COPY

707716

SDC PROPOSAL 384

U. S. Atomic Energy Commission

LINEARIZED NONLINEAR REGRESSION METHODS FOR STATISTICAL ANALYSIS OF BIOLOGICAL DATA ON EXPOSURES TO RADIATION

February 1964

System Development Corporation
2500 Colorado Avenue
Santa Monica, California

REPOSITORY OAK RIDGE/ENERGY SYSTEMS/ORNL (X-10)
OFFICE OF RADIATION PROTECTION
COLLECTION K. Z. MORGAN'S FILES (PAST DIRECTOR OF HEALTH PHYSICS)
BOX No. BLD. 4500 S. WING H ATTIC
FOLDER

A-00122
Human Subjects Project

1026159

ABSTRACT

This study is concerned with the theoretical development and practical use of methods for suitably analyzing biological data on exposures to atomic radiation. In particular, methods will be developed for analysis of the data resulting from the exposure in 1954 of persons on the Marshall Islands to fallout. Due to heterogeneity in the types of data encountered, statistical methodology with substantial flexibility for curve-fitting and an exceedingly general probability basis is needed. Appropriate extensions of the linearized nonlinear regression technique should furnish this methodology. Also, with sufficient ingenuity, foresight, and careful planning, general and feasible application of this methodology by use of a high-speed computer should be attainable. One type of method is concerned with the construction of paired comparisons between composite individuals that are approximately the same except for the characteristic being investigated. Probability models for analyzing paired comparisons of this nature will be developed. Also, methods of estimating values for missing data, of isolating important effects, of combining different sources of data, of interpolating with respect to levels of radiation dosage, of predicting outcomes for other situations, etc. should be obtainable by appropriate use of the basic methods. Confidence intervals, significance tests, and point estimates will be developed. This work will be carried out over a period of three years.

1026160

TABLE OF CONTENTS

	PAGE
TECHNICAL DISCUSSION	1
I. BACKGROUND OF STUDY	1
II. SCIENTIFIC SCOPE	3
A. Introduction	3
B. Linearized Nonlinear Regression Model	5
C. Associated Probability Model	5
D. Construction of Modified Paired Comparisons	6
E. Analysis of Modified Paired Comparisons	8
F. Missing Data Estimation	9
G. Application Difficulties	9
H. Final Comments	11
I. Tentative Schedule of Work	11
J. References	12
CORPORATE RESUME	13
I. SYSTEM DEVELOPMENT CORPORATION TECHNICAL QUALIFICATIONS	13
A. Corporate Facilities	13
B. Corporate Research	13
II. SCIENTIFIC PERSONNEL	16
Education	17
Professional Experience	17
Academic Experience	18
Books	18
Publications	19
Society Memberships	26
Additional Information	26

1026161

TECHNICAL DISCUSSION

I. BACKGROUND OF STUDY

The desire to perform a thorough and meaningful statistical analysis of some exceedingly important data on exposure of persons to radiation is the principal motivation for this study. These are the data that have been obtained, and will be obtained, with respect to persons who were exposed to fallout at the Marshall Islands in 1954. Reports on the gathering of these data through 1962, and statements of some of these data, are furnished by ref. [1], [2], and [3]. Here, for comparison purposes, data have also been obtained, and will continue to be gathered, on a control group consisting of unexposed persons who are located in the Marshall Islands.

Over-all, the persons in the exposed group differ from those of the control group with respect to values of physical and environmental characteristics that seem to be important. Thus, determination of whether differences between exposed and unexposed persons are actually due to the radiation is difficult. That is, clinical differences between the groups could be due to differences in the values of these characteristics rather than to the effects of radiation. This situation is further complicated by the presence of statistical variation. Some method that isolates the effects due to radiation and investigates these effects in a probabilistic manner (so that statistical variation receives consideration) is needed.

The statistical method that is developed for this purpose needs to be of a general nature, since specialized information about the probabilistic properties of the Marshallese data is lacking. That is, meaningful verification of specialized probability properties does not seem to be possible. Fortunately, a general nature model with a probabilistic basis and the required capabilities for isolating and investigating specified effects can be developed.

Extensions of the linearized nonlinear regression model that is outlined in the next section (also see ref. [4]) would seem to furnish methods that are satisfactory for analysis of the Marshallese data. For example, a modification

1026162

of comparisons involving pairs of persons, where each pair contains an exposed person and an unexposed person, can be developed. Here, the purpose is to investigate a specified clinical characteristic with regard to the effects of radiation. Instead of using actual persons, two composite persons are constructed for each pair (with respect to observed values of the clinical characteristic being investigated). The composite persons of a pair are the same with respect to the values of the physical and environmental characteristics that are considered to be important. Thus, the difference in the observed values for a modified pair can be attributed to the effects of radiation (to a reasonable approximation). Then, by development of suitable statistical techniques for analysis of modified paired comparison data, the effects of radiation can be separately investigated for a number of clinical characteristics.

The linearized nonlinear regression approach also has many uses in estimating values for missing data. One application for the Marshallese data is in estimating unknown ages on the basis of ages that are known and characteristics that are strongly related to age.

Another possible use of extensions of the linearized nonlinear regression model is with respect to aging studies for the Marshallese. Other uses include performing interpolations and extrapolations with respect to the levels of radiation that were received.

Although many additional difficulties would undoubtedly be encountered, one possible use of the linearized nonlinear regression approach would be to meaningfully combine the Marshallese data with that which has been obtained with respect to the atomic bombings in Japan. An over-all analysis of these combined data might furnish much worthwhile information that would not be obtained from either set of data considered separately.

The methods to be developed are not limited to uses for the Marshallese data. The generality of the linearized nonlinear regression model indicates that these methods will also be usable for other sources of data and other kinds of investigations.

1026163

II. SCIENTIFIC SCOPE

A. Introduction

With respect to investigation of a specified characteristic, each person can be represented by a vector. The first coordinate of this vector can correspond to the characteristic being investigated while the other coordinates correspond to the physical and environmental characteristics that are believed to have an important influence on the probability distribution of the first coordinate. Each characteristic has a set of possible values (or levels) and the value of the vector that occurs for a person is obtained by giving each coordinate the value of the corresponding characteristic. Notationally, a vector is of the form $(y;x_1,\ldots,x_k)$, where y corresponds to the characteristic being investigated and the x's correspond to the k other characteristics that are believed to be important with respect to the probability distribution of y. Let $y_i,x_{1i},\ldots,x_{ki}$ be the values of these characteristics that occur for the i-th person of the n persons that are considered. Then, the i-th person is represented by $(y_i;x_{1i},\ldots,x_{ki})$.

For the situations considered, the values of the x's are assumed to be fixed quantities and the value of y is considered to be a random variable. A regression approach is convenient for investigating y in such a manner that the influence of the x's receives explicit consideration. Stated in a general manner, y is expressed in the form

$$y = h(x_1,\ldots,x_k;A_1,\ldots,A_t) + e',$$

where the function h is completely specified except for the values of the "regression coefficients" $A_1,\ldots,A_t$ and e' is a random variable. A "true" set of values exists for the A's, but these values are virtually always unknown. Thus, the A's occur as unknown parameters. Ideally, h should be chosen so that the general magnitude of e' is as small as possible. The function $h(x_1,\ldots,x_k;A_1,\ldots,A_t)$ is called the regression function of y on $x_1,\ldots,x_k$. Using regression terminology, y is the dependent variable and $x_1,\ldots,x_k$ are the independent variables for the regression.

1026164

Using this representation, the observed value of y for the i-th person is expressed in the form

$$y_i = h(x_{1i},\ldots,x_{ki};A_1,\ldots,A_t) + e_i \ .$$

It is often reasonable to assume that the e_i are statistically independent. However, for heterogeneous situations such as that for the Marshallese, this is one of the few assumptions that is acceptable. There is seldom much justification for the assumptions that the e_i have zero expected values, that they are sample values from the same population, that they have the same variance, etc.

The lack of specialized information about the probability properties of the e_i is one of the fundamental difficulties encountered in analyzing heterogeneous data such as that occurring for the Marshallese. Other difficulties occur with respect to selection of the function h. If an elementary form, such as linear, is used for h, isolation of effects due to the independent variables is rather easily accomplished. For example, composite persons for paired comparisons can be constructed without substantial computational effort. On the other hand, use of an unsuitable form for h can result in poor agreement between y_i and $h(x_{1i},\ldots,x_{ki};A_1,\ldots,A_t)$ over the values of i. In many cases, the failure of a regression function to provide a good fit to the data may be largely due to the restricted curve-fitting capability of the functional form that is used for h. A basic problem in obtaining regression functions that are suitable for heterogeneous data is to greatly increase the curve-fitting capability of the regression function without substantial changes in the desirable manipulation-computation properties that occur for elementary forms such as linear. The linearized nonlinear regression model seems to furnish a satisfactory solution to this problem.

An outline of the linearized nonlinear regression model is given next. Following this is an outline of the associated probability model. A more thorough description of this approach is given in ref. [4].

1026165

B. Linearized Nonlinear Regression Model

Let the range of possible values for y be $y_L \leq y \leq y_U$. The approach for linearized nonlinear regression consists in expressing the regression function in a transformed manner. Specifically, let $g_1(y),\ldots,g_s(y)$ be specified functions of y while $g_{s+2}(x_1,\ldots,x_k),\ldots,g_t(x_1,\ldots,x_k)$ are specified functions of $x_1,\ldots,x_k$. Then, y is implicitly expressed in the form

$$(1)\quad \begin{aligned} &y+A_1g_1(y) + \ldots + A_sg_s(y) \\ &\qquad = A_{s+1}+A_{s+2}g_{s+2}(x_1,\ldots,x_k) + \ldots + A_tg_t(x_1,\ldots,x_k)+e'', \end{aligned}$$

where $A_1,\ldots,A_s$ are such that the lefthand side of (1) is a monotonic function of y for $y_L \leq y \leq y_U$. Let $h(x_1,\ldots,x_k;A_1,\ldots,A_t)$ be the solution of (1) for y when $e'' = 0$. Then, expression (1) is equivalent to an expression of the form

$$y = h(x_1,\ldots,x_k;A_1,\ldots,A_t) + e',$$

where the function h can have a substantial amount of curve-fitting capability.

The form of (1) allows linear manipulations to be used in isolating effects that are expressed in terms of the A's. Since y is a monotonic function of

$$A_{s+1} + A_{s+2}g_{s+2}(x_1,\ldots,x_k) + \ldots + A_tg_t(x_1,\ldots,x_k),$$

when statistical variation is neglected, isolation of specified linear combinations of $A_{s+1},\ldots,A_t$ is of special interest. Although the restriction imposed on $A_1,\ldots,A_s$ places some limitation on the use of linear manipulations, the computational aspects of the linearized nonlinear model seem to be at a manageable level.

C. Associated Probability Model

The dependent variable y_i for the i-th person is considered to have a probability distribution but the independent variables $x_{1i},\ldots,x_{ki}$ are considered to be

fixed. The y_i are assumed to be statistically independent (i.e., the e_i are independent) but the shape of the distribution for y_i does not necessarily have any relation to the shape of the distribution of y_j if $i \neq j$.

The key feature of the probability model is the procedure used to define what the parameters $A_1,\ldots,A_t$ represent. These definitions are intuitively meaningful and also allow useful probability results to be developed for heterogeneous situations. This procedure is a generalization of the median estimation concept.

Suppose that the total number n of persons is not too small. By some data manipulations (see ref. [4]), a few "observations" $Y(u,v)$ can be constructed that are independent, approximately continuous, and of the form

$$Y(u;v) = A_v + e(u;v),$$

where $e(u;v)$ is a random variable. Let $p(u;v)$ be the (unknown) value of $P[Y(u;v) \leq A_v]$. Then, the (unknown) value of A_v is defined by the requirement that the arithmetic average of the $p(u;v)$ over u is equal to $\frac{1}{2}$. Now, by the methods given in ref. [4], an approximate median estimate and approximate confidence intervals can be obtained for the true but unknown value of $A_v (v = 1,\ldots,t)$.

The model of ref. [4] is limited to the obtaining of probabilistic results for the individual A_v. For the paired comparison and some other applications to the Marshallese data, probability results for specified linear combinations of the A_v will be needed. This will require an extension of the probability model that is given in ref. [4].

D. Construction of Modified Paired Comparisons

In a general way, let us consider a couple of procedures that might be used to construct the composite persons for a pair that is used in a modified paired comparisons investigation of whether radiation effects are statistically significant. Under the null hypothesis, radiation has no effect and the same

1026167

linearized nonlinear regression model can be used for exposed persons and for the unexposed persons of the control group.

First, the values of $x_1,\ldots,x_k$ that are to occur for this pair are specified. Then, one problem is to construct an exposed person with these values for the x's. That is, a suitable subset of the data on exposed persons is manipulated to determine an observed value of y that, according to the linearized nonlinear regression model, corresponds to a person with these values for the x's. This value is used as if it were observed for an exposed person with these x's. The same procedure is used to determine the observed value for an unexposed person with these values for the x's. These two y values then constitute the observations for this pair. Here, nonoverlapping subsets of data are used for the various pairs that are considered. Thus, a disadvantage of the construction of composite persons is that the data for a number of persons is needed in order to construct one composite person.

One way of determining an observed value for y is to use a suitable subset of the data to individually estimate $A_1,\ldots,A_s$ and to estimate the value of

$$A_{s+1} + A_{s+2}g_{s+2}(x_1,\ldots,x_k) + \ldots + A_t g_t(x_1,\ldots,x_k).$$

Using these estimates as if they were the true values, solution of

$$y + A_1 g_1(y) + \ldots + A_s g_s(y)$$
$$= A_{s+1} + A_{s+2}g_{s+2}(x_1,\ldots,x_k) + \ldots + A_t g_t(x_1,\ldots,x_k)$$

determines an observed value for y.

A second way of determining an observed value for y is to linearly combine the data on t + 1 persons so that the weighted sum of the individual regression expressions is of the form that occurs when the specified values $x_1,\ldots,x_k$ occur. That is, weights w_i (not necessarily positive) are determined so that

$$\sum_{i=1}^{t+1} w_i[y_i + A_1 g(y_i) + \ldots + A_s g_s(y_i)]$$
$$= \sum_{i=1}^{t+1} w_i y_i + A_1 g_1\left(\sum_{i=1}^{t+1} w_i y_i\right) + \ldots + A_s g_s\left(\sum_{i=1}^{t+1} w_i y_i\right)$$

and

$$\sum_{i=1}^{t+1} w_i = 1, \qquad \sum_{i=1}^{t+1} w_i g_r(x_{1i},\ldots,x_{ki}) = g_r(x_1,\ldots,x_k), \qquad (r = s+2,\ldots,t).$$

Then, $\sum_{i=1}^{t+1} w_i y_i$ is the observed value for y.

E. Analysis of Modified Paired Comparisons

Having constructed the modified paired comparisons, the difference of the values for each pair can be formed. More generally, each value of a pair can be transformed by use of the same function and the difference of the two transformed values can be formed. These differences can be used to compare the exposed and the unexposed individuals (with respect to a specified characteristic).

Due to the rather general ways in which the modified pairs are constructed, special statistical procedures will be needed for the analysis of the differences. Even though the composite persons of a pair have the same values for $x_1,\ldots,x_k$, the observed values for y do not necessarily have the same distribution under the null hypothesis. Hence, a difference is not necessarily symmetrically distributed about zero under the null hypothesis. Also, the distribution for one difference may be greatly different from that for another difference. In spite of these difficulties, development of statistical techniques that are satisfactory for analyzing data such as these independent differences seems to be definitely possible.

1026169

One potential method, which seems to be especially suitable when the second way is used to construct the composite persons, is to investigate the value of a special parameter. For each difference, consider the (unknown) probability that the difference is less than the parameter. The (unknown) value of the parameter is defined by the requirement that the average of these probabilities is 1/2. Then, an approximate median estimate and approximate confidence intervals can be developed for the value of this parameter. In particular, procedures can be developed for testing whether this parameter is zero, which seems to be a reasonable choice for its null value in most cases.

Another potential method might be based on ref. [5] (also see ref. [6]). Here, a suitably chosen function of the medians of the distributions for the differences is investigated (for example, the arithmetic average of these population medians). By suitable modifications or extensions of the material in ref. [5], approximate equal-tail confidence intervals can be obtained for the function of medians that is investigated. These confidence intervals yield two-sided significance tests of the null value for the function of the medians. Usually, zero seems to be a suitable choice for this null value.

F. Missing Data Estimation

For estimating a missing value, the dependent variable can be the characteristic for which a value is desired. The independent variables can be characteristics whose values are known for the person considered and which seem to have an important influence on the characteristic for which the value is missing. Then, using the available data, point estimates are obtained for the parameters that occur in this linearized nonlinear regression function. Solution of the resulting regression equation (with the random effect term equal to zero) for the value of the dependent variable furnishes an estimate of the missing value.

G. Application Difficulties

In principle, the methods outlined in the preceding parts of this section seem to be reasonably straightforward. However, the practical application of these methods involves many difficulties.

1026170

For given $s,t,g_1,\ldots,g_t$, the necessary data manipulations and evaluations are nearly always such that a high-speed computer is needed for feasible applications. This, combined with the huge amount of freedom in the choices for s,t, and the functions, causes application difficulties. Namely, developing a computer program is expensive and time-consuming. Thus, only a very limited number of such programs can be developed for a study such as the one being proposed.

Fortunately, the greatest difficulty occurs with respect to the choices for s and for $g_1(y),\ldots,g_s(y)$. For given $x_1,\ldots,x_k$, the functions $g_{s+2},\ldots,g_t$ occur as numbers whose values can usually be determined without great difficulty. Thus, it is sufficient to make allowance for $t - s - 1$ numbers, where $t - s - 1$ can have a range of values.

The primary purpose of $g_1(y),\ldots,g_s(y)$ is to furnish sufficient curve-fitting capability. If y_L and y_U are finite, this can often be accomplished by letting $s = 2$, $g_1(y) = y^2$, and $g_2(y) = y^3$. The choice of $s = 3$, $g_1(y) = y^2$, $g_s(y) = y^3$, and $g_3(y) = y^4$ is available if greater curve-fitting capability is desired. It is anticipated that finite values can be used for y_L and y_U in virtually all applications to the Marshallese data.

However, even this use of cubic or quartic polynomials in y involves difficulties. That is, the conditions imposed on $A_1,\ldots,A_s$ are somewhat complicated, so that the special methods for estimating these parameters will likely also be complicated.

In addition, there are difficulties in evaluating the weights w_i that occur in the second way for developing composite persons. However, this evaluation is simplified somewhat when powers of y are used for $g_1(y),\ldots,g_s(y)$.

Hence, a substantial part of the development of the analysis methods will be concerned with computer programming. Also, some insight, ingenuity, and careful planning will be needed in selecting the form of the model that is used and the kinds of observation groupings that are to occur for the data manipulations. Although randomized grouping has advantages in some cases,

1026171

groupings which assure that the situations are of an interpolation nature may be needed for the construction of composite persons.

H. Final Comments

As described in the preceding parts of this section, the development of suitable statistical methods and the programming and use of a high-speed computer are important aspects of this proposed study. With respect to the statistical methods, fundamental developments in this general area have already been made by the Principal Investigator. Also, he should not experience much difficulty in extending these methods in the ways that are outlined. With respect to programming and use of high-speed computers, System Development Corporation is a leader in this field.

I. Tentative Schedule of Work

This proposed study is for a duration of three years. The first year will be spent making extensions of existing statistical methodology and in programming subroutines for some of the data manipulations that will be used. Except for the development of analysis methods for the differences of values from paired comparisons, the extensions of statistical methodology should be completed by the end of the second year. Programming of subroutines will begin as soon as methods are developed and the form that these methods should take for computer application are determined. All extensions of methodology should be completed by the middle of the third year and all programming should be completed at least three months before the end of the study. The final three months will be spent in developing the necessary documentation.

Some use of the methods should be available by the end of the first year. In particular, if the need occurs, small scale analyses for special situations may be made several times during the study. However, the budget may need to be increased if more than two or three special studies (involving hand-computing or special computer programming) are required.

It is anticipated that this study will result in several research papers by the Principal Investigator. Some of his time will be spent in writing these papers and in delivering them at professional meetings. This might include presenting some of this material at international meetings.

1026172

J. References

[1] Robert A. Conard et al, Medical survey of Rongelap people five and six years after exposure to fallout, Report BNL 609 (T-179), Brookhaven National Laboratory, September 1960

[2] Robert A Conard et al, Medical Survey of Rongelap people seven years after exposure to fallout, Report BNL 727 (T-260), Brookhaven National Laboratory, May 1962

[3] Robert A. Conard et al, Medical survey of Rongelap people eight years after exposure to fallout, Report BNL 780 (T-296), Brookhaven National Laboratory, January 1963

[4] John E. Walsh, "Use of linearized nonlinear regression for simulations involving Monte Carlo," Operations Research, Vol. 11 (1963), pp. 228-235

[5] John E. Walsh, "Some bounded significance level properties of the equal-tail sign test," Annals of Mathematical Statistics, Vol. 22 (1951), pp. 408-417

[6] John E. Walsh, HANDBOOK OF NONPARAMETRIC STATISTICS: Investigation of Randomness, Moments, Percentiles, and Distributions, D. Van Nostrand Co., 1962, pp. 169-170

1026173

CORPORATE RESUME

I. SYSTEM DEVELOPMENT CORPORATION TECHNICAL QUALIFICATIONS

System Development Corporation (SDC), Santa Monica, California, was organized as a nonprofit corporation in 1956. SDC specializes in the development of large information processing systems.

A. Corporate Facilities

The five buildings that comprise the Santa Monica facility have approximately 600,000 square feet of floor space. In addition to offices, these buildings house extensive computing facilities including an IBM 7090, a Philco 2000, a CDC 1604, 924 and 160A, an IBM AN/FSQ-32V, a duplexed IBM AN/FSQ-7 and several IBM 1401's.

The Philco 2000 computer serves as the central computing element for the Systems Simulation Research Laboratory (SSRL), a general purpose simulation facility of some 20,000 feet of floor space. The SSRL is used to support the Corporate research effort.

B. Corporate Research

Basic research in the System Sciences is the theme of SDC's independent research program. By policy, SDC has not interested itself in the development of hardware. Therefore, the research does not in general deal with hardware nor does it emphasize basic research in the physical sciences. Rather the emphasis is on the functional use of equipment and men in information processing systems. SDC is particularly interested in systems based upon digital computers and conducts research on the flow of information in communication systems, on the way in which this information is handled within a computer, and on the display of information for human interpretation. Closely allied to this interest is SDC's concern with man-machine systems. The research is aimed at maximizing the operational efficiency of systems of this kind and in optimizing the match between the man and the machine.

1026174

This extensive program of independent research is conducted in the Corporation's Research Directorate. The Directorate's staff of some 90 professionals includes 26 human-factors scientists, 32 information-processing scientists, 20 mathematics and operations research scientists, and 12 engineers and other specialists. Their aggregate professional experience equals 980 man-years, of which 440 have been spent at SDC. They have published a total of 480 professional publications while at SDC. To complement the resources of this staff, about 50 consultants annually are asked to contribute their specialized knowledge to the research effort, over periods ranging from one day to several months.

Also part of the Research Directorate is the Systems Simulation Research Laboratory, an advanced facility for research on man-machine systems. Built around the data-processing capabilities of a high-speed digital computer (a Philco 2000), the Laboratory is capable of accommodating a wide variety of simulation models for both real and hypothetical computer-based man-machine information-processing systems. Twenty-five different research studies have used the Laboratory, ranging from large-scale simulations of man-machine systems to use of only the computer for program development.

Among current SDC projects directly relevant to this proposal are several focused upon natural language processing and upon the solution of problems in intelligence. Among the former are Proto-Synthex and Mechano-Linguistics; the latter include applications of ideas developed by Lauren Doyle of SDC's Research Directorate. An integrated program of research is also currently being conducted by the Directorate in the following areas:

1. Man-machine Systems - Aimed at optimizing the operation of computer-based information-processing systems.
2. Artificial Intelligence - Aimed at getting computers to perform new tasks and exhibit intelligent behavior.
3. Information Processing - Aimed at investigating information processing procedures relevant to digital computing machinery.
4. Decision Processes - Aimed at investigating the nature of human decision-making in a systems context.

1026175

5. Mathematics and Operations Research - Aimed at devising mathematical techniques for the solution of operational problems.
6. Information Retrieval - Aimed at the rapid and accurate retrieval of stored information.
7. Programming Languages - Aimed at improving procedure-oriented languages for computers.
8. Automated Education - Aimed at investigating the uses of computers for instructional purposes.

The studies within each of these areas represent the most significant from among many which were proposed. Final selection was based on inherent promise of a successful outcome. An important further consideration was the applicability of the results to the aims of SDC and to the military agencies who rely on its services.

The management of SDC is well qualified to direct a research program of this sort. Corporate executives have participated in and directed military, industrial, and academic research. Their experience is supplemented by that of selected representatives from SDC's staff and line who make up the Research Committee. As further assurance of the breadth and depth of the program, an Advisory Committee on Research, composed of six outstanding scientists and advisors outside the Corporation, was formed. This Committee, which meets semiannually, provides a broad-gauge input from the scientific community to SDC's research program and makes objective reports to the Corporation.

1026176

II. SCIENTIFIC PERSONNEL

The principal investigator is John E. Walsh. Dr. Walsh will devote an average of 40 percent of his time to this study over its three year span. Following is a resume of his background and experience.

1026177

JOHN E. WALSH

Education

University of Notre Dame, Notre Dame, Indiana; 1937-1941; B.S., Mathematics

University of California at Los Angeles, Los Angeles, California; 1943-1945; M.A., Mathematics

Princeton University, Princeton, New Jersey; 1945-1947; Ph.D., Mathematics

Professional Experience

1958 Present	System Development Corporation, Santa Monica, California (Senior Scientist for Operations Research and Statistics) - Measures of effectiveness, mathematical models, and simulation procedures for operations research studies; statistical models and methodology for studies of medical data; consulting involving operations research, mathematical statistics, and applied mathematics.
1954 1958	Lockheed Aircraft Corporation, Burbank, California (Research Specialist (Military) Operations Research Division) - Measures of effectiveness, mathematical models, and simulation procedures for military and industrial operations research studies; statistical models and analysis of data; consulting involving operations research, mathematical statistics, and applied mathematics.
1951 1954	U. S. Naval Ordnance Test Station, Inyokern (Mathematician in Central Evaluation Group) - Mathematical models and measures of effectiveness for operations research studies involving weapons; statistical models and research; consulting involving operations research, mathematical statistics, and applied mathematics.
1951	U. S. Bureau of Census, Washington, D. C., (Consultant) - Consultant in mathematical statistics and linear programming
1948 1951	The RAND Corporation, Santa Monica, California, (Associate Mathematician) - Mathematical models, measures of effectiveness, and simulation procedures for operations research studies involving offense and defense systems; statistical models and research; consulting in operations research, mathematical statistics, and applied mathematics.

1026178

1947 1948	Douglas Aircraft Company, Project RAND, Santa Monica California, (Design Specialist A in Evaluation Group) - Mathematical models, measures of effectiveness, and simulation procedures for operations research studies involving offense and defense systems; statistical models and research; consulting in operations research, mathematical statistics, and applied mathematics.
1945 1947	Princeton University, Princeton, New Jersey, (Research Assistant) - Research in mathematical statistics for Office of Naval Research.
1941 1945	Lockheed Aircraft Corporation, Burbank, California (Research Analyst A in Mathematical Analysis Group) - Member of three man team that did project and consulting work in operations research, mathematical statistics, and applied mathematics.

Academic Experience

1949 1951	UCLA, Visiting Assistant Professor, Mathematics Department
1952	UCLA, Lecturer, Mathematics
1956	West Coast University, Lecturer (graduate courses), Mathematics
1958 1959	UCLA, Lecturer, Engineering Department
1962	Visiting Professor of Statistics, Stanford University
1948 1959	University of California, Extension teaching
	Taught courses in mathematical statistics, probability, and information theory.

Books

1. HANDBOOK OF NONPARAMETRIC STATISTICS: Investigation of Randomness, Moments, Percentiles, and Distributions, by John E. Walsh. D. Van Nostrand Company, Inc., 1962, 575 pages.
2. Contributions to Order Statistics. Edited by A. E. Sarhan and B. G. Greenberg, (Walsh wrote Chapter 8). John Wiley and Sons, 1962, 507 pages.
3. HANDBOOK OF NONPARAMETRIC STATISTICS II: Results for Two and Several Sample Problems, Symmetry, and Extremes, by John E. Walsh. D. Van Nostrand Company, Inc., in press, approximately 600 pages.

1026119

Publications

1. "Some significance tests based on order statistics" by John E. Walsh; Annals of Mathematical Statistics, Vol. XVII, No. 1, March 1946, pp. 44-52.

2. "Some order statistic distributions for samples of size four" by John E. Walsh; Annals of Mathematical Statistics, Vol. XVII, No. 2, June 1946, pp. 246-248.

3. "On the power function of the sign test for slippage of means" by John E. Walsh; Annals of Mathematical Statistics, Vol. XVII, No. 3, September 1946, pp. 358-362.

4. "Concerning the effect of intraclass correlation on certain significance tests" by John E. Walsh; Annals of Mathematical Statistics, Vol. XVIII, No. 1, March 1947, pp. 88-96.

5. "An extension to two populations of an analogue of Student's t-test using the sample range" by John E. Walsh; Annals of Mathematical Statistics, Vol. XVIII, No. 2, June 1947, pp. 280-285.

6. "On the power efficiency of a t-test formed by pairing sample values" by John E. Walsh; Annals of Mathematical Statistics, Vol. XVIII, No. 4, December 1947, pp. 601-604.

7. "On the use of the non-central t-distribution for comparing percentage points of normal populations" by John E. Walsh; Annals of Mathematical Statistics, Vol. XIV, No. 1, March 1948, pp. 93-94.

8. "Some significance tests for the median which are valid under very general conditions" by John E. Walsh; Annals of Mathematical Statistics, Vol. XX, No. 1, March 1949, pp. 64-81.

9. "On the 'information' lost by using a t-test when the population variance is known" by John E. Walsh; Journal of the American Statistical Association, Vol. 44, March 1949, pp. 122-125.

10. "On the range-midrange test and some tests with bounded significance levels" by John E. Walsh; Annals of Mathematical Statistics, Vol. XX, No. 2, June 1949, pp. 257-267.

11. "Applications of some significance tests for the median which are valid under very general conditions" by John E. Walsh; Journal of the American Statistical Association, Vol. 44, September 1949, pp. 342-355.

12. "Some comments on the efficiency of significance tests" by John E. Walsh; Human Biology, Vol. 21, No. 4, December 1949, pp. 205-217.

1026180

13. "On the best choice of sample sizes for a t-test when the ratio of variances is known" by John E. Walsh; Journal of the American Statistical Association, Vol. 44, December 1949, pp. 554-558.

14. "Concerning compound randomization in the binary system" by John E. Walsh; Annals of Mathematical Statistics, Vol. XX, No. 4, December 1949, pp. 580-589.

15. "On the power function of the 'best' t-test solution of the Behrens-Fisher problem" by John E. Walsh; Annals of Mathematical Statistics, Vol. XX, No. 4, December 1949, pp. 616-618.

16. "On a generalization of the Behrens-Fisher problem" by John E. Walsh; Human Biology, Vol. 22, No. 2, May 1950, pp. 125-135.

17. "Large sample tests and confidence intervals for mortality rates" by John E. Walsh; Journal of the American Statistical Association, Vol. 45, June 1950, pp. 225-237.

18. "Some estimates and tests based on the r smallest values in a sample" by John E. Walsh; Annals of Mathematical Statistics, Vol. 21, No. 3, September 1950, pp. 386-397.

19. "On the large sample distribution of mortality rates based on statistically independent lives" by John E. Walsh; Transactions of the Society of Actuaries, Vol. II, Part 3, No. 4, November 1950, pp. 228-234.

20. "Some nonparametric tests of whether the largest observations of a set are too large or too small" by John E. Walsh; Annals of Mathematical Statistics, Vol. 21, No. 4, December 1950, pp. 583-592.

21. "A large sample t-statistic which is insensitive to non-randomness" by John E. Walsh; Journal of the American Statistical Association, Vol. 46, March 1951, pp. 79-88.

22. "Some bounded significance level properties of the equal-tail sign test" by John E. Walsh; Annals of Mathematical Statistics, Vol. 22, No. 3, September 1951, pp. 408-417.

23. "Large sample validity of the binomial distribution for lives with unequal mortality rates" by John E. Walsh; Skandinavisk Aktuarietidskrift, Vol. 35, 1952, pp. 11-15.

24. "Operating characteristics for tests of the stability of a normal population" by John E. Walsh; Journal of the American Statistical Association, Vol. 47, June 1952, pp. 191-202.

25. "Concerning large sample tests and confidence intervals for mortality rates" by John E. Walsh; Journal of the Institute of Actuaries Students' Society, Vol. XI, Part 2, 1952, pp. 104-114.

1026181

26. "Some nonparametric tests for Students' hypothesis in experimental designs" by John E. Walsh; Journal of the American Statistical Association, Vol. 47, September 1952, pp. 401-415.

27. "Some tests for comparing percentage points of two arbitrary continuous populations" by A. W. Marshall and John E. Walsh; Proceedings of the International Congress of Mathematicians, Vol. 1, 1952, pp. 582-583.

28. "Some probability results for mortality rates based on insurance data" by John E. Walsh; Journal of the Institute of Actuaries, Vol. LXXIX, Part II, No. 352, 1953, pp. 213-220.

29. "Large sample confidence intervals for density function values at percentages points" by John E. Walsh; Sankhyā, Vol. 12, Part 3, 1953.

30. "Analytic tests and confidence intervals for the mean value, probabilities and percentage points of a Poisson distribution" by John E. Walsh; Sankhyā, Vol. 14, Parts 1 and 2, September 1954, pp. 25-38.

31. "Approximate salvo kill probabilities for small and medium sized targets when cumulative damage is unimportant" by John E. Walsh; Journal of the Operations Research Society of America, Vol. 3, No. 1, February 1955, pp. 69-76.

32. "The Poisson distribution as a limit for dependent binomial events with unequal probabilities" by John E. Walsh; Journal of the Operations Research Society of America, Vol. 3, No. 2, May 1955, pp. 198-209.

33. "Approximate probability values for observed number of 'successes' from statistically independent binomial events with unequal probabilities" by John E. Walsh; Sankhyā, Vol. 15, Part 3, July 1955, pp. 281-290.

34. "Bounded significance level tests for comparing quantiles of two possibly different continuous populations" by John E. Walsh; Annals of the Institute of Statistical Mathematics, Vol. VI, No. 3, September 1955, pp. 213-222.

35. "Optimum ammunition properties for salvos" by John E. Walsh; Operations Research, Vol. 4, No. 2, April 1956, pp. 204-212.

36. "Questionable usefulness of variances for measuring estimate accuracy in Monte Carlo importance sampling problems" by John E. Walsh; Symposium on Monte Carlo Methods, John Wiley and Sons, 1956, pp. 141-144.

37. A Monte Carlo technique for obtaining tests and confidence intervals for insurance mortality rates" by John E. Walsh; Symposium on Monte Carlo Methods, John Wiley and Sons, 1956, pp. 467-480.

1026182

38. "Asymptotic efficiencies of a nonparametric life test for smaller percentiles of a gamma distribution" by John E. Walsh; Journal of the American Statistical Association, Vol. 51, September 1956, pp. 467-480.

39. "Acturarial validity of a binomial distribution for large numbers of lives with small mortality probabilities" by John E. Walsh; Skandinavisk Aktuarietidskrift, Vol. 39, 1956, pp. 39-46.

40. "Estimating population mean, variance and percentage points from truncated data" by John E. Walsh; Skandinavisk Aktuarietidskrift, Vol. 39, 1956, pp. 47-58.

41. "Validity of approximate normality values for $\mu \pm k\sigma$ areas of practical type continuous populations" by John E. Walsh; Annals of the Institute of Statistical Mathematics, Vol. VIII, No. 2, 1956, pp. 79-86.

42. "An experimental method for obtaining random digits and permutations" by John E. Walsh; Sankhyā, Vol. 17, Part 4, February 1957, pp. 355-360.

43. "Nonparametric mean estimation of percentage points and density function values" by John E. Walsh; Annals of the Institute of Statistical Mathematics, Vol. VIII, No. 3, 1957, pp. 167-180.

44. "Possible motivation for a stated concept of defense" by John E. Walsh; Operations Research, Vol. 5, No. 3, June 1957, pp. 442-445.

45. "Estimating future from past in life testing" by John E. Walsh; Annals of Mathematical Statistics, Vol. 28, No. 2, June 1957, pp. 432-441.

46. "Inadequacy of cost per 'kill' as measure of effectiveness" by John E. Walsh; Operations Research, Vol. 5, No. 6, December 1957, pp. 750-764.

47. "Further consideration of normality values for $\mu \pm k\sigma$ areas of continuous populations" by John E. Walsh; Annals of the Institute of Statistical Mathematics, Vol. IX, No. 2, 1958, pp. 127-129.

48. "Nonparametric estimation of sample percentage point standard deviation" by John E. Walsh; Annals of Mathematical Statistics, Vol. 29, No. 2, June 1958, pp. 601-604.

49. "Efficient small sample nonparametric median tests with bounded significance levels" by John E. Walsh; Annals of the Institute of Statistical Mathematics, Vol. IX, No. 3, 1958, pp. 185-199.

50. "Some practical simulations of operations" by John B. Lathrop and John E. Walsh; Journal of Industrial Engineering, Vol. IX, No. 5, September-October 1958, pp. 392-396.

1026183

51. "Production schedules involving cancellation and obsolescence risks" by John E. Walsh; Operations Research, Vol. 7, No. 1, January-February 1959, pp. 123-126.

52. "A difficulty in linear programming for transportation problems" by Marvin Luther and John E. Walsh; Naval Research Logistics Quarterly, Vol. 6, March 1959, pp. 77-80

53. "Comments on 'the simplest signed-rank tests'" by John E. Walsh; Journal of the American Statistical Association, Vol. 54, March 1959, pp. 213-224.

54. "Definition and use of generalized percentage points" by John E. Walsh; Sankhyā, Vol. 21, Parts 3 and 4, August 1959, pp. 281-288.

55. "Optimum properties for defense strategy of equal attack against all targets" by John E. Walsh; Operations Research, Vol. 7, No. 2, March-April 1959, pp. 249-255.

56. "Acceptance inspection by variables when the measurements are subject to error" by H. T. David, E. A. Fay and John E. Walsh; Annals of the Institute of Statistical Mathematics, Vol. X, No. 2, 1959, pp. 107-129.

57. "Large sample nonparametric rejection of outlying observations" by John E. Walsh; Annals of the Institute of Statistical Mathematics, Vol. X, No. 3, 1959, pp. 223-232.

58. "Nonparametric mean and variance estimation from truncated data" by John E. Walsh; Skandinavisk Aktuarietidskrift, Vol. 41, 1958 (published in 1959), pp. 125-130.

59. "Exact nonparametric tests for randomized blocks" by John E. Walsh; Annals of Mathematical Statistics, Vol. 30, No. 4, December 1959, pp. 1034-1040.

60. "Method for first-stage evaluation of complex man-machine systems" by I. M. Garfunkel and John E. Walsh; Naval Research Logistics Quarterly, Vol. 7, No. 1, 1960, pp. 13-19.

61. "Nonparametric properties of some maximum likelihood estimates for median of symmetrical population" by John E. Walsh; Annals of the Institute of Statistical Mathematics, Vol. XI, No. 2, 1960, pp. 81-88.

62. "Efficiency criterion for comparing cargo aircraft in deliveries between two terminals" by Roy P. Brady and John E. Walsh; Operations Research, Vol. 8, No. 3, May-June 1960, pp. 356-361.

1026184

63. "Nonparametric linear estimation of common median of symmetrical populations from symmetrically censored observations" by John E. Walsh; Sankhyā, Vol. 22, Parts 3 and 4, June 1960, pp. 295-300.

64. "Nonparametric tests for median by interpolation from sign tests" by John E. Walsh; Annals of the Institute of Statistical Mathematics, Vol. XI, No. 3, 1960, pp. 183-188.

65. "Empirical examination of Edgeworth series" by J. R. Crawford and John E. Walsh; Annals of the Institute of Statistical Mathematics, Vol. XII, No. 1, 1960, pp. 13-26.

66. "Analyzing medical data - some statistical considerations" by John E. Walsh; IRE Transactions on Medical Electronics, Vol. ME-7, October 1960, pp. 362-366.

67. "Probabilities for Cramér-von Mises-Smirnov test using grouped data" by John E. Walsh; Annals of the Institute of Statistical Mathematics, Vol. XII, No. 2, 1960, pp. 143-145.

68. "General simulation model for logistics operation in a randomly-damaged system" by John E. Walsh; Naval Research Logistics Quarterly, Vol. 7, No. 4, December 1960, pp. 453-470.

69. "Computer-feasible method for handling incomplete data in regression analysis" by John E. Walsh; Journal of the Association for Computing Machinery, Vol. 8, No. 2, April 1961, pp. 201-211.

70. "Statistical prediction from tolerance regions" by John E. Walsh; Bulletin of the International Statistical Institute, Vol. XXXVIII, Part III, 1961, pp. 313-317.

71. "Remarks about methodology in operations research" by John E. Walsh; Journal of the Operations Research Society of Japan, Vol. 4, No. 2, January 1962, pp. 87-90.

72. "Distribution-free tolerance intervals for continuous symmetrical populations" by John E. Walsh; Annals of Mathematical Statistics, Vol. 33, No. 3, September 1962, pp. 1167-1174.

73. "Some two-sided distribution-free tolerance intervals of a general nature" by John E. Walsh; Journal of the American Statistical Association, Vol. 57, No. 4, December 1962, pp. 775-784.

74. "Loss in test efficiency due to misclassification in 2x2 tables" by John E. Walsh; Biometrics, Vol. 19, No. 1, March 1963, pp. 158-162.

1026185

75. "A general probability model for binomial events, with application to surgical mortality" by Herbert B. Eisenberg, Randolph R. M. Geoghagen and John E. Walsh; Biometrics, Vol. 19, No. 1, March 1963, pp. 152-157.

76. "Use of linearized nonlinear regression for simulations involving Monte Carlo" by John E. Walsh; Operations Research, Vol. 11, No. 2, March-April 1963, pp. 228-235.

77. "A method of emphasizing specified alternatives in sequential and non-sequential tests for symmetry about zero" by R. R. M. Geoghagen and John E. Walsh; Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, Series A, Vol. 66, No. 2, 1963, pp. 189-191.

78. "A study of surgical mortality probability for several pre-operative health conditions" by H. B. Eisenberg, R. R. M. Geoghagen and John E. Walsh; Biometrische Zeitschrift, Vol. 5, No. 3, 1963, pp. 173-187.

79. "A model for maritime transportation systems" by Paul B. Billing and John E. Walsh; Journal of Operations Research Society of Japan, Vol. 6, No. 1, 1963, pp. 1-16.

80. "Approximate determination of cost for optimum air cargo operations" by John E. Walsh; Proceedings of the 3rd International Conference on Operations Research, Oslo, Norway, in press.

81. "Approximate distribution of extremes for nonsample cases" by John E. Walsh; Journal of the American Statistical Association, in press.

82. "Simultaneous confidence intervals for differences of classification probabilities" by John E. Walsh; Biometrische Zeitschrift, in press.

83. "Bounded probability properties of Kolmogorov-Smirnov and similar statistics for discrete data" by John E. Walsh; Annals of the Institute of Statistical Mathematics, in press.

84. "Simulation of military logistics operation in an underdeveloped country" by James L. Taylor and John E. Walsh; Proceedings of Symposium on Uses of Operations Research in Developing Countries, Paris, France, in press.

1026166

Society Memberships

Operations Research Society of America (currently National Treasurer)
American Statistical Association
Institute of Mathematical Statistics
Society of Actuaries
Biometric Society
Operations Research Society of Japan
The Institute of Management Sciences
International Association for Statistics in Physical Sciences
Airline Group of International Federation of Operational Research Societies
Sigma Xi

Additional Information

Statistical consultant for Los Angeles County Tuberculosis and Health Association (1959-1963). Has been member of several national Committees for Institute of Mathematical Statistics. Referee and book reviewer for Journal of the American Statistical Association. Referee for Annals of Mathematical Statistics, Operations Research, and Biometrics. Reviewer for Applied Statistics and Quality Control, International Abstracts in Operations Research, Mathematics of Computation, Mathematical Reviews, and Zentralblatt für Mathematik. Proposal reviewer for National Science Foundation and National Academy of Science. Over 70 papers presented at meetings of technical societies. Member of Council for International Federation of Operational Research Societies (1963).

1026187

COPY

OAK RIDGE NATIONAL LABORATORY
OPERATED BY
UNION CARBIDE CORPORATION
NUCLEAR DIVISION

POST OFFICE BOX X
OAK RIDGE, TENNESSEE 37831

MAR 26 1964

March 25, 1964

Dr. A. W. Burr, Jr.
Chief, Medical Research Branch
Division of Biology and Medicine
U. S. Atomic Energy Commission
Washington, D. C. 20545

Dear Dr. Burr:

Thank you for your letter of March 5, 1964, and for the opportunity to review Dr. John E. Walsh's research proposal. Although I do not know Dr. Walsh personally, I have been told he is a very competent statistician, and judging from his list of publications and his proposal, I would be inclined to believe this. I am pleased that a competent statistician has taken an interest in this important, but difficult job of analyzing the human radiation exposure data on the Marshallese people and also has laid the ground work for analysis of the Japanese data.

Before I discuss this proposal, I should point out that I cannot be considered an expert on regression methods or statistical techniques. However, I have some experience in curve fitting and have employed statistical techniques in this work. I have only one comment to make on the regression method.

I feel almost certain that you have sent this proposal to competent statisticians for review. In the event you have not, I would suggest it be sent to one, or preferably all of the following statisticians: Dr. Paul Meier, University of Chicago; Dr. Allyn Kimball, Johns Hopkins University; Dr. Marvin Kastenbaum, ORNL.

My only comment on the linearized nonlinear regression method is that it apparently is a new and untried procedure. It has been called to my attention by someone more familiar with the statistical literature than I am that the proposed method of

1026188

analysis has not appeared in the well-known journals publishing statistical articles. Because of the fact that the problem of determining the effect of radiation exposure on man is very important and because a good deal of effort and money have been spent on gathering these valuable data, I would hope that Dr. Walsh's linearized nonlinear regression method would be tried out on some test problems before the human data were analyzed. (By test problems, I mean that a function having specified values for the coefficients could be used to generate a set of true values to which could be appended a set of random errors of different level. These data would constitute a synthetic set which would then be analyzed by the proposed linearized nonlinear regression method.) This would yield some estimates of the validity and limitation of the method and also provide some experience in its application.

Very truly yours,

S. R. Bernard
Health Physics Division

Enclosure:
Proposal

cc: M. L. Kastenbaum
K. Z. Morgan
W. S. Snyder

SRB:ss

1026189