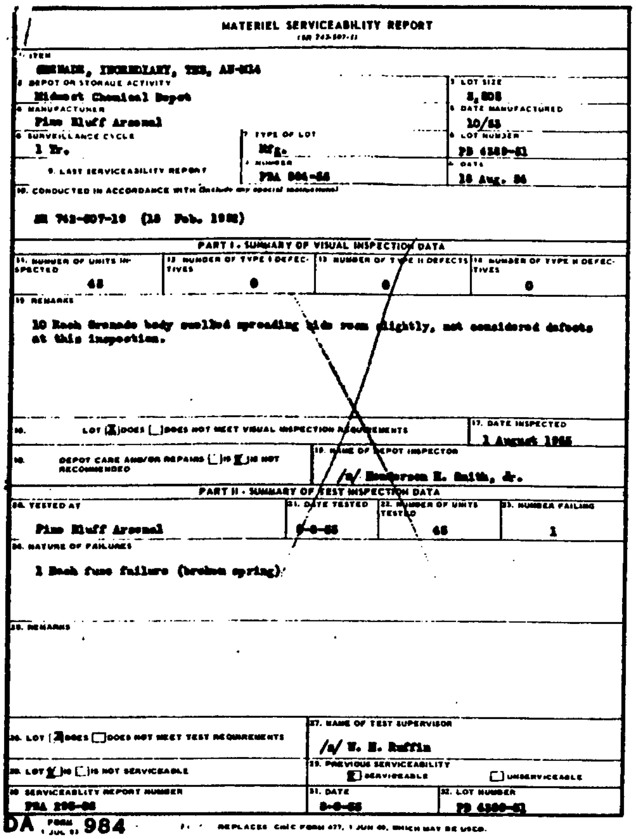
In supply planning, this report offers two excellent opportunities toward effecting a tremendous economy. The first of these opportunities is that through an analysis of the variable data contained in these reports, the degree of deterioration that has occurred in any given lot of material can be determined. After unserviceable lots are segregated and disposed of, the serviceable material will remain in supply. However, in the serviceable material there are degrees of serviceability. Using this report, one can select those lots of serviceable material with a lesser degree of serviceability and issue them prior to their deteriorating to a point wherein they would no longer conform to military characteristics. This method of issue should prove to be a very marked improvement over the existing system of first-in-first-out.

The second application in supply planning would be implementation of Supply Bulletin 3-30-1. This supply bulletin contains instructions to commanders of posts, camps, and stations, for the determination of item serviceability when normal stocks are small. Essentially, the plan recognizes the impossibility of selecting an adequate sample for destructive testing of a small group of items in order to determine serviceability. An improved procedure is that post, camp and station fragment lots be considered a part of a parent lot held in a Chemical Corps depot. Surveillance inspections would be conducted on the parent lot and the results of this surveillance would indicate to the post, camp and station commander, without his being required to actually perform surveillance tests, the serviceability of his material. This method of conducting surveillance inspections on small fragment lots is referred to as the parent-lot concept and is considered appreciably more effective than any other method yet devised.

The Chemical Corps conducts a program of environmental surveillance wherein a lot or lots of items are subjected to, and tested, under environmental extremes in Alaska, Panama, and Arizona. The purpose of the environmental surveillance program is to determine the operational suitability of thi equipment under the stresses of extreme environmental conditions. The soundness of the data generated from the environmental test program is dependent upon the typicalness of the material being tested. This report affords a method of ascertaining that the material tested is, or is not, typical.

The most significant advantage of this report, which affords a basis for analyzing surveillance data through machine records, is that any type of data analysis of compilation can be made in minutes, whereas were conventional hand, or desk methods employed, the same analysis would take weeks. Further, machines do not make mistakes.

In conclusion, the surveillance program employed in the Chemical Corps is relatively new and, as likely with any new program, is suffering certain growing pains. However, it is believed that the program has now progressed to a point wherein it serves its primary mission adequately, and in addition, it is effecting an influence toward improvement on other major programs of the Corps.



MATERIEL SERVICEABILITY REPORT
(SR 742-507-1)

1. ITEM
GRENADE, INCENDIARY, TH3, AN-M14

2. DEPOT OR STORAGE ACTIVITY
Midwest Chemical Depot

3. LOT SIZE
8,808

4. MANUFACTURER
Pine Bluff Arsenal

5. DATE MANUFACTURED
10/53

6. SURVEILLANCE CYCLE
1 Yr.

7. TYPE OF LOT
Mfg.

8. LOT NUMBER
PB 4380-81

9. LAST SERVICEABILITY REPORT — NUMBER
PBA 864-54

DATE
18 Aug. 54

10. CONDUCTED IN ACCORDANCE WITH (Include any special instructions)
SR 742-507-19 (18 Feb. 1952)

PART I - SUMMARY OF VISUAL INSPECTION DATA

11. NUMBER OF UNITS INSPECTED	12. NUMBER OF TYPE I DEFECTIVES	13. NUMBER OF TYPE II DEFECTS	14. NUMBER OF TYPE II DEFECTIVES
45	0	0	0

15. REMARKS
10 Each Grenade body swelled spreading side seam slightly, not considered defects at this inspection.

16. LOT ☒ DOES □ DOES NOT MEET VISUAL INSPECTION REQUIREMENTS

17. DATE INSPECTED
1 August 1955

18. DEPOT CARE AND/OR REPAIRS □ IS ☒ IS NOT RECOMMENDED

19. NAME OF DEPOT INSPECTOR
/s/ Henderson E. Smith, Jr.

PART II - SUMMARY OF TEST INSPECTION DATA

20. TESTED AT	21. DATE TESTED	22. NUMBER OF UNITS TESTED	23. NUMBER FAILING
Pine Bluff Arsenal	8-8-55	45	1

24. NATURE OF FAILURES
1 Each fuze failure (broken spring).

25. REMARKS

26. LOT ☒ DOES □ DOES NOT MEET TEST REQUIREMENTS

27. NAME OF TEST SUPERVISOR
/s/ W. E. Ruffin

28. LOT ☒ IS □ IS NOT SERVICEABLE

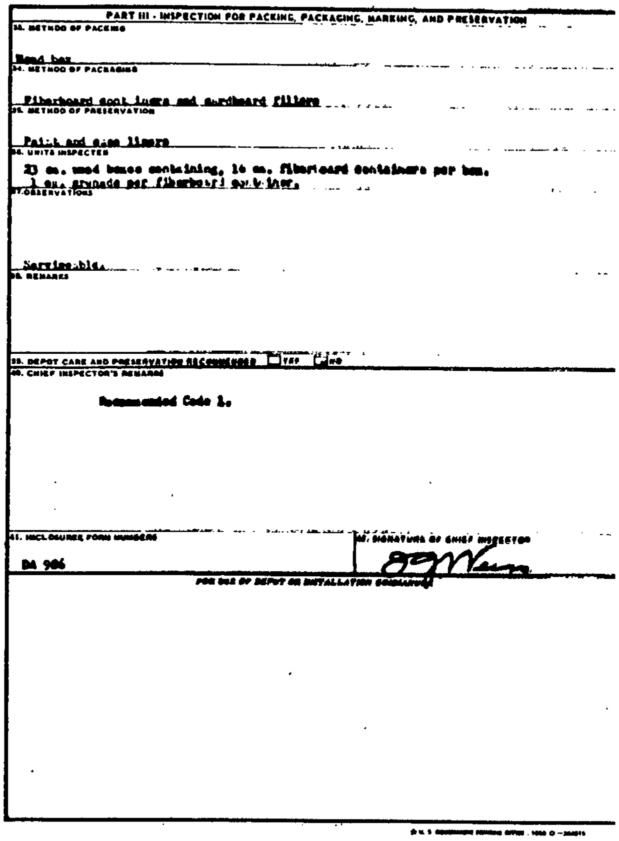
29. PREVIOUS SERVICEABILITY
☒ SERVICEABLE □ UNSERVICEABLE

30. SERVICEABILITY REPORT NUMBER
PBA 298-56

31. DATE
8-8-55

32. LOT NUMBER
PB 4380-81

DA FORM 1 JUL 53 984 REPLACES CmlC FORM 477, 1 JUN 49, WHICH MAY BE USED.



PART III - INSPECTION FOR PACKING, PACKAGING, MARKING, AND PRESERVATION

33. METHOD OF PACKING
Wood box

34. METHOD OF PACKAGING
Fiberboard cont. inners and cardboard fillers

35. METHOD OF PRESERVATION
Paint and case liners

36. UNITS INSPECTED
23 ea. wood boxes containing, 16 ea. fiberboard containers per box.
1 ea. grenade per fiberboard container.

37. OBSERVATIONS
Serviceable.

38. REMARKS

39. DEPOT CARE AND PRESERVATION RECOMMENDED □ YES ☒ NO

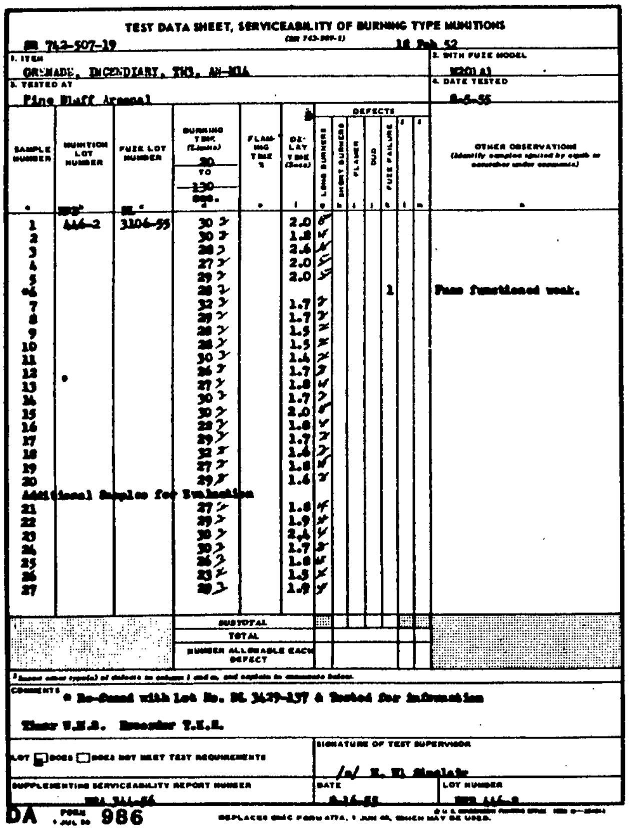
40. CHIEF INSPECTOR'S REMARKS
Recommended Code 1.

41. INCLOSURES FORM NUMBERS
DA 986

42. SIGNATURE OF CHIEF INSPECTOR

FOR USE OF DEPOT OR INSTALLATION COMMANDER

U.S. GOVERNMENT PRINTING OFFICE : 1953 O - 284815



TEST DATA SHEET, SERVICEABILITY OF BURNING TYPE MUNITIONS
(SR 742-507-1)

SR 742-507-19 18 Feb 52

1. ITEM
GRENADE, INCENDIARY, TH3, AN-M14

2. WITH FUZE MODEL
M201 A1

3. TESTED AT
Pine Bluff Arsenal

4. DATE TESTED
8-5-55

SAMPLE NUMBER	MUNITION LOT NUMBER	FUZE LOT NUMBER	BURNING TIME (Limits) 30 TO 130 sec.	FLAMING TIME %	DELAY TIME (Secs)	DEFECTS: LONG BURNERS	SHORT BURNERS	FLAMER	DUD	FUZE FAILURE	¹	¹	OTHER OBSERVATIONS (Identify samples ignited by squib or cannister under comments)
a	b MPB	c BL	d	e	f	g	h	i	j	k	l	m	n
1	446-2	3106-55	30		2.0	✓							
2			30		1.8	✓							
3			28		2.6	✓							
4			27		2.0	✓							
5			29		2.0	✓							
6			28							1			Fuze functioned weak.
7			32		1.7	✓							
8			29		1.7	✓							
9			28		1.5	✓							
10			28		1.5	✓							
11			30		1.4	✓							
12	•		26		1.7	✓							
13			27		1.8	✓							
14			30		1.7	✓							
15			30		2.0	✓							
16			28		1.8	✓							
17			29		1.7	✓							
18			32		1.4	✓							
19			27		1.8	✓							
20			29		1.6	✓							
Additional Samples for Evaluation													
21			27		1.8	✓							
22			29		1.9	✓							
23			30		2.4	✓							
24			30		1.7	✓							
25			26		1.8	✓							
26			23		1.5	✓							
27			29		1.9	✓							
			SUBTOTAL										
			TOTAL										
			NUMBER ALLOWABLE EACH DEFECT										

¹ Insert other type(s) of defects in column l and m, and explain in comments below.

COMMENTS
* Re-Cased with Lot No. BL 3429-137 & Tested for information

[illegible] U.N.S. Inspector T.E.R.

LOT ☒ DOES □ DOES NOT MEET TEST REQUIREMENTS

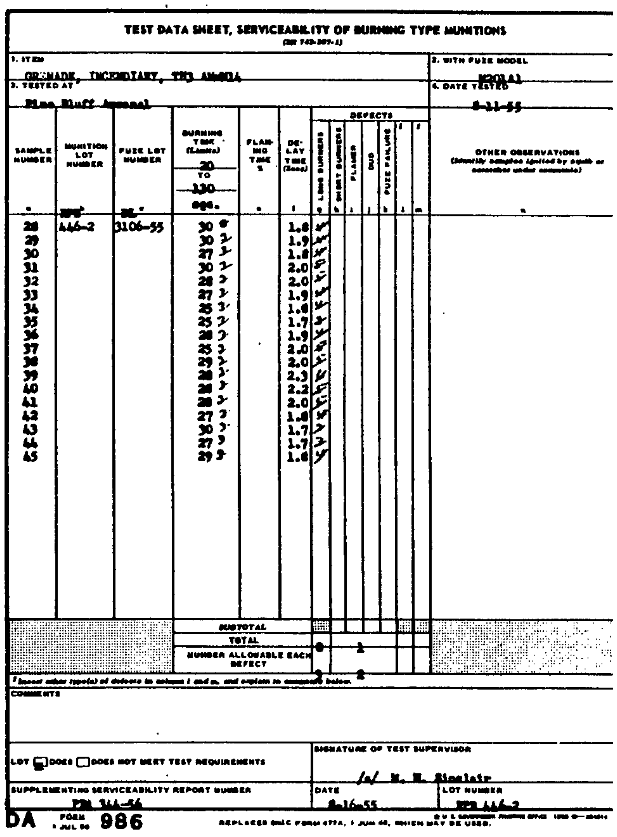
SIGNATURE OF TEST SUPERVISOR
/s/ M. W. Sinclair

SUPPLEMENTING SERVICEABILITY REPORT NUMBER
PBA 344-56

DATE
8-16-55

LOT NUMBER
MPB 446-2

DA FORM 1 JUL 50 986 REPLACES CmlC FORM 477A, 1 JUN 49, WHICH MAY BE USED.



TEST DATA SHEET, SERVICEABILITY OF BURNING TYPE MUNITIONS
(SR 742-507-1)

1. ITEM
GRENADE, INCENDIARY, TH3 AN-M14

2. WITH FUZE MODEL
M201A1

3. TESTED AT
Pine Bluff Arsenal

4. DATE TESTED
8-11-55

SAMPLE NUMBER	MUNITION LOT NUMBER	FUZE LOT NUMBER	BURNING TIME (Limits) 20 TO 130 sec.	FLAMING TIME %	DELAY TIME (Secs)	DEFECTS: LONG BURNERS	SHORT BURNERS	FLAMER	DUD	FUZE FAILURE	¹	¹	OTHER OBSERVATIONS (Identify samples ignited by squib or cannister under comments)
a	b MPB	c BL	d	e	f	g	h	i	j	k	l	m	n
28	446-2	3106-55	30		1.8	✓							
29			30		1.9	✓							
30			27		1.8	✓							
31			30		2.0	✓							
32			28		2.0	✓							
33			27		1.9	✓							
34			25		1.8	✓							
35			25		1.7	✓							
36			28		1.9	✓							
37			25		2.0	✓							
38			29		2.0	✓							
39			28		2.3	✓							
40			28		2.2	✓							
41			28		2.0	✓							
42			27		1.8	✓							
43			30		1.7	✓							
44			27		1.7	✓							
45			29		1.8	✓							
			SUBTOTAL										
			TOTAL			0		1					
			NUMBER ALLOWABLE EACH DEFECT			3		2					

¹ Insert other type(s) of defects in column l and m, and explain in comments below.

COMMENTS

LOT ☒ DOES □ DOES NOT MEET TEST REQUIREMENTS

SIGNATURE OF TEST SUPERVISOR
/s/ M. W. Sinclair

SUPPLEMENTING SERVICEABILITY REPORT NUMBER
PBA 344-56

DATE
8-16-55

LOT NUMBER
MPB 446-2

DA FORM 1 JUL 50 986 REPLACES CmlC FORM 477A, 1 JUN 49, WHICH MAY BE USED.

LOT NUMBER	LOT SIZE	TYPE	QTY	VISUAL DEFECTS						DEPOT	PKG COND	CL	PACKAGING DEFECTS				SAMPLE SIZE	DATE INSP	DATE MFG	DATE RENOV
				TYPE	QTY	TYPE	QTY	TYPE	QTY				TYPE	QTY	TYPE	QTY				

LOT NUMBER	TEST DEFECTS				GROUP I TEST DATA						GROUP II TEST DATA						TOTAL QTY. DEF.	SAMPLE SIZE	DATE INSP	DATE MFG	DATE RENOV
	TYPE	QTY	TYPE	QTY	1	2	3	4	5	6	1	2	3	4	5	6					

```
                                                                                     136 mo
      159   24  H^251 Dud                                    3 8                 50 172  36
      159   24  TO    3           2 13  7                                    3   25 172  36

    361 Grenade, Red, M18

                                                                                      24 mo
  PB  4366  17  M^3500 Fuze failure                          5 8                 86 175 151
  PB  4366  17  TN 10 TF Flamer 1   6 60 16  4                2 44 38  2    11   86 175 151

    363 Grenade, Yellow, M18

                                                                                      32 mo
  PB  2111  22  L^1500 Fuze failed to ignite starter mix     3 8                 45 166 134
  PB  2111  22  TN 3 TH Fuze failure 5   28 13  1                   38       8   45 166 134

                                                                                      17 mo.
  PB  3395   1  L^1500 Long burner                           3 8                 20 166 149
  PB  3395   1  TA  5               1  8  2  4  5                   20       5   20 166 149

                                                                                      17 mo.
  PB  3395   3  L^1500 Fuze failed to ignite starter mix     3 8                 45 166 149
  PB  3395   3  TN 3 TA Long burner 5   3 23  9  2  5               44       8   45 166 149

                                                                                      16 mo
  PB  439   72  P^37,500 Dud                                 3 8                450 166 $50
  PB  439   72  TC  3 TA Long burner 10   5 19  7  1 10             42      13   45 166 $50

                                                                                      13 mo.
  PB  4391  15  L^1500 Fuze weak causing dud                 3 8                 25 167 154
  PB  4391  15  TD 3 TH Fuze failure 1   16  1  1  4           1 15  9       4   25 167 154

                                                                                      11 mo.
  PB  4391  57  M^3500                                       3 8                 20 166 155
     Next page for Cards 2&3
```

PUNCHED CARD COMPUTING OF ANALYSES OF VARIANCE*

Clifford J. Maloney
Fort Detrick, Frederick, Maryland

Introduction. The development of punched card machines for census work first in 1890 by Dr. Herman Hollerith[1] and later by James Powers and others[2] for use in the United States Census Bureau is well known. The "Analytical Engine" of Charles Babbage which traces back to 1820 was card controlled.[3] It is not so well known that a "Tabulating Device" developed by Colonel Seaton was used in the 1870 census for which Colonel Seaton was paid $15,000 by the United States Government for perpetual rights to the use of the device.[4] This may or may not have been a bigger bargain than the first census UNIVAC.

The first non-census statistical use of punched cards is not known but there was an application to meteorological records in Austria (or at least a projected one) in 1922 and again in the United States Department of Agriculture following the 1920 census. In 1924 Bradford B. Smith devised a method for obtaining correlation tables on a punched card tabulator and sorter[5]. Computations were completed by hand by the "method of grouping"[6]. Shortly afterward punched card equipment was applied to statistical computations by Brandt[7] and Snedecor[8] at Iowa State and by Eckert[9] and Mendenhall and Warren[10] at Columbia University. The method was expanded at the Institute of Statistics on its formation at the University of North Carolina in 1944. This work and most later work in this country has employed the equipment manufactured by the IBM Corporation[11].

An exposition of the methods in use for computing analysis of variance by punched card methods was reported by Monroe[12]. Complete calculation of analysis of variance by punched card methods, including the formation of all tables, sums of squares, and corrections for the mean, has been done on our Sperry Rand equipment at the U. S. Army Chemical Corps Biological Warfare Laboratories, Fort Detrick, Frederick, Maryland[13], since August 1950.

Sperry Rand punched card equipment employs a 90 column punched card shown in Figure 1. This card is prepared from the original records on a card punch shown in Figure 2. (See page 105). They are arranged in any order desired on the sorter shown in Figure 3. In Figure 4 is shown a combination collator-reproducer, which permits the use of punched card tables of functions such as squares, square roots, logarithms, powers, exponential, trigonometric functions, and others with one card pass, and without the necessity of merging the decks or disturbing the completeness of the detail deck or the completeness or order of the table deck. It is, however, necessary to rearrange the detail deck twice. The UNIVAC 120 computer is shown in Figure 5. Figure 6 shows a tabulator with attached summary punch. Details of the equipment may be obtained from Sperry Rand[14]. Figure 7 shows a locally developed system for input to the computing equipment[15]. A great deal of the data treated by us consists of bacterial number counts. The input system is designed to count these data automatically and at the same time prepare cards with the resulting total counts for later processing in the system. Finally, a special column by column card reader for controlling an electric typewriter developed by Fort Detrick[16] is shown in Figure 8. This device is used to produce finished statistical tables of the results calculated by the unit.

* Cleared for open release as paper number BWL 1661, 19 April 1956

Computation Procedure. Statistical analyses take a variety of forms which vary from application to application, each depending to a considerable extent on the subject matter field. In the analysis of planned, controlled experiments, the usual form is that of Analysis of Variance and Covariance. Those not familiar with these techniques are referred to one of the textbook expositions now available. However, it may be said that analysis of variance consists of a great deal of rearranging and adding, and a certain amount of squaring of the resulting sums, which in turn are then added together. Finally, a "correction for the mean" must be applied to the resulting "sums of squares". The calculation of the original sums of the data values is an obvious application of the tabulator, and the results are summary punched for subsequent repetition of the process as desired. Before acquisition of the UNIVAC 120 the squares were obtained by a "table" operation on the multi-control reproducing punch. Finally the "correction for·the mean" was obtained by the use of the tabulator.

Acquisition of the UNIVAC 120 computer has rendered obsolete the use of most punched card tables of squares, logarithms, and so on, as well as permiting calculation of analysis of covariance, simultaneous linear equations, curve fitting, and other standard statistical calculations. Two applications of our equipment have been made that so far as we know have not been done elsewhere. One consists of the calculations necessary to conduct a quality control program on laboratory processing of test results[17]. The other involves machine calculation of bioassay results, using either the logit or probit technique[18]. Research is underway to devise methods of determining observed F and t values by calculation on the UNIVAC 120, so that the choice of appropriate error terms may be based on any selection of pooling rules.

As work comes to the section it will have already received the attention of a professional statistician. In general, the data will have resulted from a properly planned experiment. Even so, there will in some cases be a considerable unavoidable loss of data. Procedures for dealing with such cases have been worked out, but will not be discussed here. Each set of related data constitutes a job, which may include a number of analyses. In general in our practice, analysis is performed in logarithms of the percentage of an original bacterial count which is still present and viable after a certain process of a certain period of time in storage or adverse treatment. The local expression "value summing" refers to the formation of the usual two, three, and n-way tables, preliminary to computing sums of squares. The average experiment receiving analysis of variance contains some 500-1000 readings. These are consolidated to 100-300 by the time the analysis of variance stage is reached. Table I shows the times for the various steps of the process for data received from one using group. In any computing installation provision for insuring accuracy is of great importance. In our practice all data are checked for correctness before receipt in the computing section as indicated by the initials of the experimenter on the data sheets. It is next scanned, punched twice, and the two decks compared on the collator-reproducer or MCRP. The various summing operations are checked by comparing with their common total. The UNIVAC 120 automatically checks every calculation. Before running the problem, the machine is checked out with a check deck. Finally, the results of the calculation are examined for general reasonableness, by a computer check program, and by hand. Even with all these precautions absolute accuracy is not obtained, and some errors are detected only by the professional statistician in his examination of the final tables.

Performance records indicate that each job involves about 8 separate analyses of variance and takes one hour or slightly longer per analysis of variance to compute. This includes a number of operations to insure accuracy, converting all readings to logarithms, and carrying out a factorial analysis of variance. Those familiar with such work on desk calculators can compare their speed with that given here. It is our estimate that the punched card unit is about four times as productive on an employee basis and, as equipment cost is about as much as the salaries of the people who operate it, about twice as economical on a dollar basis. In making these estimates allowance is made for the fact that the personnel do not spend all o their time operating the equipment, and also that the equipment is not always performing useful work, but may be undergoing repairs or be idle.

Computing Unit Capacity Estimation. Punched card computing installations and still more, electronic computers are composed of relatively few high capacity devices. Whether work is done on one machine at a high rate or on two machines each working at half the rate of speed is of no consequence so far as actual production is concerned. It does have an effect, however, on the amount of time spent in waiting if the several jobs are of variable length and/or arrive in the computing section at erratic intervals.

The question of estimating the performance of facilities working under random demand or conversely of estimating the extent of facilities required to give service of a given standard under these conditions seems first to have arisen in the telephone field[19]. Rather complete study of the problem has been made of this application[20] and sporadic and limited application has been made in other fields. Under the heading of "Queueing Theory" the application has been extended to vehicular traffic at toll gates, checking counters at chain stores, parking lot facilities, air traffic service, ticket lines, and many others[21]. Of course, all authors have realized otherwise, but the applications cited suggest a limitation of the method to situations in which demand consists of a very large number of very brief requests. The application in our computing room involves relatively few operations but of somewhat longer relative duration. It was with some diffidence therefore that an investigation of the applicability of this theory in investigating the capacity of our computing facilities was attempted.

Statement of the Problem. It will first be desirable to explain just what the problem is. If computing work were tended to the section only when all previous work had been completed and the total volume of work offered were less than the total capacity of the system, then no work would be delayed in the sense of having to wait to receive attention. However, if the work comes in at random, even though the total volume is less than the capacity of the unit, work will be delayed in many cases though at other times the equipment stands idle. On the other hand even if work comes in at regular intervals but takes a variable length of time for processing, the same situation develops. At times the work is waiting, at others the system is idle. Finally, if work comes in at random and requires random processing time, the situation is aggravated. The four cases are shown in Table II. Figures 9, 10, 11 and 12 show these four cases pictorially.

In the ideal case of random arrivals and random service times, formulas were worked out some time ago giving the average delay to be expected expressed in terms of the average time required for processing[20]. Applying these equations, the theoretical delay of one of our jobs because of computing room congestion was computed for varying conditions, Figure 13. Delay is expressed in units of service time. Traffic density is referred to as percent utilization of the equipment and is simply the ratio of hours of actual use to potential hours of use. The curves were computed for 1, 2, and 3 channels, or like units of equipment, and the number of sources, or persons submitting work requests, was assumed to be infinite. This assumption is hardly realistic, but additional calculations indicated that even when the number of sources was assumed to be 5 and the channels 1, the results corresponded very well to the case with infinite sources and 1 channel. Examination of the figure reveals some startling conclusions. If the work requirement is such that the system is kept busy only 50 percent of the time, each job will nevertheless be delayed for a time equal to that required for its processing if only one "channel" is available. If the work load goes up to 66-2/3 percent the delay will actually be twice as long as the time required for servicing despite the fact that the system will be idle 1/3 of the time. For 80 percent utilization, the delay is four times the service time, and for 90 percent utilization, quite enormous. A great improvement is introduced when two channels are used if delay is expressed in terms of service time. The further improvement introduced by a third machine, while large, is not quite as large. This same trend would continue for a greater number of channels. However, if the additional units are purchased at the cost of slower individual operation, the situation is reversed, Figure 14. Here, the time unit of the y-axis is expressed in terms of the single high capacity channel. Case B represents two channels but of one-half capacity, C represents three units, now of 1/3 capacity.

Basis for Theoretical Delay Calculations. In order to calculate theoretical delay curves a "mathematical model" of the system is set up. This consists in selecting certain assumptions as descriptive of the real process. The first of these is that the distribution of job arrivals at a particular machine is Poisson, i.e., distributed individually and collectively at random. Records were not available which would permit a study of the distribution characteristics of arrivals but the assumption appears reasonable. The second assumption was that the service times on a particular machine are exponentially distributed. This was examined, as discussed later. In the pictorial situation this situation is that of Figure 12. It should be emphasized that these particular assumptions are chosen only for simplicity of calculation.

The machines routinely employed in most computing sequences are (1) keypunch, (2) reproducing punch, (3) sorter, (4) UNIVAC 120 electronic computer, and (5) tabulator. The order of listing is not necessarily the order in which the equipment is used. Further, there may be some projects which require the use of one or more of the machines at several stages in the computing sequence, while some jobs may bypass one or more of the machines completely. Records of the service times for a thirty-day period for all of the above machines except the sorters were examined and tested to determine whether or not an exponential distribution could be used to describe them. The chi-square test for goodness of fit was applied. The available data indicated that the exponential could be used as a distribution function for the tabulator, the reproducing punches, and the computer, but not for the keypunches. The computer was chosen as the machine of major interest inasmuch as it was an essential link in the computing sequence of the jobs later referred to. The distribution of computer service times is shown graphically in Figure 15, along with an exponential curve fitted to this data.

Comparison of Observed and Theoretical Delay. To see how well the actual and theoretical delays agreed, actual performance figures for the output of the computing section for a period of one month were examined. In making this study the jobs submitted by only one operating division were used, as daily records on their progress were available. During this period twenty-six jobs were completed. Each of these jobs contained an average of 8.7 analyses of variance and required on the average 8.3 working days for completion. Of these 8.3 days, approximately 3.3 were consumed by the keypunch and verification operations. Since the keypunch service time distribution was not exponential, it was felt that this phase of the process should be omitted in attempting to compare actual delays with expected. Of the remaining 5 days, approximately 1 day was consumed by an independent review of the output of the unit for accuracy and compliance with instructions at several stages in the computing process, leaving about 4 working days, or 32 hours, spent in the computing room. Unfortunately, during this period it was not possible to ascertain how much of these 32 hours was actually spent on the several machines because of additional processing with the same cards for other purposes and our inability to separate easily the total machine time into the analysis of variance effort (our concern here) and the other semi-related computations. The estimates given earlier indicate, however, that slightly in excess of one hour is required on the average for the calculation of one analysis of variance, excluding keypunching. It would appear, therefore, that approximately 9 to 12 hours on the average were required in the actual processing of each job, i.e., 8.7 analyses at one and a fraction hours per analysis.

Percent utilization for this period for the computer was 65 percent, for the tabulator, 49 percent, and for the reproducing punch, 39 percent. For the moment, however, let us assume that the percent utilization on all of this equipment was 65 percent. This may be defended by considering the heaviest loaded machine type as a "bottleneck" or "master rate." Referring to the delay curves of Figure 13 and reading upward from the point of 65 percent utilization to the curve for one channel (since there was but one of each of the three machines in use at this time), a delay ratio of approximately two is obtained. This then says that for every unit of time spent on a machine, approximately two units are spent in waiting for that machine to become available, with the assumed overall traffic density of 65 percent. This would compare favorably with the actual performance figures which indicates that of the 32 hours spent in the computing room, roughly 9 to 12 were required for actual operation of equipment, with the remaining 20 to 23 hours presumably spent in waiting for one machine or the other to become available.

Study of our records has shown that despite the fact that our equipment was only two-thirds loaded, work was spending twice as long in waiting as in being processed. It was gratifying to discover that this is just what the simplest congestion theory indicates.

REFERENCES

1. Hollerith, Herman. "The Electrical Tabulating Machine". J. Roy Stat. Soc., Vol. 57 1894, pages 668-689.

2. Mahaney, John J. "Machine Tabulation". 18 October 1945, Unpublished Manuscript.

3. Bowden, Bertram V. "Faster Than Thought". Pitman, 1953.

4. Walker, Francis A. "Report of the Superintendent of the Ninth Census".

5. Smith, Bradford B. "The Use of Punched Card Tabulating Equipment in Multiple Correlation Problems". 24 pages, USDA, October 1923.

6. Snedecor, George W. "Statistical Methods". 4th ed., 1953, art. 8.9.

7. Brandt, A. E. "Use of Machine Factoring in Multiple Correlation". J. Am. Stat. Assoc., Vol. 23, No. 163, pages 291-295, September 1928.

8. Snedecor, George W. "Uses of Punched Card Equipment in Mathematics", American Mathematical Monthly, Vol. 35, No. 4, pages 161-168, April 1928.

9. Eckert, W. J. "Punched Card Methods in Scientific Computation". 1939.

10. Warren, Richard and Mendenhall, R. M. "The Mendenhall-Warren-Hollerith Correlation Method". Col. U. Stat, Bureau, Doc. No. 1, 1929.

11. Bibliography on "The Use of IBM Machines in Scientific Research, Statistics, and Education". Watson Scientific Computing Laboratory, 1947.

12. Monroe, Robert J. "The Application of Machine Methods to Analysis of Variance and Multiple Regression". Proceedings of the IBM Computation Seminos, December 1949, pages 113-116.

13. Maloney, Clifford J. "Statistical Computation by Punched Cards". 1954, Unpublished.

14. 1615 L Street, N.W., Washington 6, D. C.

15. Maloney, Clifford J. "Automatic Bacterial Density Assessment". Paper given at 1954 meeting of Maryland Section of the Society of American Bacteriologists.

16. Walker, Douglas and Maloney, Clifford J. "Development of a Punched Card Reader". Interim Report 103, Biological Warfare Laboratories, Fort Detrick, Frederick, Maryland.

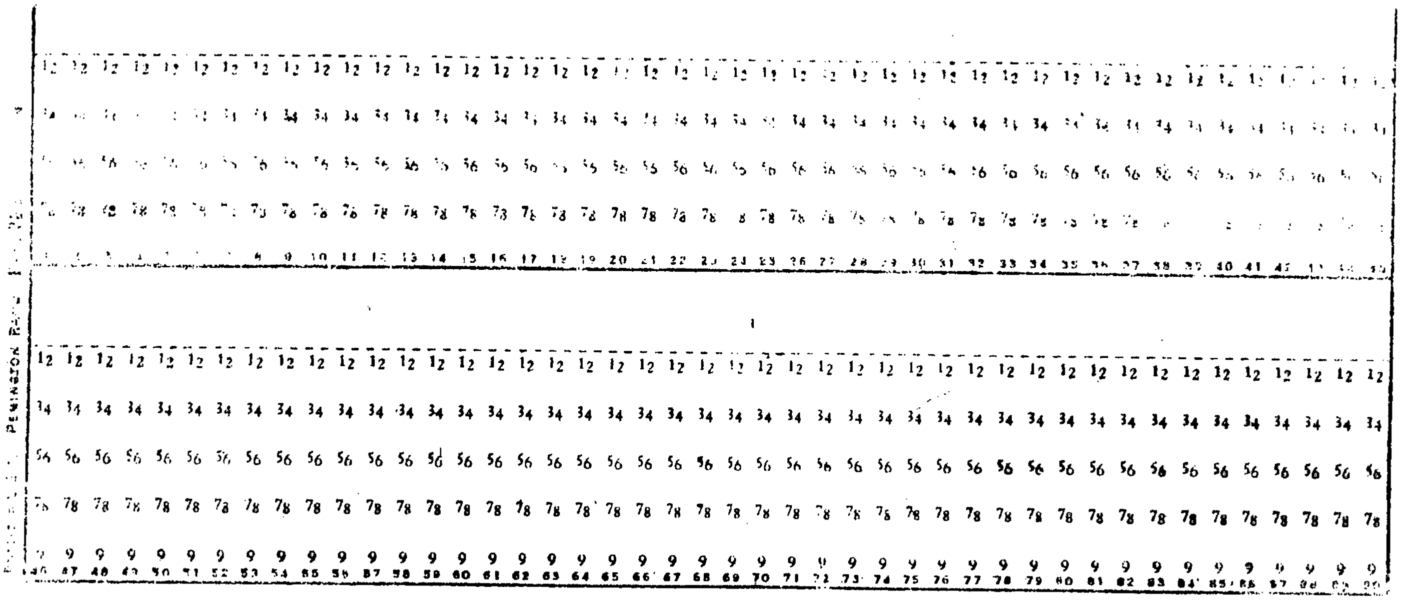
17. Maloney, Clifford J. "Quality Control Applied to Routine Counting of Bacterial Plates". Paper read at the spring 1953 meeting of the Virginia Academy of Sciences.

18. Maloney, Clifford J. "Statistics of Bioassay". Appendix to Minutes of the Ninth Meeting Subcommittee of Animal Reservoirs and Vectors of Diseases, Committee on Sanitary Engineering and Environment, National Academy of Sciences, National Research Council, 1 December 1953.

19. Johannsen, F. The Post Office Electrical Engineers Journal, October 1910, page 244.

20. Fry, Thornton C. "Probability and its Engineering Uses". Van Nostrand 1928.

21. Bartlett, M. S. "Introduction to Stochastic Processes with Special Reference to Methods and Applications". Cambridge University Press, 1955.


REMINGTON RAND

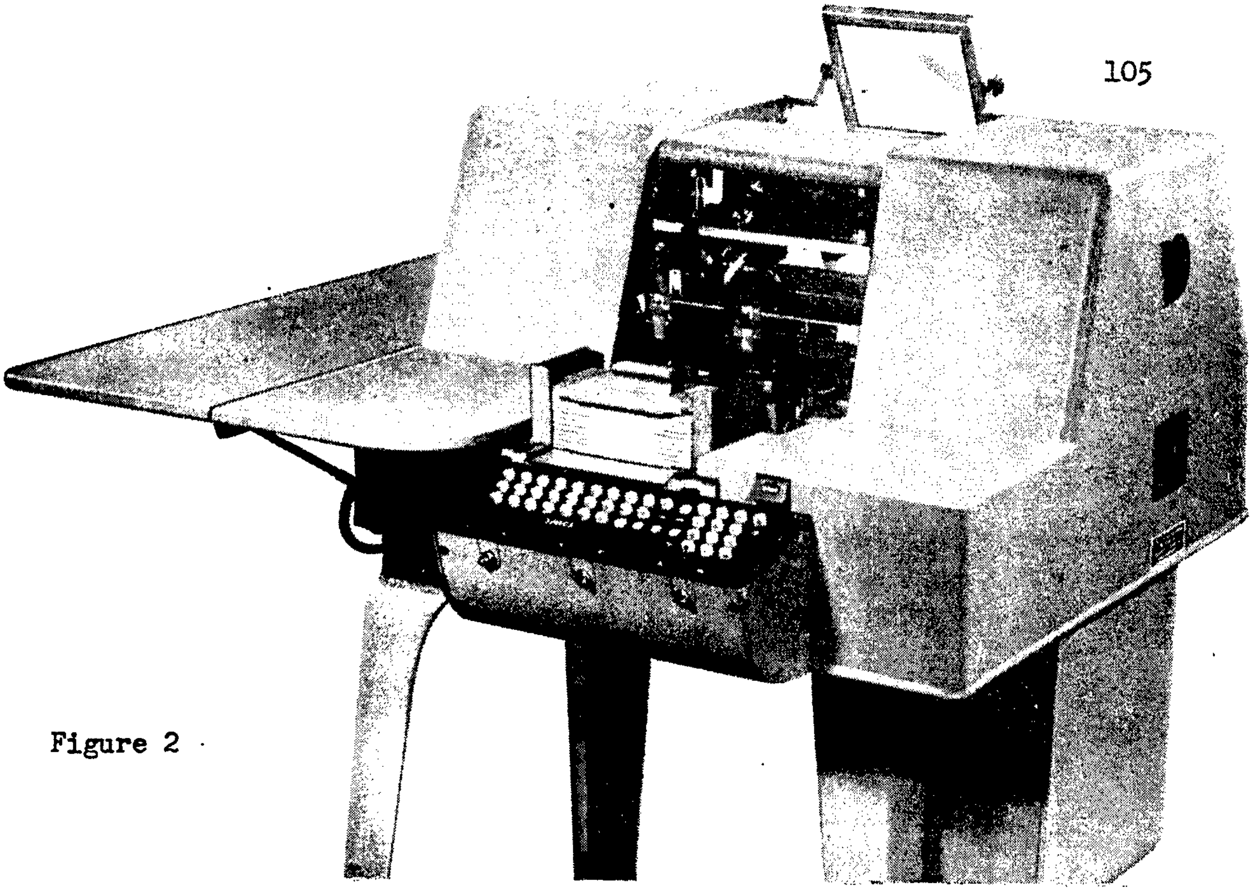


Figure 2

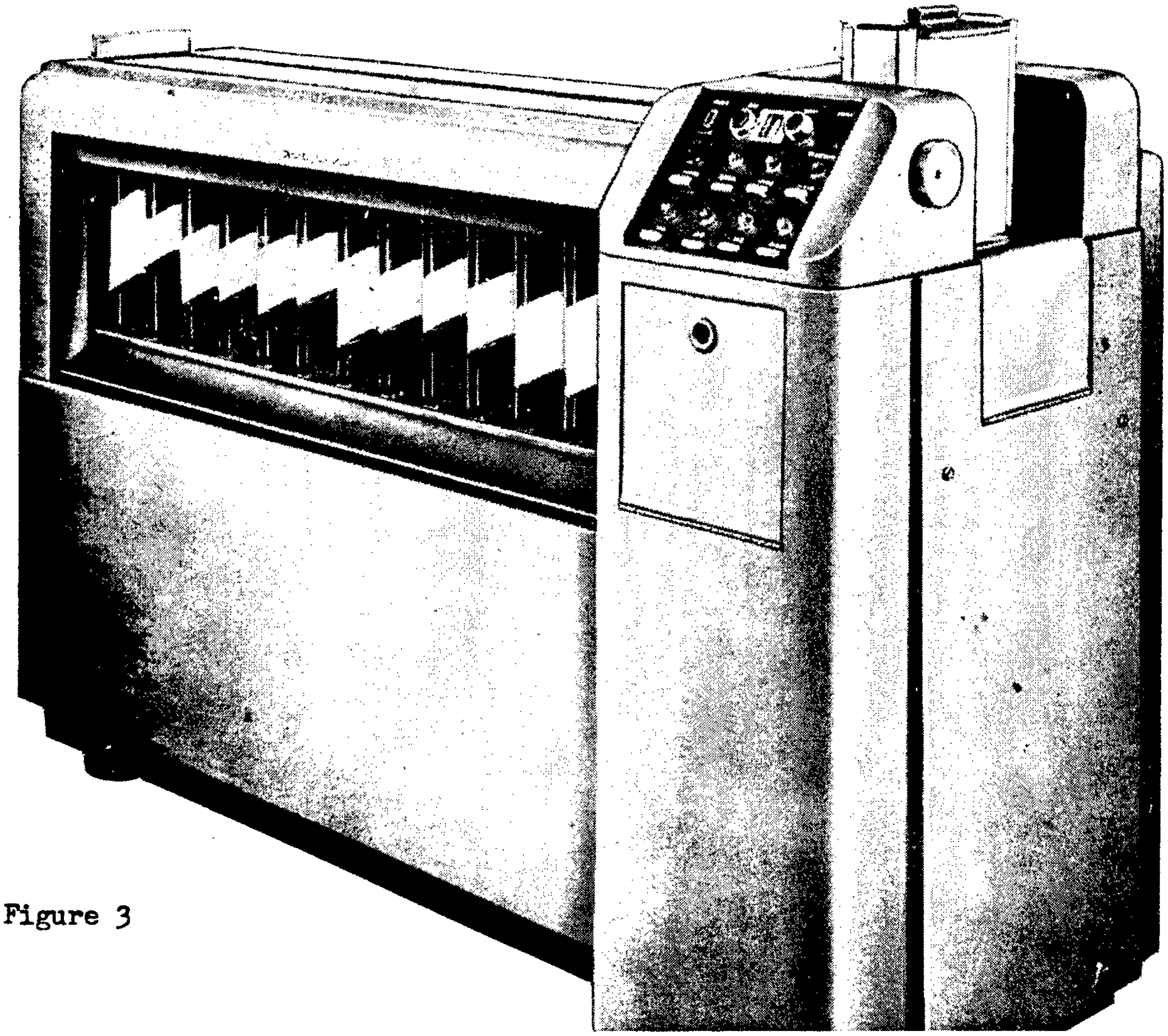


Figure 3

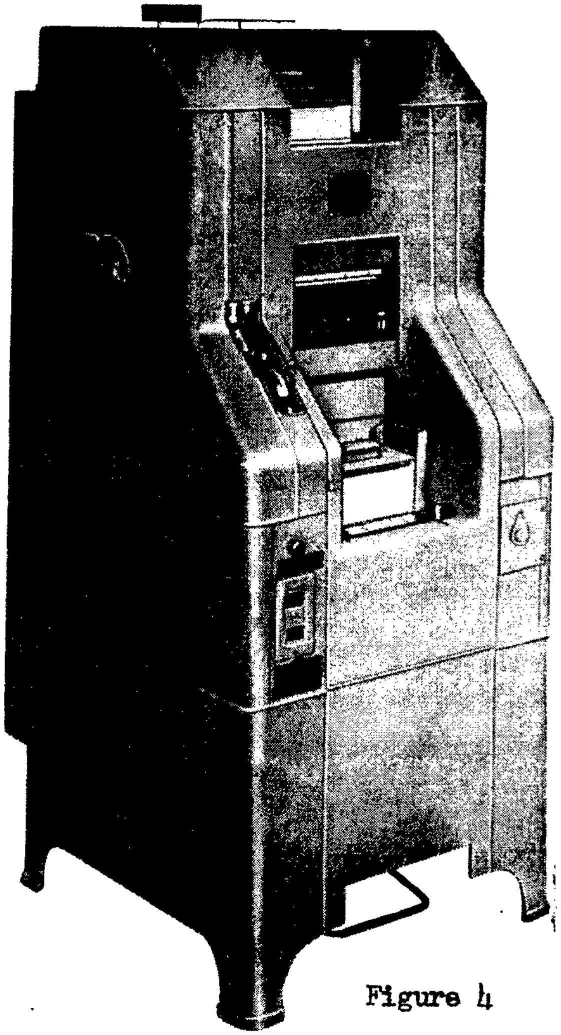


Figure 4

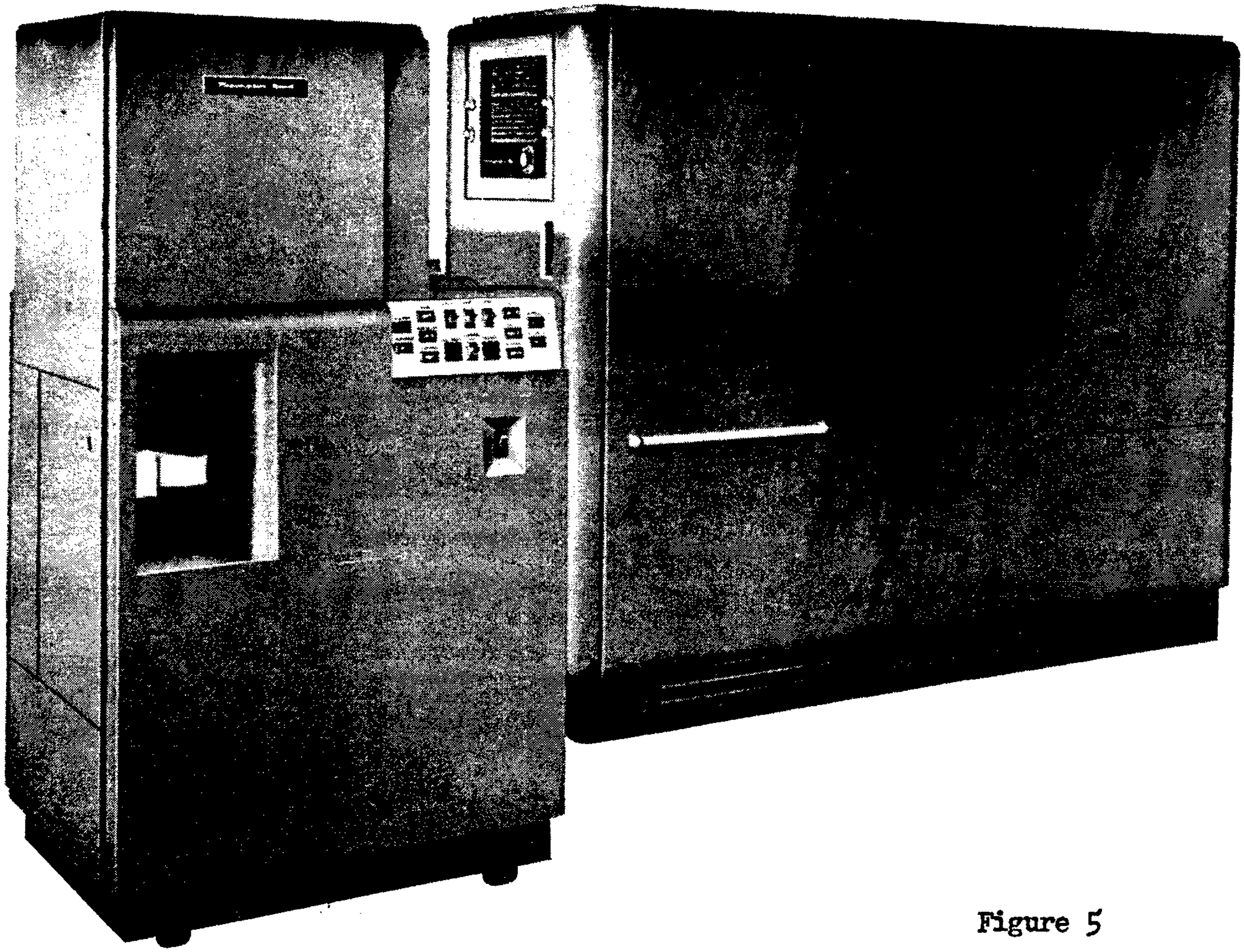


Figure 5

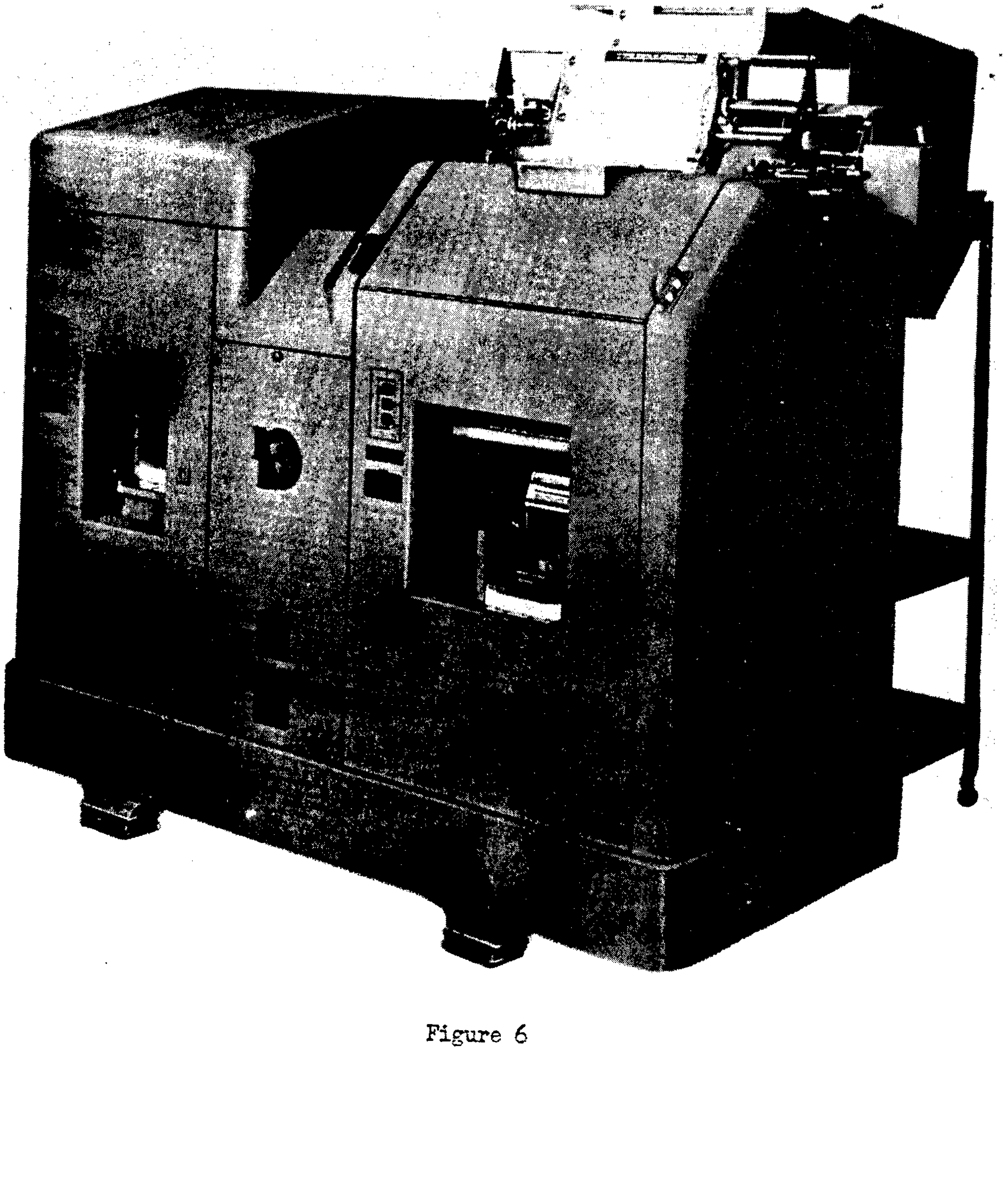


Figure 6

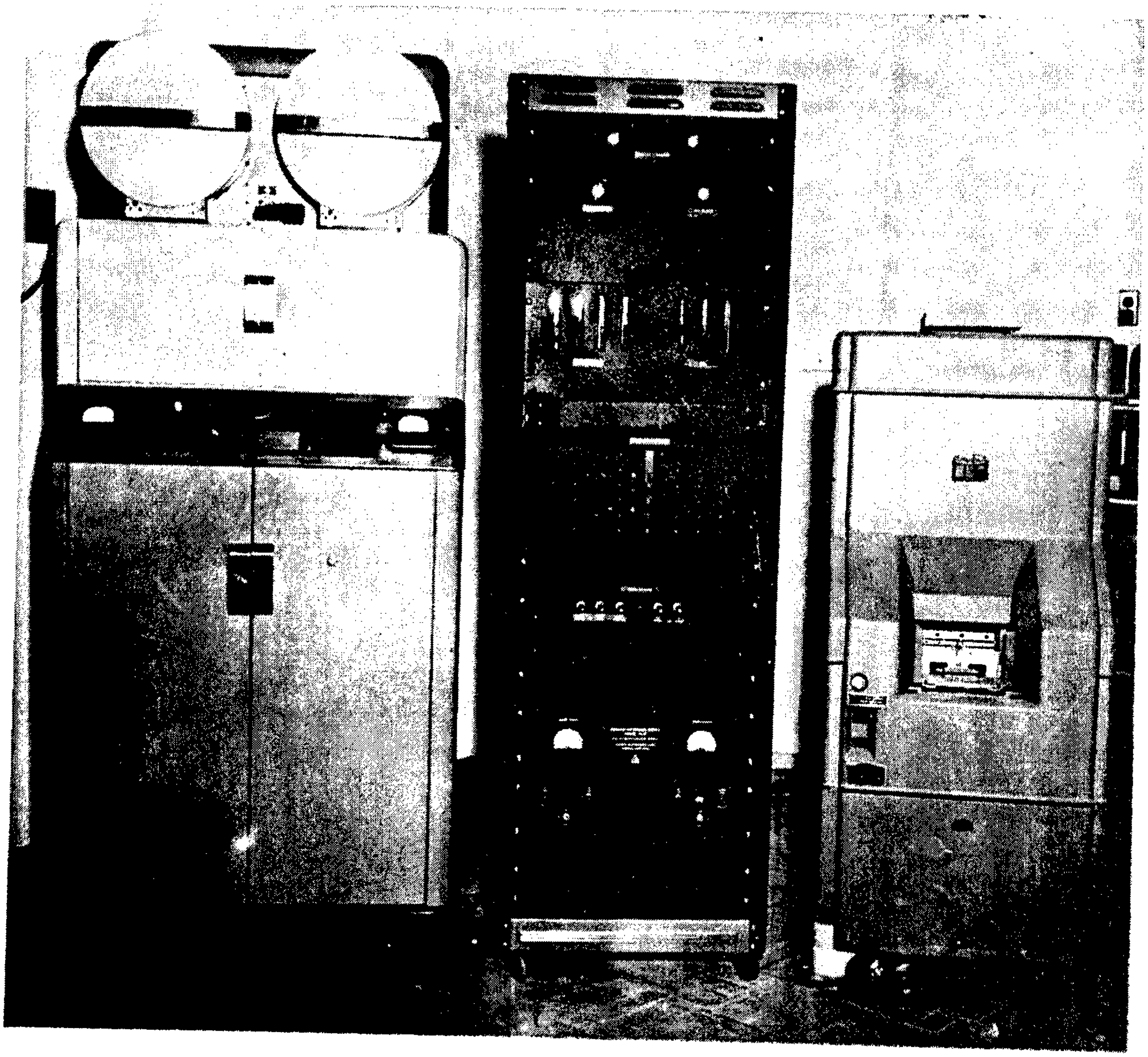


Figure 7

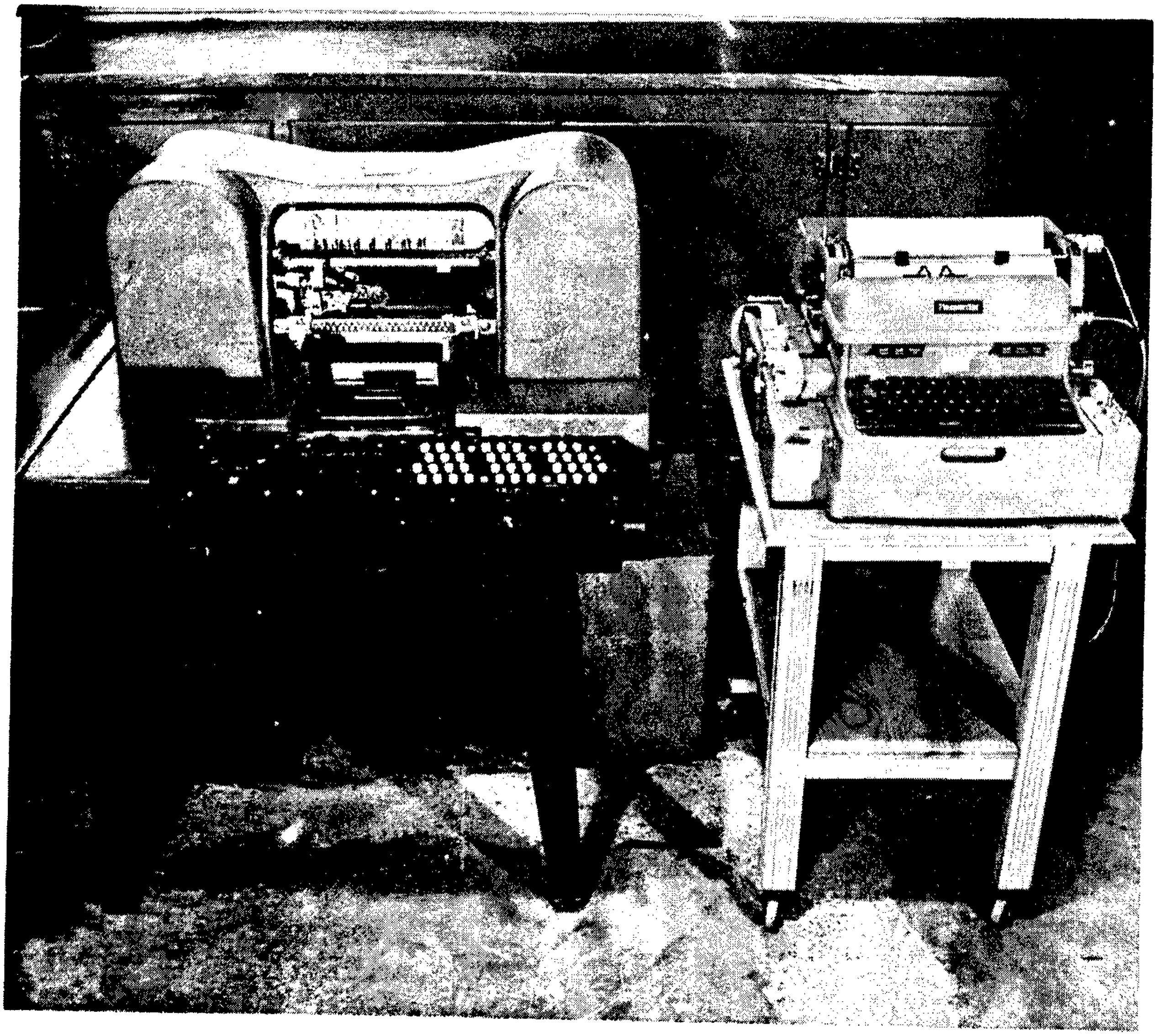


Figure 8

STATISTICS BRANCH PERFORMANCE
ROUTINE PROCESSING OF ASSESSMENT DIVISION DATA

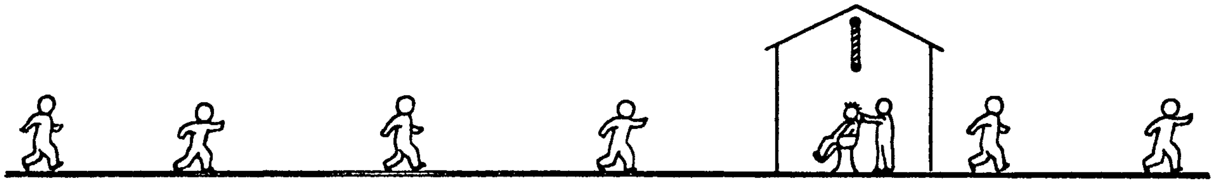
TIME CONSUMED (WORKING DAYS)		1955			
FROM	TO	MAY	JUNE	JULY	AUGUST
1. DATA RECEIVED	DATA PUNCHED	1.7	2.4	2.4	2.7
2. DATA PUNCHED	PUNCHING VERIFIED	0.3	0.1	0.1	0.4
3. PUNCHING VERIFIED	CONTROLS PUNCHED	0.4	0.1	0.1	0.2
4. CONTROLS PUNCHED	% REC. COMPUTED	1.1	1.0	1.0	1.1
5. % REC. COMPUTED	% REC. CHECKED	2.2	1.8	1.8	0.8
6. % REC. CHECKED	DECKS PREPARED	3.3	0.9	0.9	1.7
7. DECKS PREPARED	ANALYSES COMPUTED	1.5	0.8	0.8	1.3
8. NUMBER OF JOBS COMPLETED		21	18	25	26
9. AVERAGE NUMBER OF ANALYSES PER JOB		6.4	6.1	8.3	8.7
10. AVERAGE NUMBER WORKING DAYS REQUIRED		10.8	8.6	15.9	8.3

SEVERAL ARRIVAL — SERVICE TIME COMBINATIONS FOR CALLS DEMANDING A SERVICE

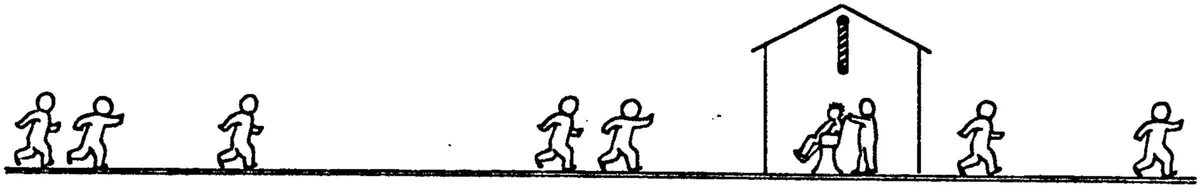
TYPE	ARRIVAL TIME	SERVICE TIME
I	FIXED INTERVAL	CONSTANT
II	"	RANDOM
III	RANDOM INTERVAL	CONSTANT
IV	"	RANDOM
IV-A	POISSON	EXPONENTIAL

FACTORS AFFECTING SERVICE QUALITY

FIXED INTERVAL | TYPE I | CONSTANT SERVICE TIME

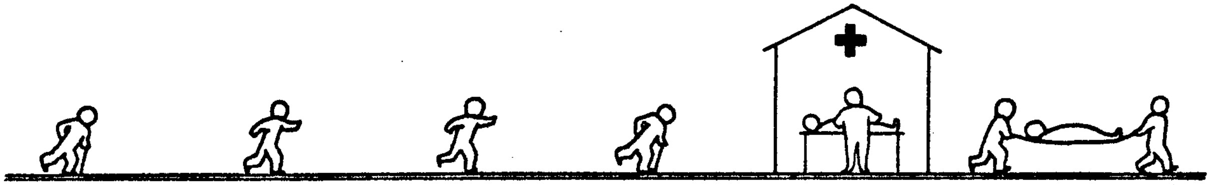


RANDOM INTERVAL | TYPE III | CONSTANT SERVICE TIME

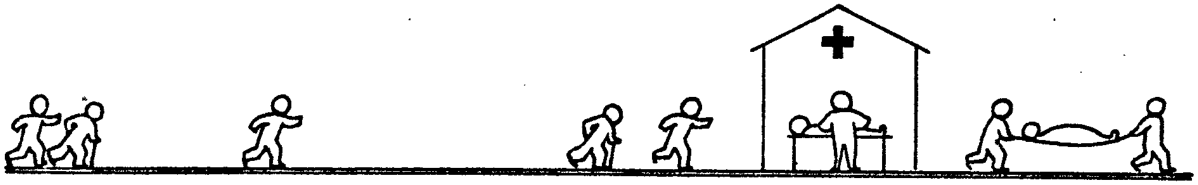


FACTORS AFFECTING SERVICE QUALITY

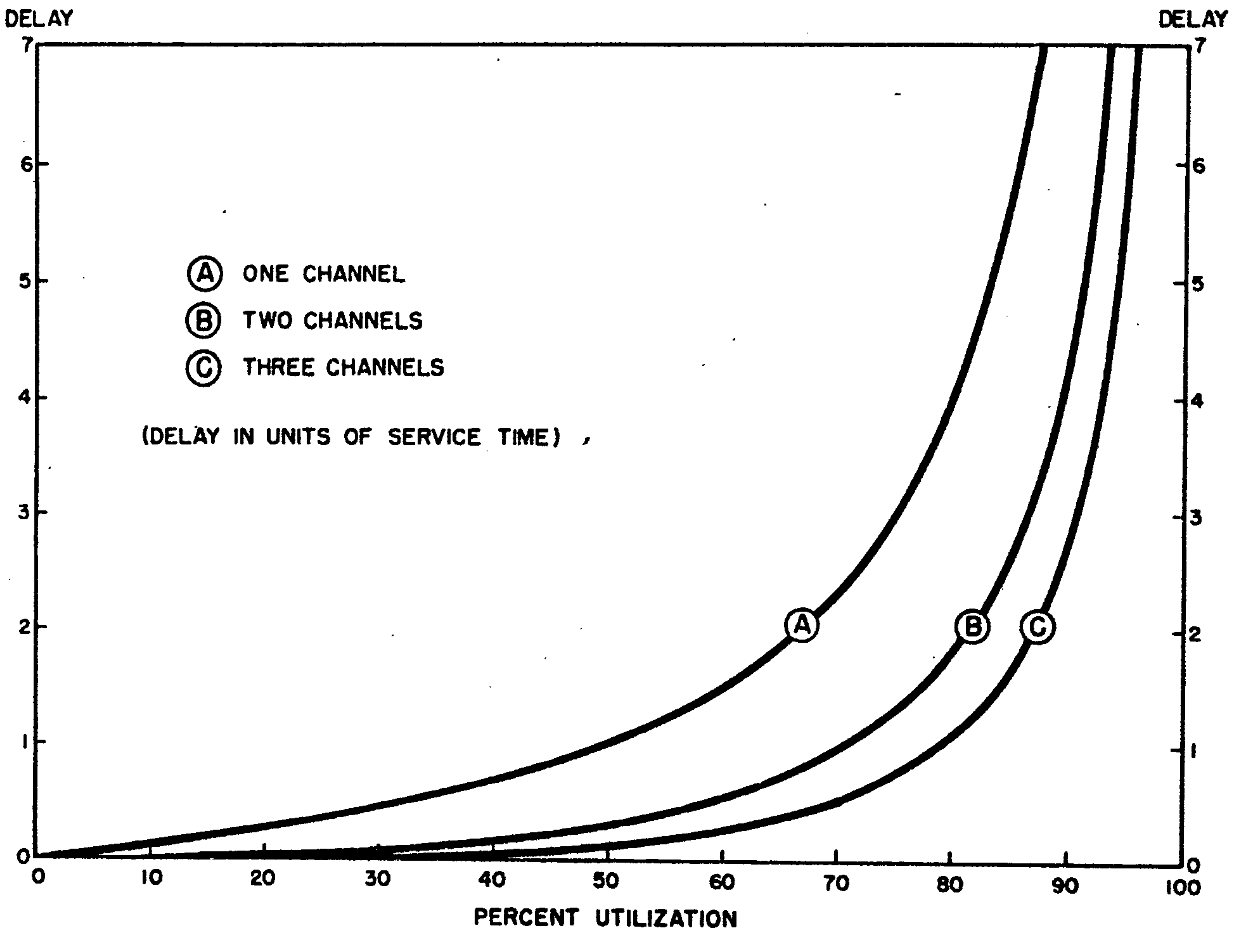
FIXED INTERVAL — TYPE II — RANDOM SERVICE TIME



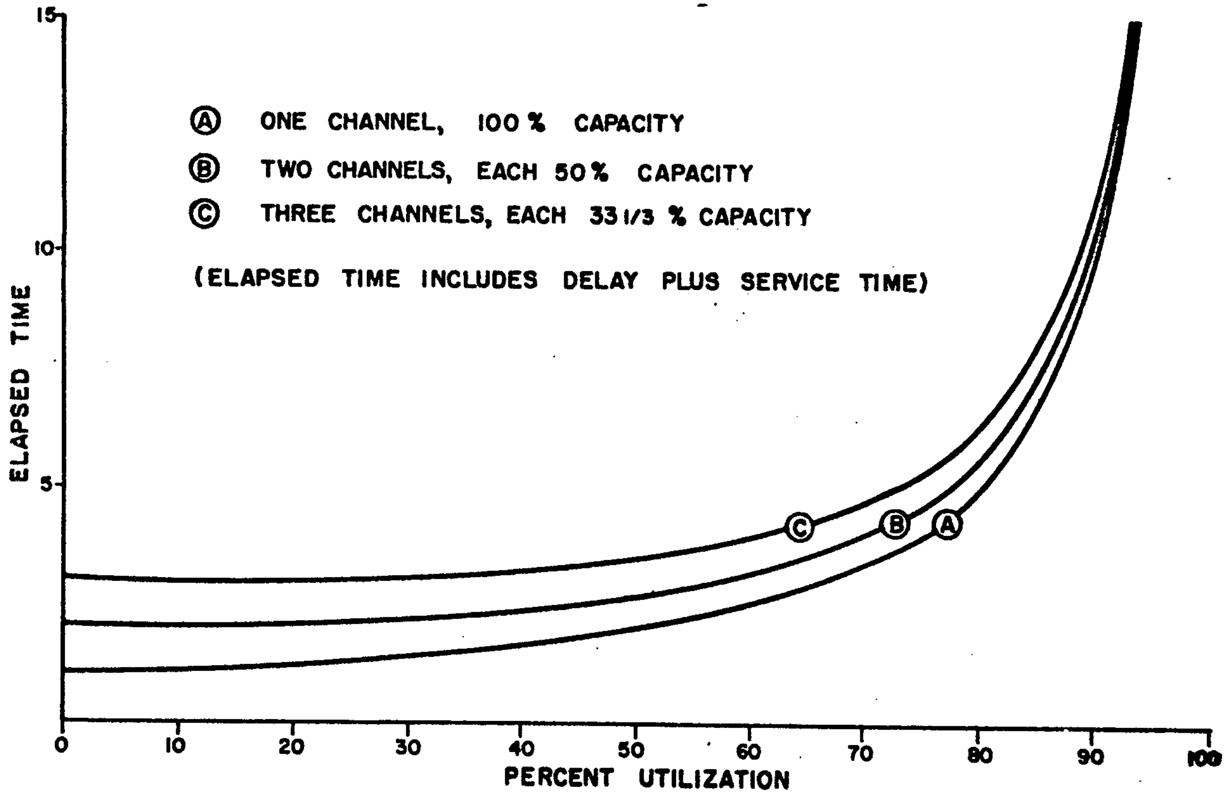
RANDOM INTERVAL — TYPE IV — RANDOM SERVICE TIME



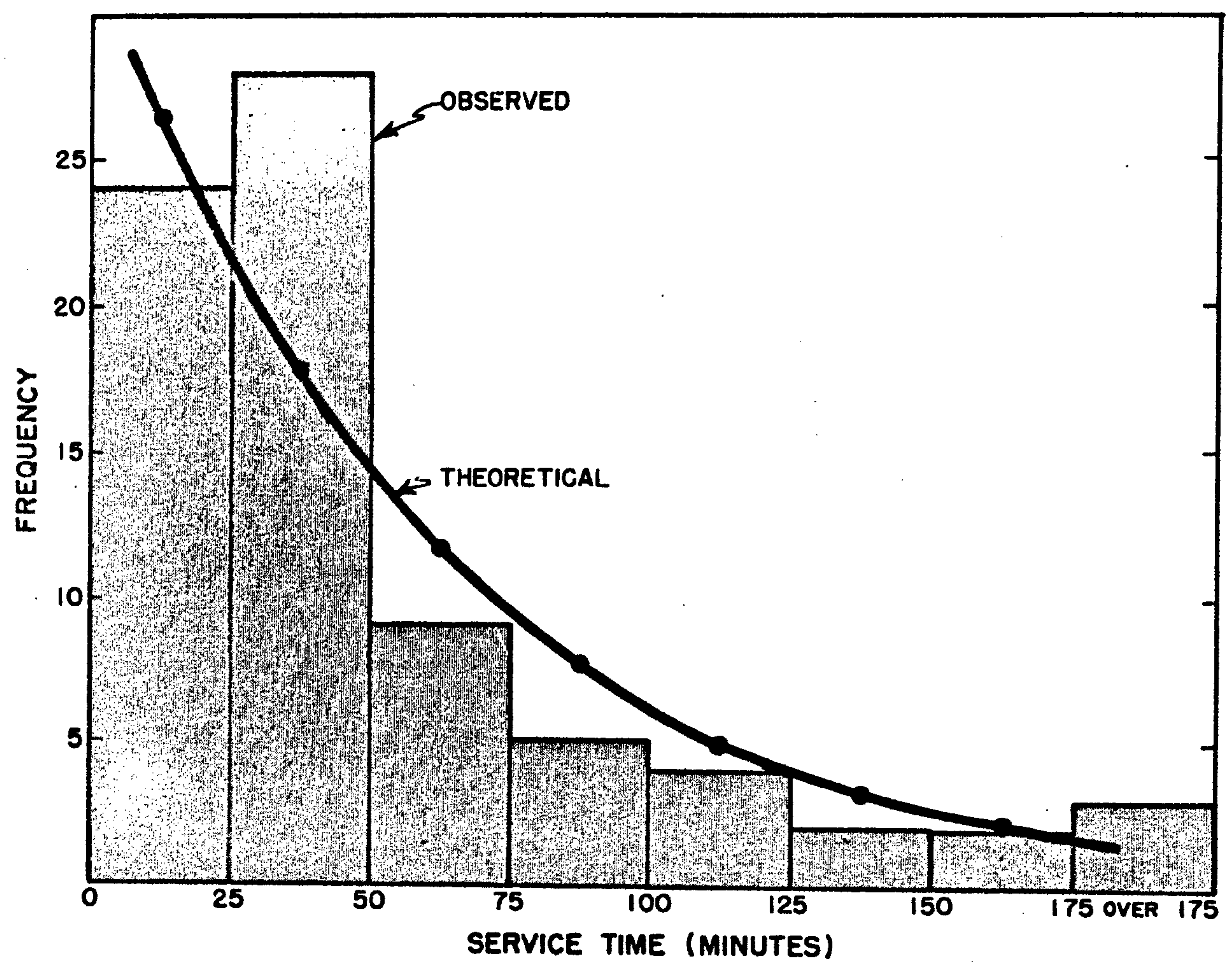
DELAY VERSUS PERCENT UTILIZATION



ELAPSED TIME VERSUS PERCENT UTILIZATION
FOR CHANNELS OF DIFFERENT CAPACITY



DISTRIBUTION OF COMPUTER SERVICE TIMES



SOME EXAMPLES OF THE USE OF HIGH SPEED COMPUTERS IN STATISTICS

J. M. Cameron
National Bureau of Standards

Introduction. Electronic computing machines have been available to statisticians for routine analysis only for two or three years. Already there is developing a body of literature on the efficient utilization of these machines for statistical calculations and a number of problems relating to methods of computing have developed. This paper gives a short summary of the types of problems for which high speed machines have been used and, by means of some examples, shows the type of problems facing a statistician wishing to use these machines.

Types of problems programmed for computers. Considerable publicity has been given to the use of computers in data processing particularly in census, business inventory and accounting problems and other areas in the social sciences. Such applications can be expected to produce prodigious saving in time and corresponding increases in the amount of useful information as output -- information that might otherwise be economically impossible to obtain.

I chose in this talk to consider computers, not from the point of view of a data processing system, but rather as a tool for the statistician in solving his own problems -- as a sort of super desk calculator. The following is a brief summary of the types of problems on which computers have been used or on which their use would be expected to be fruitful:

1) Table making. It is almost economically mandatory to use high speed computers for table making -- examples of such use are legion.

2) Empirical sampling. The determination of properties of statistical distribution and of the power function of statistical tests of hypotheses can in some cases be done only by empirical sampling methods. For example: (a) Teichroew (4) computed (on SWAC) the power curve for the records tests for trend. (b) The operating characteristics curve for mixed variable and attributes acceptance sampling plans can be worked out on high speed computers to whatever accuracy one is willing to pay.for. Previously one had to be content with rather widely separated upper and lower bounds to the OC curve.

3) Simulation of complex phenomena. This approximation of a physical or mathematical process by means of a stochastic model (called the Monte Carlo method) is discussed in (3, 6, 13, 16, 17). The simulation of bombings, of engagements in which objects are fired at moving targets and of complex problems in physics have been made on computers. Computers are especially useful for such simulation because they can generate their own random numbers internally and work fast enough so that sufficient repetitions can be made to achieve the desired accuracy in the final answers..

4) Construction of experiment designs. (a) The Institute for Numerical Analysis did some work on the creation of an orthogonal set of 10 x 10 Latin squares. (b) R. S. Gardner at the Naval Ordnance Test Station,

Inyokern has developed a program for the construction of designs having certain prescribed contrasts estimable. Other results are reported in (5).

5) Analysis of data

a) Analysis of variance. Nearly all installations have developed general programs to handle the analysis of a wide variety of standard designs. It requires too much coding time to prepare, for all types of designs, a separate program tailormade for each particular design. The calculating time on the machine is small in comparison to data preparation time, so that the saving of a few seconds by having specialized programs is meaningless. The fact that general purpose programs are in use in almost every major installation is a clear indication of utility of high speed computers for this work.

b) Least squares analysis. The fitting of polynomials, multiple regression analysis, and similar situations involving linear estimators can be handled by the usual matrix methods or by a general linear program (2).

c) Non-linear systems. A numberof installations have been working on programs for estimating the constants of functions such as $y = a_0 + a_1e^{a_2x} + a_3e^{a_4x}$. John O. Tilly of NOTS Inyokern reported on a general method for curve fitting any arbitrary function (ACM meeting, Philadelphia, September 1955). These programs involve some differential correction and the accuracy and speed of convergence depend on the correctness of initial values used to approximate the unknown parameters.

d) Miscellaneous. Certain order statistic methods, ranking methods and other non-parametric methods involving enumeration or ordering have not received much attention.

Examples of problems in computing.

1) Random numbers. First a word about random number generation which the machine must do internally for efficient use of Monte Carlo or empirical sampling methods. A convenient method of generation is to create the sequence $r_n = r_0 \, r_{n-1} \bmod a^k$ where $r_0 = 5^{17}$ for binary machines (or $r_0 = 7^{13}$ for decimal machines) where $a = 2$ (or 10) and k is the number of digits the machine uses. Properties of such sequences are discussed in (8, 9, 12). Both of these types of sequences have been tested for randomness and no significant deviation from randomness has yet been reported. Other methods of generation involving only addition have been tried but none have as yet proven satisfactory (15).

2) Fitting polynomials. It is not always sufficient to program for electronic computers the same methods that have proven best for

desk calculators. A serious study has to be made of the effects of rounding error. Consider the least squares fitting of a polynomial $y = \sum_i a_i x^i$ by different programs using 8 digits (with floating decimal point). As a specific example I have chosen x = -10(1)10 and 4 sets of coefficients, a_i. Thus, we have 21 paired values (x, y), without random error, and we wish to compare the a_i estimated from this data with the known a_i with which we started. Note that in the usual matrix inversion method we will have an element $\Sigma\, x^{10} > 10^{10}$ so there will be inevitably a loss of at least two significant digits in one element of the matrix. (For a set of test matrices for use in checking the accuracy of proposed matrix calculation see (11).)

An alternative method using an ortho-normalization procedure is decribed in (2). The following Table gives a comparison of matrix method and the ortho-normalization method.

Although these examples do not permit any conclusive statement to be made, they illustrate the need for some careful study of the effect of round-off errors on the accuracy of results.

TABLES

Error in estimates of the coefficients of a polynominal as determined by method of matrix inversion compared with estimates determined by ortho-normalization process.

Error in Coefficients

(parts in 10^8)

Coef. to be estimated	Matrix	Ortho-Normal
a_0 = 1.0	1	0
a_1 = .1	280	1
a_2 = .01	163	1
a_3 = .001	114	1
a_4 = .0001	17	0
a_5 = .00001	921	4
avg.	249	1

Error in Coefficients

(parts in 10^8)

Coef. to be estimated	Matrix	Ortho-Normal
a_0 = 1.0	36164	50
a_1 = 1.0	21168	4317
a_2 = 1.0	332	48
a_3 = 1.0	7362	18
a_4 = 1.0	36	0
a_5 = 1.0	5	2
avg.	10844	739

Error in Coefficients

(parts in 10^8)

Coef. to be estimated	Matrix	Ortho-Normal
a_0 = 100.	0	0
a_1 = 1.	213	3
a_2 = .01	3226	375
a_3 = .0001	8610	14051
a_4 = .000001	29510	2542
a_5 = .00000001	7035516	105903
avg.	1179489	20479

Error in Coefficients

(parts in 10^8)

Coef. to be estimated	Matrix	Ortho-Normal
a_0 = .00001	630 2412	4453 9237
a_1 = .0001	8275 5841	9112 3584
a_2 = .001	2 6710	6 0620
a_3 = .01	302	5009
a_4 = .1	0	96
a_5 = 1.0	5	5
avg.	1484 7545	2262 1425

3) Analysis of variance. H. O. Hartley (7) had developed a general program for the analysis of variance applicable to all standard designs. His method calls for three "operators" and what he calls "rearrangement." His method also has the desirable feature that it computes the individual deviations of the observations from their expected values. This will be of special value as an aid in the interpretation of complex experiments where the analysis of variance table is not very enlightening in arriving at an understanding of the data.

For a standard factorial experiment an alternative method is available. The techniques of Yates (24) by which all the individual degrees of freedom are computed is readily programmed. For example for a 2^n factorial the method calls for forming at each of n steps 2^{n-1} sums by pairs followed by 2^{n-1} differences between elements of consecutive pairs, the data having been entered in a standard sequence. The squares of the elements of the final column obtained, appropriately divided, give the individual degrees of freedom in the analysis of variance. This method appears to be optimal for electronic computers for this particular example. The method is easily generalized for factors at more than 2 levels so that the necessary divisors for, and regrouping of, individual d.f. are done by machine. However, the comparison with Hartley's method this alternative has little to recommend it for general problems in this class, but it does serve to point up the variety of approaches that one has to choose from.

Remarks. The availability of high speed computers will make it possible for the statistician to provide difficult analyses heretofore not attempted because of the time and cost involved. One would look for an increase in the effectiveness of statisticians thus relieved of computational burden.

REFERENCES

(1) Brown, J. A..C., Houthakkar, H. S. and Prais, S. J., "Electronic computation in statistics", *Journal of the American Statistical Association*, **48**, (1953), 414-429.

(2) Davis, P. and Rabinowitz, P., "A multiple purpose orthonormalizing code and its uses", *Journal of the Association for Computing Machinery*, **1**, (Oct. 54), 183-191.

(3) Davis, P. and Rabinowitz, P., "Some Monte Carlo Experiments in computing multiple integrals", *Mathematical Tables and Other Aids to Computation*, **10**, No. 53 (Jan. 56), 1-8.

(4) Foster, F. G. and Teichroew, D., "A sampling experiment on the powers of the records tests for a trend in a time series," *Journal of the Royal Statistical Society*, Series B, **17**, (1955), 115-121.

(5) Hall, Marshall and Swift, J. D., "Determination of Steiner Triple Systems of Order 15," *Mathematical Tables and Other Aids to Computation*, **9**, No. 52 (Oct. 55), 146-152.

(6) Hammersley, J. M. and Morton, K. W., "Poor man's Monte Carlo", *Journal of the Royal Statistical Society*, Series B, **16**, (1954), 23-38.

(7) Hartley, H. O., "Programming analysis of variance for general purpose computers," to appear in *Biometrics*.

(8) Johnson, D. L., "Generating and testing pseudo-random numbers on the IBM Type 701," *Mathematical Tables and Other Aids to Computation*, **10**, No. 53 (Jan 1956), 8-13.

(9) Juncosa, M. L., "Random number generation on the BRL high speed computing machines," Report 855, Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland, 1953.

(10) Lipton, S., "A note on the electronic computer at Rothamsted," *Mathematical Tables and Other Aids to Computation*, **9**, No. 50, (Apr. 55), 69-70.

(11) Lotkin, M., "A set of test matrices," *Mathematical Tables and Other Aids to Computation*, **9**, No. 52 (Oct. 55), 153-161.

(12) Moshman, J., "The generation of pseudo-random numbers on a decimal calculator," *Journal of the Association for Computing Machinery*, **1**, (Apr. 54), 88-91.

(13) National Bureau of Standards, "Monte Carlo Method," *Applied Mathematics Series*, **12**, Government Printing Office, 1951.

(14) Powell, J. G., "The analysis of a factorial experiment (with confounding) on an electronic calculator," *Journal of the Royal Statistical Society*, Series B, **16**, (1954), 242-246.

(15) Taussky, O. and Todd, J., "Generation and testing of pseudo-random numbers" to appear in Proc. Gainesville Symp. on Monte Carlo.

(16) Teichrcew, D., "Numerical analysis research unpublished Statistical tables," *Journal of the American Statistical Association*, 50, (Jan. 55), 550-556.

(17) Teichroew, D., "Distribution sampling with high speed computers," Ph.D. Thesis, University of North Carolina, 1953.

(18) Tocher, K. D., "The design and analysis of block experiments," *Journal of the Royal Statistical Society*, Series B, 14, (1952), 45-100.

(19) Tocher, K. D., "The application of automatic computers to sampling experiments," *Journal of the Royal Statistical Society*, Series B, 16, (1954), 39-60.

(20) Tocher, K. D., "The application of automatic computing machines to statistics," Symposium on Automatic Digital Computation, Her Majesty's Stationery Office, (1954), 166-178.

(21) Votaw, D. F. and Rafferty, J. A., "High speed sampling," *Mathematical Tables and Other Aids to Computation*, 5, No. 33, (Jan. 51), 1-8.

(22) Weil, H., "Reduction of runs in multiparameter computations," *Journal of the Association for Computing Machinery*, 2, (Apr. 55) 99-110.

(23) Woolf, B., "Calculatior and interpretation of multiple regressions," *Journal of the Royal Statistical Society*, Series B, 13, (1951), 100-119.

(24) Yates, F., "Design and analysis of factorial experiments," Imperial Bureau of Soil Science, Tech. Comm., No. 35, Harpenden, 1937.

COMPARATIVE CHARACTERISTICS OF MEDIUM-PRICED FULLY AUTOMATIC COMPUTERS FOR STATISTICAL APPLICATIONS*

M. E. Stevens and S. N. Alexander
National Bureau of Standards

1. INTRODUCTION. Continued progress in scientific research, development, and testing is increasingly dependent upon the capacity to analyze large volumes of experimental data accurately and rapidly. New concepts in the design of experiments and new and improved methods for the analysis of data are already providing greater accuracy and reliability of results. In addition, new computing tools are now available that offer the advantages of truly high-speed processing. These new tools are the automatic digital computers, operating internally at electronic speeds, that have beem designed and built in the past ten years.

The idea of automatic computers for data processing is, of course, not new. What is new is a technology that makes high-speed computing devices operationally effective. Three principal ingredients are blended in this new technology. The first is the telegraphic communication of information, making possible the transfer of information from one place to another by the transmission of electrical energy. The second is the ability to transfer this information from a sequence of electrical signals into a physical storage medium by such means as magnetic recording or punching holes in paper tape, and then to regenerate selectively the electrical signals whenever the information is needed. The third is the ability to process the information in accordance with rules of arithmetic and elementary logic. These ingredients were first successfully embodied in electromechanical devices just before World War II, in relay devices during the war, and in electronic devices as the war came to an end. It is, in fact, particularly appropriate to discuss the characteristics of some modern computers at a conference sponsored by the Office of Ordnance Research, because the Ordnance Department was responsible for the very first electronic computer, ENIAC, which was designed and built at the University of Pennsylvania under Army contract, and completed in 1946.

2. GENERAL CHARACTERISTICS OF DIGITAL COMPUTING SYSTEMS. In the digital computer of today, the three technological ingredients are combined in a machine system that is able to communicate, store, and process information automatically, reliably, and at fantastic speeds. The system is a general-purpose tool in that the information that is received and processed may be census statistics, pay roll records, test results, data for mathematical and engineering calculations, reports of issues and receipts in accounting systems, or any of a very wide variety of other types of data. As a system, it has input-output devices that enable communication to and from the outside world; internal storage (or "memory") where original data, operational instructions and intermediate results are stored; an arithematic-logical unit where operations such as addition, multiplication, and logical comparison are performed on designated data, and a control unit where a pre-planned sequence of operations is decoded and executed in proper order.

This logical organization gives the modern computer flexibility and auto-

* This work has been supported in part by the U. S. Army Chemical Corps, Fort Detrick, Frederick, Maryland.

maticity of operation not found in earlier types of computing tools. For example, in a desk calculator, control of the sequence and kind of operations desired is exercised by the human operator who also enters the data to be operated upon, one at a time, by depressing the proper keys; and who reads the intermediate and final results from the values standing in the output register of the machine. For the automatic computing system, a problem planner lays out the necessary sequence of operations to be performed on designated data, and the sequence is translated into a machine language consisting of a series of machine instructions which the control unit can decode and execute. Both the data and the control information are entered via an input device into internal storage. Then, under the supervision of the control unit, the proper data are routed at the proper times to the arithmetic-logical unit where they are operated upon in accordance with the pre-planned sequence of instructions, and, when all designated operations have been performed, the final results are made available via an output unit.

During this processing, the computer operates automatically in the sense that it carries out a long and complicated series of varied operations on various data without the need for human intervention once the data and instructions have been read in. It is able to make very elementary "Yes-No" decisions and to follow different courses of action in accordance with these decisions. Thus it is able to select alternate courses of next action in accordance with results obtained and to carry out the same operations repetitively for a specified number of variables, proceeding to the next step when and only when the iterative cycle has in fact been completed.

The automatic computer carries out arithmetic and logical operations with high speed and high reliability. Typically, a digital computing system adds or compares two 10-digit numbers at rates ranging from 500 to 15,000 operations a second. Multiplication and division operations are usually about 10 times slower than addition and subtraction, but are still performed at very high speed. A machine that is capable of 4,000 additions or 400 multiplications each second is able to turn out completed computations in about 15 minutes that would take a man with a desk calculator one whole month to carry out, working 8 hours a day, 5 days a week. A variety of checks can be provided to assure the accuracy and reliability of results. It is not uncommon for automatic computing systems to perform between 10,000,000 and 100,000,000 arithmetic operations without a single error being detected.

The information that is received, stored and processed by the computing system is expressed internally as a machine code or "language" that is comprised of patterns of electrical signals. The system is able to convert information to and from the hole no-hole code patterns on punched cards or punched paper tape, or the impulses received from modified typewriter keyboards, or records made on magnetized wire or tape. Thus it can accept data from a variety of sources, recorded on a variety of media. The input and output devices of the system carry out this conversion between the language of the external world and the internal language, and in addition effect a transformation in the time scale, so that information entering the system at rates compatible with the external world is delivered to the central processing unit at the tremendously increased rates necessary for proper utiliation of the high speed of internal operation.

The internal storage devices used in automatic computing systems typically provide capacity for between 10,000 and 100,000 digits of information at one time, which can consist of any desired combination of coded machine instructions, data to be processed, constants to be used in the computations, and storage space for intermediate results. The system provides for automatic, high-speed selection and retrieval of the information so stored.

These general characteristics of automatic self-sequencing in operation, decision-making ability, high speed and reliability of operation, ability to communicate via electrical signals to and from a variety of media and over distances, large-capacity storage, and high-speed retrieval of stored information together give this important new tool adaptability to a wide variety of computing and data processing applications. For this reason, both the rate at which computers have already been applied and the rate at which technological improvements for even more versatile systems have been developed are phenomenal.

Today, over 250 fully automatic digital computing systems that are commercially available are in operation in the United States, and 1,000 or more additional systems are on order from the score of manufacturers who offer production models. These installations are of equipment that ranges in price from $50,000 to more than $2,000,000, with differences in versatility, speed, and capacity directly related to the differences in purchase cost.

3. CHARACTERISTICS OF MEDIUM-PRICED AUTOMATIC COMPUTERS. The first of the fully automatic digital computers were large-scale installations whose counterparts today cost a million dollars or more to purchase. Subsequently, design efforts for the development of less expensive systems were directed to computers that differ from the large-scale computers primarily in slower speed of operations, e.g., several hundred operations per second versus several thousand, and in slower and less versatile imput-output devices, e.g., having effective data read-write rates of 500 decimal digits per second as compared with 10,000-15,000 digits per second that can be achieved with available magnetic tape input-output devices. The first of these less expensive computing systems to become avialable used magnetic drums for internal storage and were limited to keyboard and punched paper tape devices for imput and output, but cost less than $100,000 to purchase. They were designed primarily for those scientific applications where there is a small volume of data to be entered, a large number of calculatidns to be performed on these data, and a small output of final results.

More recent developments in magnetic drum computers have led to a variety of systems in the price range of $100,000 to $300,000 that are increasingly versatile and provide for multiple input-output units capable of receivir information from both punched cards and magnetic tape. Thus they are more flexible and powerful tools for statistical applications where considerable input of data is required.

The operating speed of an automatic computer is usually governed by the access time to the instructions and operands stored internally. The internal storage devices used in large, medium, and small systems vary considerably in the access times, and this factor is closely related to the cost of the storage device. The drum devices are used for internal storage provide large capacity (10,000 to 500,000 digits at any one time) with reasonably fast access (5 to 25 milliseconds for operands up to 10 decimal digits in length) at relatively low cost.

The drum storage device is a rotating cylinder whose surface can be magnetized. Information is recorded by means of electromagnetic heads in parallel channels or tracks, with each channel carrying the recorded information arranged in a character-by-character string. As the drum rotates at high speed (typically, between 3,600 and 12,500 rpm) it passes one or more of the read-write heads, usually one for each track. Access is direct to the proper track, but sequential in terms of the information arranged on the track, so that the device is a form of circulating storage. This means that, on an average, one-half a drum revolution will be completed before a specific single operand can be located and read. However, in the more powerful of the drum computers, a few tracks are provided with more than one set of heads, thereby reducing the average access time of about 10 milliseconds to between one-fourth and one-tenth of that time. Information such as a series of instructions and operands can be transferred in blocks between the channels with one head (usually termed "main memore") and those with several heads ("fast access loops") as it is needed in the course of the machine program. The access time for instructions and operands when actively in use is thus brought into closer balance with the speed of the arithmetic unit.

Technical features that may differ in different computers include such details as the word length, which is the fixed or variable number of characters in an ordered set that is stored, transmitted or operated upon as a unit within a particular computer, and tha availability of buffers, which are storage devices used to compensate for differences in rate of flow of information, for example, to and from input-output units and the computer. Differences in instruction mode relate to whether the sequence of machine operations is controlled by an explicit designation of the source of the next instruction in each instruction or whether the source of the next instruction is determined by the settings of a control counter, as well as to the number of addresses used in any one instruction. Differences in ease of programming may relate to the availability of such devices as the B-register which can be used to modify systematically the storage addresses to which the instructions refer or to terminate an iterative seqence of operations upon completion of a designated number of repetitions.

The fully automatic computers that range in price from $100,000 to $300,000 typically have magnetic drums for internal storage, provide punched card input-output devices as well as other means for input or output, and use a binary-coded decimal language. Usually, magnetic tape units for auxiliary storage or input-output are available at extra cost. Computing systems in this class are available on either a rental, lease with option to buy, or outright purchase basis. Typical rental rates are from $3,750 to $4,500 per 8-hour shift per month, with maintenance usually provided by the supplier. The commercially available computers in this class that are already in productive operation may be briefly described as follows:

<u>Datatron</u>. This machine is produced by the ElectroData Corporation,

Pasadena, Califronia. Over 25 of these computing systems are currently in operation. The system uses a fixed word length of 10 decimal digits plus sign, or 5 alpha-numeric characters. The instruction mode is implicit, 1-address. The drum capacity is 40,000 decimal digits with 800 additional digits of fast-access storage. Average access time to fast-access loops is 0.85 millisecond, giving average operating rates of 1.7 to 2 milliseconds for addition. Input devices include a punched paper reader operating at up to 540 characters a second and punched card readers at 100, 200, or 240 cards perminite, or magnetic tape at 5,000 digits per second. Output is available via punched paper tape (60 digits per second), punched cards (100 cards per minute), line printer (150 lines per minute), and magnetic tape (5,000 digits per second). Up to 10 magnetic tape units may be used with the system. Buffered punched-card input is available with any combination of 7 inputs and outputs capable of being fed simultaneously. Special features include aids to programming such as B-registers and floating-point as well as fixed-point arithmetic.

Elecom 120, 120-A, and 125. These computing systems have been designed and produced by the Electronic Computer Division of the Underwood Corporation. At least five Elecom 120 computers are in operation. The work length is fixed and provides for 8 decimal digits plus sign. The Elecom 120-A is a later version with word length of 10 decimal digits or 5 alpha-numeric characters plus sign. The 120-A computer uses a 2-address, automatically sequenced next address (implicit) instruction mode. Magnetic drum internal storage is available in increments of 10,000-digit capacity up to a maximum of 100,000 digits. Fast-access storage available ranges from 100- to 1,000-digit capacity. Using the fast-access loops, an average addition time for two operands of 3.5 milliseconds is achieved. Input-output is available via punched paper tape (400 characters per second in, 60 characters per second out) and magnetic tape (400 digits per second in and out). Punched card tie-in equipment can also be provided. Optional features include B-register and floating-point arithmetic. The Elecom 125 computer, also produced by Underwood, with 40,000-digit memory and providing additional instructions in the reper-toire for input and output editing, is replacing the Elecom 120-A. The 125 has magnetic tape units with input-output speeds of 2,000 digits per second. Fast access memory is available with capacity of either 500 or 1,000 digits. The 125 computer was designed for use with a special device, the Elecom 125 File Processor, that carries out independently operations such as automatic sorting and merging for file maintenance activities at the rate of 6,000 digits on magnetic tape with buffered input and output. The complete 125 system is currently in operation at the manufacturer's plant.

IBM-650. This drum computer of the IBM computer series is in extensive use, with over 300 installations currently in operation. It is available with either 10,000 or 20,000-digit internal storage capacity, and a fixed word length of 10 biquinary-coded decimals is used. The 650 is explicit, 2-address in instruction mode, so that the programmer may so place instructions in memory that drum revolution time is minimized. Operating times for addition maybe as fast as 0.77 millisecond using such optimum programming ranging to 5.57 millisecomds for sequential programming. Primary input-output at present is by punched card tie-in, with rates of 200 cards per minute in, 100 cards per minute out, or tabulator output of 150 lines per minute. Howeve magnetic tape units can be ordered that will be compatible with tapes for IBM 700-series computers. Transfer of information between tapes and drum storage will be through magnetic core storage of 600-digit capacity which can also be used for fast-access storage except when the tape units are reading or writing. Up to the present time, the 650 has been available only on a rental basis, but will be made available on a purchase basis on or before January 24, 1957.

Miniac. The Miniac drum computer is produced by Marchant Research, Inc., of Oakland, California. Three or more are now in operation. The fixed word length provides 10 binary-coded decimal digits and the instruction mode is implicit, 1-address with provision for B-register as an option. Internal storage consists of 38,400-digit capacity for main memory, with 2,560-digit additional fast-access capacity. Average access time to information in fast-access storage is 1.25 milliseconds, yielding operating times for addition that range from 1.8 to 6.2 milliseconds. Input-output is my means of either punched paper tape or special magnetic tape package devices, with read-write rates of about 5,000 digits per second.

NCR-CRC-102D. This drum computer is a binary-coded decimal modification of an earlier model, the CRC-102A computer, of which about 23 machines were produced by the Computer Research Corporation, which has since become a part of the National Cash Register Corporation. Several of the 102D computers are already in operation. The fixed word·length will accommodate 10 decimal digits or 6 alpha-numeric characters without sign, or 9 decimal digits with sign. Instruction mode is implicit, 3-address. Drum capacity is 10,240 decimal digits without sign plus 80 additional digits of fast-access storage. Average operation time for 3-address addition is 9.8 milliseconds. The principal means of input-output are punched paper tape and magnetic tape, with reading rates of 200 digits per second and 600 digits per second, respectively. Punched cards may be used at a rate of 100 cards per minute. Special features include provision for computer-controlled tape search that can proceed concurrently with other operations.

Readix. This is a drum computer produced by the J. B. Rea Company of Santa Monica, California. One installation is in operation, with several others in order. The fixed word length provides 10 decimal digits plus sign, and the instruction mode is 1-address, implicit. Drum capacity is 40,000 digits plus 1,600 digits of fast-assess storage. Operating times for addition range between 0.84 and 9.44 milliseconds when instructions and operands are stored in fast access. Both floating point arithmetic and B-registers are included. Input-output devices include punched paper tape, punched cards, and magnetic tape with read-write rates of 1,000 digits per second. Independent tape search can be carried out concurrently with other operations.

In addition to these medium-priced computers already at work on varied applications, several computing systems of generally similar performance characteristics and cost are either under development or can be provided by the manufacturer on a custom basis, tailored to the purchaser's problem requirements. These include the following systems: Monrobot-VI and Monrobot-MU, Monroe Corporation; UDEC, Electronic Instruments Division, Burroughs Corporation; and the UNIVAC File Computer, Remington Rand UNIVAC Division of Sperry-Rand Corporation.

The Monrobot-VI is a small machine used for scientific calculations, while the Monrobot-MU is based on a multiple-unit concept such that various combinations of input-output units and of magnetic drums of various capacities up to 500,000 digits can be linked to basic calculating and control units. Operating times, however, may range from a maximum of 135 milliseconds down to about one-sixth of this rate.

The UDEC systems are exemplified in an installation maintained by the manufacturer. This machine has a fixed word length of 9 decimal digits plus sign and uses a 1-address instruction mode. Drum capacity is 5,300 decimal digits with average operating time of 9.12 milliseconds for addition. Input-output for this installation is punched paper tape.

The UNIVAC File Computer, currently under development, is also based on a building block concept that can be tailored to customer needs. It has unusual features of multiple input and output from varied types of devices, combines plugboard programming with stored program operation, and provides a variety of magnetic drums of varying capacities, access times, and record lengths. The hierarchy of internal storage units offered includes input-output (buffer) storage of 120-character capacity (or 110 digits with signs) for each of up to 32 input-output units, intermediate storage of 240-character capacity, high-speed general storage in either 2,280 roll,880-character size with average access time of 2.5 milliseconds, and one or more large capacity storage units with space for 180,000 alpha-numeric characters each. Operating times, using the faster storage, are expected to be between 2 and 8 milliseconds for an addition

Slightly below the price range for the computers mentioned above are digital computing systems using a binary machine language, so that either manual or programmed binary-to-decimal and decimal-to-binary conversion are necessary for the solution of many problems commonly arising in statistical applications. The price range for these binary systems is $30,000 to $80,000 for the basic machine. For this price, input-output is limited, and either no magnetic tape units are provided or those offered have read-write rates of less than 500 characters per second. In this class of binary drum computers are the following, all of which have at least one installation currently in operation: ALWAC III, Logistics Research, Inc., Redondo Beach, California; Bendix-G15, Bendix Computer Division, Bendix Aviation Corporation; Circle, Hogan Laboratories, New York; and LGP-30, Librascope, Inc., Glendale, California.

4. COMPARATIVE EVALUATION. For effective evaluation of different automatic computers the comparative features, including cost factors, of the various systems that are available must be balanced against the actual processing requirements in a particular proposed application. Among the major factors are the relationship of both operating speed and storage capacity to cost, and the flexibility of the system for later expansion and the extension of the use of the system to additional types of problems. Some fo the other characteristics that would be considered include the number and variety of instructions available, extent of compatibility with other equipment in use, the kind and extent of checking features, aids to programming and maintenance, engineering reliability of the equipment and components, and power, space, and air conditioning requirements. Obviously, the relative advantages and disadvantages of any one computing system as against those of any others must

be appraised in the light of detailed performance specifications based upon careful analysis of the typical processing requirements to achieve the best balancing of equipment characteristics for a specific application.

The two proceding papers in this Technical Session have dealt with methods and tools for analysis of the results of designed experiments, with particular reference to the work of the Research and Development Program at Fort Detrick, Maryland. The need for additional statistical processing facilities at Fort Detrick poses a number of interlocking considerations regarding the characteristics of medium-priced fully automatic computers, compatibility with present and project workloads, and performance required for processing of typical problems. This proposed application illustrates the close interrelationships between the problem characteristics and the evaluation factors appropriate for use in establishing performance requirements. For example, the method of Yates for the analysis of variance of 2^n factorial experiments minimizes the number of machine multiplications. The choice of this method therefore significantly reduces computer operating time, since medium-priced digital computers can typically perform 10 or more additions or subtractions in the time required for one multiplication. Again, where the analysis of variance of a 2^{11} factorial design is to be computed in minimum time, an internal storage capacity of at least 24,000 decimal digits is indicated, a word length of 10 decimal digit plus sign is desired, and indication of the occurrence of overflow, with provision for double-precision operations in case of overflow, is required.

The three papers in this Session have thus dealt with the use of punched card techniques for computing results of designed experiments, analytical methods that are well suited to the use of high-speed computers for statistical applications, and the preformance characteristics of the new automatic computing tools now available for this purpose at moderate price. That such new tools can solve the 2^{11} design in 20 minutes or less, in contrast with a comparable number of hours by punched card techniques, is indeed evidence that the automatic computers can make significant contributions to continued progress in the statistical analysis of experimental data.

AN APPLICATION OF THE DESIGN OF EXPERIMENTS TO THE SURVEILLANCE OF AMMUNITION

Jerome R. Johnson
Ballistic Research Laboratories

The Design of Experiments is a subject of considerable interest to the Surveillance Branch of the Ballistic Research Labcratories at Aberdeen Proving Ground, since this organization must frequently design and analyze various investigations and tests of ammunition. The Surveillance Branch is primarily concerned with the inspection and testing of ammunition after it has passed its initial acceptance tests and has been placed in storage in the various Field Service installations. The Surveillance Branch is concerned with a number of different types of programs such as malfunction investigations calibration studies, depot tests and classificaticn investigations. In all of these investigations, consideration must be given to the statistical design of the program in order that valid conclusions can be reached without expenditure of excessive amounts of ammunition and test effort.

The program I am going to discuss in some detail this afternoon is a classification investigation. In classificaticn investigations, samples are selected from a group of lots of a given type of ammunition and these samples are subjected to various tests, usually of a destructive type in which the item is actually functicned either by firing it from a weapon or in a simulated functioning test in the laboratory. On the basis of the results of these tests the ammunition lot is generally assigned one of three grades, Grade I, Grade II or Grade III. Grade I ammunition is ammunition which is considered to be as good as new and is usually suitable for "long term" storage. Grade II ammunition is ammunition which is still considered serviceable but of lower quality than Grade I. This ammunition is given priority of issue. Grade III ammunition is ammunition which is considered unserviceable and must be either renovated or scrapped.

The particular design I am going to describe was used in part of a classification investigation of 60mm mortar ammunition. In classification programs for this item a sample of forty (40) rounds is drawn from each lot to be tested. It is considered desirable to test this item with both its maximum and minimum charge, because some characteristics such as fuze functioning is given its most severe test at the minimum charge while other characteristics such as flight stability is given its most severe test at the maximum charge. Also, by testing the item at two charges, a better picture of the expected performance of the ammunition at all possible charges is obtained than would be possible by testing at a single charge. Therefore, twenty (20) of the forty (40) rounds from each lot are fired with charge the minimum propelling charge, and twenty (20) rounds are fired at charge 4, the maximum propelling charge. All of the rounds are visually inspected prior to firing. When fired, the muzzle velocity and range of the round are measured and the flight and functioning characteristics of the round are observed.

For our purposes here I plan to deal primarily with the testing of the lots for range. From the twenty (20) rounds fired for range with each charge, it is desired to obtain estimates of the average range and the round to round dispersion of the range for each lot tested. In order to gain some insight into the difficulties that may be encountered in obtaining these estimates let us examine the results of some range firings for this item conducted for another purpose that are listed on page i of the table. This data consists of the results of range firings for five lots, each lot being tested on two different days. Assuming these days are random samples from the population of days the assumptions of the model of the so called two-fold hierachal classification[1] seem to be satisified and this model was used in the analysis of this data. The results of this analysis are summarized in the analysis of variance table on page i. On testing the day to day variation in range against the within day variation this day to day variation is found to be highly significant. The expectation of the mean squares of the analysis of variance table are listed. If these expectations are equated to the computed mean squares of this table an estimate of the among day variance, σ_b, may be computed. The estimated value of this variance was 3743 which is larger than the round to round variance within days. Although the amount of data in this analysis is small it does indicate there can be important sources of variation in range due to factors other than the ammunition. Other studies of range firings with 60mm Mortar Ammunition have shown that this day to day variation is highly correlated with the meteorological conditions at the time of firing, particularly with wind and air density. There are also many other factors that effect the range of this item. Efforts are made to control as many of the sources of variation as possible. The rounds are temperature conditioned prior to firing and the rounds are fired from a well emplaced mortar, the base plate of the mortar frequently being mounted in concrete. However, variation in meteorological conditions and other sources of variation can not be controlled and consideration must be given to this uncontrollable variation when selecting the experimental design to be used in the test.

One procedure that has been used in an effort to minimize the effect of this day to day variation is to fire a sample of rounds from a single lot together with a sample of rounds from a reference lot. The reference lot is a lot from which samples have been fired on a number of different days and from a number of different weapons. By averaging the results of all these firings the average range of the reference lot under more or less average firing conditions is found. Therefore, by correcting the average range obtained on a given occasion for the difference of the average range of the reference rounds fired on that occasion and the long term average of the reference lot, it is possible to make a correction for day to day variation. However, for the particular surveillance test under consideration this method appears to be unsuitable since this test is to include samples from twenty-four test lots of ammunition and the test would consequently require the expenditure of a very large number of reference rounds. When it is attempted to reduce the number of reference rounds used by firing the samples from more than one test lot with one series of reference rounds, the procedure soon becomes unsatisfactory as the number of test lots is increased because the conditions are not the same at the time the samples from the test lots are fired as the conditions at the time the reference lot is fired. Meteorological conditions are particularly subject to change and can undergo considerable change in a single hour. Assuming that meteorological conditions can be considered to be relatively constant for periods of only 30 to 40 minutes and with a rate of fire of about one round per minute, the number of rounds that can be tested under relatively constant conditions is thus limited to from thirty to forty.

[1] Kempthorne, O. The Design and Analysis of Experiments John Wiley & Sons, N.Y. 1952

If the average level was the only characteristic of range required, the use of a randomized block design would provide a simple and efficient solution to the problem. One round from each lot could fired in each block of the design which would give twenty (20) blocks of twenty-four (24) rounds each. Since these blocks of twenty-four (24) rounds should require less than thirty minutes to fire, test conditions within a block should be fairly homogeneous. The order in which lots are fired would of course be randomized for each block. A reference wound could also be fired in each block so that the final results could be reduced to a more or less absolute level.

The use of this randomized block design, although it would be excellent from the standpoint of obtaining estimates of the average ranges of the lots, would be unsuitable for our purposes since it would not provide estimates of the round to round dispersions of range for the lots. Thus in order to obtain estimates of both the average range and the round to round dispersion of range for each lot it will be necessary to fire more than one round from each lot in the block. Since blocks containing 50 rounds would be produced if only two rounds from each lot were fired in each block, the use of a randomized block design does not seem promising since this block size would probably be excessively large. Also this design would provide only 10 degrees of freedom for estimating the round to round variance of range for each lot.

Since it appears that samples from all the lots can not be fired in a single block the use of an incomplete block design seems necessary. Again since it is desired to reduce the results of the test to a more or less absolute basis, it is desirable to fire a sample from a reference lot along with the lots to be tested. Having twenty-five lots there are several incomplete block designs available. However, the one that appears to be best suited for our purposes is the Repeated 5 x 5 Simple Lattice Design. The arrangement of the lots in this design is illustrated on page ii of the handout. The numbers from 1 to 25 are used to identify the twenty-five lots.

Two replications of the design are shown. The rows of the two replications are the blocks of the design. Thus block (c) of Replication I would contain samples from lots 11, 12, 13, 14, and 15. It will be noted that the blocks of Replication II contain the lots that are together in the columns of Replication I. In the design used for our test of mortar ammunition the simple lattice was repeated so that there is a total of four replications. Replication III will contain the same grouping of lots together in blocks as Replication I and Replication IV will contain the same grouping of lots as Replication II. Five rounds from a lot will be fired as a group in each block that contains the lot and thus the twenty rounds from a lot will be fired in the four replications of the design.

For the actual firing sequence of the test the twenty-five (25) lots are assigned at random to the numbers from 1 to 25 thus indicating which lots are to be fired in each block. The order of firing lots in a block should be at random and the order of testing the different blocks of each replication should be randomized also. This randomization is necessary in order that no lot shall be favored and thus an unbiased estimate of error may be obtained. The randomization can be carried out by use of a table of random numbers or by other means. The firing order to followed in the test should be explicitly written out before the firing is started. Instructions should be given that all rounds required for a block should be fired without interruption of the firing program, although different blocks may be fired at different times. The firings of each replication should be grouped as closely together in time as possible. Any major change in the test procedure should be made between replications if possible.

On page iii of the tables the results of the firings are tabulated. The sequence of lots within blocks and blocks within replications have been ordered to facilitate the computations. In each cell the average range for the five rounds fired from the lot is recorded and it is with these average ranges that the analysis of the design will be carried out. The number in parentheses is the lot identification and this is followed by the average range for that lot less 1600 yards. Each row corresponds to a block in the design and the sum of the five (5) lot averages for a block is recorded at the end of the row under block total. At the bottom of the page are recorded the lot totals and the adjusted lot totals. The lot totals are merely the sum, over the four replications, of the average ranges for each lot. For example, the lot total for lot (6) is the sum of 271.2 from the first replication, 208.0 from the second, 321.0 from the third and 279.2 from the fourth. The adjusted lot totals listed below the lot totals are obtained from the lot totals by using the correction factors μC that are listed around the edge of the table of lot totals. Thus the adjusted lot total for lot (6) is obtained from the lot total for (6) by adding to this total the μC correction factors of the row and column in which lot (6) is located. Thus the value of 1111.0 contained in the table of adjusted lot totals is obtained by adding -5.47 and +37.08 to the lot total for (6) of 1079.4.

The fact that the lots, must be adjusted for block differences is one of the features of the incomplete block designs that distinguishes them from the complete block design such as a randomized block and the Latin Square designs. This adjustment is of course necessary since all of the lots are not fired in the same block. Actually the computations of these adjustment for the simple lattice is not difficult and this is one of the reasons why it is one of the most attractive of the incomplete block designs. The computations of these correction factors are given on the next two pages of the tables and a detailed discussion of the method of carrying out the computations for any lattice design is given in Experimental Designs by Cochran & Cox. In order to compute these adjustments it is necessary to compute the sums of squares for the anlaysis of variance on page v.

All these sums of squares are computed in the usual way, except that for blocks adjusted for lots. In the repeated lattice design the block sums of squares consists of two components. Component (a) is estimated from the differences between the totals of pairs of blocks containing the same set of five lots. The compoment (b) is estimated from the sums of pairs of blocks containing the same set of lots. The sum of squares computed for these two components are added and this is the sum of squares for blocks. The intra-block error sum of squares is computed by subtracting the sums of squares of the other factors from the total sum of squares. All of these sums of squares are collected together in the analysis of variance table on page v. We also have listed there the mean squares for blocks and for the intra-block error term. The significanceof the blocks of our design can be tested with these mean squares with a resulting F of 4.46 with 16 and 56 df. which is highly significant. The mean squares of blocks and intra-block error will be used to compute the weighting factor which is used in computing the block adjustments. The formula for μ is given on page v and its value turns out to be .13951 for this test. This value is then multiplied by the C values and these values of μC are the adjustments used in obtaining the adjusted lot totals on page iii.

Looking again at page v we see the overall estimate of round to round error variance computed from a weighted average of the round to round variance of each cell in the design i.e. from the one hundred five-round groups. If this variance is divided by 5 an estimate of the error variance for the average of five observations is obtained. This error variance for the average of five observations should have the same expected value as the intra-block error unless there is an interaction of Blocks and Lots. Therefore, an F test of the intra-block error term ovdr the error variance of the average of five observations provides a test of this interaction and for our data turned out to be not significant and thus we have no evidence of an interaction between blocks and lots.

The analysis presented provided for the recovery of inter-block informatio which is used in the adjustment of the lot means. With this type of analysis the design cannot be appreciably less accurate than a randomized complete block design that would be obtained if only the replications in our lattice design were considered. This would not be true had the analysis not provided for the recovery of inter-block information. Of course, if there are significant differences among the blocks of the design, as there were in our test, the lattice design provides considerably more precision than the randomized block design.

The average range adjusted for block differences and the standard deviations of range estimated from the within cell variation of the four five-round-groups fired from each lot are listed on page vi. As indicated earlier a sample of reference rounds were also fired in the design together with the test lots. After the lots were adjusted for block differences a correction was made for reference. The final average range used for grading purposes is the average range corrected for reference.

In this discussion I have presented an application of a repeated 5 x 5 simple lattice design to a program concerned with obtaining estimates of the average range and dispersion of range for twenty-four lots of mortar ammunition. The experimental design used was of the incomplete block type. This type of design is very useful when a large number of treatments are to be investigated and the size of the homogeneous blocks available is relatively small.

Observed Ranges for Shell, HE, M49A2 for 60mm Mortar Fired with Charge 4 at an Elevation of 45°
(Range measured in yards)

Lots	MA-1-53		MA-1-82		MA-1-363		MA-1-604		MA-1-613	
Date	24 Oct 52	31 Oct 52	17 Dec 52	12 Jan 53	19,20 Aug 53	8,9 Sep 53	30 Dec 54	10 Jan 55	20 Jan 55	31 Jan 5
	1890	2057	1925	2112	1988	1932	1967	1980	2000	2110
	1863	2028	1903	2083	1876	1862	2021	2025	1769	1983
	1927	1964	2043	2096	1874	1863	2014	1983	1885	2098
	1830	1955	1957	2078	1914	1927	2019	1862	2004	2084
	1803	1976	1946	2031	1882	1907	2002	2041	1904	2015
	1951	1996	1940	2084	1774	1763	2128	2001	1865	1978
	1995	2057	1916	2017	1872	1841	1949	1970	1927	2098
	1967	1979	1967	2035	1822	1914	1904	2053	1972	2124
	1934	2010	1958	1978	1891	1837	2029	1978	1886	2077
	1897	2037	1879	2045	1855	1911	1989	1940	2019	2036
	1869	2013	1995	2002	1809	1866	2052	1980	1884	2060
	1847	1990	1980	2078	1894	1797	2042	1921	1990	2141
	1882	2015	1934	2118	1870	1983	1835	2002	1884	2075
	1965	1975	1990	2017	1910	1873	1970	1969	1938	2074
	1954	1934	1906	2107	1970	1907	2028	1849	1950	2003
	1973	2074	1985	2005	1980	1923	2000	2030	1999	2077
	1870	2071	2000	2094	1885	1962	2041	2006	1987	2110
	1894	1993	1951	1993	1775	1859	2033	1999	1823	2061
	1927	1943	1979	2020	1871	1993	1996	2006	1951	2099

Analysis of Variance Table Under Model for 2-Fold Hierarchal Classification[1]

Source of Variation	SS	df	MS	Expected Value-MS
Among Lots	369,254	4	92,314	$\sigma^2+19\sigma_b^2+38\sigma_a^2$
Among Days within Lots	369,941	5	73,988	$\sigma^2+19\sigma_b^2$
Within Days	517,069	180	2,873	σ^2
TOTAL	1,256,264	189		

Among days within lots, F=73,988/2,873=25.75, sig. at .01
Among lots F=92,314/73,988=1.25, not sig. at .05

[1] Kempthorne, O. The Design and Analysis of Experiments; John Wiley and Sons, New York 1952.

Arrangement of Lots in 5x5 Simple Lattice Design

Rep I

Block (a)	1	2	3	4	5
Block (b)	6	7	8	9	10
Block (c)	11	12	13	14	15
Block (d)	16	17	18	19	20
Block (e)	21	22	23	24	25

Rep II

Block (a)	1	6	11	16	21
Block (b)	2	7	12	17	22
Block (c)	3	8	13	18	23
Block (d)	4	9	14	19	24
Block (e)	5	10	15	20	25

Average Ranges (less 1600 yds.) for Cells of a Repeated 5x5 Simple Lattice
(Number in Parentheses is Lot Identification)

		Rep I			Block Totals
(1) 277.0	(2) 252.0	(3) 257.0	(4) 267.4	(5) 254.2	1307.6
(6) 271.2	(7) 270.2	(8) 254.5	(9) 170.2	(10) 240.0	1206.1
(11) 260.6	(12) 238.2	(13) 278.2	(14) 268.4	(15) 279.2	1324.6
(16) 198.4	(17) 159.0	(18) 268.6	(19) 256.0	(20) 309.2	1191.2
(21) 229.6	(22) 291.2	(23) 240.6	(24) 249.2	(25) 210.6	1221.2
					6250.7

		Rep II			Block Totals
(1) 234.4	(6) 208.0	(11) 245.2	(16) 176.0	(21) 221.0	1084.6
(2) 252.8	(7) 260.8	(12) 219.0	(17) 174.8	(22) 245.6	1153.0
(3) 344.0	(8) 302.0	(13) 296.0	(18) 323.6	(23) 320.8	1586.4
(4) 319.8	(9) 221.8	(14) 280.2	(19) 308.2	(24) 349.0	1479.0
(5) 341.2	(10) 324.8	(15) 374.8	(20) 359.3	(25) 318.4	1718.5
					7021.5

		Rep III			Block Totals
(1) 299.4	(2) 333.6	(3) 316.2	(4) 320.5	(5) 313.2	1582.9
(6) 321.0	(7) 297.0	(8) 345.0	(9) 271.8	(10) 315.6	1550.4
(11) 272.4	(12) 301.2	(13) 388.4	(14) 311.8	(15) 331.6	1605.4
(16) 246.8	(17) 239.0	(18) 342.4	(19) 293.6	(20) 340.7	1462.5
(21) 272.5	(22) 357.5	(23) 293.4	(24) 287.4	(25) 278.4	1489.2
					7690.4

		Rep IV			Block Totals
(1) 303.8	(6) 279.2	(11) 190.3	(16) 221.8	(21) 303.4	1298.5
(2) 318.8	(7) 288.8	(12) 255.0	(17) 222.2	(22) 337.8	1422.6
(3) 302.2	(8) 323.5	(13) 342.8	(18) 264.8	(23) 308.5	1541.8
(4) 297.8	(9) 210.6	(14) 295.0	(19) 302.6	(24) 301.8	1407.8
(5) 306.8	(10) 297.8	(15) 339.0	(20) 284.4	(25) 318.5	1546.5
					7217.2

			Lot Totals (4 reps)			μC
	(1)1114.6	(2)1157.2	(3)1219.4	(4)1205.5	(5)1215.4	+18.29
	(6)1079.4	(7)1116.8	(8)1225.0	(9) 874.4	(10)1178.2	- 5.47
	(11) 968.5	(12)1013.4	(13)1305.4	(14)1155.4	(15)1324.6	-12.93
	(16) 843.0	(17) 795.0	(18)1199.4	(19)1160.4	(20)1293.6	- 2.23
	(21)1026.5	(22)1323.1	(23)1163.3	(24)1187.4	(25)1125.9	+43.86
μC	+ 37.08	+ 22.78	- 20.08	- 26.58	- 54.73	

		Adjusted Lot Totals		
(1)1170.0	(2)1193.3	(3)1217.6	(4)1197.2	(5)1179.0
(6)1111.0	(7)1134.1	(8)1199.4	(9) 842.4	(10)1118.0
(11) 992.6	(12)1023.2	(13)1272.4	(14)1115.9	(15)1256.9
(16) 877.8	(17) 815.6	(18)1177.1	(19)1131.6	(20)1236.6
(21)1107.4	(22)1298.7	(23)1187.1	(24)1204.7	(25)1115.0

Statistical Analysis of Repeated 5x5 Simple Lattice Design with Recovery of Inter-Block Information

Block Totals					
I	III	Diff.	Sum	C	μC
1307.6	1582.9	-275.3	2890.5	+131.1	+18.29
1206.1	1550.4	-344.3	2756.5	- 39.2	- 5.47
1324.6	1605.4	-280.8	2930.0	- 92.7	-12.93
1191.2	1462.5	-271.3	2653.7	- 16.0	- 2.23
1221.2	1489.2	-268.0	2710.4	+314.4	+43.86
		-1439.7		+297.6	

II	IV	Diff.	Sum	C	μC
1084.6	1298.5	-213.9	2383.1	+265.8	+37.08
1153.0	1422.6	-269.6	2575.6	+163.3	+22.78
1586.4	1541.8	+ 44.6	3128.2	-143.9	-20.08
1479.0	1407.8	+ 71.2	2886.8	-190.5	-26.58
1718.5	1546.5	+172.0	3265.0	-392.3	-54.73
		-195.7		-297.6	

Total Sum of Squares

$$(277.0)^2+(252.0)^2+\ldots\ldots+(318.5)^2-\frac{(28,179.8)^2}{100}=$$

$$8,162,481.22-7,941,011.28=221,469.94$$

Sum of Squares for Replications

$$\underline{(6250.7)^2+(7021.5)^2+(7690.4)^2+(7217.2)^2}-\frac{(28,179.8)^2}{100}$$

$$7,984,117.62-7,941,011.28=43,106.34$$

Sum of Squares for Lots (ignoring blocks)

$$\underline{(1114.6)^2+(1157.2)^2+\ldots+(1125.9)^2}-\frac{(28,179.8)^2}{100}$$

$$8,056,709.71-7,941,011.28=115,698.43$$

Sum of Squares for Blocks (eliminating lots)

Component (a)

$$\frac{(275.3)^2+(344.3)^2+\ldots+(172.0)^2}{10}-\frac{(1439.7)^2+(195.7)^2}{50}$$

$$57,368.888-42,220.692=15,148.196$$

Component (b)

$$\frac{(131.1)^2+(39.2)^2+\ldots+(392.3)^2}{20}-\frac{(297.6)^2+(297.6)^2}{100}$$

$$21,731.639-1,771.315=19,960.374$$

Analysis of Variance Table

Source of Variation		Degrees of Freedom		Sum of Squares	Mean Squares
Replications		3		43,106.34	
Lots (unadj)		24		115,698.43	
Blocks within replications (adj)		16		35,108.57	2194.29
Component (a)	8		15,148.196		
Component (b)	8		19,960.374		
Intra-block error		56		27,556.60	492.08
TOTAL		99		221,469.94	

The Weighting Factor

$$\mu = \frac{p\,(E_b - E_e)}{K\left[(r-p)\,E_b + (p-1)\,E_e\right]}$$

$$\mu = \frac{2\,(2194.29-492.08)}{5\left[2(2194.29) + 492.08\right]} = .13951$$

Round to Round Error Variance (within Cells)

$$\hat{S}_e^2 = \frac{708,860.5}{388} = 1826.96$$

Error Variance for Average of Five Observations

$$S^2 = \frac{1826.96}{5} = 365.39$$

Approximate Test for Significance of Block X Lot Interaction

$F = \dfrac{492.08}{365.39} = 1.3467$ with 56 and 388 df Not significant (.05 level)

RANGE
Shell, HE, M49A2 for 60mm Mortar, Charge 4, Elevation 45°

		Av. Range	Av. Range	
Lot Number	Lot Ident.	Adjusted for blocks yd.	Corrected for ref. yd.	St. Dev. Range yd.
WC-32-9	1	1892.5	1931.0	40.55
WC-94-242A	2	1899.6	1938.1	55.44
WC-2-116A	3	1904.4	1942.9	31.05
WC-94-288A	4	1899.3	1937.8	27.37
WC-94-290A	5	1894.8	1933.3	38.11
WC-2-101B	6	1877.8	1916.3	54.48
Wc-2-22A	7	1883.5	1922.0	45.71
WC-6-255B	8	1899.9	1938.4	52.95
KOP-110A	9	1810.6	1849.1	38.85
WC-32-7	10	1879.5	1918.0	40.22
WC-2-128A	11	1848.2	1886.7	42.83
WC-6-192B	12	1855.8	1894.3	32.62
WC-38-135	13	1918.1	1956.6	52.12
WC-2-128B	14	1879.0	1917.5	52.19
WC-2-28B	15	1914.2	1952.7	49.72
KOP-147A	16	1819.5	1958.0	35.37
KOP-113A	17	1803.9	1842.4	22.10
WC-2-18B	18	1894.3	1932.8	29.03
WC-32-8	19	1882.9	1921.4	48.52
WC-6-255A	20	1909.2	1947.7	58.26
WC-6-876A	21	1876.8	1915.3	41.27
WC-2-172A	22	1924.7	1963.2	25.26
WC-2-120B	23	1896.8	1935.3	47.92
WC-2-23A	24	1901.1	1939.6	47.85
KOP-64A	25	1878.8	1917.3	28.01

THE DESIGN OF EXPERIMENT IN STABILITY TESTING

by

James W. Mitchell

Frankford Arsenal

INTRODUCTION

A frequent problem in ordnance research and testing is measurement of the deterioration of material and equipment with time. This is maybe known as stability testing and concerns deterioration while in storage in contrast with deterioration or wearing while the item is in actual use (Life Testing). Storage affects many of the more obvious physical and chemical properties of materials. Stability testing is also applied to many less obvious changes in complex assemblies which often cannot be identified with any specific physical property or set of properties, but only with some performance aspect of the assembly.

A naive assumption is often made that it is possible to predict from the stability test the safe maximum storage life of an item. For many ordnance engineers this may seem to form a sufficient and workable objective for a stability test. It is hoped that this paper will create in the reader an awareness of the difficulties inherant in such a broad objective and the virtual impossibility of predicting the complete storage life of a military item. This will require development of a definition of stability and reasonable objectives for stability testing as well as a discussion of measurements, applicable statistics and interpretation of results.

A DEFINITION OF STABILITY

For the following definition the writer is indebted to Ernest Rechel, Director of the Chemistry Research Laboratory of the Frankford Arsenal (1). In this definition the stability of an object is identified with the stability of its attributes. The term attribute is used to denote any or all qualities, properties or modes of behavior of an object. Normally the attributes of an object are known or measurable single-valued functions of time and the environment. Objects and classes of objects are distinguished by their sets of respective attributes. It follows therefore to define stability of an object in terms of the stability of its attributes.

A definition of stability should meet certain formal demands. First, stability is always associated with change in properties with time. The definition must therefore appear as some time rate of change of the attributes. Also the definition should be independent of sign, i.e. whether

(1) References appear at the end .

the attributes are increasing or decreasing. And finally it would be convenient if stability were defined in dimensionless terms. These demands are met by the following function. If k represents the value of the k-th attribute at time t and S_k its stability.

$$\frac{1}{S_k} = \left| \frac{1}{k} \cdot \frac{dk}{dt} \right|$$

Here the stability S_k appears as a function of the rate of change of the attribute; it is independent of the sign of dk/dt; and it is independent of the units in which the attribute is measured. The reciprocal relationship provides that as dk/dt increases, S_k decreases.

Since every object possesses a large, in fact practically infinite number of attributes, a, b, c, ... k . ., its total stability may be expressed by

$$\frac{1}{S} = \left| \frac{1}{a} \cdot \frac{da}{dt} \right| + \left| \frac{1}{b} \cdot \frac{db}{dt} \right| + \left| \frac{1}{c} \cdot \frac{dc}{dt} \right| + \cdot \cdot \cdot \cdot \cdot + \left| \frac{1}{k} \cdot \frac{dk}{dt} \right| + \cdot \cdot$$

This is perhaps the simplest function of all the attributes that continues to satisfy the intuitive demands. It is not difficult to accept the concept that an object has an infinite number of attributes since there are an essentially unlimited number of environments - and an object can be expected to behave differently in each environment. In practice we must necessarily deal with a finite number of attributes. A specific set is chosen which are sufficient to define the class under study and the rest are ignored, or in effect treated as zero.

The above definition of stability is formally correct and should find practical application. It will be apparent that if all the attributes are constant S will be infinitely large. If, however, at least one of the attributes exhibits a rate of change not equal to zero, S will take on finite positive values. Therefore the smaller S becomes, the lower the stability of the object.

OBJECTIVES OF A STABILITY TEST

A complete objective for a stability test based on the above generalized definition would require that the entire future life and environments of an object be known before the test could be planned, - a most unusual condition in ordnance to say the least. It is therefore the first step in a stability test to recognize and define a more limited and attainable objective. It is suggested that an acceptable and meaningful objective would be to determine the useful life of an item in a specific selected environment chosen from the following categories:

1. Selection of one specific environment as arbitrarily representative. This may be a natural environment or a simulated laboratory condition such as a constant temperature or a salt spray. Storage life is defined thereby for only this condition.

2. Selection of some maximum condition matching the severest condition in the field and determing storage life under these conditions. The condition will usually be simulated in the laboratory and maintained continually, although cyclic condition might be selected to simulate a natural cycle such as night and day or the tides. This gives a minimum but never a most probable storage life.

3. Field surveillance of marked items for durability under a wide variety of naturally occuring storage conditions. The disadvantages of this method are the obviously difficult and expensive examination of objects in the field or their return to the home laboratory for test, the time involved and the integration of data from many sources into a single estimate of storage life. This latter difficulty appears to be surmountable only by the use of some arbitrarily selected set of weights for the different environments. On the other hand this method is intuitively the most convincing and is widely used commercially on new products.

4. Selection of a specific environment but at an artificially high level above that of the maximum of the field for the purpose of accelerating the rate of failure and shortening the time of test. This is the so called accelerated test.

5. Comparative tests in which a standard item is tested along with one or more to be evaluated. Through long use and past evaluation, the storage life of the standard is approximately known and at least acceptable. Any of the previous specific environmental conditions may be selected for storag

Any of the above permit the definition of a constant or uniformly variable condition under which the storage life of an object may be defined as the change in the selected set of attributes with time. The limited objective then becomes the life of the item in this limited and defined environment. The above seems to be straightforward in its application to most storage problems except for the accelerated test condition. This special case will be dealt with later on.

Selection of the set of attributes to be measured and the measurement to be used on them is the next step in planning a stability test. The material can vary tremendously in complexity. There is usually no great problem in this area of measurement for simple items such as small components or engineering materials. However when a highly complex item involving mechanical, electrical and chemical elements must be considered, stability is usually approached by study of only the newer untested or obviously less stable components or materials. The stability problem then becomes like that for the simple item. Of course complex items such as guided missles or automotive equipment are given exhaustive service tests but this is not stability testing.

The measurement to be used should meet certain demands. These may be enumerated as follow:

1. Responsive to changes in the use attribute being evaluated.

2. Provides quanitative data

3. Permits replication

4. Reproducible, inexpensive and nondestructive if possible

There is usually some lattitude in the selection of attributes to be observed in the test and in means of measuring them. Selection from among these possible measurements can then be made to best meet the demands given above. The advantages and disadvantages of a number of classes of attributes are discussed.

1. Observation of the major use function. This is frequently best and unambiguous in its interpretation if quanitative data can be obtained at reasonable cost.

2. Overall qualitative evaluation of an object can be used where the use function is passive and thus not measurable in terms of performance. This often results in adjective ratings or ranking values which are nonquanitative and therefore difficult to treat so as to obtain estimates of error.

3. A specific attribute identified solely or closely associated with the major use function. This is probably the most frequently employed type of measurement and is usually quite satisfactory. Difficulties might arise in establishing the responsiveness of the use function to the measured attribute and visa versa, although this should be evident from the knowledge leading to its choice.

4. Several attributes identified with the use function, Sometimes performance cannot be described in terms of a single number but requires expression of the values of a number of attributes. Since it is usually impossible to combine and reduce these to a single value, the stability function cannot be treated as a single one. The experimenter must work out some compromise to suit the problem.

5. Observation of success or failure of the item to some stimulus. Data of this type is rendered quanitative by testing a number of items at each time and observing the number or fraction failing. This kind of data can also be treated as an increased severity test to obtain an estimate of

the stimulus causing 50% failure and standard deviation. The change of this 50% point is then treated as the stability variable (2).

6. Study of the basic deterioration mechanism of the use property by refined physical or chemical research. The advantages of this approach are obvious if time and the nature of the object permit.

STATISTICAL TECHNIQUES IN STABILITY TESTING

Many of the elements of an experimental design will have become apparent by the time satisfactory objectives have been chosen and the responsive attributes of the item and methods for their measurement selected. At this period of an experiment it is possible to look ahead and anticipate the several possible outcomes of a storage program and then to decide what kind of data is needed to make a statistically significant choice between these possible outcomes. This implies the selection and use of statistical procedures. It is of course always possible to handle data graphically and where large differences occur, significant conclusions can be drawn from graphical treatment. However it is much preferable to obtain estimates of error so that differences of any magnitude can be compared by the few general techniques mentioned below It is not intended to go into any detail regarding these methods since they are well covered in numerous textbooks.

The most useful statistical procedures are those of curve fitting. Stability curves are usually, empirical in the sense that there is usually no theoretical reason for the curve to fit a specific mathematical function. However if a theory of decomposition does exist, such as based on a known chemical reaction, then the data should be fitted to the mathematical equation derived from theory. For convenience of fitting, equations of other kinds should be transformed into a linear form where possible for a least squares analyses. Where no theory exists to indicate a specific mathematical form, it is usually best to attempt the fitting of a polynomial of first, second or higher degree. The method of least squares or orthogonal polynomials may be used to fit the equation. The advantage of the orthogonal polynomials method as developed by Fisher(3-4), is that the fitting may be carried through in successive stages, the success of fitting terms of higher degree being observed and tested for significance at each stage. Both methods permit estimation of error. For a discussion of tests of significance between two curves see the last portion of this article.

An experimenter may wish to become fancy and attempt a correlation between the attribute being measured and several environmental variables as well as time. This is clearly a case for a multiple regression analyses and may be worth the extra trouble in a storage program if a natural environment is being employed and uncontrollable natural variables come into play and are measurable. However such treatment is bordering on an attempt at discover of the mechanism of deterioration. If this is the object rather than the time rate of change in the natural environment, it might be better to employ controlled laboratory conditions arranged so that the data may be treated analytically to obtain an expression of the effect of each variable. When a chemica process is involved this is often possible.

If sufficient prior knowledge on variability both of the material and the measurement, is available and the objectives require specific comparison against an upper service requirement, it is possible to design a sequential technique to efficiently detect a change in trend. This would detect a break in a uniform slow rate of change indicative of the end of safe storage life(5)

A word might be added about sampling error and product variability. As mentioned previously, some knowledge of error is essential to testing the difference between any two stored items or establishing confidence intervals for experimental results. When only a limited sample is available without prior information on the variability of the product or testing method, and only a single item can be tested at a time interval, error can be estimated from the regression itself. However this error will contain the product and test variability and also changes (usually increases) in product variability caused by the storage conditions. It is very desirable to be able to separate the latter since increased variability is a common outcome of storage. In fact severe storage conditions seem to have an effect like an increased severity stimulus causing some items in a sample of apparently identical items to change rapidly or fail in a short time while others may last for a long period. The net effect of this is a large increase in variability. Product variability should be measured prior to the storage if possible. Furthermore, if the number of items permit, replicate tests should be made at each withdrawal time. By fitting the average of these replicates, a much better fit should be obtained with a reduced standard error of estimate. It is also possible then to directly determine the change in product variability with time.

ANALYSIS OF RESULTS

The limited objectives given previously fall into three categories -

1. Estimation of storage life under a defined natural or simulated condition

2. Accelerated storage testing

3. Comparative testing against a standard

Each of these call for somewhat different treatment of results to arrive at valid conclusions and will be discussed separately.

The first or storage life test clearly requires a regression curve fitted to the data from which quality and variability of quality of the observed attribute can be determined at any period of time. Conversely if a minimum deterioration level of the attribute is specified or set by use requirements, the regression curve and observed standard error can be used to estimate the point at which a certain proportion of failures will occur. This is the storage life.

Extrapolation of the fitted curve beyond the last observation point (in cases where insufficient material was stored to run out to complete failure) is extremely questionable and not to be relied on. However in the special case where basic studies have disclosed the physical or chemical processes of deterioration, refined measurement of this process may successfully yield a natural law as of a chemical reaction. Such a law when shown to fit the experimental data well, may be cautiously extrapolated as far as the researcher's initiative permits him to stick out his neck.

An example of the latter is the deterioration of magnesium powder in moist air. The kinetics of the reaction have been successfully worked out permitting complete description of change in magnesium content with time (6).

The accelerated storage test seeks to provide an estimate of storage life on long lived material in a short time under artificially severe or elevated conditions. If such a test can be successfully designed, a very desirable objective is attained. However many difficulties beset this type of test. The problems are less severe when an accelerated test is applied to the comparison of two materials or a new item with a standard, as will be considered later. The main difficulties with the accelerated stability test stem from the usually unknown multiplication factor relative to service life. Since there is no standard condition of field storage, there can be no single multiplication factor. A further difficulty is caused by the many uncontrolled and unknown variables which may intrude into the field condition at one time and not at another. A more hopeful correlation is that between the accelerated condition and a simulated environmental condition. Such correlation would yield a multiplication factor and be reproducible but still offers no more guarantee of true or meaningful storage life under use conditions than the straight storage test.

Another difficulty is incurred if a large multiplication factor is obtained by the use of a very elevated condition. Under such conditions, extreme sensitivity or insensitivity of the measured attribute may exist unknowingly for some materials and not others, thus rendering use of the multiplication factor in later work very uncertain. Again the environmental factor chosen for accelerating the deterioration may not be continuously related to the use property or fail to respond above certain limits. These difficulties practically demand research on the accelerated test condition before its use in actual testing.

A way out of the above difficulties with accelerated testing lies in running the storage test at a number (three or more) levels of the accelerated test condition ranging from the maximum down to a value in the upper range of actual use conditions. This permits exploration of the response surface of the item with time and the stimulant. Discontinuities can be found and the accelerating effect of the stimulant studied. An example is the use of several temperatures and calculation of the temperature coefficient of the decomposition reaction.

Another method of circumventing difficulties with the multiplication factor is to employ the accelerated test only for comparison of several items or against a standard material. This use is still subject to the difficulties of extreme sensitivity or insensitivity as mentioned above.

Finally we can consider the comparitive stability test. Probably this is the most convincing and satisfying test and is obviously applicable to any of the natural, simulated or accelerated storage conditions described. The comparisons would be between two or more similar objects to see which has the longer life or between a newly developed item and a standard of known and acceptable quality. Thus one gives up any objective of predicting storage life except relative to the standard. Since only one or a limited number of storage conditions would be used in the comparative test, a small residue of doubt may exist that the relative position of the several items in the comparative test may change in other environments. It is not easy to resolve this doubt.

Comparative stability tests will result in two or more regression curves which must be compared to detect real differences. Too frequently stability curves are judged only on the basis of relative position of the curves (slope) without proper consideration of dispersion of the storage test results. A description of statistical tests of significance applicable to all types of regression curves would be too extensive for inclusion in this paper. In fact methods are lacking for some of the more involved combinations of regression curves. The following generalizations may be helpful to the engineer faced with a statistical comparison but reference to any of the modern textbooks on general statistics is recommended before actual statistical analysis is started.

1. Many textbooks describe methods of comparing the slope or intercept coefficients in linear regression. The variance of the dependent variables may be estimated from single observations used in the calculation of the regression curves or from the several variances obtained when a number of samples are tested at each observation interval to yield both the mean and variance of the observation.

2. When independent estimates of variance are obtained for each observation it is possible to detect some of the non-random changes in variance mentioned previously. When these changes assume the form of a regular increase or decrease, it is reasonable to associate this with change in product variability and smooth out the random fluctuations by fitting a regression to variances. This would relate change in variance with change in the dependent variable, time. Other large but non-uniform fluctuations in variance can be attributed to lack of control in the test method or sampling procedure or poor control over environment.

3. When the variances are found not to be constant or when one or both of the regressions are found to be quadratic or of higher degree, the simple significance tests of slope and intercept used in the linear case no longer apply. It is possible to calculate a standard error of estimate for specific calculated values of the dependent variable obtained by the regression corresponding to values of the independent variable, time. Corresponding points of the dependent variable on two regression curves may then be tested for significant difference by the Student t test. After a number of points have been tested, it will usually be possible to state that after a given time the

difference in stability of the two objects became statistically significant, or the negative of this. The method of curve fitting by orthogonal polynomials is particularly well suited to providing error estimates of individual values of the dependent variable.

SUMMARY

The design of experiment in stability testing is important to insure valid results and conclusions for experiments which have required long periods of time to conduct. Design of experiment in stability testing may be summarized as follows:

1. Choice of reasonable limited objectives

2. Selection of the proper attributes for which there exist quanitative measurements and which are responsive to the important use function of the object being tested.

3. Anticipation of the several possible outcomes of the stability test and deciding what kind of data is needed to make a statistically significant choice between the several possible alternates.

4. Use of theory where possible to select the regression function and otherwise applying the best available theory and experience in fitting a regression curve to the data.

5. Proper appreciation of the several source, of variation which may occur in the storage test and appropriate measurement of variation and its use in testing significance between curves or establishing confidence intervals for any predicted storage life.

6. Application of past experiences and scientific horse sense in arriving at the final conclusions.

REFERENCES

(1) Rechel, E. R., Pitman-Dunn Laboratories, Frankford Arsenal, Philadelphia 37, Pa., unpublished paper

(2) Churchman, C. W. and Epstein, B. "Tests of Increased Severity", J. Amer. Stat. Assoc., Vol 41, 567-590, (1946).

(3) Fisher, R. A., "Statistical Methods for Research Workers", 8th Ed pp 140-149, Oliver and Boyd, Edinburgh, (1941).

(4) Anderson, R. L. and Houseman, E. E., "Tables of Orthogonal Polynomial Values Extended to N =104", Research Bulletin 297, Agricultural Experiment Station, Iowa State College, April 1942.

(5) Villars, D. S., "Statistical Design and Analysis of Experiments for Development Research" Wm. C. Brown Co., Dubuque, Iowa, (1951).

(6) Regan, J. E. and Silverstein, M. S., "Kinetics of the Corrosion of Atomized Magnesium in Moist Air", Report R-1092, Frankford Arsenal (July 1952

SENSITIVITY TESTING

A. Bulfinch
Picatinny Arsenal

Picatinny has joined the quest for better tests of increased severity. In the explosives industry tests of this type have wide application, such as impact and friction sensitivity, functioning rates of electrical detonators, minimum initiating and detonating charges in explosive trains and the laboratory sand test, functioning rates of fuzes, minimum explosive charges for ejection mechanisms, and even the drop test for packing containers.

The dilemma in this problem involves the properties of asymptotic curves. That is, in most practical applications we are interested in the two extremes of such curves. In these portions of the curves one variable of course changes very rapidly while the other variable changes very little. Unfortunately the independent variable in this problem is the insensitive one - that is, it changes very slowly with large changes of the dependent variable. The only known solution to this dilemma is the use of very large samples and therein lies the problem.

Most of the work done on this problem in the explosives field for sensitivity tests and in the biological field for dosage tests has been in the area of the 50% point of the curves. From a statistical point of view this is desirable since reliable data can be obtained at this point with minimum sample sizes. However, the explosives engineer insists he is not interested in the 50% point. From safety considerations he is interested in the lower extreme of the curve and from functioning considerations he is interested in the upper extreme of the curve.

Methods such as the "Up-and-Down" method and the "Run-Down method are designed to measure the mean and standard deviation at the 50% point. The characteristics of these methods are as follows:

Advantages

1. The mean at the 50% point contains the least error.

2. A measure of the standard deviation makes it possible to quantitatively evaluate observed differences.

Disadvantages

1. The 50% point is of little practical value.

2. The validity of the results obtained from these methods depend upon assumptions made concerning the form of the frequency distribution of the parent population.

Work at Picatinny has shown that there is reason to doubt that the sensitivity data of all explosives have the same frequency distribution. In laboratory-scale tests the tail of the curve for Comp B was found to deviate from the normal curve in one direction while the curves for the RDX and Tetryl were found to deviate in another (See Table I at the end of this manuscript). These deviations were determined by actual measurements with large sample sizes in the lower tails of the curves. The results of the actual measurements were found to differ significantly from those obtained by extrapolation from the 50% point using the assumption of normality. Of the two methods used which are based on an assumption of this type the "Run-Down" method appeared to give the better estimate although not an acceptable one.

At the 50% point RDX was found to be similar to TNT, which is contrary to experience. But in the area of the one percent point RDX was found to be similar to Tetryl which agrees with experience in the use of explosives.

This comparison brings out an important requirement for any laboratory method. The results of laboratory tests should be of use in predicting functioning characteristics.

Methods such as the Picatinny Arsenal Method and the Naval Powder Factory Method are designed to measure the ten percent point. The characteristics of these methods are as follows:

Advantages

1. They are not dependent upon assumptions of normality.

2. They obtain results near the tails of the curves instead of the 50% point.

Disadvantages

1. The variances of these methods are not known.

2. The size of the sample used is not large enough to obtain reasonably precise results.

The lack of a known variance is a distinct disadvantage. Without a measure of dispersion quantitative evaluation of observed differences is not possible. However, this type of sequential approach to a particular percentage point may be a possible solution to the problem created by the need for large samples. This is yet to be determined. The choice of the ten percent point appears to be an unfortunate one. Work at Picatinny shows that the impact sensitivity (using the PA apparatus and the "Run-Down" method) of Comp B, RDX, TNT, and Tetryl are all equal at the seven percent point (See Figure I following Table I at the end of this paper). In addition the uniqueness of the sensitivity characteristics that were found occurred in the area of the one percent point.

From the work done to date the following method has been derived and is presented here for consideration as a partial solution to the subject problem:

1. Collect the data in a manner similar to that described in the "Run-Down" method with the following modifications:

a. Any portion of the curve which is of no interest can be omitted.

b. For those portions of the curve which are of interest use the sample size required for the desired precision. These sample sizes can be obtained from tables for the confidence limits of the binomial distribution in the published literature.

2. Calculate the confidence interval (from tables such as those referred to above) for the proportion of explosions obtained at each height level used.

3. Plot the terminal values of these confidence limits versus corresponding height values on probability paper.

4. Graphically determine the confidence lemit for the height value associated with any desired percentage point.

5. Report the confidence interval for the height value as the impact sensitivity of the explosive.

6. To evaluate observed differences between two or more explosives or lots of the same explosive and determine whether the confidence intervals overlap. If they do overlap the difference is not significant, if they do not overlap the difference can be considered significant.

Characteristics of this method are as follows:

Advantages

1. It is a general method applicable to any test of increased severity in which the observed results are attribute-type data.

2. It is simple to conduct.

3. It is completely flexible for determining any desired percentage point with any desired predetermined precision.

4. It is free of all assumptions concerning the form of the underlying distribution.

5. The results are simple to calculate.

6. It includes a simple method for quantitatively evaluating observed differences.

Disadvantages

1. Large sample sizes are required in the tails of the curves to attain a reasonable precision.

2. The procedure for evaluating observed differences is based on a graphical method for calculating the confidence interval of the height.

3. The actual standard deviation is not determined.

Recent work such as that being conducted at Wayne University for Frankforc

Arsenal has shown that much can be learned from sensitivity tests about the mechanism of initiation and the propagation of initiation of explosives. Instruments which can accurately measure energy input and output and methods that are based on known properties of explosives such as crystal state, thermodynamic properties, and other physical properties are required. Laboratory methods and instruments currently being used for the determination of the impact sensitivity of explosives may soon be completely antiquated by methods and instruments of this type. However, methods of increased severity of the type in current use will always be required for determining functioning rates of fuzes and detonators and minimum explosive charges of all kinds. For this reason further work to improve these methods is desirable.

Part of the work conducted at Picatinny on this problem has been confined to impact testing of primary explosives at the 50% point and below. Other work conducted at Picatinny and for Picatinny at Franklin Institute has been on the functioning rate of electrical detonators at the 50% point and above. Mr. Fred Lawrence of our Fuze Laboratory will describe this latter portion of the problem.

Additional work is contemplated to develop a non-parametric method with known variance. The object of this effort is to effect economies in collecting valid data in this type of testing. This can be accomplished in either or both of the following ways:

1. Reduce the sample size required for a given precision.

2. Increase the precision for a given sample size.

Private Ehrenfeld of our Ammunition Research Laboratory will describe a proposal for the application of the Monte Carlo Method to this problem.

If the impact sensitivity of explosives is of any value and its use to be continued, it would be very desirable to have a standard method and instrument established for use throughout the Ordnance Corps. It would also be desirable to establish a standard explosive which can be supplied through a central source for use in calibrating instruments. A step in this direction has already been taken by the Physical and Chemical Properties Subcommittee of the Blast and Incendiary Committee. If efforts to improve sensitivity results is to be continued then a coordinated program for the Ordnance Corps should be established.

Your comments are invited.

TABLE 1

Probability[a] of an Explosion at a Height of Eight Inches For Various Explosives

		Binomial		Extrapolation from 50% Point Using Assumption of Normality	
Explosives	No. of Trials	No. of Explosions	Actual Measurements[b]	Run-Down Method[b]	Up-and-Down Method
HMX	50	8	5.7 - 23.3%	10.7 - 29.1%	-
RDX	350	6	0.63 - 3.75	0.71 - 1.25	less than 1.0 x 10^{-6}
Tetryl	1000	4	0.11 - 1.02	0.05 - 0.10	-
Comp B	1000	1	0.00 - 0.56	1.66 - 2.28	-
TNT	50	-	Estimated[c] less than 0.5	0.15 - 0.75	-

[a] At the 95% confidence level

[b] Confidence intervals for the PA apparatus

[c] Estimated from observed data

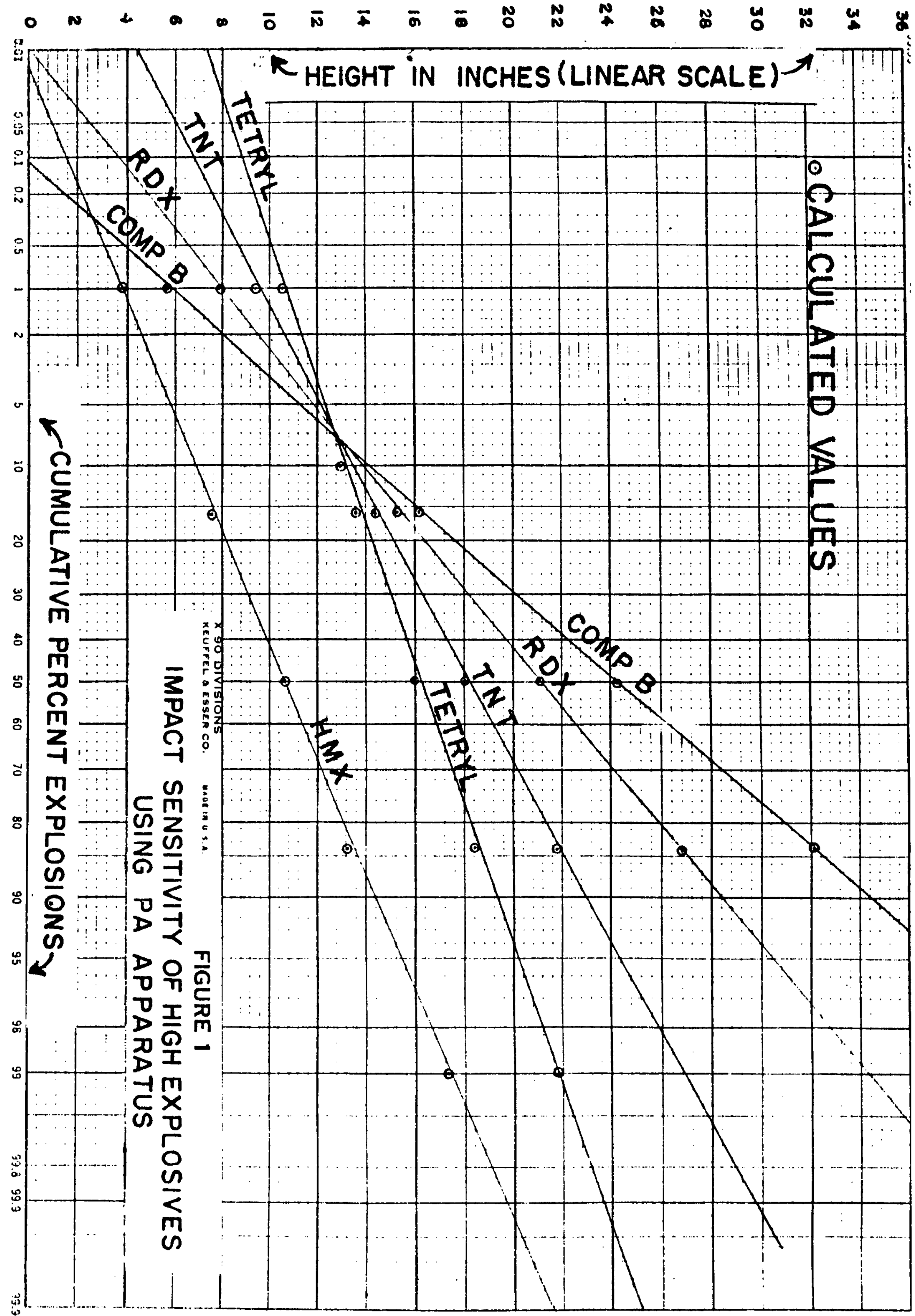


FIGURE 1
IMPACT SENSITIVITY OF HIGH EXPLOSIVES
USING PA APPARATUS

STATISTICAL APPROACH TO THE EVALUATION OF ELECTRIC INITIATORS

F. N. Lawrence
Picatinny Arsenal

I INTRODUCTION

A. Importance of Initiators in Explosive Round
B. Need for High Reliabilities
C. Complexity of Requirement Demands
D. Need for Applying Statistical Method

II SENSITIVITY CHARACTERISTICS OF INITIATORS

A. Definition of and Need for Sensitivity Analysis
B. The Perfect Initiator - Slide #1
C. The Actual Initiator - Slide #2
D. The "Population"

III THE NORMAL CURVE

A. Brief Description of Mean and Standard Deviation - Slide #3. Slide #4
B. Apparent Normality of Sensitivity Distribution
C. Importance and Uncertainty of Tails of Distribution - Slide #5

IV THE BRUCETON STAIRCASE METHOD

A. Application to Sensitivity Analysis
B. Description of Method - Slide #6
C. Disadvantages of Method
D. Method as Used for Sensitivity Curves - Slide #7

V COMPARISON OF INCREASED SAMPLE SIZE

A. Reasons for Doubting Sensitivity Curves
B. Design of Experiment
C. Results of Experiment - Slide #8
D. Conclusions

VI THE ACID TEST

A. Binomial Test
B. Conclusions

I. Introduction. Considerable attention is currently being focused on the evaluation of electric initiators. As the name implies, the initiator is the first element in an explosive train of a fuze attached to artillery shell, guided missile warhead, bomb, etc. Before the munition can be expected to function on target reliably, it must be ascertained that the fuze will function reliably and before the fuze can be expected to function reliably the initiator must be known to be capable of functioning reliably. Thus it can be seen that the very heart of the most complicated piece of munition is the initiator.

The Ordnance designers developing fuzes are therefore naturally desirous of learning as much about the characteristics of initiators which play such an important role in the fuze. Initiator designers have undertaken a program designed to analyze the capabilities of initiators. The task is made even more difficult by the fact that as modern warfare becomes more complex and costly, the consequences of a dud become more disastrous and therefore the functioning reliability required of initiators more closely approaches perfection. At the same time, the global tendency of modern warfare places added responsibilities on the evaluator. For now the item must be evaluated for all kinds of environmental conditions encountered in the various areas and climates of the earth.

Intelligent use of an item requires an understanding of the item. To obtain this understanding in the case of electric initiators, a series of tests are undertaken in an effort to determine what the item is capable of doing. This series of tests is referred to as the evaluation of the initiator. An attempt is made to investigate every aspect of the initiator, particularly under conditions and situations similar to those to which the round is expected to be subjected. The initiator may be expected to be subjected to extreme cold, extreme heat, dry or damp atmosphere, sand dust and many other environmental conditions. They are required to withstand approximately 20 years of storage and still be serviceable and to undergo varied forms of transportation and rough handling by personnel without becoming hazardous to use.

If each item could be tested to determine whether or not it could be expected to meet all of these requirements one could still not get a satisfactory answer since the effect of the test on the tested item would still have to be determined. In addition testing each item would involve tremendous expense in time and money. It has been shown that total testing involves a large enough human error to make sampling procedure much more accurate, and by virtue of the fact that it is less expensive, much more desirable. There is however one factor which eliminates any choice on our part. That is the fact that many of the tests involved are destructive tests. We therefore have no alternative but to use sampling procedure, or to be more exact, statistical methods.

II. Sensitivity Characteristics of Initiators. This paper attempts to discuss the application of statistical methods to the evaluation of electric initiators. I cannot hope to discuss all of the statistical applications in the short time allocated for this paper but will endeavor to analyze one of the more important areas; sensitivity.

Electric Initiators, as the name implies requires some form of electrical energy for operation. If the item is to perform reliably, the

the proper amount of electrical energy must be assured. However, in order to assure the proper amount of energy to the initiator one must first determine what this proper amount is. A coexistent problem stems from the fact that knowledge of the minimum amount of energy necessary for operation of the initiator will allow one to build safety features into the item. This characteristic of the item to require a certain amount of input energy for operation is known as its sensitivity.

A student in scientific fields soon becomes accustomed to that approach which starts with the discussion of ideal situations and perfect objects and then comes back to reality with situations somewhat less than ideal and objects that leave something to be desired. Using this approach one can begin by imagining the perfectly reproducible perfect initiator which requires exactly "v" volts at a given level of capacitance in order to function and fails to function if even a fraction less volts is supplied. Since it is perfectly reproducible every initiator functions exactly as any other of the same type. Such an initiator could be represented by the simplest type of functioning curve.

SLIDE I (Page 183)

With voltage on the abscissa and probability of functioning as the ordinate, the curve takes the form of a straight vertical line showing that for "v" or more volts the item has a functioning probability of 100 percent and for anything less than "v" volts its functioning probability is zero. For such an item the statistical problem would be simple indeed.

This is however a far cry from the real situation. To begin with, the true functioning curve of any specific initiator is unknown and since perfect reproduction is still a utopian dream no two initiators can be expected to have the same functioning curve. While these factors on the one hand complicate the statistical problem, on the other hand, they make statistical methods the only known practical means of handling the situation. Due to the fact that each item can be expected to have its own functioning curve we are forced to use probabilistic language in describing the functioning of them. Thus we say that the probability is such and such that the initiator will function if supplied with so much electrical energy at a particular level of capacitance. The functioning curve that we are forced to use to illustrate this actual situation deviates from the straight line curve shown above.

SLIDE II (Page 185)

This curve shows that there is some voltage v_1 below which the probability of functioning is zero and some other voltage v_2 above which the probability of functioning is 100 percent. Between these two values the probability of functioning increases gradually as the voltage is increased. At varying times we may be interested in any one of the several sub-sets of populations which may be described. Particularly of interest in this instance are the sub-set populations which may be described as; (1) all initiators made from a given design, (2) all initiators made from a given design by a given manufacturer or, (3) all initiators made from a given design by a given manufacturer within a specific time period.

For many reasons connected with design, production, acceptance inspection, safety and others, it is necessary to know the energy levels in terms of voltage and capacitance at which practically all of a given population, as the term is used above, will (1) function and (2) fail to function. Apparently the best way of determining these levels would be to first determine the functioning distribution of the items and then read off the functioning levels desired.

Determination of the functioning distribution curve of electric initiators falls into that class of analysis known as quantal responses or sensitivity data. It is characterized by items which are altered each time they are impulsed. Thus once an initiator has been impulsed at a given energy level it can only be observed whether or not it responded favorably or otherwise and then must be discarded.

Several methods of analysis have been used for this type of data. Success of these methods vary depending on what is desired and on the size of sample available.

III. The Normal Curve. The normal curve, one of the most common distributions in statistical theory is also one of the most loosely used distributions. Slide No. 3 and Slide No. 4 (cumulative). Much use is made of the fact that its symmetry about the mean permits statements concerning the percentage of the population to be expected within stated limits, based on the mean and standard deviation. Often overlooked is the fact that the method of moments used for deriving the parameters has an efficiency of 80 percent or more only within a relatively narrow range near the normal form in which Beta one;(B_1) the measure of symmetry does not exceed .1 and Beta two; the measure of kurtosis, lies between 2.65 and 3.42.

In problems such as this where the points to be determined are definitely in the tails of the distribution it comes almost naturally to look for a resemblance to the normal curve. In practice the distribution of the logarithm of the voltage necessary to cause detonation of initiators seems to closely approximate that of the normal curve. Unfortunately, because of the nature of the analysis necessary or because of the large numbers of initiators required to give even reasonably accurate results in the tails of the distribution, this statement of resemblance is based only on the middle portion of the curve.

It has been shown by Dr. Carl Hammer of the Franklin Institute that several curves, which will pass the X^2 test for goodness of fit, can be passed through the middle portion of this curve.

SLIDE V (Page 191)

This represents a family of curves all having a common mean and central distribution but showing the possibility of vast differences in the tails. However as pointed out above our entire analysis is directed toward determination of the all-fire and no-fire points which are in the tails.

IV. The Bruceton Staircase Method. The Bruceton Staircase method is one of the most widely used methods for determining all-fire and no-fire points.

Use of the method for finding these extreme points necessarily assumes normality of the distribution. In this method the first item is tested at a point near the expected mean. Successive items are tested at an increment higher or lower than the preceding item depending on whether or not the preceding one failed or fired.

SLIDE VI (Page 193)

A plot of a Bruceton test using X's for detonations and O's for failures might show that an initiator fired at the point expected to be the mean, was followed by a failure at a point one increment lower, then a failure at the expected mean and a fire at one increment above the mean, and so on.

This method has certain obvious weaknesses. Probably the most important of these is the one already touched upon, the assumption of normality. Others are; (1) the non-randomness of the choice of test levels - each test level depends upon the result of the last test, (2) the need for preliminary estimation of the mean for starting point and (3) the need for the preliminary estimation of the standard deviation for determination of the increment. When an item to be tested is known to be similar to one already tested a satisfactory estimate of the starting point and increment can be made based on the known item. When this is not possible a small sample of about fifteen or twenty items can be tested for an approximation prior to the real test. The precision with which these preliminary estimates are made and thus the effect on the efficiency of the test in this respect depends upon the test designer. In addition the effect of the dependence of one item upon the result of the previous item diminishes as the number of items is increased. A method based on the X^2 test for goodness of fit is used to test for normality of the distribution. However as has been pointed out before, this test is concentrated about the mean of the distribution. Very few items if any are ever tested in the tails. (SLIDE V) However, as has been shown by Dr. Hammer very little can be learned about the behaviour of the item in the tail from its actions in the area immediately around the mean.

From the results of a Bruceton test a mean functioning level, the voltage at which fifty percent of the item would be expected to function, and an estimate of the standard deviation is computed. Assuming a normal distribution the all-fire and no-fire points are then computed by a formula $V_a = \overline{v} + ks$. It can be noted that any error in the estimate of the standard deviation s will be multiplied by k, which in the case of all-fire and no-fire points (99.9 percent and .1 percent) is 3.09. In practice while estimation of the mean is fairly accurate even for small sample sizes, estimation of the standard deviation is quite erratic and depends on sample size as well as the increment used. For an accurate estimation of the two end points then, it is essential, if the Bruceton test is to be used, that the choice of interval size be good and that the sample size be adequate.

SLIDE VII (Page 195)

The current sensitivity curves of electric initiators showing all-fire and no-fire voltages for varying capacitances are based on Bruceton tests of samples of forty initiators. The computed means are plotted on logarithmic graph paper on which capacitance is varied on the abscissa and voltage on the ordinate (SLIDE VII). A smooth curve is then drawn through the points. The computed standard deviations are checked for trend and after it is seen that there is no correlation between capacitance and magnitude of standard deviation, that is, the standard deviation neither tends to increase of decrease as the capacitance increases or decreases, they are pooled. The resulting average standard deviation is then used in the formula $V_a = \overline{v} + ks$ to compute 99.9 percent and 0.1 percent points. These two points represent respectively the voltage necessary to insure that 99.9 percent of the initiators will detonate and that no more than 0.1 percent of the initiators will detonate.

V. Comparison of Increased Sample Size. Doubt concerning the accuracy of these curves had been aroused by several instances involving excessive failures above all-fire points and detonations below no-fire points. While these inconsistencies could have been due to other factors such as; (1) deterioration of the item due to time, or (2) differences in the firing systems used, it was felt that it should be investigated. The derivation and analysis of the Bruceton method indicate that while the method estimates the mean fairly accurately even with small sample size, the standard derivation is estimated very poorly with small sample sizes. Experience has shown this to be true.

It was decided to check the accuracy of the curves statistically. There were on hand some fifteen hundred electric initiators from a known lot, which had been analyzed and graphed in the Electric Initiator Handbook. Since the analysis on the lot had been made a year prior to this check, it was realized that any difference between the current analysis and the existing curves could possibly be due to deterioration. Another uncontrolled factor was the testing device since the Initiator Test Set on which the original tests were made was not available. However since there were those who claimed that the test set incorporated losses in energy and since the testing device used in this test was much simpler in design and calculated to eliminate most of the energy loss, this one test had the potential of answering still another question.

The level of capacitance called for in Arsenal Specifications, .0022 microfarads, at which no testing had been done in setting up the curves, and two points, .001 microfarads and .01 microfarads, which bracketed the first point and at which Bruceton tests had been run in setting up the curves were chosen as the levels at which the check tests should be run. It was proposed that a Bruceton Staircase Test of one hundred detonators be run at each level and these results compared with results gained from the Brucetons of forty initiators previously tested.

When this was accomplished it was found that the mean values compared very well. There was no significant difference between those computed on the basis of one hundred and those computed on the basis of forty (SLIDE VIII). This was as anticipated.

SLIDE VIII (Page 197)

However, the standard deviations presented quite a different picture. One of the two usable points passed the F-test, the other point showed a significant difference between the two values, and the pooled value corresponding to the way in which it is used in presenting the curve also showed a significant difference. The standard deviation as computed from samples of one hundred initiators gave a significantly larger estimate of the population standard deviation. This resulted in a higher value for the all-fire point and a lower value for the no-fire point.

The fact that the means compared so favorably seems to indicate that the cause of difficulty is not attributable to either deterioration of the item or differences in firing devices. The fact that the estimate of the standard deviation obtained from the samples of one hundred were consistently larger than those from samples of forty substantiates the belief that decrease of the sample size results in underestimation of the standard deviation beyond that which would be expected from normal curve reasoning.

VI. The Acid Test. It was then determined to check these all-fire points by firing a large enough sample to assure 90 percent confidence using the binomial method for attribute testing. To be 90 percent confident that an item would function with 99.9 percent reliability would demand zero failures from 2300 items. This large number of items was not available; therefore, it was not possible to conduct this test. It was considered satisfactory if the point proved to be at least the 99.5 percent point. Four hundred and fifty items giving zero failures would give 90 percent confidence that the functioning of the item was 99.5 percent reliable. Two samples consisting of 450 each were tested at .001 uf and .01 uf resulting in 11 failures and 23 failures respectively. It did not require a very careful analysis to see that these could not satisfy the 99.5 percent functioning reliability and certainly not 99.9 percent functioning reliability. The actual reliability of these two points was computed to be 97.4 percent with a 90 percent confidence band of from 96.3 percent to 98.3 percent with a confidence band of from 93.3 percent to 96.0 percent.

This meant that if the input energies obtained from the curves had been employed, 190 volts would have been supplied with a capacitance of .001 microfarad or 70 volts with a capacitance of .01 microfarads believing that no more than one failure would occur in 1000 items. Actually in the first instance anywhere from 2.7 to 4.7 failures in 100 or 27 to 47 failures in 1000 could have been expected.

Assuming that the distribution was actually normal, use was made of the values of the means, which seemed to be well approximated by all the tests. Substituting these mean values ($\bar{v}$), the value of the points computed by the binomial tests (V_a) and the Z constants (k) associated with the points, in the equation $V_a = \bar{v} + ks$, the value of the standard deviation (s) was solved for. These solutions proved to be extremely consistent, .15974 and .16482 giving an average value of .16228.

The value of the standard deviation was substituted back into the same equation with the value of the mean and the Z constant (3.09) for the 99.9 percent point and the 99.9 percent points then solved for.

SLIDE IX (Page 199)

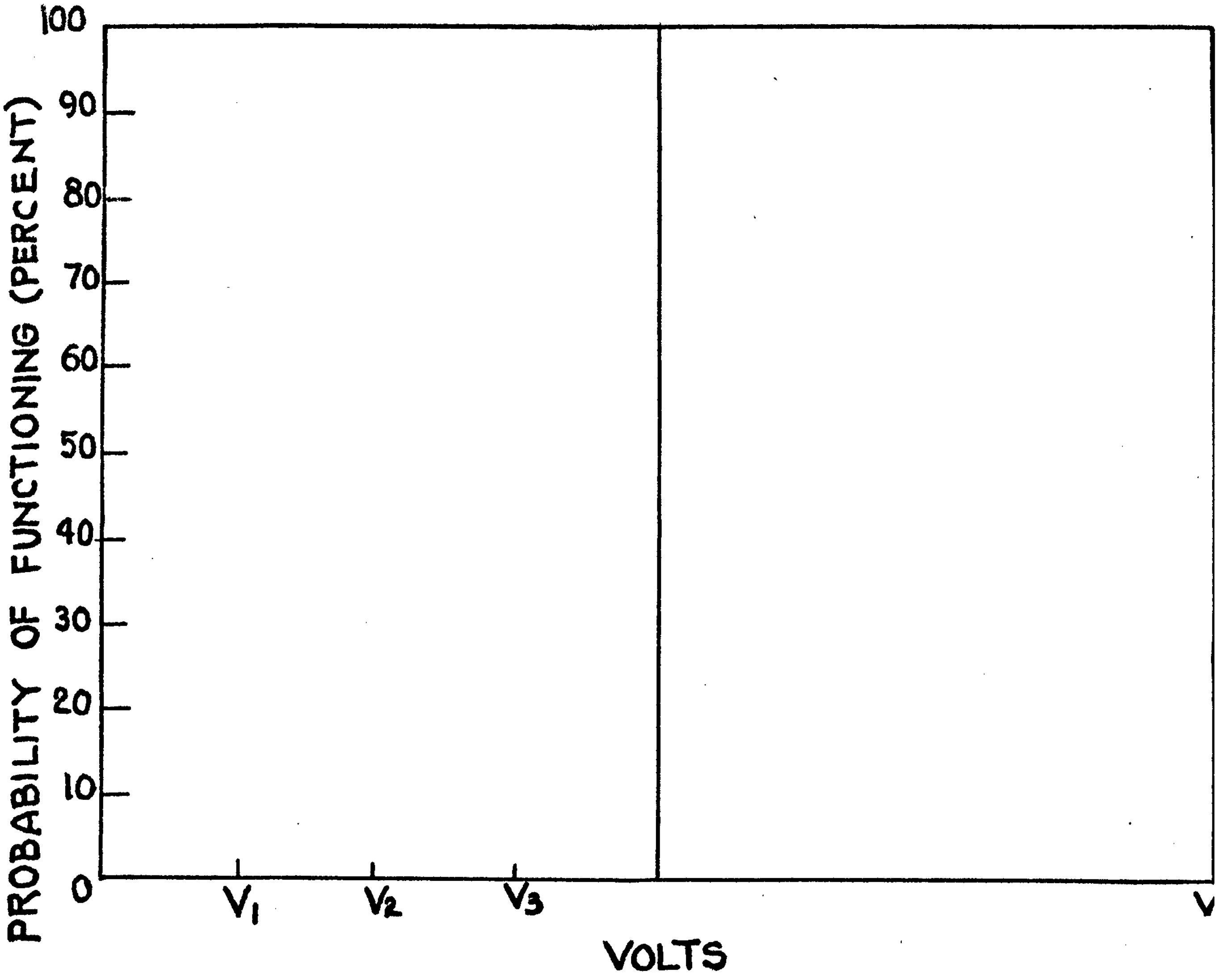
These values turned out to be 308 volts at the .001 microfarad capacitance level as compared with 240 volts given by samples of 100 and 190 volts given by samples of 40; and 117 volts at the .01 microfarad level as compared with 88.7 volts from samples of 100 and 70 volts from samples of 40.

If the values computed by the latter method are correct the error in the functioning curves are in the neighborhood of a 40 percent underestimation, actually 38 percent and 40 percent for the two levels of capacitance.

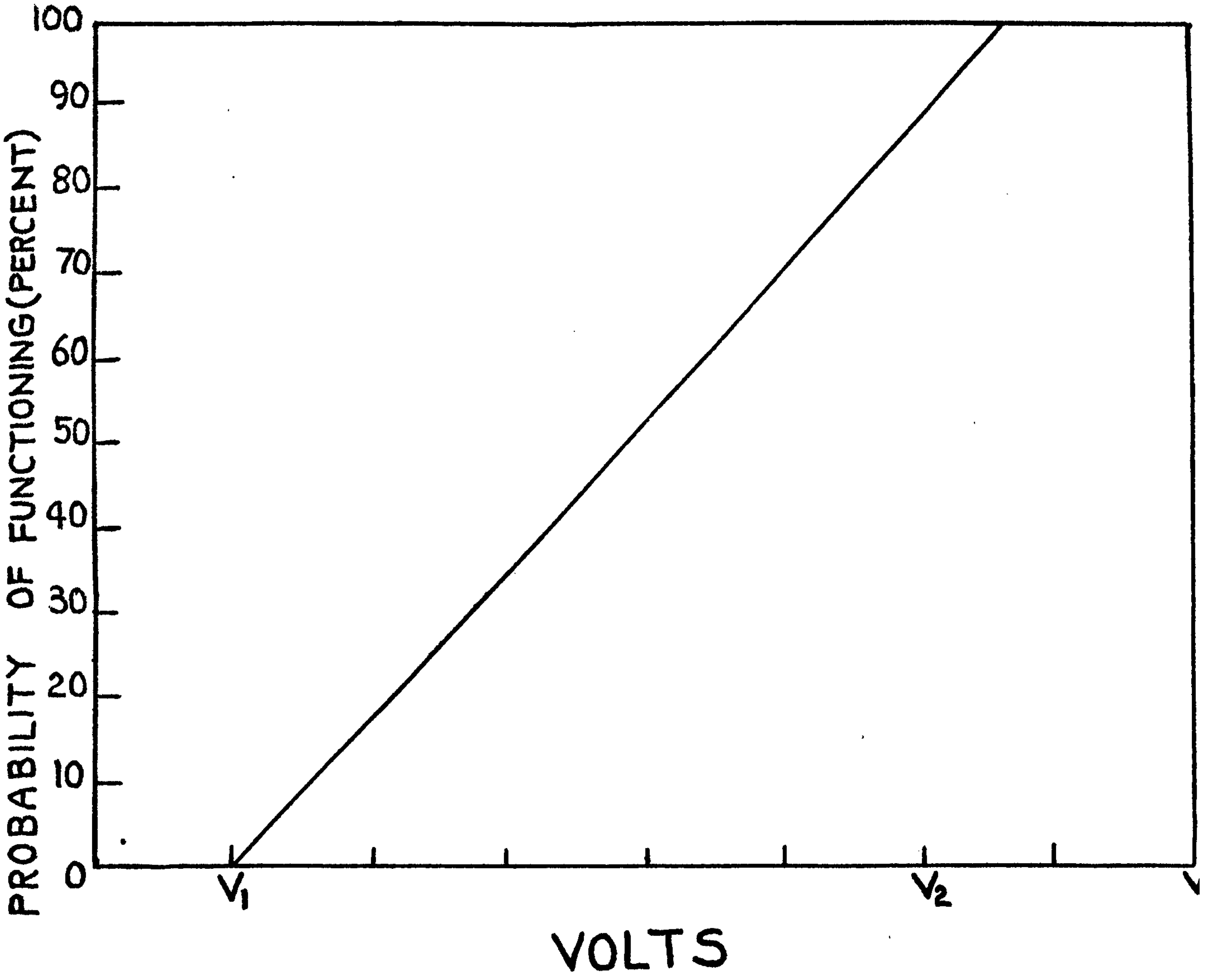
Any doubt regarding the validity of the results of this experimental design must be based on the degree of normality of the initiator distribution.

However the results of two Bartlett tests recently completed by Dr. Hammer of Franklin Institute indicate that the distribution is fairly normal. Most of the deviation from normalcy seems to be explained by the effect of duds.

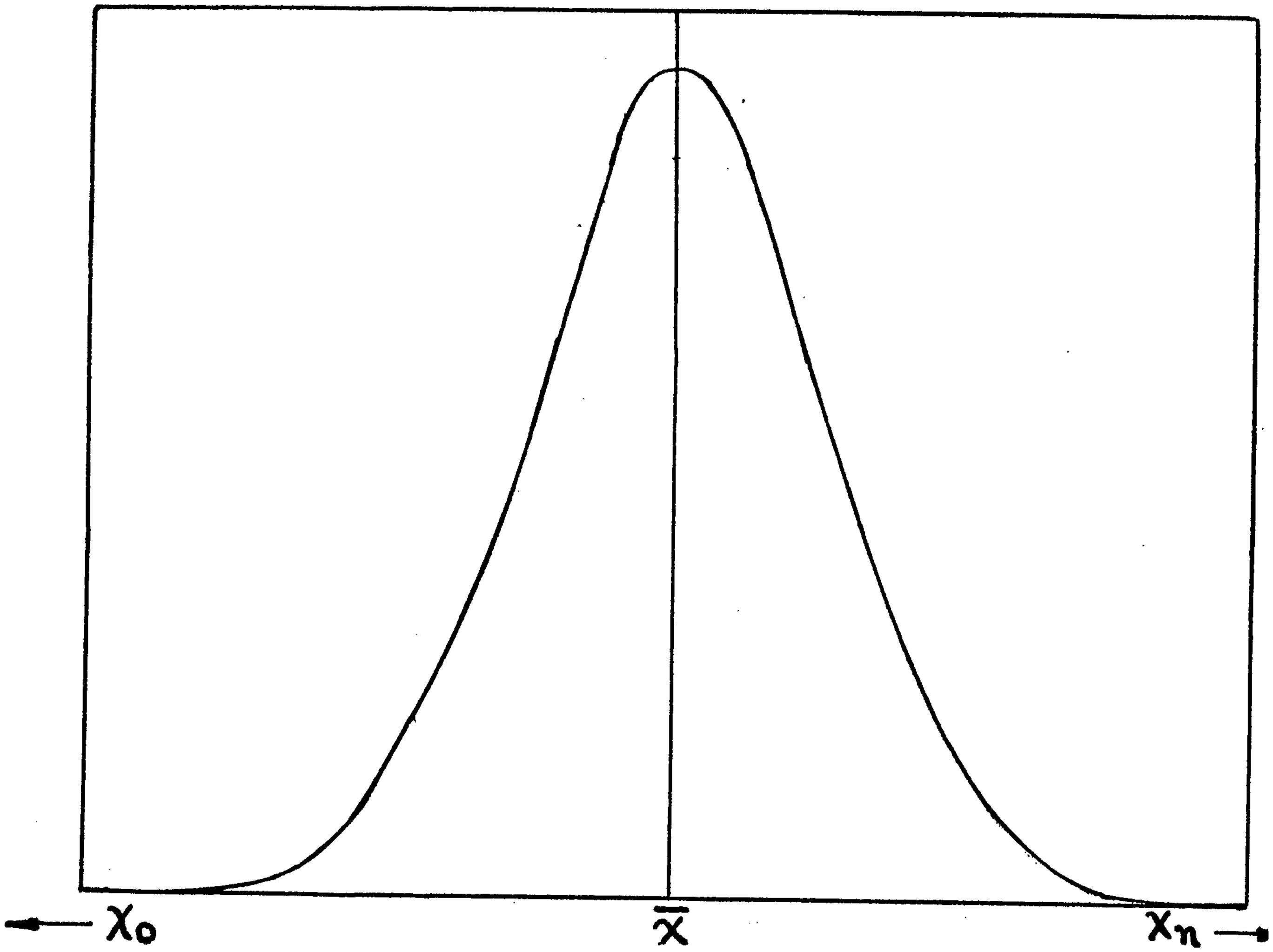
The problem of correcting the functioning curves and of obtaining a method of deriving future curves without expending prohibitive numbers of initiators is already underway and the outlook for satisfactory results is optimistic.


PROBABILITY OF FUNCTIONING (PERCENT)
100
90
80
70
60
50
40
30
20
10
0
V1
V2
V3
V
VOLTS

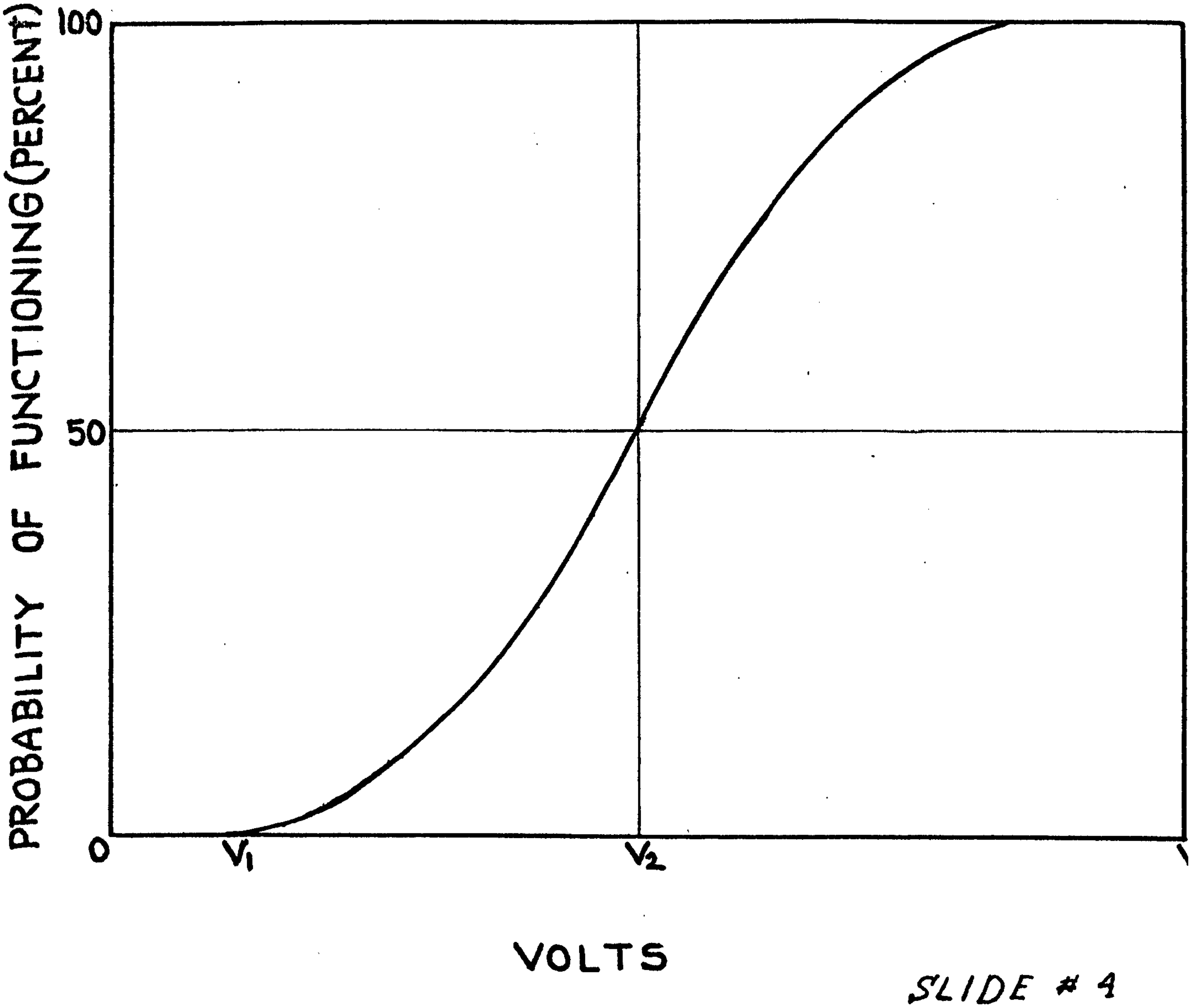
SLIDE #1


PROBABILITY OF FUNCTIONING (PERCENT)
100
90
80
70
60
50
40
30
20
10
0
V_1
V_2
V
VOLTS

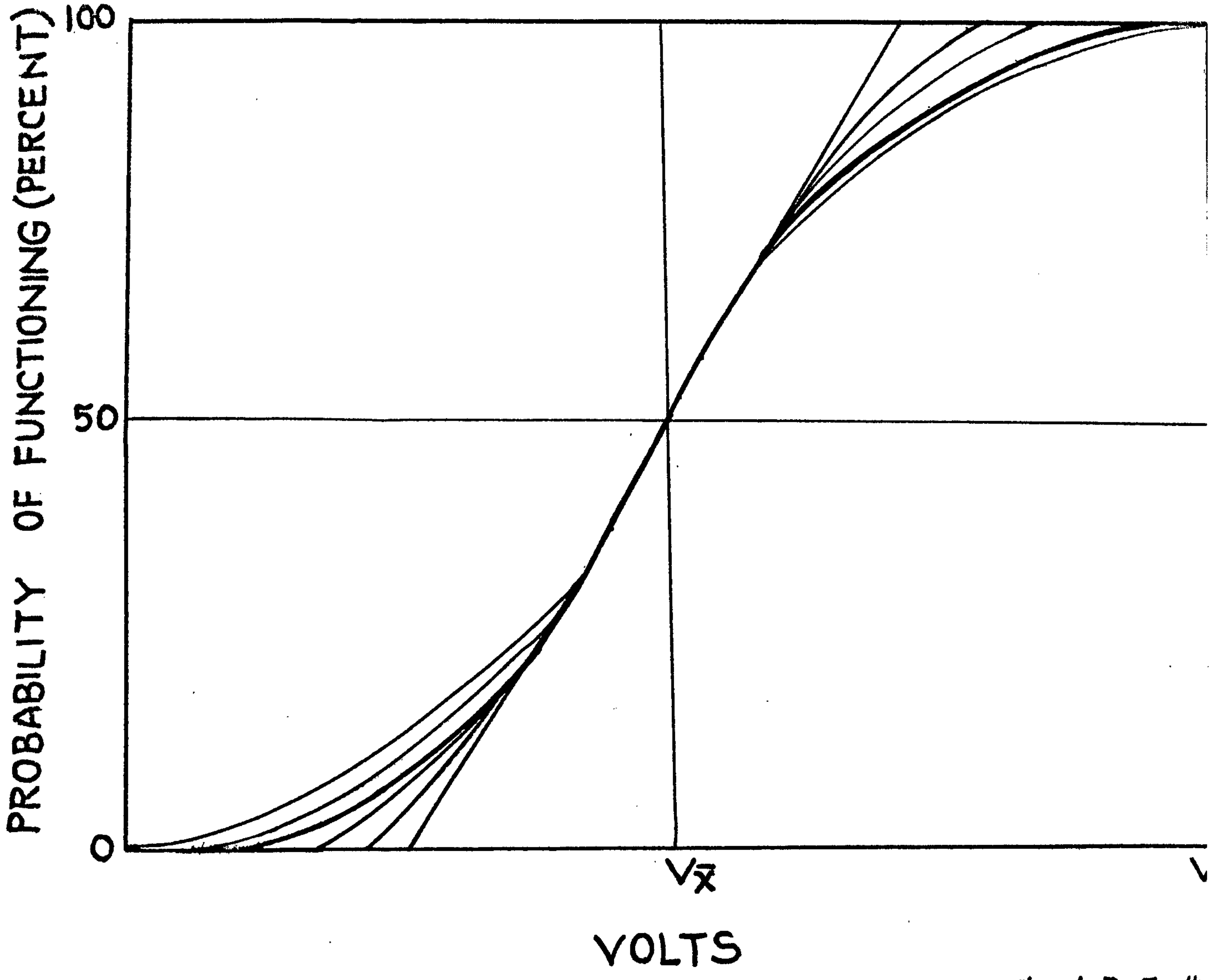
SLIDE # 2



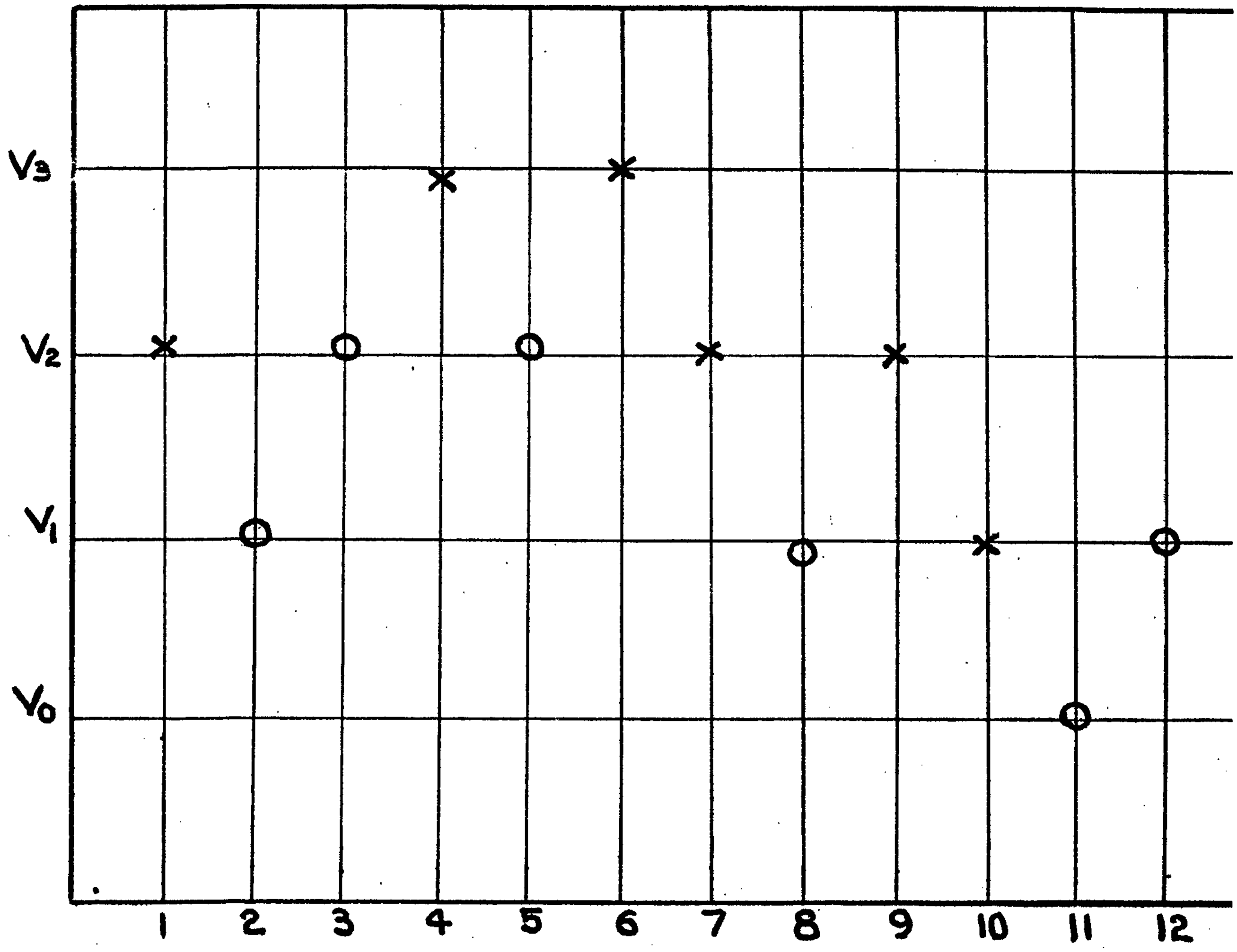
SLIDE # 3



SLIDE # 4



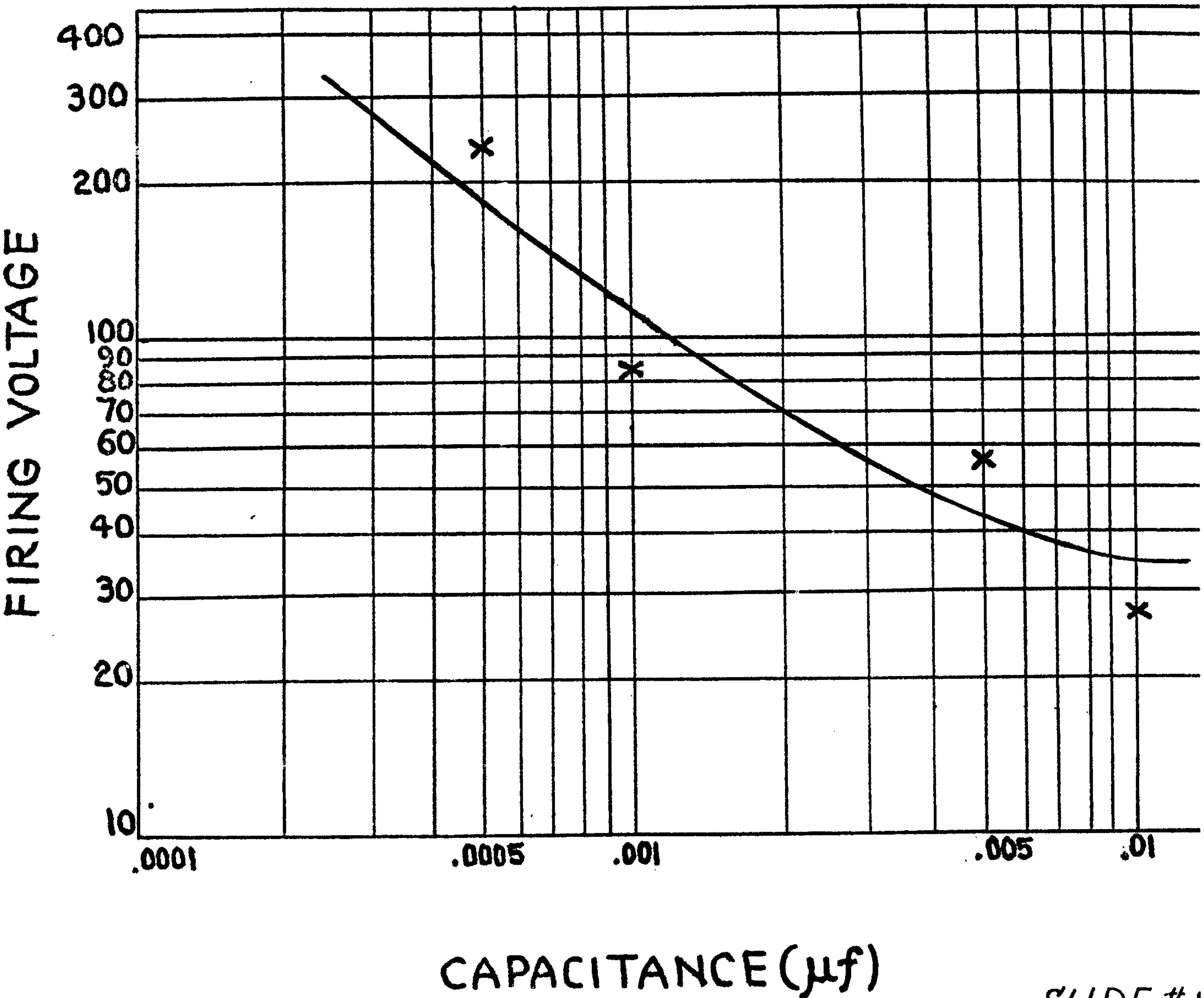
SLIDE #.



STAIRCASE RESULTS

X-FIRED
O-FAILED

SLIDE # 6


FIRING VOLTAGE
400
300
200
100
90
80
70
60
50
40
30
20
10
.0001
.0005
.001
.005
.01
CAPACITANCE(μf)

SLIDE #:

CAPACITANCE (μf)	MEANS (VOLTS) N=40	N=100	S.D. N=40	LOG. VOLTS N=100
.001	97	90	.101	.11689
.0022	70	57	--	.12832
.01	37	33	.052	.16856
		AVERAGE S.D.	.09281	.13792

	N=40	ALL-FIRE POINTS (VOLTS)		NO-FIRE POINTS (VOLTS)		
		N=100		N=40	N=100	
	POOLED	IND	POOLED	POOLED	IND	POOLED
.001	190	207	240	50	39	34
.0022	130	143	153	37	23	21
.01	70	110	89	19	10	12

EFFECT OF SAMPLE SIZE SLIDE # 8

CAPACITANCE μf	COMBINATION	N=100 (VOLTS)	N=40
.001	.308	240	190
.002	—	153	130
.01	.117	89	70
STANDARD DEVIATION	.16228	.13792	.09281

ALL FIRE POINTS

SLIDE # 9

SENSITIVITY TESTING

by

Sylvain Ehrenfeld

Engineering Research Section
Samuel Feltman Ammunition Laboratories

This talk is divided into two parts:

(I) Use of the Monte Carlo Method for comparing methods of Sensitivity Testing.

(II) A possible "quick and dirty" method for obtaining the standard deviation of estimates.

I. Use of the Monte Carlo Method for Comparing Methods of Sensitivity Testing

It is clear from the material presented by the previous speakers, as well as from most of the work done in this field, that the approach to the investigation of the comparitive efficiencies of different methods of sensitivity testing includes the following characteristics:

(1) The use of actual items for testing.
(2) The fact that the "true" characteristics of sensitivity curves are unknown.

Among the obvious disadvantages of the above are the following:

(a) Cost of items.
(b) Various methods cannot be compared to "true" situation.
(c) Uncontrollable physical factors might enter.

It is proposed to overcome some of the disadvantages by use of The Monte Carlo Method. The Monte Carlo approach is not new in this field,(1,2), but further acquaintance with an application of the method could be very useful and economical.

The Monte Carlo Method is partly based on the following theorem in probability theory: A series of observations from any known distribution or series of known distributions can be simulated by a table of random numbers.

To see how this might be done, let us examine how one could simulate a series of observations from a random variable X with cumulative distribution function (well behaved) F(x) namely;

$$\text{Prob}\{X \leq x\} = F(x)$$

Let U have a uniform distribution over the unit interval, namely,

Prob $\{U \leq u\} = u$, where $0 \leq u \leq 1$. The random variable U can be simulated by a random number table. Letting

$$S = F^{-1}(U) \quad , \quad F(S) = U$$

it can be shown that S has the same distribution as X

The proof of the above statement can be seen from the following;

$$\begin{aligned} \text{Prob}\{S \leq s\} &= \text{Prob}\{F^{-1}(U) \leq s\} \\ &= \text{Prob}\{U \leq F(s)\} = F(s) \end{aligned}$$

Thus, if we let $u_1, u_2, \cdots$ be values from a random number table, then,

letting $x_r = F^{-1}(u_r)$, $r = 1, 2, \cdots$

it is seen that $x_1, x_2, \cdots$ is a series of observations of the random variable X .

The idea indicated above can be generalized in many ways. The Monte Carlo technique can be applied for generating data which might come from the application of various sensitivity methods, such as the up-and-down method, and various other staircase methods. The methods described might, therefore, be very useful for answering various questions in sensitivity testing. Some of these questions are the following:

(1) How do many of the small sample size sensitivity methods compare for estimating percentage points?
(2) How sensitive are the various methods to distribution?
(3) How are the estimates for the various methods distributed?

The advantages of the Monte Carlo approach include the following:

(1) Various methods can be compared to a known value.
(2) No actual items are used.
(3) Various methods can be compared as to sensitivity to assumption of distribution.
(4) No uncontrollable physical factors enter.

II. A possible "quick and dirty" method for obtaining the the standard deviation of estimates

An important problem occuring in the application of various sensitivity methods comes from the fact that the standard deviation of the methods are often not known. Furthermore, there is in many cases no known way of estimating the standard deviation of the methods.

The foregoing considerations make it impossible in many cases to make tests of significance for the comparison of items.

A possible method for estimating standard deviations of methods might be to repeat the method several times, and use some function of the range to estimate the standard deviation.

Suppose x_1 and x_2 is a sample of size 2 from a population P.

Let the range R be defined by the following:

$$R = |x_2 - x_1|$$

In reference (3), a class of populations of the following form is considered:

$$f(x) = f\left(\frac{x-\mu}{\sigma}\right)/\sigma$$

where $f(x)$ is the density function, and μ and σ are the mean and standard deviation respectively. For population having densities of the above form, the standard deviation is proportional to the expected range. ($\sigma = k_P\, E(R)$). The subscript P in k_P indicates the dependence of the constant on the population P. The values of the constant for some populations are the following:

P		K_p
Rectangular	-	.866
Triangular	-	.875
Normal	-	.886
U-Shaped	-	.904
Parabolic	-	.807
Skewed	-	.875

The value of K_p does not seem to depend on p very much. Thus, a "quick and dirty" unbiased estimate of σ, when p is unknown, might be, $\hat{\sigma} = (.9) R$.

Thus, to estimate $\sigma_{\bar{x}}$ (where $\bar{x} = \frac{x_1 + x_2}{2}$), the value of $\hat{\sigma}_{\bar{x}}$ might be used, where

$$\hat{\sigma}_{\bar{x}} = \frac{(.9) R}{\sqrt{2}}$$

A similar procedure might be carried out for more than two observations.

In general, the procedure for comparing populations, using a method M, might be to repeat the method two times. Then $\bar{x}_1$, $\bar{x}_2$ and

$$\tilde{t} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\hat{\sigma}^2_{\bar{x}_1} + \hat{\sigma}^2_{\bar{x}_2}}}$$

are computed. Finally, a standard t-test, with appropriate degrees of freedom, is used. The Monte Carlo Method might be used to compare the distribution of $\tilde{t}$ with the t-distribution, and to find what degrees of freedom would best approximate the t-distribution.

The above method should not be used indiscriminately since further work is necessary to justify the methods. Furthermore, care must be taken to insure that other variations (e.g., the between-lots variation) do not affect the result.

References:

(1) "Staircase Methods of Sensitivity Testing", by T. W. Anderson, P.J. McCarthy, and J. W. Tukey. NAVORD Report 65-46, March 1946.

(2) "Effect of non-normality of Staircase Methods of Sensitivity Testing", by D. F. Votaw, JR. Statistical Research Group, Princeton University, May 1948.

(3) "Estimation of the Mean and Standard Deviation . by Order Statistics", Parts I and II, Vol. 25 (1954), pp 317-328 and Vol. 26 (1955), pp 505-511.

AN APPLICATION OF ANALYSIS OF VARIANCE TO THE EVALUATION OF THE EFFECT OF TEST VARIABLES AND REPRODUCIBILITY OF A NEWLY DEVELOPED LABORATORY APPARATUS

Kurt R. Fisch
Frankford Arsenal

Summary. An example is presented of the application of statisitical method to test the reproducibility of new laboratory equipment.

A mechanism has been developed which is capable of producing a grease film of uniform and reproducible thickness on steel rods. Apparatus and procedure of application are described. The device produces a film 35-70 microns thick depending on the grease. Measurements were carried out using six different types of grease, and the results were subjected to statistical analysis. The conclusions of the statistical results are given.

Apparatus. A schematic drawing of the device is shown in Figure 1. The top part of the cylinder (c) consists of a concial section (A), ending in the orifice (B). The top of (A) is a threaded portion (d) into which the cap (E) is screwed. The steel rod (F) is inserted into the specimen holder (G) and fastened by means of a recessed set screw.

In order for the device to function properly the following must be kept within tolerances:

(1) Fit between holder and cylinder

(2) Exact concentricity of the holes in (B) and (C)

(3) Rod diameter

Film Uniformity. The diameter of the rod was determined to the nearest .0001 inch (2.5 microns). The rod was then coated with the grease. The greas was completely removed from one side of the rod and the diameter of the rod plus the remaining grease coat was then determined with a traveling microscope (40X). This technique eliminates errors due to compensating non-uniformities in coat thickness on opposite sides of the rod.

Film thickness measurements were made on six greases (Table I. Tables be found at the end of the paper.), using four (4) different rods for each can grease. Four measurements were made on each rod at approximately 3/4 inch intervals. The rods were then cleaned, recoated, and three additional measurements were taken on each rod. The data are given in Table II. Numbers 1 to 4 represent the first set of measurements and numbers 5 to 7 the second set. The frequency distribution of the data is given in Table III.

In order to study the effect of variables in method and materials, the data were subjected to analysis of variance. Specific variables studied were:

a. Improper functioning of the mechanism, e.g., misalignment of the rod.

b. Differences in the greases, e.g., texture, consistency, and

c. Irregularities in the rod diameter.

Determinations 1 to 4 and 5 to 7 (Table II) were analyzed separately and the results are given in Table IV. The hypothesis H_o used was that the main effects and interactions had no significant effect on the result. The probability of a Type I error - the error of rejecting H_o when H_o is actually true - was set at .05.

The following conclusions may be drawn from the data in Table IV.

(1) The mechanism is capable of producing a uniform and reproducible film regardless of the grease or rod used.

(2) The thickness of the film may vary depending on the grease (i.e., texture, consistency, etc.), but without affecting the uniformity.

(3) The grease thickness may also be affected by the size of the rods, e.g., a grease which yields a thick film per se may produce a still thicker film when used in conjunction with an undersized rod.

There is, however, an approximate physical upper limit for the film thickness. This limit depends on the total clearance available for the grease (i.e., the difference in diameters between the orifice (B) and the steel rod). In the model used B = 9.233 ± .002 mm, and the average rod diameter varied between 9.100 - 9.134 mm, allowing a maximum of .045 - .067 mm (45 - 67 microns) for the grease coat. The measured total range (Table III) was 35 - 70 microns, with a superimposed measurement error of 2.5 microns.

TABLE I. Test Greases

Grease No.	Main Components	Specification
180	Li Soap - diester	U.S. Army 2-134
186	Li Soap - mineral oil	U.S. Army AXS-637
290	Bentone - diester	Experimental Sample
399	Li Soap - mineral oil/diester blend	MIL-G-3278
514	Silica gel - diester	Experimental Sample
540	Na soap - mineral oil	MIL-G-2108

Table II Grease Coat Thickness[a]

Grease No.	180				186				290				399				514				540			
Rod No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Determination No.																								
1	39	49	61	55	55	64	61	58	42	39	44	41	51	53	46	47	63	59	56	64	61	55	61	62
2	44	47	60	50	70	61	64	68	45	35	48	44	47	57	45	46	64	54	55	62	62	53	54	58
3	49	50	53	55	64	64	70	61	47	39	43	43	49	57	49	50	61	60	61	58	64	59	55	53
4	45	48	55	48	57	70	69	63	37	41	42	38	48	57	48	48	59	60	55	63	60	56	58	58
5	45	53	45	58	63	59	64	55	41	41	37	40	53	51	51	48	62	60	62	57	63	53	60	60
6	41	51	48	55	57	58	61	60	47	43	44	40	50	50	48	50	53	61	60	61	61	51	55	62
7	47	48	52	52	55	63	59	66	51	38	41	43	48	53	48	51	54	55	57	60	55	58	56	51
Mean per column $\bar{X}_c$:	44	49	53	53	60	64	64	62	44	39	43	41	49	54	48	48	59	58	58	61	61	55	57	58
Mean per grease $\bar{X}$:	50				62				42				50				59				58			
Grand mean $\bar{\bar{X}}$:	54																							
Overall Standard deviation $\bar{\bar{S}}$:	8																							

[a] - All measurements are in microns

TABLE III. Frequency Distribution of Grease Thickness Measurements

Thickness	Grease No.							
(Microns)	180	186	290	399	514	540	Total	%
31 - 35	-	-	1	-	-	-	1	0.6
36 - 40	1	-	8	-	-	-	9	5.2
41 - 45	5	-	15	1	-	-	21	12.5
46 - 50	10	-	3	17	-	-	30	17.9
51 - 55	9	3	1	7	6	10	36	21.4
56 - 60	2	6	-	3	11	10	32	19.0
61 - 65	1	12	-	-	11	8	32	19.0
66 - 70		7	-	-	-	-	7	4.2
Total	28	28	28	28	28	28	168	100.0
Range (Microns)	22	15	16	12	11	13	35	
Total range	35-70 microns							
Average range	15 microns							

TABLE IV. ANalysis of Variance

a. Determinations 1 to 4

Contribution*	Sums of Squares of dev.	Degrees of Freedom	Mean Square Deviations	F	(0.05) Significance
X	24.41	3	8.14	0.859	Not significant
Y	5140.45	5	1028.09	108.51	Significant
Z	22.16	3	7.39	0.780	Not significant
XY	161.97	15	10.80	1.140	Not significant
YZ	156.76	9	17.42	1.838	Not significant
YZ	827.22	15	55.15	5.821	Significant
XYZ	426.36	45	9.4747	1.000	------
Total	6759.33	95			

b. Determinations 5 to 7

X	21.00	2	10.50	0.741	Not significant
Y	2949.80	5	589.96	41.64	Significant
Z	18.16	3	6.05	0.427	Not significant
XY	110.83	10	11.08	0.782	Not significant
XZ	41.89	6	6.98	0.493	Not significant
YZ	348.25	15	23.22	1.639	Not significant
XYZ	425.00	30	14.1667	1.000	-------
Total	3914.93	71			

*X - measurements along rods

Y - measurements on different greases

Z - measurements on different rods

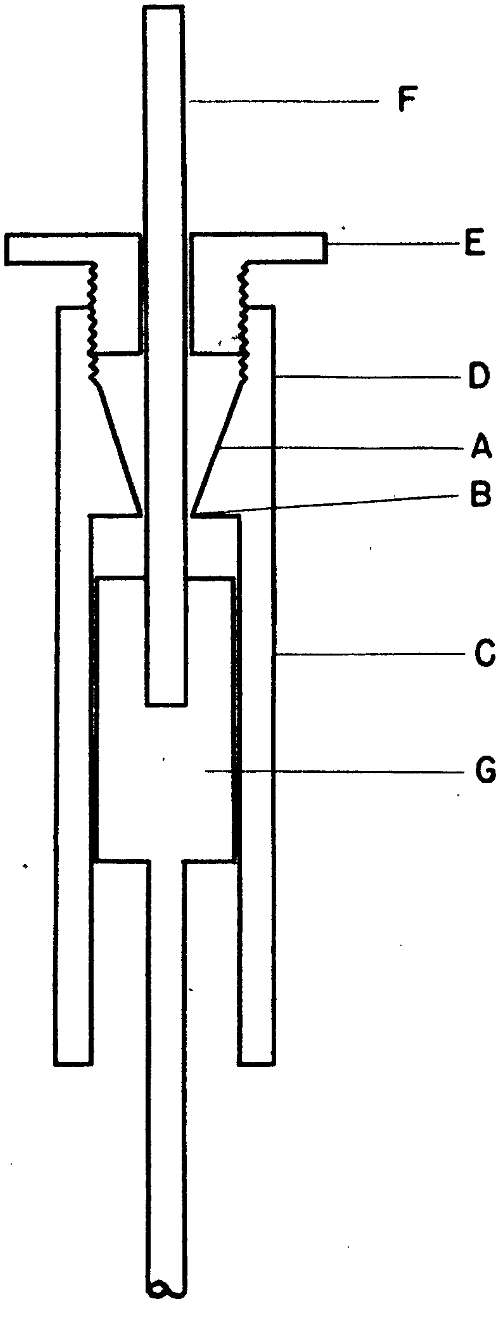


Figure 1. Grease Coating Device

A. Cone
B. Orifice
C. Cylinder
D. Flange
E. Cap
F. Rod
G. Holder

The Problem of Grouped Firing

Paul C. Cox
White Sands Proving Ground

The following problem is submitted not because the solution presents any difficulty, but because it has extensive application, and to the writer's knowledge the technique of analysis is not found in print and does not appear to be widely known.

The technique of firing missiles and rockets in groups (or pairs) is extremely important because a comparison can be made of the variation within groups and the variation among groups. This is useful to observe how much of the variability is a result of day to day variation, including effects of weather and other metro conditions, and how much is variability which is apparently due to unknown or uncontrollable causes. This is especially important if firing tables are used, because a comparison of the within and among variability should give some idea of how well the firing tables do their job.

These concepts have fairly wide application. For example: (1) Comparing twins with brothers or sisters who are not twins; (2) Comparing products of chemical mixes from the same batch with products from different batches which were mixed under the same conditions.

The technique of analysis will be illustrated by the following example: Suppose a certain rocket program calls for firing at three nominal slant ranges and with three levels of propellent temperature. The dependent variable is the azimuth coordinate of miss distance and three groups of three rounds each will be fired for each set of conditions. (Data from the same group of 3 rounds are placed together in a vertical column.)

	SR_1			SR_2			SR_3		
PT_1	-10	-22	-9	-5	-17	-4	11	-10	1
	-13	0	7	-9	6	13	-5	10	20
	14	-5	12	21	0	20	22	6	24
PT_2	-15	-25	-15	-14	-3	14	-9	8	14
	-17	-5	2	15	-1	5	-3	-2	18
	7	-11	5	-11	-20	-10	20	-15	-2
PT_3	-21	-26	-15	-18	-8	0	13	-5	-8
	-23	-8	-5	5	-26	-13	-9	-18	3
	0	-10	0	-10	-10	3	-13	-3	12

Table 1. Azimuth Component of Miss Distance for 3 groups of 3 simultaneous firings at each of 9 treatments (3slant ranges and 3 propellent temperatures.)

Treating this as a simple 3x3 factorial with 9 replications will give the following analysis:

Sources of Variation	D/F	SS	MS
SR	2	1540	770
PT	2	1568	784
SR X PT	4	24	6
Error	72	9586	133
Total	81	12718	

Table 2. Analysis of Variance of data in Table 1, assuming 9 replications.

To complete the analysis, the error term should be broken down into two parts, one to represent the within variation, the other to represent the among group variation. The procedure here is to take each cell and solve an analysis of variance, giving sums of squares for: (1) total; (2) within groups; (3) among groups. The sums of squares are then added for all cells giving the results listed in table 3. (Note that the total sum of squares in table 3 is the error sum of squares in table 2.)

Sources of Variation	D/F	SS	MS
Among Groups	18	1874	104
Within	54	7711	143
Total	72	9586	

Table 3. Division of the error term into within and among variation.

It is concluded that the firing tables appear to be doing a good job in keeping the variability down as a result of metro conditions and other day by day variation (i.e., variability from day to day is of the same order of magnitude as variability from simultaneous firings). On the other hand, there are significant differences in the value of the X coordinate as a result of both slant range and propellent temperature, suggesting certain biases exist under certain firing conditions, and these should be carefully investigated.

The most frequent situation will be groups of only two. When this is true the analysis will be greatly simplified as will be illustrated by the example of table 4. Here again each group is placed in a vertical column.

	SR_1	SR_2	SR_3	Totals
PT_1	20 18 18 19	23 24 21 25	31 30 32 27	288
PT_2	16 15 18 13	20 17 21 18	23 26 20 23	230
Totals	137	169	212	518

Table 4. Results of firing 2 pairs of rounds at each of 3 slant ranges and 3 Propellent Temperatures.

Treating this as a simple 2 x 3 factorial with four replications will give the analysis of table 5.

Sources of Var.	D/F	SS	MS
SR	2	254	177
PT	1	140	140
SR X PT	2	16	8
Error	18	66	3.7
Total	23	576	

Table 5. Analysis of Variance of the data of table 4. assuming 4 replications.

It is possible to find the sum of squares within pairs by taking one half the sum of squares of the difference of the two numbers in pairs, thus: $1/2 \left[2^2 + 1^2 + \cdots + 3^2 \right] = 24$. It is then possible to obtain the sum of squares between pairs by subtraction. However, since there are but two pairs in each cell, this could be computed by taking one fourth the sum of squares of the differences between the sums of pairs in each cell, thus: $1/4 \left[1^2 + 5^2 + 6^2 + 6^2 + 6^2 + 6^2 \right] = 42.50$.

The breaking up of the error term is demonstrated by table 6. Here it is seen that between pair variation is barely significant at the 5% level, indicating a difference between firing together or separately.

Sources of Var.	d/f	S.S.	M.S.	F.
Between Pairs	6	42	7.00	3.5*
Within Pairs	12	24	2.00	
Total Error	18	66		

Table 6. Breaking the Error Term of Table 5 into Between and Within Pairs (* Indicates significance at the 95% Level).

One final comment is that, in general, the appropriate error term to use to determine whether there exists a significant difference between slant ranges and propellent temperature is the within pair variation which would be 2.00 for the M. S. with 12 degrees of freedom in the last example.

APPLICATION OF SEQUENTIAL ANALYSIS TO CATAPULT TESTING

L. E. Stout, Jr.
Frankford Arsenal

Introduction. A survey of the acceptance testing technique used for experimental catapult systems indicated that a sequential testing procedure could be developed for use in catapult test programs.

The system in use consisted of firing 10 tests and evaluating each as a "go" or "no go" with regard to both velocity and acceleration. If all 10 were acceptable, the lot was accepted as good. The most information which could be obtained from these 10 firings was that there should be no more than 1.5 rejects per 100 firings if the 10 test catapults were taken from a sample consisting of 40-60 units.

A sequential test was developed, assuming an unknown mean and known sample variance, in which the testing efficiency was increased.

Using data on the M 1 catapult, a test was devised in which the expected test length was 7 runs and the acceptance limit was no more than one faulty unit per 1000 units. The details of the test procedure are presented in the paper as well as an example of its use.

Experimental Program. In order to evaluate the relative efficiencies of various statistical techniques, some representative experimental data were needed. No applicable data were available, so the writer consulted with Pfc A. Hess on the conduction of a test on the Ml catapult system. This test was primarily conducted for the purpose of evaluating various experimental acceleration measuring devices. However, the experiment was designed to give a quantative measure of the reproducibility of the entire testing procedure - as well as an indication of the experimental variance of each measurement. The results of the analysis of the various acceleration measurements is discussed in the Cad status report Mar-55 - 31 May 1955. The pertinent data are given in table I in the appendix. Three runs of 10 tests each were performed on 3 different dates. The data are given in table I. Analysis revealed that equivalent results were obtained on each of the 3 dates. This indicated that the entire test procedure was consistent and stable. This information was necessary before attempting to make any comparison of tests made on different days.

Since the effect of day to day variation was insignificant, the data were pooled to obtain an estimate of the population variance with 29 degrees of freedom for the various types of measurements. The following table contains the results:

Measurement	Variance	Std. Deviation	Degree of Freedom
Max acceleration (piezo gage)	0.308g	0.555	29
Max acceleration (thrust)	0.606g	0.780	29
Average velocity	1.83 ft/sec.	1.35 ft/sec.	29

The present test procedure for lot acceptance involves the firing of 10 system. If the 10 samples meet the specifications on minimum velocity and maximum acceleration, the lot is accepted. This type of test is termed a "go-no go" test. Each unit is either good or bad. Such tests reveal less information than tests involving the use of quantative measurements.

According to MIL-STD-105A pll to test that no more than one unit per Thousand is bad by "go-no go" sampling, 150 units should be tested out of a lot of 500-800 units. If one bad unit is discovered the lot should be rejected. This indicates the inadequacy of the present method in which 10 units out of small lots (less than 100 in some instances are tested).

A sequential analysis type test was devised, based upon methods presented in the book "Sequential Analysis" by A. Wald. If no more than 1 sample per 1000 is to exceed 20 g acceleration, the average of a sample should equal $20.0-3.08\sigma$.

Using the value obtained from the test of 30 runs σ- 0.55. Therefore, in order to meet the specifications that no sample have an acceleration greater than 20 g the mean of any lot should not exceed $20 - 0.55(3.08) = 18.3$ g.

Using equations of sequential analysis the Graph in Figure I (at the end of the paper) was made. The accept line was drawn at the 0.001 confidence level. This means that the actual system being tested will be acceptable only 1 time out of 1000 if the average value of g exceeds 18.3. The derivation of the equations used to calculate the lines on the graph are presented in the appendix. If the average value of g equals 17.8 the expected average number of tests required to complete a test is 7.

The graph in Figure I is used in the following way to test a group of catapults for acceptable performance with respect to the acceleration requirement. If no catapult is to exceed 20.0 g, the average of a test should not exceed 18.3 g. The actual test to be used is a test that the average value of g should not exceed 18.3. After each round is fired the term Σg is calculated by summing all the g values obtained up to and including the last test. This sum is plotted against round number, as shown on Figure I after each round. Using data in Table I for rounds 8-17

Round	Piezo g (acceleration)	Σg	
8	15.7	15.7	
9	15.0	30.7	
10	15.5	46.2	
11	14.6	60.8	
12	15.2	76.0	
13	15.9	91.9	
14	15.5	107.4	end -accept lot

Note that the point equivalent to Σg for 7 runs has crossed the accept line. The test is therefore terminated with this run and the lot of catapult units can be accepted.

A similar sequential analysis type of tests was derived for determining whether a lot of catapults meet a given minimum velocity specification. It is presented in the appendix of this report.

In conclusion, it can be pointed out that 7 test catapults, runs, when analyzed in a sequential statistical method yielded as much information about the lot meeting specifications for maximum allowable acceleration as 150 runs would yield when analysed in the "go-no go" method. Therefore the adoption of sequential analyses techniques is strongly recommended by the author.

TABLE I

Date	Round Number	Piezo Acceleration
11/9	8	15.7
11/9	9	15.0
11/9	10	15.5
11/9	11	14.6
11/9	12	15.2
11/9	13	15.9
11/9	14	15.5
11/9	15	14.3
11/9	16	15.6
11/9	17	15.4
11/10	18	14.5
11/10	19	14.7
11/10	20	14.7
11/10	21	15.9
11/10	22	14.4
11/10	23	15.4
11/10	24	15.0
11/10	25	15.6
11/10	26	15.2
11/10	27	15.1
11/15	28	13.3
11/15	29	14.8
11/15	30	15.6
11/15	31	14.8
11/15	32	15.1
11/15	33	15.0
11/15	34	15.1
11/15	35	15.8
11/15	36	15.1
11/15	37	15.3

APPENDIX

Acceleration Specification

Problem: calculate if $\theta < \theta'$ (any specified value)
θ represents any specified variable - acceleration in this case.
The preference for accepting a lot increases as θ decreases when $\theta < \theta'$.
It is now possible to find two values, θ_o and θ_1 ($\theta < \theta'$ and $\theta_1 > \theta'$) such rejection of the lot is considered an error if $\theta \leq \theta_1$ and acceptance of the lot is considered an error if $\theta \geq \theta_1$. For values of θ between θ_o and θ_1, no decision is made.

After the values of θ_1 and θ_o have been chosen, the tolerable risks can be expressed in the following way:

The probability of rejecting the lot when $\theta \leq \theta_o$ should be $\leq \alpha$
The probability of accepting the lot when $\theta \geq \theta_1$ should be $\leq \beta$
Let $(x_1 x_2 \ldots x_n)$ be a series of observations on X
The probability density of the sample if $\theta = \theta_o$ is given by:

$$P_{Om} = \frac{1}{(2\pi)^{m/2}\ \sigma^m} \quad \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{m} \quad (x_i - \theta_o)^2\right]$$

If $\theta = \theta_1$, the probability density is given by

$$P_{1m} = \frac{1}{(2\pi)^{m/2}\sigma^m} \quad \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{m} \quad (x_i - \theta)^2\right]$$

The probability ratio P_{1m}/P_{Om} is computed after each run. Additional runs are taken if

$$B < \frac{P_{1m}}{P_{Om}} < A$$

Testing is ended with acceptance of the lot if $\frac{P_{1m}}{P_{Om}} \leq B$

Testing is ended with rejection of the lot if $\frac{P_{1m}}{P_{Om}} \geq A$

A and B are given by the following approximate formulas

$$A = \frac{1 - \beta}{\alpha}$$

$$B = \frac{\beta}{1 - \alpha}$$

These were obtained in the following manner:

Let H_o be the hypothesis that $\theta < \theta_o$

H_1 be the hypothesis that $\theta > \theta_1$

(1) the probability of accepting H_1 when $\theta < \theta_o$ is α

(2) the probability of accepting H_o when $\theta < \theta_o$ is $1 - \alpha$

(3) the probability of accepting H_o when $\theta > \theta_1$ is β

(4) the probability of accepting H_1 when $\theta > \theta_1$ is $1 - \beta$

The hypothesis H_o is accepted when

$$\frac{P_{1m}}{P_{om}} \leq B$$

for convenience, dropping the subscript m

$$P_1 < BP_o$$

from condition (3) above P_1 in this case equals β

from condition (2) above P_o in this case equals $1-\alpha$

therefore $\beta \leq B\ (1-\alpha)$

$$\text{or} \quad B \geq \frac{\beta}{1-\alpha}$$

Similarly, H_1 is accepted where

$$\frac{P_1}{P_o} \geq A$$

from condition (4) P_1 in this case equals $1 - \beta$

from condition (1) P_o in this case equals α

therefore $\frac{1 - \beta}{\alpha} \geq A$

By taking the logarithm and simplifying, the equations become

$$\ln \frac{\beta}{1-\alpha} < -\frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_i - \theta_1)^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_i - \theta_o)^2 < \ln \frac{1 - \beta}{\alpha}$$

$$\text{or} \quad \ln \frac{\beta}{1-\alpha} < -\frac{1}{2\sigma^2} \Sigma (x_i^2 - 2x_i\theta_1 + \theta_1^2) + \frac{1}{2\sigma^2} \Sigma (x_i^2 - 2x_i\theta_o + \theta_o^2) < \ln \frac{1-\beta}{\alpha}$$

$$\text{or} \quad \ln \frac{\beta}{1-\alpha} < \frac{\theta_1 - \theta_o}{\sigma^2} \Sigma x_i + \frac{m}{2\sigma^2} (\theta_o^2 - \theta_1^2) < \ln \frac{1 - \beta}{\alpha}$$

Adding $-\frac{m}{2\sigma^2}(\theta_o{}^2 - \theta_1^2)$ to inequalities and dividing by $\frac{\theta - \theta_o}{\sigma^2}$

$$\frac{-\frac{m}{2\sigma^2} + \ln\frac{\beta}{1-\sigma}}{\frac{\theta_1 - \theta_o}{\sigma^2}} < \Sigma x_i < \frac{-\frac{m}{2\sigma^2}(\theta_o^2 - \theta_1^2) + \ln\frac{1-\beta}{\alpha}}{\frac{\theta_1 - \theta_o}{\sigma^2}}$$

or $$\frac{\sigma^2}{\theta_1-\theta_o}\ln\frac{\beta}{1-\alpha} + m\,\frac{\theta_o+\theta_1}{2} < \Sigma x_i < \frac{\sigma^2}{\theta_1-\theta_o}\ln\frac{1-\beta}{\alpha} + m\,\frac{\theta_o+\theta_1}{2}$$

therefore if $\Sigma x_i < \frac{\sigma^2}{\theta_1-\theta_o}\ln\frac{\beta}{1-\alpha} + m\,\frac{\theta_o+\theta_1}{2}$, we accept lot

and if $\Sigma x_i \quad \frac{\sigma^2}{\theta_1-\theta_2}\ln\frac{1-\beta}{\alpha} + m\,\frac{\theta_o+\theta_1}{2}$, we reject lot

A graph such as Figure 1 can be obtained from these equations.

The slope of the lines = $\frac{\theta_o + \theta_1}{2}$

The intercepts are $\frac{\sigma^2}{\theta_1 - \theta_o}\ln\frac{\beta}{1-\alpha}$ and $\frac{\sigma^2}{\theta_1 - \theta_o}\ln\frac{1-\beta}{\alpha}$

SEQUENTIAL ANALYSIS

Illustration

Sequential analysis

$\alpha = 0.05 \qquad \theta_1 = 18.3$
$\beta = 0.001 \qquad \theta_o = 17.3$

using values for acceleration $\sigma_{av}^2 = 0.308$

$$s = \frac{\theta_o + \theta_1}{2} = \frac{17.3 + 18.3}{2} = 17.8$$

$$L_o \text{ intercept} = \frac{\sigma^2}{\theta_1 - \theta_o}\ln\frac{\beta}{1-\alpha}$$

$$= \frac{0.308}{1}\,2.3\log\frac{0.001}{.95} = -2.109$$

L_1 intercept (reject line) $= \frac{\sigma^2}{\theta_1 - \theta_2}\ln\frac{1-\beta}{\alpha}$

$$= \frac{.308}{1}\,2.3\log\frac{.999}{.05} = 0.922$$

for simplified case, $E_{s(n)}$, the expected number of runs needed to complete the sequence (if $\theta = \frac{\theta_1 + \theta_o}{2}$) is

$$\frac{L_o \times L_1}{\sigma^2} = \frac{(2.11)\ (0.922)}{0.308} = 6+ \text{ or } 7$$

Velocity Specification

A similar deviation was made for the case when, knowing σ it is desired that

$\theta > \theta'$ (any specified value).

Define the region of indifference $(\theta_o < \theta' < \theta_1)$

Accept the lot of samples if $\theta > \theta_1$

Reject the lot of samples if $\theta > \theta_o$

Define α = the probability of rejecting the lot if $\theta > \theta_1$

Define β = the probability of accepting the lot if $\theta < \theta_o$

$$P_{om} = \frac{1}{(2\pi)^{m/2}\ \sigma^m} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_i - \theta_o)^2\right]$$

$$P_{1m} = \frac{1}{(2\pi)^{m/2}\ \sigma^m} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_i - \theta_1)^2\right]$$

If $B < \frac{P_{1m}}{P_{om}} < A$, an additional run should be taken.

If $\frac{P_{1m}}{P_{om}} \leq B$ reject the lot

If $\frac{P_{1m}}{P_{om}} \geq A$ accept the lot

Let H_o be the hypothesis that $\theta < \theta_o$

Let H_1 be the hypothesis that $\theta > \theta_1$

(1) The probability of accepting H_o when $\theta > \theta_1$ is α

(2) The probability of accepting H_1 when $\theta > \theta_1$ is $1 - \alpha$

(3) The probability of accepting H_1 when $\theta < \theta_o$ is β

(4) The probability of accepting H_o when $\theta < \theta_o$ is $1 - \beta$

H_1 is accepted when $\frac{P_1}{P_o} \geq A$

In this case $P_1 = 1 - \alpha$ from condition (2)

In this case $P_o = \beta$ from condition (3)

therefore $\frac{1-\alpha}{\beta} \geq A$

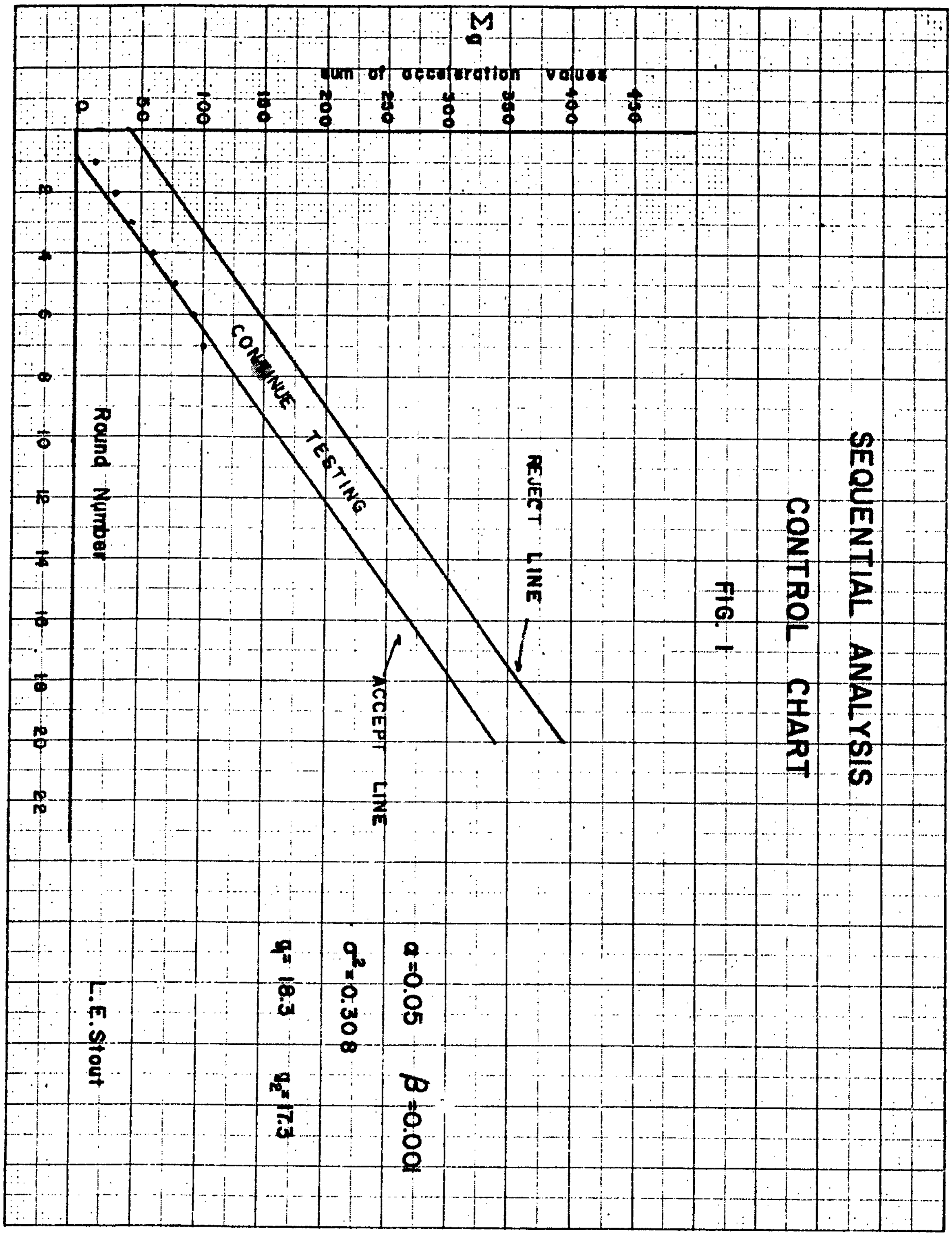
H_o is accepted when $\frac{P_1}{P_o} \leq B$

From condition (1) $P_1 = \alpha$ in this case

From condition (4) $P_o = 1 - \beta$ in this case

therefore $\frac{\alpha}{1-\beta} \leq B$

From this information it can be seen that the rest of the derivation will be identical to the previous one presented for the case where $\theta < \theta'$. The graph of this case will obviously have the accept and reject lines reversed.



SEQUENTIAL ANALYSIS CONTROL CHART

FIG. I

Estimating An Average Or Standard Trajectory

Paul C. Cox
White Sands Proving Ground

Let us assume that a sample of N rockets are to be fired under conditions which are controlled as much as is reasonably possible; and if non standard conditions creep in, every effort will be made to strip out the effects of such conditions. The problem to be presented is, what would be the best way to amalgamate the data so that an average or standard trajectory may be estimated. (It is to be assumed the trajectory will be estimated with empirical data. This particular problem is not concerned with computing the trajectory from the equations of motion of the rocket).

This same idea has a great many other applications, for example: (1) Take the data from several successive years and estimate average climatic conditions or average business conditions over a certain season; (2) Take the amount of wear or fatigue from several machines of the same type, and from this data estimate the expected wear for that type of machine or equipment.

The following suggestions are offered for discussion and consideration as possible methods of attack:

(1) One technique would be to take a set of points from each trajectory, put all points from all trajectories together, and compute a polynomial, or a sequence of polynomials, by accepted methods. Intuitively this method does not seem right because of the dependence which exists among points of the same trajectory and the independence of the points from different trajectories.

(2) A second technique would be to find the mean value of the sample of trajectories at certain points, then fit a curve to these average values. Unfortunately, it is doubtful whether the known coordinates for the trajectories in the sample of missiles will all have the same reference points. This would require some form of interpolation to obtain the desired coordinates.

(3) A third technique would be to compute a curve for every flight by orthoganal polynomials, or by some similar plan, arbitrarily select a set of points for the dependent variable, and compute the corresponding values for the independent variable for each trajectory. For each point along the abscissa, estimate the mean value for all trajectories, and fit a curve to these mean points.

(4) A fourth technique would be to compute polynomials to the same degree for all flights, and then compute the mean value for all polynomial coefficients to obtain the average trajectory.

The problem may now be expressed as the determination of that technique from the four mentioned above which has the most merit, or perhaps the formulation of another technique which is still more desirable. We are interested , among other things, in obtaining confidence bounds on the average trajectory or on the coefficients of the polynomials. It would be desirable to know which of the appropriate methods, if there is more than one, would give the smallest valid confidence region for the trajectory.

The solution offered here represents the combined suggestions of both the panel members and the participants in the clinical session. Before discussing the solution I would like to mention a comment made by Dr. Churchill Eisenhart in which he drew an analogy between this problem and the wear on automobile tires. In particular, Dr. Eisenhart pointed out that tire wear will probably be smooth and even until at some time when brakes are applied abruptly, there will be a large and instantaneous increase in the wear.

The basic solution is primarily the work of Dr. John Tukey. The steps in the solution are largely as follows:

(1) Restrict the study to a portion of the trajectory. It is hoped this portion will be nearly homogeneous.

(2) After each portion has been studied and suitable estimates made, a study should be made of the connecting links between the successive portions.

(3) Certain abrupt changes in a trajectory actually may occur as a result of either external or internal conditions which affect the rocket. These should be expected and a study should be made to determine their causes.

(4) In the region under consideration, points should be selected dividing the abscissa into equal intervals. At each of these points, the value of the trajectory should be ascertained. If these values can be obtained from the raw data, that would be most desirable; If not, perhaps the third technique which was discussed early in this presentation should be used.

5. Using this data an analysis of variance should be worked, and the linear, quadratic, cubic, etc. effects should be removed (If a polynomial does not seem appropriate, then modify the approach accordingly.). By these methods, it should be possible to obtain some idea of the curve which the trajectory follows in this region, and also to obtain an estimate of the variance. This variance should apply to the entire region because this region was chosen sufficiently small that the variance should be nearly homogeneous. This variance will be useful in estimating what limits should contain K% of the trajectories. It will also be useful in obtaining a confidence bound for the estimated curve in the region. This confidence bound should be very closely related to the confidence interval for the overall mean in the analysis of variance, inasmuch as the variance should be very nearly homogeneous and the effects of curvature should be removed from the analysis.

6. It is inevitable that the question of independence will be raised, inasmuch as this is one of the principal assumptions of analysis of variance, and quite obviously successive points on a trajectory will not be independent of one another. It is believed, however, that if the linear, quadratic, cubic, etc. effects are removed the lack of independence will not be serious. Dr. Tukey warned about the danger of extending this idea to an analysis of variance to study the effect of time upon some variable in which months are used as one set of treatments and years as the other set. The difficulty with this is that December and January are actually very close together, but they are at opposite extremes in the analysis of variance.

LONG TERM EXPOSURE TESTS OF VARIOUS ORDNANCE MATERIALS

S. L. Eisler
Rock Island Arsenal

During the development of many new Ordnance materials the experimenter must, at one time or another, seek the answer to one or more of the following questions:

a. What correlation can be made between service conditions and accelerat laboratory tests?

b. How does the new material compare with conventional types as to aging resistance or protection afforded?

c. May the new material be used in combination with other materials?

The answers to the above questions can only be obtained by a series of long term exposure tests. These tests may vary in the type of exposure used but have in common the long time factor of from one to ten years. This factor alone makes it essential that the series of tests be so planned as to provide the maximum amount of information which may in turn be statistically analyzed.

This type of problem is common to several sections of the Rock Island Arsenal Laboratory including the rubber, rust preventive, packaging material and metal finishing sections. Therefore, we are very anxious to develop a standard plan which may be used as a pattern for all long term exposure tests to accomplish the purpose outlined above.

Let us take the following packaging material problem as an example, since it is typical of the problem where the experimenter desires to study all variables under all conditions and generally decides that the number of tests required is too great. Such a problem might have the following variables:

4 types of barrier material used in the form of bags

2 weights of polyethylene film in each type of barrier material

3 types of vapor corrosion inhibited papers used as liners in the bags

3 exposure conditions

6 exposure periods

3 replications

It may readily be seen that a total of 1296 samples would be required for this problem alone. However, when one considers that there may be other types of materials or other combinations which should also be investigated the number of samples increases tremendously. In addition, when one considers that from three to five different tests are conducted for each sample the amount of work involved becomes excessive.

Therefore, it is hoped that a simpler method such as a two level factorial experiment may be used as a preliminary screening test prior to designing a complete factorial exposure experiment. This problem has not received a great deal of emphasis up to the present time and is being presented at this time in the hope that some of you may have had some experience in designing such experiments. We certainly are interested in any plan which will reduce the number of tests without any resultant decrease in the significance of the results.

DETERMINING THE EFFECTIVENESS OF CUTTING OILS IN REDUCING MACHINE TOOL WEAR

S. L. Eisler
Rock Island Arsenal

This work is a continuation of work previously done at R.I.A. to determine the feasibility of using radioactive tools to determine tool wear. The method, as developed, utilizes the activity of the tool wear products transferred to the chips during the machining operation as a measure of the tool wea

This particular experiment was designed to evaluate six cutting fluids using only one type of stock. Six tools of the same composition and design which had been made radioactive in the nuclear reactor at Oak Ridge were available for use. Each tool had four cutting edges which could be used for the experiment. In addition, it was planned to study the effect of cutting speed upon the efficiency of the oils.

This problem was discussed with Dr. E. H. Jebe of the Statistical Laboratory, Iowa State College, who was assigned to R.I.A. on temporary duty as a Reserve Officer for a short time last summer. He suggested a latin square design since there were to be six tools and six oils used in the experiment. The oils, as the principal treatment of interest would be randomized in the rows and columns of the square. The six columns were designated tools 1 to 6. Since only four edges were available on each tool it was necessary to designate one edge for each cell and a half of the square. See Fig. 1. The latin square arrangement selected was randomized for both tools and edges.

The four speeds were randomized within each cell. This resulted in 144 tests for the entire experiment. The speeds are to be considered as a split plot within the latin square design.

The analysis of variance developed from the data obtained is shown in Figure 2. It will be noted that both oils and tools showed a significant difference. This was to be expected between oils but not between tools. Since the tools were all cut from the same stock, with the same angles etc., and irradiated together for the same time, a difference in tools is difficult to explain. However, if the experiment had not been designed to provide this analysis, a difference in tools would not have been detected. On the basis of this test of significance, tools must be considered a possible source of variation for all future experiments.

The principle objective of this work was to rate the six oils as to their efficiency in reducing tool wear. Unfortunately this was not possible with the data obtained. Three oil - water mixtures provided total values from 7805 to 8356 while the three undiluted oils provided total values from 4486 to 4974. (The higher the value the greater the amount of wear.) Efforts to determine significant differences between the means in one group or the other proved unsuccessful.

This indicates that the test error is too large to be able to measure the small differences between similar oils.

Perhaps, we have overlooked some other method of analysis, and if so, we would certainly appreciate hearing about it at this time.

Tool No.

		1	2	3	4	5	6
Edge 1	Z Y W X	Oil C	Oil D	Oil E	Oil F	Oil B	Oil A
	W Z Y X	Oil E	Oil F	Oil C	Oil A	Oil D	Oil B
Edge 2	Y W X Z	Oil A	Oil C	Oil D	Oil B	Oil F	Oil E
Edge 3	Z X W Y	Oil B	Oil E	Oil F	Oil C	Oil A	Oil D
	Z X Y W	Oil D	Oil A	Oil B	Oil E	Oil C	Oil F
Edge 4	W Z Y X	Oil F	Oil B	Oil A	Oil D	Oil E	Oil C

Latin Square Arrangement

Figure 1

ANALYSIS OF VARIANCE

LATIN SQUARE

	S.S.	d.f.	M.S.	F.
Tools	241,593	5	48,319	9.03***
Edges	23,161	5	4,632	.87
Oils	735,407	5	147,081	27.63***
Error (L.S.)	106,976	20	5,349	
Total (L.S.)	1,107,137	35		

SPLIT PLOT

Speeds	4,117,348	3	1,372,449	316.7***
O x S	288,133	15	19,209	4.43***
Error (b)	390,030	90	4,334	
Total	5,902,648	143		

Figure 2

SEQUENTIAL BIOASSAY?

G. Stanley Woodson
Medical Laboratories, Army Chemical Center

The Medical Laboratories at the Army Chemical Center are faced with a situation, which, although not unioue with us, has become of definite interest. We have an animal colony which is large enough for us to generally assume a given reaction to be approximately constant when we draw consecutive samples for experimentation purposes. We are constantly working with "treatments" (be they agents, conditions, or others) that have had little or no experimental work done with them, and the various parameters of the Dose-Response relationships with their accompanying variances are known only within relatively wide limits. Consequently, it is difficult to determine in advance the numbers of animals that will be required in any one investigation. Therefore, we start with small numbers and, if our results are not fairly precise and stable, we then run additional animals and combine the results.

For instance, suppose we are working with a "treatment" which has and ED_{50} that is known to lie in the interval X_1 to X_2 (X_1 X_2) i.e., X_1 ED_{50} X_2. We desire to know the ED_{50} and, with a stipulated probability, it's confidence interval. We also desire our results to be in a form that will allow direct comparisons with other "treatments" which may be other experimental conditions or so-called "standards". In general we accomplish this by using probits 11 , or Logits 2 , for our analytical technique and a log-dose as our dose metameter. Consider the following:

Given four doses: x_{-2}, x_{-1}, x_1, x_2: we run, for example, four animals at each dose with the following results.

DOSE	RESPONSE
x_{-2}	1/4
x_{-1}	2/4
x_{-0}	2/4
x_{-1}	3/4
x_{-2}	4/4

Obviously our results are not satisfactory. Consequently four more animals per dose are run and we combine the results from the two runs.

DOSE	RESPONSE
x_{-2}	1/8
x_{-1}	3/8
x_{-0}	5/8
$x\ 1$	7/8
$x\ 2$	8/8

Again, the results are not too satisfactory on the lower end of the distribution, so we make another run and combine.

DOSE	RESPONSE
x_{-2}	2/12
x_{-1}	5/12
x_0	8/12
x_1	11/12
x_2	12/12

Now we feel that we have stabilized the regression enough for our purposes and proceed with the analysis. In later phases of our research, in connection with other points of interest, the above "final" regression may be duplicated several times, and any discrepancies noted would of course be subjected to investigation.

In discussions with representatives of other installations, as well as with representatives of groups outside the structure of Government research and development, I have found that this approach is far from beimg unique with us. As a matter of fact, it is quite widely used.

The point that I would like to make from this is that we are using, in practice, what is obviously a sequential sampling technique. We are, at the same time, retaining the analytical techniques of classical bioassay. In so doing we are making the following assumptions:

(a) The variation of the response that we are interested in studying is, for all practical purposes, constant in our animal population during the period of time covered by the experiment.

(b) The selection of animals from the colony at any phase of such an experimental procedure is completely random.

(c) Variation in experimental conditions, (the "treatment", the technique, the weather, etc.) does not influence the experimental results.

(d) Continuation of further replications beyond our stopping point would not have materially altered our results.

(e) And finally, based on these four assumption, we proceed to make the overall assumption that the classical analytical techniques are applicable.

We have, as of this presentation, found no dramatic errors occurring as a result of making these assumptions. However, COST and EFFICIENCY are constant reminders that prompt us to periodically reconsider our techniques in a constant search for a "better" methodology, and they have led us to consider the question of whether or not this experimental approach can be improved. So, let us take a closer look at the five assumptions I have mentioned.

Assumption (a) states that over relatively short periods of time, the variaticn in the biological responses from our experimental animals can be considered as a constant value. Upon closer examintation this breaks down into two interlocking questions: "May we, in all actuality, consider the magnitude of the variation of a given biological response as being relatively constant over a period of time, say θ?" and "what is the value of θ?". It is well known that changes in the level of tolerance of the population frequently make it impossible to relay on assays of materials carried out singly for purposes of estimating relative potency [11] [12] . This would apply if the potency were being extimated relative to another material, or to a previously run assay on the same meterial. Fortunately, our experimental work has shown that we can derive techniques which will allow us to detect significant variations of experimental subjects or samples from the mean of the group [7] However, we are limited in the applications of this approach in that we have not made a thorough enough study of the question of the constancy of responses in our animals. From our experience we can only mention that we have not demonstrated reason to doubt it's existence, we simply haven't quantified it and studied it.

Assumption (b) makes the stipulation that we have random selection of animals from our colony. This obviously cannot be enforced without interrupting the continuity of the colony's breeding program and consequently influencing the "constant variation" noted under assumption (a). Thus we face a seeming dilemma. For, if we desire to satisfy assumption (b) then we must ensure that the sires and dams (the very ones we desire to keep separate for breeding to maintain assumption (a)) have a probability of bein selected which is the same as that of any other animal. Actually, this could be very simply resolved by considering as our population those animals not pre-selected for breeding purposes.

Assumption (c) obviously must be evaluated for each experiment. Chemicals may vary from lot to lot or from day to day; the precision with which "treatments" may be reproduced may vary; weather changes must be evaluated and their biological implications assessed; etc. These, however, are the very things which must be closely watched in any experiment, and are not unique to our problem.

Assumption (d) offers us two questions of merit, and these become of immediate interest: First, the assumptions made in (a), (b). & (c) must be valid before we can assume (d); Second,;we assume that we can arbitrarily designate a point beyond which the addition of further experimental groups contributes little or nothing. Here then is our problem; The evolution of a process whereby we may designate, under certain established risk functions, an arbitrary point of terminating the procedure. Since the choice is now arbitrary, we cannot, under current analytical procedures, attach any a priori probability statements to our results. What we have been doing is attaching an a posteriori probability statement to our conglomerate results by making assumption (e).

Actually, since our doses are divided by equal increments, we can derive a fairly accurate approximation to a sequential analysis. Let us suppose that we set ourselves a goal: what we want is an estimate of the ED_{50} with a confidence interval of size $\hat{\ell}$ such that $\frac{1}{2}\ell$ will be equal to or less than, say, 25% of the ED_{50}. Now if we attach a number i to each of our doses, letting our lowest dose be zero, the next higher be one, the next two, and on up to calling our highest dose 'four' (since we are only using five doses), and if we work with the advancing differences in the number responding, say 'm', we find that a relatively simple estimate of the ED_{50} is

$$\mu' = (x_i - x_{i-1}) \left(\frac{\Sigma\, im}{\Sigma\, m} - 1/2 \right)$$

and an estimate of the standard deviation of the response distribution is

$$\sigma' = (x_i - x_{i-1}) \left\{ \frac{(\Sigma\, m)\,(\Sigma\, i^2 m) - (E\, im)^2}{\Sigma\, m^2} \right\}^{1/2}$$

Now letting

$$1/2\,\ell = t\sigma' / \sqrt{N + K-1}$$

where t = value from Students distribution for (N + k - 1)df.
N = Total number of responses
K = Total number of non-responses

we can continue our sampling until $1/2\,\ell \leq .25\,\mu'$ and, upon doing a full analysis of our data, we will find that our criterion has been fully satisfied.

For an example I will work through the series of data I have already presented.

Run No. 1

dose	observed response	"working dose"	working response		
		i	m	im	i^2m
x_{-2}	1/4	0	1	0	0
x_{-1}	2/4	1	1	1	1
x_0	2/4	2	0	0	0
x_1	3/4	3	1	3	9
x_2	4/4	4	1	4	16
			4	8	26

$$\mu' = (1)\ \left[(8/4)-1/2\right] = 1.5$$

$$\sigma = (1)\ \sqrt{\frac{4(26) - 8^2}{4^2}} = 1.58$$

$t_{.05}(df = 19) = 2.09$

$.25\ \mu' = .375$

$1/2\ \ell = (2.09)\ \ (1.58)\ /\ \sqrt{19}\ = .758$

$.375 \ngeq .758$ so we continue by adding another series of observations.

Run No. 2

dose	Cumulative observed response	i	m	im	i^2m
x_{-2}	1/8	0	1	0	0
x_{-1}	3/8	1	2	2	2
x	5/8	2	2	4	8
x_1	7/8	3	2	6	18
x_2	8/8	4	1	4	16
			8	16	44

$$\mu' = (1)\ (16/8 - 1/2) = 1.5$$

$$\sigma' = (1)\ \sqrt{\frac{(44) - 16^2}{8^2}} = 1.22$$

$t_{.05}(df = 39) = 1.96$

$.25\ \mu' = .375$

$1/2\ \ell = (1.96)\ (1.22)\ /\ \sqrt{39} = .383$

$.375 \ngeq .383$, so we continue by again making a series of observations and combining the results with that we have.

Run No. 3

dose	Cumulative observed responses	i	m	im	i^2m
x_{-2}	2/12	0	2	0	0
x_{-1}	5/12	1	3	3	3
x_o	8/12	2	3	6	12
x_1	11/12	3	3	9	27
x_2	12/12	4	1	4	16
			12	22	58

$$\mu' = (1) \quad (22/12 - 1/2) = 1.33$$

$$\sigma' = (1) \quad \sqrt{\frac{12(58) - 22^2}{12^2}} = 1.21$$

$$t_{.05}(df = 59) = 1.96$$

$$.25\mu' = .333$$

$$1/2\,\ell = (1.96) \quad (1.21) \quad / \sqrt{59} = .308$$

$.333 \geq .308$, therefore we terminate the observations.

Now, at any point in such an analysis, we can compare our results back with the previous run and note any dramatic variations which would indicate that our analysis was invalid. Actually, the formula given for the estimation of the ED_{50} can give us an error of up to 25 percent of an interval, and this should be kept in mind when it is used. Under most experimental conditions this error will not exceed 10 percent.

The final observations have been analyzed using Probits and a comparison between the Probit results and our results is rather interesting:

	Probit	"Sequential"
μ'	1.21	1.33
σ'	1.20	1.21
$1.96 \ SE_{\mu'}$	.38	.31

The above method has been used for determining the parameters of dose-response curves with considerable success. Considering that it was derived merely as an approximation tool to allow a truncation procedure when a desired precision was obtained, this has been pleasantly surprising to us.

A relatively simple modification of the above allows us to terminate observations as soon as a potency ratio between two "treatments" reaches a level of predetermined significance. However, we have as yet found no way to determine a stopping point if the potency ratio is not different from one. The same comment applies to the analysis of a single curve, we have no method of determining if $1/2\lambda$ is actually greater than our criterion, regardless of the number of observations taken. So far we have used merely a rule of thumb, whereby a lack of change for three successive series constitutes a reason to halt.

There is an alternative approach to the question of sequential bioassay which has been given some impetus lately. This has been to consider the question of the comparison of two treatments on the basis of pairs of observations. Under this approach the pairs of observations become the units of analysis. For a simple question obtaining estimates of the parameters, this has some merit. However, in most experiments we are interested in obtaining as much information as possible, and consequently it becomes necessary to search for a technique such as the one suggested in this paper which will allow the estimation of the various parameters. Unlike most experimental situations, we cannot sacrifice some information to gain the advantages offered by current sequential procedures. Fully sequential procedures have been shown to be applicable to approaches to composite hypotheses [5,9,10,13,14], but as yet no successful application has been made to the field of bioassay other than the method suggested by Dixon the Mood [8]. Bross [3] has worked out some sequential medical plans with proper truncation techniques, but these are based on the paired comparison method and do not answer the needs of bioassay.

It has been remarked [4] that some experimenters might not trust the results of an experiment that terminated very promptly according to the rules of the sequential plans, and that these same experimenters report that their professional colleagues would certainly not trust reported results from such procedures. Now, although the sequential approach generally offers the possibility that a smaller number of observations, on the average, will be needed, it also offers a safeguard against terminating the observations before a meaningful conclusion can be reached. Thus sequential procedures are like a double-edged sword that can work <u>for</u> the experimenter in more ways than one.

I have attempted to present an outline of the problems facing us in our attempts to apply quasi-sequential procedures in our experiments. If there are any questions or comments on the presentation, I would be most happy to hear them - - especially if they can present me with an answer of how to do a correct analysis.

APPENDIX

Let us consider the experimental approach that has been outlined in the body of the paper as a theoretical model. Our interest will center in a relatively narrow range of doses in which the response or non-responce of a subject is a matter of probability, such that at the upper limit a response is very likely, while at the lower end of the range a non-response is very likely. Above and below this range response or non-response becomes a matter of practical certainty. The range over which response or non-response is indeterminate will be defined as the "Critical Range", and the point within the range at which the probabilities of response and non-response become equal will be defined as the ED_{50} (the dose which is expected to produce a response in 50 percent of the subjects).

Suppose we know that for our "treatment" the logarithms of the doses within our Critical Range, when plotted against the proportion responding (p), form a cumulative normal distribution. Now, letting x = log dose, we desire to estimate the mean (μ) and the variance of the distribution (σ^2). If we perform our experiment by selecting a dose x_0 near where we expect to find the mean (μ), and selecting four other doses (x_{-2}, x_{-1}, x_1, x_2) such that they will divide the expected Critical Range into six equal parts, the doses will then be so spaced that the transformed variate is equally spaced. Now, if we are correct in our selection of x_0, the total number of responses will be approximately equal to the total number of non-responses.

If we let

$$N = \text{total number of responses}$$

and if we let n_{-2}, n_{-1}, n_0, n_1, n_2 denote the number of responses at the corresponding doses, we have then that

$$\Sigma\, n = N$$

Correspondingly, if we let

$$K = \text{total number of non-responses}$$

and if we let k_{-2}, k_{-1}, k_0, k_1, k_2 denote the number of non-responses at the corresponding doses, we have then that

$$\Sigma\, k = K$$

Now it can be seen that at x_i we have n_i responses and k_i non-responses, and the likelihood of (n_i, k_i) is

$$P(n,k)\, x) = C \prod_i p_i^{\,n_i} q_i^{\,k_i} \qquad (1)$$

where

$$Pi = \int_{-\infty}^{x_1} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y-u}{\sigma}\right)^2\right) dy = 1 - q_i \qquad (2)$$

and where C is not a function of μ and σ.

Now, if we define

$$m_i = n_i - n_{i-1} \text{ , } \Sigma m = M$$

We find that we have a situation which is somewhat analogous to the "up-and-down" method of Dixon and Mood (8).

Now, letting

$$\alpha = \Sigma \, im \text{ , and } \beta = \Sigma \, i^2 m \qquad (3)$$

where i is now defined as the ordinal of the interval from the lowest dose showing a response (the dose range may rightly, and should be, extended upwards and downwards enough steps to insure that the Critical Range is completely covered).

we see that

$$\mu' = x_{i=o} + (x_i - x_{i-1}) \left(\frac{\alpha}{M} - 1/2 \right) \qquad (4)$$

and that

$$\sigma' = (x_i - x_{i-1}) \left\{ \frac{M\beta - \alpha^2}{M^2} \right\}^{1/2} \qquad (5)$$

A good approximation for a confidence interval, interval for u, can be obtained from

$$\mu + t\sigma' / \sqrt{N + K - 1} \qquad (6)$$

where the value of t is given by the "t" distribution for (N + K - 1) degrees of freedom.

Now let us assume that we desire to estimate our mean (μ) with a confidence interval such that we are γ percent certain that we have estimated μ within $\wp$ percent. In other words we desire a confidence interval (ℓ) with a confidence coefficient of γ, such that

$$t_\gamma \, \sigma' / \sqrt{N + K-1} = 1/2\ell \leq (\wp \text{percent}) \, (\mu') \qquad (7)$$

It will be found that the probability that μ lies in the interval $\mu' + 1/2\,\ell$ is actually less than γ. However, it has been shown [1] that the true probability can be approximated by

$$\gamma - \frac{0.176\ell^2 \; g(t)}{4 \, \sigma^2 \, t} \qquad (8)$$

where g(t) is the ordinate of the frequency function of a standard normal variable when the abscissa is t.

Under the foregoing considerations, our experimental procedure becomes as follows:

(a) Observations are taken in small sub-groups until the inequality (6) is satisfied.

(b) we calculate $\mu' \pm 1/2 \ell$

(c) the probability that μ lies between these limits is now given (approximately) by equation (8).

BIBLIOGRAPHY

1. Anscombe, F. J., "Fixed-sample-size analysis of sequential observations," Biometrics, Vol. 10 (1954), pp. 89-100.

2. Berkson, J., "Maximum likelihood and minimum chi-square estimates of the logistic functions," J. Amer. Stat. Assn., Vol. 50 (1955), pp. 130-162.

3. Bross, I. D. J., "Seq. Medical Plans", Biometrics, Vol. 8 (1952), pp. 188-205.

4. Cochran, W. G., "Some aspects of experimental design", an address before the Statistical Engineering Symposium, Army Chemical Center, Md., April 1955.

5. Cox, D. R., "Sequential tests for composite hypotheses", Proc. Cambridge Philos. Soc., Vol. 48 (1952), pp. 290-299.

6. Cramer, H., Mathematical Methods of Statistics, Princeton Univ. Press (1946).

7. DeArmon, Frazier, Ludlow, and Wayne, Chemical Corps Medical Laboratories Research Report No. 174, March 1953.

8. Dixon, W. J., and Wood, A. M., "A method for obtaining and analyzing sensitivity data", J. Amer. Stat. Assn., Vol. 43 (1948), pp. 109-127.

9. Dvoretzky, Kiefer, and Wolfowitz, "Sequential decision problems for processes with continuous time parameter, testing hypotheses", Ann. Math. Stat., Vol. 24 (1953), pp.254-264.

10. ----------, "Sequential decision problems for processes with continuous time parameter, problems of estimation", Ann. Math. Stat., Vol. 24 (1953),pp.403-415.

11. Finney, D. J., Probit Analysis, Cambridge Univ. Press (1952).

12. ----------, Statistical Method in Biological Assay, Griffin & Co., London (1952).

13. Johnson, N. L., "Some notes on the application of sequential methods in the analysis of variance", Ann. Math. Stat., Vol. 24 (1953), pp. 615-623.

14. Perron, C., "Uber das verhalten einer ausgearteten hypergeometrischen Reihe bei unbegrentzen wachstum eines parameters", J. fur reine und anq. Math., Vol. 151 (1921), pp. 63-78.

15. Stein, C.,"A two sample test for a linear hypothesis whose power is indepedent of the variance", Ann. Math. Stat., Vol. 16 (1945), pp 245-258.

16. Wald, A. Sequential Analysis, Wiley & Sons, New York (1947).

ESTABLISHING HYPOTHESES WHEN THE EXPERIMENTAL FACTORS ARE NOT AT PRESENT UNDER THE CONTROL OF THE EXPERIMENTER

K. R. Wood
Quartermaster Food and Container Institute

Below is described a basic problem that is rather frequently encountered. And, insofar as we are aware, is one which has no satisfactory solution.

Given a matrix of n rows and k columns, in which values in the jth column represent observed values of a variate x sub j, and values in the ith row represent observations on the ith object (sample, individual, or item).

Assuming that there exists another matrix of rank r(r less than k) of n rows and r columns, in which the values in the jth column represent "true" values of a parameter (variate), t sub j, for the n objects.

Assuming that except for uncorrelated, normally distributed errors with constant (unknown) variance, x sub j is, say, a general second degree function of t sub 1, t sub 2,......t sub r, how does one approximate this function? How does one estimate the coefficients in the function for x sub 1, x sub 2,... for x sub k?

On the above problem, Paul Meier, Johns Hopkins University, suggested trying the "response surface" approach, but in general, our experience shows the x sub j's to vary considerably in their interdependence. They are merely observed - not under the researcher's control, and he is seeking (not testing) hypotheses. Once he establishes some hypotheses, a rigorous experiment can perhaps be designed for their testing.

PANEL DISCUSSION ON HOW AND WHERE DO STATISTICIANS FIT IN*

Chairman of the Panel:	John Tukey, Princeton University and Bell Telephone Laboratories
Members of the Panel:	Cuthbert Daniel, Private Consultant
	Besse Day, Bureau of Ships
	Churchill Eisenhart, National Bureau of Standards
	M. E. Terry, Bell Telephone Laboratories
	S. S. Wilks, Princeton University

* This interesting phase of the program was recorded. Unfortunately several parts of it were not clear. Unfortunate also is the fact that several of the members of the audience who formulated some of the questions and added many points to the Discussion could not be identified.

Tukey: I was originally asked to speak on this subject this afternoon, but being a consultant I am used to putting other people to good use. After all: - "A consultant is a man who thinks with other people's brains." And so instead of speaking, I am here to chair a roundtable discussion - the brains you see assembled at the table. I don't think introductions are needed but from the far side to the near side - Mr. Eisenhart, Miss Day, Mr. Daniel, Mr. Wilks, Mr. Terry represent our stellar team of statisticians with diversified experience on where and how statisticians fit in. We've done as well as we could in providing one of each type. (The panel has been bothering me from time to time about the question as to how this session will be conducted. I've told them at various times, including this morning, that I didn't know.) I propose to give the audience the chance to provide us with some provocative questions. If they don't provide us with enough provocative questions, we may start issues with members of the panel - so the situation is in your hands, ladies and gentlemen. Does anyone have a question?

Unidentified person from the audience: I would like to address a question to the Chairman. What are the five kinds of statisticians? (Laughter)

Daniel: Mr. Chairman, I have just worked that out. There is one lady, one gentlemen, one statistician, one administrator, and one genius. (laughter)

Tukey: I think that this is an excellent answer by the general appearances, but I would suggest that there is no indication that each person over there represents only one kind of statistician, since Miss Day is now holding down two jobs, and reporting to two different officers. This is complete proof that a lot more than five kinds may be represented.

Do we have any other questions about where statisticians fit in in general, rather than where these five particular people fit in?

Unidentified voice from the audience: I would like to raise a question concerning organization. Too many times our statisticians have been brought into engineering problems.

Tukey: Just for the benefit of the panel and myself and possibly some of the audience what organization are you thinking of? A military research and development organization?

Voice: Yes.

Tukey: All right, now while the panel is thrashing their collective heads, could I ask you what you mean by being drawn into an engineering project?

Voice: Sometimes a statistician working very closely with project engineers suddenly finds for all practical purposes he is acting like a project engineer. The problem is not so much one of what to do or not to do, but how to retain his individuality as a statistician. How can he keep from being drawn so far into the problem that he loses his identity?

(Incomplete translation)

Tukey: Gentlemen, I am sure you have had experience with such situations.

Terry: I think you have got to have both (kinds of statistician). Actually, we at the Laboratories have three kinds. We have engineering statisticians who have major responsibility for a project, or a group of projects. If they are clever they try to keep away from becoming a project engineer. (They have got enough on their neck as it is, without taking over his prerogatives, but I'll admit at times it gets very close to that.) Then backing them up must be a team at the development level, who are one step removed from the project and who can do a little lateral investigation and thinking. And then behind these again must be a group of mathematical statisticians — engineering statisticians who are paid to do research on the statistical methodology and particular problems that these people may select and bring to them. (There will usually be within this top group, people who, by temperament, do more consulting than actual research and people who do more research than actual consulting). This is essentially the system we have at the Bell Telephone Laboratories.

Tukey: Well, let me suggest that one cause for this difficulty was that the statistician was not being kept busy enough. If he had been involved in about three projects to begin with, he would have been so busy that he would not have thought of trying to take up project engineering.

Is there another question?

Member of audience: I have a question. That is, how can the statistician encroach upon the prerogatives of the engineer without rubbing him the wrong way? (laughter)

Day: Should be easy.

Member of audience: If I may have a minute I should like to suggest that a temporary bridge be built at times between the engineers and statisticians. What does the Panel think?

Tukey: Who on the Panel would like to lead out on this question? (Long pause)

Wilks: Listen, the Chairman is not immune, you know.

Tukey: I know, I know. (laughter)

I know that the Panel is doing its best to get back at me. Well, what do you mean by building a bridge momentarily between the engineers and statisticians? Why should we think we should ever be able to get along without it?

Member of audience: (Tape not clear)

Tukey: Thank you. What you are saying is that if you are careful about it now, perhaps the bridge will be big enough from now on so that you will not have to worry about it carrying the load.

Member of audience: Right.

Tukey: Well, it seems to me there has to be a bridge. It seems to me, in response to the question at the rear of the room, that no statistician can take over an engineer's prerogatives without making the engineer feel bad. The statistician has to work by infiltration and cooperation. And after he has done this with a particular group for a certain length of time, instead of wondering about taking over prerogatives and wondering if he can get away with it, he will, instead, probably be trying to get out. R. L. Anderson was discussing a paper at one of the Washington meetings a year or two ago and brought out the problem of statisticians who get called into a group, and eventually get asked to make more and more decisions that are non-statistical. At what stage does his conscience start to bother him? How does he manage to stay out? But even this is a question of being called in for advice - being asked for advice from time to time rather than any question of taking over prerogative. I don't think in the long run that you can impose good statistics upon people, but you can expose people to good statistics. I don't think you can make people like it or take it, in the long run except by building up the record. Does this partly answer the question you had, Doctor?

Tukey: Are there any others on the Panel now ready?

Wilks: John, I've got a few questions - I mean a few remarks to make on this subject. I think a very important point has been raised here; namely, how do we help these people? I think the situation as far as engineers receiving training in statistics is concerned is new. It really started since World War II. There is not much of it yet, but my opinion is that eventually we will have to develop the training of our engineers to the point where they can handle most of the routine problems that are now being handled by the consultants.

This reminds me of some experiences I have had over the last twenty or twenty-five years with one or two organizations. One of them is the College Entrance Examination Board and its successor the Educational Testing Service. When I first started on some of their problems, I found most of them were of the routine type. And my feeling, even at that time, was that sooner or later they would have to have people to handle them. As things have developed over the last 15 or 20 years, they have gradually brought people in who can handle all the problems that I used to get asked about as a consultant. In fact, the situation has gotten to the point now, where the real problems are stinkers and so difficult that I can't touch them! On the other hand, there ought to be statistical consultants available at the current stage of their operations who can handle these frontier problems.

Many of the problems that come up in engineering are the kind that the engineer ought to be able to handle himself, and could handle with a reasonable amount of statistical training -- that means one or two good courses in college. Then they could handle a lot of the routine problems and save the consultants for the newer kinds of problems that arise.

The statistical talent that is available for these situations is extremely limited. At Princeton we get requests all the time -- several a week. They write in, and they call in by telephone, saying 'we need somebody'. This goes on and on. Today the tasks to be done by the available statisticians are so numerous that we simply can't spread over all of them.

Of course, this brings us to the question of what are the interim measures for the next 10 or 15 years until we get the engineers trained in some of these things. This is a long story in itself, and perhaps this is not the time to get into it. But I feel that there is a need for short courses, evening courses, perhaps for short summer institutes of two or three weeks in length, in which engineers and scientists in industry can get together and pick up some of the main methods and philosophy of modern statistics. I think the quality control people have done a very good job in this direction. (You remember that that group started only a few years ago.) They have a lot of these short courses even now. If you look at what they are doing in their courses, conferences, etc., now and compare it with what they were doing ten or fifteen years ago, you will see there has been a tremendous change.

Back to the question of routine type problems. My feeling is that we have to train people in statistics, both people in engineering and those in the sciences, so they can handle most of these problems. There will always be frontier problems on which the statistical frontiersman--the consultant--must be brought in. But he should be relieved, more and more, from having to deal with the routine problems, by further statistical learning on the part of engineers and scientists.

Day: John, I'd like to comment along that line. I think there is an area of training or indoctrination that comes before the one Sam was talking about. That is the acquainting of engineers or physicists or scientists wit the fact that here is a tool they can use. I think very often a great many people do not understand that statistics can be of help to them - even befor they start using it themselves. At the laboratory I used to work in, we found the short courses - the indoctrination courses - were very helpful in getting people interested in using statistics. In them they found out what we were trying to do, and what kind of a tool this was and where it could be used. I think the same thing is true in regards to management. It has to be understood that this is a bright new tool. If there is an understanding, and if there is an appreciation for what it can do, then they will use it. Of course, I think we are farther along now than we were say 10 years ago, or 5 years ago, because a great many engineers are beginning to appreciate statistics. But I think that takes a different type of training than the type of thing you had in mind, Sam.

Eisenhart: I would like to offer one other suggestion along that line: that at universities and other places, people who have been through a particular program might very well get close to the teachers of engineers, physicists, etc. and audit the first course in physics, in engineering, or what have you, and then just casually slip in some statistical methods that are appropriate to the subject being taught. In this way the students arrive quite naturally at the use of statistics in their regular program. I did this about a decade ago in two courses, one in psychology and one in agronomy, and it was lots of fun just to see how much I could get built into the other fellow's course.

Tukey: I think the gentleman has the floor back there.

Member of audience: Well, personally you raise in my mind just what the Panel is talking about. Seems like the nettle (to be grasped) here is the field of statistics, itself. First, you've got a period of training in which we have to train management and administrators to appreciate statistics, and then a period in which we train engineers to use statistics. Wonder what it all leads to? Are we going to get away from the general field of statistics - or are we going to specialize it - both of which take a certain amount of work? In other words in 10 years, or maybe 15, in a meeting like this, will we have a Panel consisting of 5 people who can all talk about the general field of statistics or will it be comprised of people talking about certain phases of research and certain development problems?

Tukey: You are thinking again about specializations in the direction of application? Terry, is that what you started to speak on?

Terry: Yes, I think so. Every engineer we'll assume, knows how to use a slide-rule, but not every engineer knows how to use a SEAC. But he knows the principles of computing, and he would, with a little training, be able to go further in this area. I think what we are claiming is that statistics must be a basic part of the engineer's training. It's a new field and he isn't, in general, getting it at the university. It is something he has to get subsequent to his formal engineering training, but is becoming more and more an absolute necessity for a good research engineer.

Tukey: Or a good development engineer.

Terry: Or a good development engineer.

Tukey: Possibly even more for the development engineer. I would like to challenge the Panel on the grounds as to whether or not they are now saying where the statistician ought not to fit in, rather than where he should fit in. They're saying if these engineers only knew enough statistics, we wouldn't have to go quite so far down the line. I know this is interesting to report on, but it is just a little bit off the edge of our subject. So I am going to try to divert the discussion for a while. Are there more questions as how the statistician does fit in?

Mr. P. C. Cox (White Sands Proving Ground): I would like to ask for comment. If the statistician is always worrying about taking over an engineer's job, what kind of a character does a statistician have that an engineer never takes over a statistician's job?

Tukey: Well, my understanding is that the Panel members are looking for, and are very pleased to find, the engineers who are willing to take over their jobs. (Laughter) The sort of jobs they have to do now! This would leave the statistician free to do the jobs that they are even more interested in doing. (But I don't want to spell out the Panel's comments myself).

Eisenhart: John, you may recall that at a meeting in Montreal I read a most provocative memorandum written by a geologist who was working for an oil company, on the acquisition and function of a staff statistician in an industrial laboratory. His company granted me permission to read the letter in that connection. I feel that it might enlighten the discussion if I were to read it again now.

Tukey: I would think that this is definitely in order. It is clearly about how the statisticians fit in.

Eisenhart: A little over a year ago I was visited by Melvin A. Rosenfeld, Senior Research Geologist at the Magnolia Petroleum Company's Field Research Laboratories in Dallas. He came to ask where he could find a statistician. He not only asked, but he also showed me a memorandum that he had written entitled "Acquisition and Function of a Staff Statistician in an Industrial Laboratory." It was truly remarkable. It read (with a few deletions) as follows:

> "By all odds the most important consideration is the fact that the function of the statistician is as an adjunct to experimental work. At no time must the idea of "Statistics for statistics sake" become supreme to the experimentation; statistics, in one sense, has been defined as the mathematics of experimentation and, for our present purposes, it should remain as such. Experimentation is our primary work and statistical applications are to strive for better and more efficient experimental techniques. The two are inextricably welded together; no experiment is better than its design and statistics are worthless without data.
>
> .....For this reason it is strongly urged that the statistical effort be not devoted to advancing mathematical research. We do not need new statistics, we do need applications.
>
> "The above paragraphs are not to be construed as meaning that the staff statistician should not, if the need or occasion arises, develop new theories and techniques. Rather it is intended to mean that this is not to be his major effort and to indicate that, most likely, there are enough statistical techniques in existence and being developed daily to last a good long while. If he is to operate efficiently as an integral part of experimental work he will have to, in effect, rub elbows with the technologist.
>
> "This liaison between technologist and statistician may present some initial difficulties but these, I am confident, will be readily overcome. There is a high probability the statistician obtained will be weaned from some phase of biological science. A wholly unfamiliar field of terminology and operations will be showered upon him and,

despite the fact that a statistician works only with sets of numbers, there is no doubt that the better he understands the problem the better he will be able to assist in its solution. From the viewpoint of the technologist the liaison may be equally difficult. It has been my experience that the ability to ask a question amenable to answering in an unequivocal manner is one of the most difficult techniques to master. It is in this field of asking precise questions that the technologist and statistician must come together and they will do so to the benefit of both, provided that the initiative comes from the technologist......

"One other matter for consideration—and this is akin to pure statistical research--is the question of whether the statistician has long range projects of his own. It is highly doubtful whether, at the outset of the activity, it will be valuable to engage in this type of effort. It would be unremunerative to toss the "sampling problem" to the statistician and request a solution. In one sense there is no "sampling problem" that can be solved by a statistician working alone. It is unquestionable that there are "sampling problems" but I am inclined to believe that these problems that lack only data for their solution. No "sampling problem" can be met squarely unless the technologist is capable of knowing precisely what the sample is for and can furnish some preliminary estimates of variability. From this point the statistician will derive a particular sample design for a specific area of work. This, again, is a case where the technologist and statistician will have to work closely together in obtaining at minimum expense the information necessary for efficient design.

"The remarks in the above paragraph are prompted by my fear that there may develop a tendency to foist upon the statistician problems that properly belong to the technologist. This fear may be groundless but I wish to reemphasize that, unless there is a joint attack on problems, the acquisition of a statistician will not serve the purpose originally visualized. The statistician is not to take the load off the technologist; to the contrary there will be some cases where, with statistical advice, the technologist will be required to do <u>more</u> work than he intended. In the long run, however, the proper application of statistical technique will lead to minimum expense---maximum information experimentation. If, at a later date, the statistician be given a long range problem of his own it should be, I hope, with the understanding that frequent and lengthy interruptions in the interest of current experimentation be expected.

"It is evident from the foregoing that a very special sort of person is required as the staff statistician. In my limited experience I have had contact with three different kinds of statisticians, two extremes and a composite. These models are based upon actual living persons with whom I have had courses and/or occasion to consult.

1. <u>The pure mathematician.</u> A professor of mathematics who is capable of deriving from scratch any formula used in statistics. He is thoroughly grounded in the theory of probability and, given time, can likely find some exact solution to a technological problem although it may not be the most efficient in practice. It is very unlikely that this person has ever seen a physical experiment in progress or that he is abreast

of current statistical practice. He is very apt to be a purist and thoroughly unfamiliar with the vagaries of actual data and may experience some difficulty in speaking a language that the technologist understands. He does not have the facility based on experience to say, "Hell, it doesn't matter that we lost a sample" - or - "Let's make an approximation here similar to one I used years ago on another experiment". It is highly doubtful that this type of statistician would be of benefit.

2. The "rote" statistician. This character can do all of the operations as given in standard textbooks and is reasonably aware of current developments. He is likely to be an excellent agronomist or biologist and may find the transition to petroleum work difficult because of his intense basic training. He is not primarily a statistician but, because statistics is imperative in his field of work, he has taken a large number of courses to obtain his Ph.D. He may teach elementary courses in statistical method as applied to his particular field. There is little likelihood that he has a very thorough knowledge of basic theory or can derive even the simplest of equations. He will design efficiently, implement experimentation and analyse data to definite advantage in pursuit of his studies. This type of statistician is very useful to fit in (for more work of the same kind) to an already established operating statistical laboratory, but is not recommended as the initial member of a statistical group.

3. The genuine statistician. A man who combines the best qualities of both of the above types. He has had a thorough grounding in basic theory and concurrently or subsequently has some experience in practical experimentation. He is familiar with current literature and is capable of making applications to problems which he has not previously encountered. It is also likely that he has contributed in some way to the literature of statistics, and will have the ability to converse in a language understood by the technologist. In problems where the first two statistician types have failed entirely to help me I can state that a consultation with this third type, a member of a well known Agriculture Experiment Station, has never failed to be remunerative. Usually in a short session he would wrestle from me a precise statement of a question I was trying to frame--and this without his having much knowledge of the subject matter.....

"The stock pile of good statisticians is not likely to be overflowing; it may be that an intensive search will become necessary. I can say, with confidence, that the opportunities in an industrial laboratory are a gold mine of new and interesting applications that should arouse the interest of a statistician even if he is employed elsewhere in a different line of work at the present time."

I was really amazed to receive this from a geologist, because I thought that geology was one of the areas where they didn't understand statistics.

Tukey: Does anyone else on the Panel want to comment on this subject?

Daniel: I can sympathize with the geologist who wants a statistician, but we are here to look at a bigger problem. We are here to look at a pair of populations, each one interlocking in some degree with the other. There is a whole spectrum, of course, of abilities and interests among statisticians and among persons who call themselves statisticians, and they go all the way from real experts to self-confessed statisticians who have just learned how to pronounce the word. There is a whole range of abilities and experiences. There is also a whole range of abilities and attitudes among engineers whom statisticians presumably can help. The attitudes I want to speak of are spread all the way from engineers who hope that statisticians will solve all the problems that they are now concerned with (and solve them, in fact, more or less by graphical methods), to engineers of more sophistication, and to engineers with more strongly negative attitudes towards what a statistician might be expected to do for them. But, when I am asked "where do statisticians fit in", I have to respond in terms of my own feeling, which is that even if you put them in as a monomolecular layer there would not be anything like enough to go around. Then I have to ask, not so much how do statisticians fit in, but how do we get the job done that now has to be done? We now have to translate this into the job that you would like to have done, that industry would like to have done, that the Army would like to have done, in statistics. Those jobs, frankly, will not be done as well as their leaders would like them to be done. They will not be done because the statisticians are not available, because some attitudes are so very unsympathetic, becasue of the friction of the human relations, because of the usurping of prerogatives and because of a dozen other lacks of efficiency. These jobs will not be done and we have, I believe, to move on to another question which is: 'How will statisticians fit in? What are the prospectives for getting effective use of statistics without too much usurpation, without too much domination, and so on?'

While I am speaking I want to answer Dr. Cox's question, because the man he says doesn't exist, the engineer who has taken over a statistician's job, is speaking. (Laughter) There are some advantages and some disadvantages, and I speak as a man who is mediocrely prepared in both fields, so I don't feel I offend either side when I speak of a statistician. The expert - I mean now - the serious expert who has had a lot of experience and, of course, has a full technical background, has to spread himself very thin indeed and should usually not spend his time giving Lesson One. I have given Lesson One. It is one of the lessons I feel moderately capable of giving, not only because I have given it 100 times, but because that is about as far as my real statistical education went. The experts should not be required to give the lesson 200 more times. That is not the way for statisticians to fit in. There is a group, however, a big group of fairly well-prepared statisticians and the question now comes up about how they should be organized and how their work should be used. They should be used mainly in teaching. We are at the teaching level still - I don't mean the way a consultant teaches a man to solve a particular problem in front of him. I mean that the particular problem must be viewed by the teaching statistician as a tool, not only as a problem to be solved. The problem to be solved is how to get the mathematics of experimentation in the hands of engineers and how to use this general tool.

One must be careful not to offend the engineer. This happened to me only the other day. The engineer I was working with felt slighted because I did not go downstairs and see his particular tensile testing machine on which he was breaking his own miserable little 'dumb-bells' (excuse me 'dog biscuits') of material. 'How can you possibly help me if you don't see the equipment on which I am working', was the rather offended question that he asked. I only refused to go downstairs because he asked me this question at 5:15. But the important thing is that he needed to be told there are things in our field which correspond to 1 + 1 = 2. One does not need to know whether one dog biscuit plus one dog biscuit = 2 dog biscuits or not. One doesn't need dog biscuits to find this out. I am talking about unbroken dog biscuits. (Laughter) The point I am trying to make is that we are in a situation of such disequilibrium that to talk about establishing equilibrium with either the present stock-pile and/or the potential stock-pile of statisticians is to take far too short a view. Our job is still one of training statisticians to remove a major economic scandal since the ratio of demand to supply of statisticians is 20 or 100 to one. (Wilks just said so!) It is common knowledge that there is a defect in the rate in which the law (of supply and demand) tends to equilibrate itself. Industry and the Army apparently have no way - no effective way and that is what is important - of equilibrating these two. There is only one way they could do this and that is to decrease the demand. It can't be done by increasing the supply. There is no effective way in which industry can demand of universities a training which the university would then supply with reasonable lag. We sit here on the Panel and predict that, say 10 or 15 years from now, the situation will be different. I heard Panels like this 10 years ago and exactly the same prediction was made. I drew my own conclusions - I became an industrial statistician and this turned out to have been, if not good for industry, at least remunerative. (laughter) This solved my problem, but does not solve yours. The fact that there are now thirty or forty times as many statisticians of medium competence as there were ten years ago does not relieve the situation at all. The demand has increased in the same ratio as the supply. So the problem - the real problem before us - is how to use the statistical abilities that we have. We have to make it clear that this demand must now approximate the supply, and we must increase this supply by training programs such as the short course Wilks speaks of, and the in-training service programs, the courses given by outside consultants, and by the self-taught statistician. All of these things have to be pushed to the limit. There are no either-ors and there are hardly even questions of emphasis. The questions are, "What do we have strength for?" and "How clearly can we recognize that the problem is one of supply?"

(Applause)

Day: John, I'd like to say something.

Tukey: All right.

Day: About the problem of increasing supply - I sat here and thought as Mr. Daniel was talking - there is one way of increasing supplies somewhat. And that is that during some of this indoctrination, or somewhere along the line, it should be impressed on the people who are taking training in statistics that statistics is a service. We have too many trained statisticians who are unwilling to get their hands dirty and therefore are not being used. Because they take the attitude that (John, maybe I will have some tomatoes thrown at me, or something like that here) since they know some statistics, they have been put on a pinnacle -- way up in the sky. Within the last two months I have had two experiences with mathematical statisticians - both of them supposedly have very good backgrounds - but they can't be used effectively because they are living on a pinnacle and don't know how to make their contacts. Somewhere in their training they didn't get the right indoctrination. The use of statistics can contribute greatly, but there is an attitude that should accompany it, and there is a feeling of how the statistician can best work in the laboratory. Of course, we have to have our statisticians well grounded in the theory - more so now than 10 years ago because the problems are getting harder. (It won't be very long until I'll be out of a job because I will not be able to handle the hard problems!) But, at the same time, I think there is a tremendous need in the universities to train the men to realize that statistics is a service, it isn't just a useful display of mathematics for its own self. And if you're going to build a house you have got to have more than the foundation. It is sort of an indictment, Sam. I hope you will not hesitate to answer it. I don't mean it necessarily for Princeton, but it's a very sad situation.

Wilks: I agree we have to train our statisticians well in both theory and applications, and I agree that the development of a good attitude on their part toward statistical problems which arise from various fields is very important. We are trying to do all of this at Princeton.

Prof. Boyd Harshbarger (Virginia Polytechnic Institute):

I would like to repeat what several other people have said here. Industry, I feel, has discovered what a tremendous influence they have at colleges. I want to tell you what happened at VPI, and I want to tell you how people who come to interview us can change the entire atmosphere of a college campus. About three or four years ago people coming down to visit our engineers began to ask the heads of the departments, "has this individual had a few courses in statistics?" You know - that did more than all we could do to interest people in statistics. The result is that three or four of our departments are now requiring undergraduates to take statistics when they had never done it before. It is also working out to another advantage. A few of the young men in engineering, who by the time they get through college, have been indoctrinated to such an extent that they are interested in going on with graduate work in statistics. I think we can work it up if we can get the support of industry to go on asking the engineering schools, "Do you have an engineer who has had work in statistics?" If they do that, the schools are all going to put in a department of statistics or at least courses in this field.

Tukey: Before I recognize one or two others, I think I want to point to something which puts the clamp on things for the industrial administrators. I don't know whether it puts a clamp on Army administrators or not. I was thinking of a telephone call I had not so many weeks ago with a friend of mine who works for one of the drug houses. I was asking him how things were going. He said things were fine. He had finally managed to persuade them to get a girl on the computing machine. He would now have a little time to get caught up with the job, and to get around to find out more of what his company was doing and get going generally. I think this is a most serious issue, if we all admit to the shortage of statisticians, and so far I haven't heard anyone deny it.

We need to make maximum utilization of all these people with varying degrees of statistical background. For some of them this means giving them help - or giving them some time on the computers. The statistician who's hired to have one hand on the Monroe calculator and one hand on the tape adding machine - he really must be a good statistician if he escapes and ends up somewhere else.

In the case of this intermediate class that we've heard a little about, I don't include just the statistician that's in the intermediate class, but we have engineers at various stages of statistical training. If we admit that statisticians are in short supply, I think we have to look forward to letting such people spend a little more time doing statistics than one might like from the point of view of the pure engineer. Unfortunately, it's going to be true that a substantial fraction of these people that start to drift toward statistics will be good engineers. (The boss may feel bad if he loses the chance to have them work on a specific project just because they are getting to be sort of a local consultant for the people on the next five levels.) It seems to me that one of the most important ways to meet the training problems, to meet the shortage of professional or semi-professional, or sub-professional statisticians, is to let these people who are picking up a little statistics spend a little time helping their neighbors. If they pick up some more statistics let them spend even more time - and perhaps a little time helping other people who have picked up just a little. There has got to be a whole graded sequence! The groups that Terry was talking about earlier are by no means all there are to the chain. There are a lot more links between the statistician farthest from mathematical statistics he mentioned and the engineer who is getting some help out of a quick course or a word of advice.

I think I will recognize Joe Cameron who had his hand up earlier. Do you still want to say something, Joe?

Mr. J. M. Cameron (National Bureau of Standards):

I just want to add a few remarks to what has already been said. The psychologists and the education people seem to recognize that statistics should be made a part of their training. The psychologists have long recognized their needs in this direction, and the question is, why isn't the engineering group as much aware of their needs for statistics? (The remaining remarks by Cameron were not distinct on the recording).

Terry: Last year the American Society for the Teaching of Mathematics to Engineers, or something reasonably close to that --

Tukey: Let's see, it was probably the American Society for Engineering Education.

Terry: That is the one. (Laughter).....was injudicious enough to invite Ellis Ott, Professor of Mathematics and Statistics at Rutgers, and E. B. Ferrell and myself from Bell Telephone Laboratories to come down and talk with the mathematicians about our concept of the teaching of statistics and mathematics at the engineering school. We split the audience right down the middle, half of them protected us until we got out of the building, while the other half were ready to tear us limb from limb. Our attitude, I think, was essentially this.--- The mathematics department of a good engineering school is responsible for keeping its courses alive to the engineering needs of its students. And unless it does continue to grow in numerical analysis, statistics, and in other mathematical developments pertinent to modern engineering, then the engineers will set up their own departments. Mathematics will go back to its ivory tower of playing with Horner's methods for the extracting of roots, world without end, and will cease to be a function of the modern university. This view was received with great joy by mathematicians who believed the way we did and with considerable feeling with those who did not.

Day: It might be a good idea for us.

Terry: But I think, Joe, that at the present time getting good English brought into the engineering curriculum, good mathematics and good statistics is a pious wish and it is going to take a long time. English has disappeared - one man said that the multiple choice question is God's gift to the teaching profession and a curse to the students. The multiple choice test is easy to correct but the engineer comes out untrained - we get bright young ones who cannot read an English sentence nor can they write it. As long as the question is in an equation form or that of multiple choice, they have enough strength to find the right answer. (Laughter)

Tukey: It seems that there are a lot of chemical engineers who are running into more and more of these problems. Of course, I see this at Princeton, where chemical engineers have now put in a semester course in their senior year in statistical methods and where somewhere between one-half and two-thirds of their majors take a course in statistics. This is because chemical engineers on the whole have been pretty well awakened to the need of statistics. But even if you look at mechanical engineers, or civil engineers, and compare the percentage of their professional organizations that deal with statistics in their reports with the percentage of the teachers in colleges of these fields who teach some statistics you will find the latter percentage higher, even though it is still pretty small. The engineering faculties seem to be somewhat ahead of the professional groups, but it seems to be hard for them to work in much statistics.

Tukey: I'll take Cuthbert first.

Daniel: I would like to speak a moment about short-sightedness and about some of the people who have this property, in particular, directors of industrial research and Army officials interested in getting statistical work done. I want to speak only about the short-sighted ones. The type of short-sightedness that I speak of is called practicality. Let's be practical, they say. By this they mean: let's see, who we can get to answer this for us by Tuesday -- next Tuesday? This type of practicality is a form of suboptimization and it just occurred to me that it is really disastrous in its effects on teaching and its effects on engineering. The administrator persuades the head of the engineering department, in a weaker moment over a few drinks, that a course in statistics would be useful to his engineers. As soon as the head of the department sobers up, he realizes he has fifty-some other demands. This demand is number fifty-seven. There is a course in surface chemistry, another course in ceramics for chemical engineers, etc. that are on his list of urgent courses - or courses that he has consented to put on his list of urgent courses several years back, or even recently - so there are already many courses competing for positions on an already crowded curriculum. This says something is wrong with the curriculum. I'd like to subscribe to that Princeton Dean's opinion as it was reported in Chemical and Engineering News last week. It was suggested that engineering theory is what has to be taught. Engineering practice and know-how and all of these things which are just so fearfully practical should be taught by the people who are interested only in things that are practical. What this particular dean said - he did not really say it but recognized it very clearly - was that what gets applied is theory. And that the effective way to teach applications is to teach theory with the emphasis in mind that it is theory which gets applied. Not all theory get applied, and of course I don't think that all theory should be taught, but the criterion that decides what theory should be taught is - that which is applicable. Until we get to this point of view with engineering and statistics we are not going to be able even to jam the courses in except by winning power struggles inside the universities. And this is not the way to proceed. The way to proceed is at a different level. We need engineering theory broadened and we need statistical theory broadened so that both of these can be learned by the men who will be engineers.

Tukey: Would you like to say something, Besse?

Day: I just wanted to make the point that a major thing is, and I think it is major in a number of cases, that the administrators think that because they have got a mathematician, they have gotten a statistician. The two disciplines are very different. We have a sad case of that in our laboratories. I know two persons who are fine mathematicians and the heads of the laboratories think that they are well prepared to do statistics. That is far from being true. Of course, I am still harping on the same idea, but statistics is a special discipline, and it takes more than a mathematician. It takes a different training, and I don't like to see it (statistics) in a mathematics department unless it is headed up by somebody who is very broad minded.

Tukey: I have my eye on two hands that have been up here -- what I am going to do is declare a ten minute recess then we will come back together and I will start to recognize those hands.

Tukey: Let us get started again. I will recognize the question right here.

Member of audience: Well gentlemen, you have a heretic in your midst. You have a gentleman here who has asked for the answer on Tuesday - on Friday. I just want to say that you have in your midst a heretic. By heretic I mean a representative of research and development management. (Laughter) A person who Mr. Daniel says wants things by Tuesday and I mean next Tuesday. I came down to this meeting primarily to find out just what is being discussed here this afternoon. And that was to find out just how do statisticians, mathematicians and others of that ilk operate. (Laughter)

Tukey: A word of correction - those ilks. (Laughter)

Member of Audience. The various facets that have been presented here are very interesting. However, I think you can benefit by an objective viewpoint, which I think I can furnish. Primarily I don't think that you should try to make an engineer into a statistician. I can't go along with that. You might devote a little time to making an engineer a better engineer. I don't think you should convert statisticians into engineers either. Dr. Thrall and I both concurred on the fact that we are already short of engineers, and to make one category a little bit more numerous by robbing one that is already short benefits nobody. At least all, or most people, have to worry about getting both of these types of people. So I think what you need to have, from what I could gather these last few days, is - what the lady on my right was talking to me about - I think you need to teach your engineers, shall I say, mathematical and statistical appreciation, so that they can recognize the qualities of this tool that is handed to them. A good many of our engineers do not have this information. A friend of mine in the audience - who I will not name - has already run into this particular thing of teaching engineers statistics and he says that the end product was really something. You come up with a statistician who has all the wrong answers and books to prove it by. So you don't get anywhere with that line of endeavor. I have just two suggestions to make. One is that you do teach them what the, shall I say, areas of limitation are in your particular field. I think the engineer should know that. And the other thing I would suggest is that these statisticians, and mathematicians as well, should go a little bit out of their way to educate the rest of the proletariat that is represented by people like me. Thank you.

Tukey: Does anyone on the Panel want to comment on that?

Terry: Yes. The Laboratories have the following training program for entering engineers and physicists below the doctor's level. (Anyone they hire at the doctor's level is considered to be a specialist and is not given the training program.) In his first quarter of 14 weeks he receives two lectures weekly and two recitations weekly on what I would call elementary statistics and the analysis of data. The latter you may say is quality control at the engineer's bench and on the engineering measuring devices. They also get the same amount of basic physics of wave and basic mathematics. Now of course we know that he has a diploma which proves he has taken these courses, still we find that he finds areas of novelty. And we find it is excellent use of his time. In the next two parts of the year he learns other things - information theory, switching logic, solid state physics, more mathematics, and I think circuit theory. In the next year he gets as an elective the statistical design of experiments, and a course in traffic statistics. At the present time I have a group of about ninety in design of experiment. Some of the lectures and most of the classes are handled by two electrical engineers, one of whom has come up through our training program. In about three years he has come from being a straight electrical engineer, to an electrical engineer who spends about thirty percent of his time serving as a statistician. I think both his supervisor and his subdepartment head would agree that the effectiveness of that whole subdepartment has risen by this one man's additional training. He has saved them from making two blunders, where classical engineering techniques alone would not have sufficed. So I think I would take issue a little bit, and say that you can wisely make statisticians out of engineers. If you get an engineer who is a good half-time statistician, he can increase the effectiveness of a group of ten engineers by at least ten percent per engineer. Which means you get your engineer back - free - plus half another. (Laughter) We have found this so effective that there is no manifest feeling anywhere in the laboratories that this part of the program should be reduced. Indeed, we have other local consultants who are giving "out of hours" courses whenever they can find the energy to do it. And there is always a waiting list of senior engineers, subdepartment heads, and management people who would like to take these courses. Unfortunately we do not have enough time to give as many as are demanded. We have found that it makes money for us, and nothing pleases any engineer more than making money.

Member of audience: I believe I would like to hear from Dr. Thrall over there on that subject, if he would make a few comments.

Tukey: We still have one hand up here at the Panel. Cuthbert, you have the floor next.

Daniel: I want to take issue just a little with the heretic from industry. In the first place, on the grounds of arithmetic, to take a few engineers and make statisticians out of them penalizes the engineering profession indetectably, because there are thousands of engineers graduated every year. But ten more statisticians a year than are now produced would add a very large percent to the available pool of statisticians. So there is not a disproportionate loss. If you add the proportions up you find Terry is only half right - and he often is. (Laughter) Then the case is made against the heretic here.

Tukey: Thrall, you have your hand up, and have been called upon.

Prof. R. M. Thrall (University of Michigan): Well, I would like to comment on this and several other points that have come up. I'm speaking now as a teacher. (The next few remarks were not caught by the tape recorder). One point that was made several times already concerns the relationship between mathematics and statisticians. This is apparently a crucial one. I have had some experience in this connection, because we have made a study of this recently in my own university. We have learned that in the major institutions of this country, which are giving substantial work in statistics at the doctoral level, all but one or two, out of the leading fifteen, fall in one of the following two categories. Either they have created a separate Department of Statistics in the period since 1940 or earlier - most of them since 1940 - or they are seriously considering it now. The other class consists of small universities where the size of the department is such that it doesn't make much difference how it is organized. And I just heard during intermission that one of these is considering separating. Is that right, Sam?

<u>Wilks:</u> You are talking about Princeton?

<u>Thrall:</u> Yes.

<u>Wilks:</u> I don't know. (Laughter)

<u>Thrall:</u> So we are facing here an educational - I won't call it revolution - I think it's an evolution; statistics has changed its nature and seems firmly placed in the educational structure. And, I think it not at all unlikely that in another ten or fifteen years the trend and the rule will be various degrees of separation.

The second point I would like to make to reply to is this business about how the social scientists and educational psychologists teach their statistics. They do teach their statistics at a very early level, and unfortunately, in many cases, at very much the rote learning level. It's just a matter of this formula does this, that formula does that. And the students in these fields, who really need to make use of statistics as a research tool, have to come back and take what we call mathematical statistics on top of the statistics they have already had. So I don't think the people in those fields would consider their courses were entirely successful so far as training in the graduate fields are concerned. However, these courses are viewed as just one stage in the process, and the better students continue with the courses provided by the mathematical statisticians. Of course, this is one reason for the recommendation that every person who takes a bachelor degree in psychology take a course in mathematics before going to graduate school.

The next point gets back a little closer to the one we were just discussing about the role of the engineer in statistics. Here, I agree a little bit with both sides. It is certainly true that an engineer, who learns statistics becomes a very effective statistician, because he can communicate with the engineer. He already possesses the basic engineering background. The man who comes in cold from the outside doesn't have this advantage. But I think that such men must be used, in view of the vast deficiency of engineers with statistics. For we cannot expect to educate, at the undergraduate level, each engineer into an accomplished statistician. The most we can hope for is for the general engineering student to have a speaking acquaintance with statistics, or at least that he will know enough statistics so he will know when he needs a statistician. This would be quite an achievement.

Now I would like to quote the Dean of our Engineering School, who happens also to be a chemical engineer. He has been heard to say that he doesn't care what his students take in high school provided they include both mathematics and English all four years. He has also been heard to say that he now considers engineering as a branch of applied mathematics. In line with this, the University of Michigan has set up a new program, called the basic sciences program, which a student can go through and get the degree of bachelor of engineering without being a specialist in aeronautical engineering or mechanical engineering or other kinds. He will get the basic science tools that he will need in industry, in government laboratories, or wherever he is going in to use them, to learn their specific techniques. This is happening to a number of engineering schools, so I think we ought to mention that there is some progress, although it isn't clear sailing anywhere. There are too many empires sitting around to expect to get away with doing this all at once. But the trend, I think, is in the right direction.

Now, I would like to raise one more question. What does the Panel think should be the proper relation of the statistician in the type of team problems which come up in what is termed Operations Research, or Operations Analysis, etc.? What should be the relationship between the statistician and the operations analyst, and the other people working with them?

Tukey: Who wants to take this operations question?

Wilks: Well, I'll start that. Of course, operations research teams started during the war. I had some connections with one back in 1942, and I remember very distinctly the theory of setting up such a group at that time. I don't know whether it still holds or not. The operational research people who know the present position will have to speak to that because I have lost track of the precise organization of these groups and how statistics fit into it, etc. As I remember, when the Navy group started in 1942 - it started on a particular problem, namely, anti-submarine warfare - the whole concept was that this needed to be a well-balanced group of scientists, a statistician, a physicist, a radar expert -- something like a total of ten or twelve people in the various fields who could tackle the various aspects of the problem. The statistician was brought in to deal with such problems as studying depth charge patterns, optimum search procedures, and the statistical information obtained in all sorts of search and attack effort. He was part of the team. I don't know what the situation is now -- whether they still visualize a team operation for attacking a problem with all the necessary skills, including statistical skills, required by the problem. I assume this is still true. Perhaps someone else could speak on that.

Tukey: Let me ask Sam a question. These problems that you are raising here, that these people were brought in for, were they really statistical problems or probability problems?

Wilks: I would say it was a mixture.

Tukey: In the examples you mentioned - it would be mostly probability.

Wilks: Well, I would say it was something like 50-50. Some of them were very much on the probability side because a lot of work was done at the very earliest stage and before much data or information was available. But as the War went on the statistical people used more and more of the statistical data of military experience in making their studies.

Thrall: I raise this question because of its connection with the very first question raised in the meeting today, how do you expect the statistician to determine his role with the engineer? One of the possibilities here is to have - when you don't have resources of the Bell Telephone people who have statisticians backed by statisticians - your limited research team organized into some group inside the large system that we call operations research, or just research group, or whatever you want to call it. Then that group can serve as a service group and consult with temporary attachments to individual problems. This way they may preserve their identity as statisticians, mathematicians, and so on, which is worth considering.

Day: John, I would like to say a few words.

Tukey: All of us would.

Day: You don't have to say anything.

Tukey: (A few remarks here were not picked up by the tape.)

If you want to begin by combining statisticians with linear programming and game theoretic mathematicians who have not had close touch with the problems, than I am against it. I think the immediate effect of trying to do it right off the bat would be to take the statisticians out of their direct contact with a lot of engineers. I think there is a danger if you put the statistician in with the quasi-modern quantitative techniques. It is going to take him a little too far from the problems. But I certainly hope that statistics will get stirred into those groups by statisticians who have been exposed to it while they walk down the laboratory corridors. Besse?

Day: Well, what I was trying to say is I think the position of the statisticians - though it may depend somewhat on the kind of organization, but in ordinary government testing fields or laboratories, or industrial laboratories - should be at the staff level. They should have the support and confidence of top management. They should be at the staff level for two reasons. One, it gives prestige to their work, and two, their movements are more fluid. They should be at a level with the major units in the laboratory, so they are free to circulate all over the laboratory and so that they will know what's going on. They should be paid - you didn't ask me about that, but I feel very strongly on this subject - they should be paid out of a special fund. Some money should be set aside, so that they are not paid by the projects they work on. If they get paid by the project they work on, even the engineer who is most in favor of them, when things get pretty tight, would be prone to save on the statistical score. They should'nt be paid out of M and O money, because that's so often juggled around. They should be independent. They should be able to work. If they were paid by the project, then the project engineer would be the one to say how much statistics he was going to get on this problem after the statistician has been called in. And I think that would be very bad, because only the statistician knows how much is needed.

Does that answer your question?

Tukey: Churchill, do you have something to say?

Eisenhart: I would second everything Bess has said and comment on this and also the laboratory angle. I am interested in the statistics that Professor Thrall has given us, because one of the difficulties, I think, of teachers of statisticians in universities have been, in the past, the same sort of difficulty experienced by teachers in all borderline subjects. Let us say the statisticians are in a mathematics department, then when questions about promotion in pay come up they kind of get compared with mathematicians who are not exactly their peers in what they do. In fact, until recently, the mathematics department, it seems to me, is likely to be the department least qualified to know how good they are really doing their job. Because until Monte Carlo and empirical methods of solving problems and numerical analysis came in, the mathematics department would likely be the one department on the campus that the statisticians never helped. Unless it was to help them with their grades. (And that's no joke, because, if you have a lot of sections, the problem of getting the grades on an even keel is quite a problem.)

On the matter of pay, in an industrial laboratory they should be paid out of some central fund - certainly not out of a project fund. It would be helpful, I think, in the university if some of those who are working would be paid partly out of some central type of research funds to cover their time when they are helping people in the other departments. This has been done in some places, and I think it is safe to say it is a healthy trend. The best procedure is to set up separate departments and let the statisticians be judged by their peers. We can imagine what a terrific commotion there would be if physicists were under the mathematics department - what the mathematicians would say about Dirac's work, for example, with regard to rigor and things of that sort.

One more point concerning universities. I have been disturbed by the effect, one effect, of government contracts in the field of statistics in that it has tended to keep students working at the same university all during the year. Whereas in the olden days you went to college in the winter time then went out and worked somewhere in the summer. I don't know whether this idea would be feasible, but it seems to me it might be explored on an experimental basis. Some means whereby an ONR or OOR or what-have-you project at a particular university would not only be identified with the investigator but also with some of the graduate students that are working on it. Then they would - I don't want to say be compelled - but be urged to shop around a little bit in the summer time - in particular the unmarried ones. So that they could get a little broadened experience in their field. (Laughter).

Coming on to the government laboratory, Besse covered things quite well. In the two government laboratories I have been in, namely the Wisconsin Agricultural Experiment Station and here at the National Bureau of Standards, we have kept our statisticians, for the most part, in a central pool. From which we report on request and help with whatever comes up. This has a number of advantages, we feel, in that it does give you some opportunity to pick and choose among the problems which are there to work on. The ones you take are the ones you feel are supported by a combination of the man's needs and your ability to really help. If you are in a particular group, and are paid by that group, I anticipate that you will be obliged sometimes to spend some of your time on things that if you could escape them you could more profitably spend your time elsewhere. We have in the audience, or at least did have in the audience, a man who has never been in my laboratory but who is at the Bureau, who represents the other school. But I believe the difference is only in degree.

Tukey: Cuthbert!

Daniel: I want to talk about a section of our society where statisticians do fit in, and will fit in more and more, and that is industrial statistics. I am not sure that this is the main focus of interest of the audience. Industry is going to do certain things - there's no doubt about it - about statistics. Right now it's doing them, and all we can do really is encourage it to do it a little bit more, because what it is doing is roughly right. It's encouraging men who are not statisticians to go into statistics: men who are through college; men who were hired for some other purpose; men who have competence; really completely opposite to what our heretic recommended. Industry has got to do this more and more because it isn't getting anything like as many statisticians under the present arrangements as it needs.

Industry needs to give men time to think in this field, and that doesn't mean time to think about what they are going to do next, but just time to think. This means time to read books. A scandalous thing, a man sitting there doing nothing - he's reading a book! This is not a generally permitted practice in industry.

It is sometimes a matter of policy not to let a man think, he is supposed to do things. He can do his thinking someplace else: ....do it at night or something. He doesn't get time to sit still somewhere where there are neither four computers nor three typewriters nor five telephones, and think about what he ought to be doing, or to read a book, or to take a course on the company's time at full rates. None of these things are being done by industry - done very much by any industry. They will be done more and more. The Bell Laboratories is clearly a place in which these things are carried forward, but this isn't where I looked when I made my list.

It is clear that a great many more concessions have to be made. They have the effect right now of doubling a man's status before the demand begins to reach the supply. The universities will play some part in this, and by knowing where to look some can tell which way the very mild wind is blowing inside the universities. The wind I speak of is a 10 inch wind that is really blowing and will be taken care of from the other side, so to speak, by industry getting away from its short sightedness in wanting practical results, which means results tomorrow. You can get results - you can always get results tomorrow - but they are not good enough. That's why we need this kind of training. But take a little longer view of it, and that view includes even breaking down some of the matters of policy that are not quite rigid. Don't make men take courses at night. Don't say: men can take courses at night if they want courses. Don't say: make them study at home if they want to study; if they want some peace and quiet let them get it at home. Most men's homes have more bedlam than the laboratory. (Laughter) The laboratory could change this - the homes can't it seems. All these things I want to suggest to you are beginning to be done, and one of the main jobs, in my view, and indeed one that forms the main emphasis in my own work is to encourage some managers to do more of this.

Then we will have statisticians that fit in, and the problem of how they fit in - such organizational questions as should they be staff or line, should there be one in each group, or should they stand vertically or horizontally when they do their work - all of these things will get settled, after we have some statisticians to talk about.

Tukey: Well, you aren't going to object to the philosophy, Cuthbert, that says you do what you can to make sure that they have a reasonable amount of flexibility and that the individual's services are used.

Daniel: No.

Tukey: You're going to try to make efficient use of an individual. Whether he decides to go along or not may be something else. If you have a scarce commodity, it's going to pay to move it around.

Daniel: If you've got a small commodity that's self-reproducing - rather self-propagating - the most efficient way to use it is to make it propagate and not to consider how to get the last drop of blood out of the men you have - in statistical output.

Tukey: One of the good ways to propagate is sometimes through consultants. That is the way some of the engineers get started to be converted. The consultants will train some engineers and they in turn will work with other engineers.

Eisenhart: There are just two points I want to make about efficient use of statisticians. I think that all who have had any experience in applying statistics - even Cuthbert who said he didn't go down to see this dog biscuit machine - that you do gain a great deal of benefits from going over to see the particular set-up where you are going to apply it. And the second advantage is that if you are over with the man who is consulting with you, then you can decide when to leave, but if he is in your office you can't always get him out tactfully. (Laughter)

Tukey: At least you're convinced that these people can talk. Jay, you had your hand up a while ago?

Professor Emil H. Jebe (Iowa State College): I would like to speak to a point that Thrall brought up about how statisticians fit into an Operations Research group. I am a member of an OR standby unit. I am the statistician. I'd like to talk a little about the kind of experience I've had coming in as a statistician among people that were physicists and engineers. First they wanted me to give them a series of lectures. Well, I tried that for several meetings. Now all these fellows were individualists and after we had gone over a few of the elementary ideas, we were pretty soon thrashing things around and arguing. I thought that the only way they can get ahead now would be to start reading books, like Cuthbert suggested. The next thing that happened was that we were asked to give a paper or something of that sort. Well I had a little bit of a project running myself on which I had made some analysis of the data, and I turned this over to them and they argued some. The results were moderately interesting. Perhaps they didn't think much of this data anyway. But we did this another time, passed the data around, wrote up a little report on it, and this got passed around. Finally we got some better data on the same subject that they had looked at earlier, and we got an opportunity to present this in a meeting. Then this got to be considered as hot stuff, at least they thought so, and we became more popular. As more projects came along, they finally got to the point where they didn't want me to decide what they should do. Instead they would ask what does a statistician think about what should be done. I don't know whether that helps answer your question, but it does give a little indication of how a statistician may fit into one of these Operations Research groups.

Tukey: What I don't understand is why you broke off the arguments. I always thought people learn more by arguing than any other way.

Jebe: Well, I said these were pretty strong individualists, and after all I was outnumbered. (Laughter)

Tukey: Are there any more questions? Seeing none, since Wilks has asked for a chance to make a brief statement, I shall turn the meeting over to him.

Wilks: I would like to take this final minute to express the thanks of the Army Mathematics Advisory Panel and the Office of Ordnance Research to all of the participants in this program. We also want to thank the National Bureau of Standards and the Diamond Ordnance Fuze Laboratories for serving as our hosts. In particular we should like to thank Mr. John Wheeler who has carried the load of making the very excellent local arrangements for this conference. (Applause) As you know, our intention is to have all of these papers, or at least most of them - all we can get, let's say - brought together and issued in a Proceedings. I think this is all unless someone else has an announcement. So I'll adjourn the meeting. (Applause)