SOME STATISTICAL USES OF LARGE COMPUTERS

W. J. DIXON
UNIVERSITY OF CALIFORNIA, LOS ANGELES

1. Introduction

The analysis of many medical research problems requires a great amount of computation including types for which high-speed computing machinery is necessary. Many medical research problems remain in the discussion and planning stages because of the complex nature of the analysis. Progress on some of these problems can proceed only with the use of a large electronic computer. The following are examples of these types of problems (not necessarily representative examples).

1.1. *Scientific computations.*

(1) Construction and testing of simulation models. These models as used in genetics, epidemiology, and psychology and models constructed for analysis of physiological and biochemical systems, in order to be realistic, are necessarily complex. Although simple mathematical expressions are sometimes sufficient to describe a system, general solutions require numerical treatment even when these solutions may be obtained in closed mathematical form.

These models often employ systems of differential or difference equations, which may be linear or nonlinear. These equations are conditioned with various types of restrictions or boundary values and always involve many variables. Also, the boundary values may be determined statistically instead of being given constants. Solutions are not possible without large computers.

These simulated systems are evolved from known relations concerning, for example, homeostasis and feedback mechanisms, diffusion and reaction kinetics, rates of growth and decay.

(2) Analysis of experimental data.

(a) Statistical analysis of data from clinical and laboratory research requiring complex and/or extensive computation, for example, multivariate analysis or data screening.

(b) Spectral and correlational analysis for continuously recorded observations, for example, analyses for various EEG leads related to behavioral and physiological phenomena and to each other.

1.2. *Data storage and retrieval.*

(1) Access to patient records. For direct immediate access of records for clinical use and to provide increased availability of records for research use.

(2) Analysis of patient records. Analysis of large amounts of data for clinical or laboratory research.

(3) Assistance in diagnosis by automatic coordination of laboratory and clinical findings.

(4) Maintenance of active registry (with descriptions) of chemical compounds, for example, for study of relation between structure, chemical properties, and pharmacological activity.

In spite of many articles that describe these applications, no information was found concerning a computer installation solely devoted to the development of medical applications.

2. The role of a medical computing laboratory

The activities comprise the educational, research, and service objectives as follows: (1) development of methods of analysis appropriate to research in basic medical science and to clinical research; (2) research into methods of numerical analysis basic to programming medical data; (3) development of methods of scaling and coding medical data; (4) educational and training programs; (5) consultation on computing problems; (6) consultation on statistical design and analysis; (7) preparation and maintenance of a library of programs; (8) data handling; (9) computer operation.

As a first step in tackling this large problem of assistance to medical research, it seems desirable to make available to the medical research worker those techniques of statistics which have been of value in other areas. Availability implies programs which: (1) can be used by someone not specially trained in programming or in mathematics; (2) are of great flexibility of input and output; (3) recognize the special problems of medical data such as (a) incompleteness of data; (b) special distribution forms; (c) large number of variables per individual case.

The series of computer programs thus far written have been developed with the following criteria or goals in mind.

2.1. *Need.* It should be possible to choose a particular kind of analysis which seems appropriate without concern for the complexity of the numerical operations. This implies also that one can comparatively easily perform a variety of analyses if one wishes on the same general problem and can redo the computation easily in a modified form, as by repeating the computation using transformation of the original observations. For example, a multivariate computation could easily be repeated with, say, the logarithmic transformation made on several of the variates, or a computation may be performed on various different subsets of the original sets of variates.

2.2. *Flexibility and generality.* Programs should be flexible in that the user

may specify various modifications on a general type of computation so that the analysis will suit the particular problem at hand. For example, the research worker will wish to have the same program operate for problems with varying numbers of variables for varying numbers of observations, for varying numbers of digits, and for varying magnitudes of each observation.

Programs should be general so that problems of widely varying size and structure may be handled without writing many special programs.

2.3. *Form.* Programs should be constructed so that they may be used by research workers with limited knowledge of computers or computer language. However, they should be written in a standard computer language so they may be modified easily by a computer expert. Ideally, programs should be available in subroutine form so an analysis can be constructed for each problem by writing a few commands and assembling the appropriate subroutines. At least a partial solution is available here by writing subroutines in a standard language like Fortran and also preparing several standard packages of these subroutines with a users' manual or description.

2.4. *Adaptability.* So far as possible, programs should be written to facilitate adaptation to a wide variety of models and makes of computers. As an example here, it may be possible to avoid commands which will compile only on larger computers. Also, it may be possible with a little care to write programs within the capacity of a computer with an 8000 or 16000 word high-speed memory rather than requiring one with a 32000 word memory.

2.5. *Experimentation.* Programs should be written for performing some of the computations now a part of the "art" of statistics, programs which perform some of the work a statistician may do in preliminary analyses on data: plotting data in various forms, investigating linearity of response, scanning columns of data for "peculiar" values, checking on associated observations when observations are missing or obviously erroneous, and making various other checks on the data which the knowledge of a particular field of application may give the research worker. Two "scanning" programs to be discussed later are examples of preliminary analyses.

2.6. *Internal decisions.* Programs should be developed which will make internal decisions based on the outcomes of earlier portions of the analysis. An example of a program of this sort is that written by Efroymson for stepwise addition and deletion of variables in a linear regression problem.

2.7. *Multiple problems.* Programs should provide for performing a series of problems in the same machine run.

3. The work of a computing laboratory

When the Western Data Processing Center was established on the University of California campus at Los Angeles, it became possible to use a large computer on these problems, and beginning last summer a number of seminars were organized under the direction of Professors F. J. Massey, Jean Dunn, and myself.

As a result of these discussions, a series of problems were programmed by John Tauchi and Kenneth Ferrin. This work was performed under the sponsorship of contract SA-43-PH-3039 of the National Chemotherapy Service Center of the National Cancer Institute and other research grants.

The present list of programs cover problems in the following areas (see table I):

TABLE I

AN OUTLINE OF BIMED PROGRAMS

* indicates that transformations available are $x^{1/2}$, $x^{1/2} + (x + 1)^{1/2}$, $\log_{10}(x + c)$, arcsin $x^{1/2}$, arcsin $[x/(n + 1)]^{1/2}$ + arcsin $[(x + 1)/(n + 1)]^{1/2}$.

BIMED	Groups	Independent Variables	Dependent Variables	Sample Size	Transformations
1. Life tables, 50 time intervals					
2. Component analysis	1	25		150	
3. Regression on components	1	25	20	150	
4. Discriminant analysis, several groups	5	25		150 each gp.	
5. Discriminant analysis, two groups	2	25		150 each gp.	$\log_{10}$
6. Multiple regression No. 1	1	29	30 (same set)	5,000	*
7. Multiple regression No. 2	28	29	30 (same set)	5,000 each gp. 32,000 total	*
8. Polynomial regression	1	10th degree	1	500	
9. Stepwise regression	1	59	1	100,000	*
10. Regression subroutine	1	30	1		
11. Analysis of variance No. 1 (factorial)	1	8 up to 999 cat.	1	20,000	*
12. Analysis of variance No. 2 (factorial)	1	14 up to 999 cat.	1	20,000	*
13. Analysis of covariance	1	6 anal. of var. up to 999 cat. 8 covariates	1	2,000	*
14. General linear hypothesis	1	60		10,000	*
15. Data screening No. 1	1	50		2,000	*
16. Data screening No. 2	defined by variables	30		650	*

regression, BIMED 6, 7, 8, 9, 10; analysis of variance and covariance, BIMED 11, 12, 13; general linear hypothesis, BIMED 14; discriminant analysis, BIMED 4, 5; component analysis, BIMED 2; regression on components, BIMED 3; life table analysis, BIMED 1.

Data handling covers: cross tabulation, BIMED 15; tests of normality, BIMED 15; search for outliers, BIMED 15; description of strata, BIMED 16.

4. General features of the written programs

4.1. *Input.* The Fortran *programs* can easily be modified to a wide variety of punched card or type input forms. The assembled packages are in most cases written for a maximum of six digits in each observation. Therefore each punched card may carry 12 variates in 72 card columns. Additional variates continue on successive cards. However, each new case starts on a new card. *Scale cards* are prepared to accompany the program specifying the power of ten for locating the decimal point for each variable.

Transformations are called by code number for each variate: 01 $x^{1/2}$; 02 $x^{1/2} + (x + 1)^{1/2}$; 03 $\log_{10} x$; 04 $\log_{10} (x + c)$ with c specified; 05 arcsin $x^{1/2}$; 06 arcsin $[x/(n + 1)]^{1/2}$ + arcsin $[(x + 1)/(n + 1)]^{1/2}$.

The specifications of a particular problem are indicated in a *control or problem card.* For example, in the use of the regression program a problem card will include problem number, total number of variables, number of observations, call for transformations, if any.

A *subproblem card* will include problem number, subproblem number, specification of readout information desired, which may include call for a list of residuals, call for range of residuals, call for product of matrix and its inverse, variate number chosen as dependent variable, number of variables to be excluded from independent variables, list of variables to be excluded.

The computing center also requires an *identification card.* Cards are then assembled for the computer as follows: (1) identification card; (2) program cards (198 binary cards); (3) problem card for problem 1; (4) scale cards; (5) transformation cards, if used; (6) data cards; (7) subproblem cards (any number)—repeat of 3, 4, 5, 6, 7 for problem 2, and so on (any number); (8) end card (code provided by computer center).

4.2. *Readout.* The readout includes the following: sum, sum of squares, mean, standard deviations for each variate; cross product sums, cross product of deviations sums and correlation coefficients for each pair of variates; inverse of correlation matrix; regression coefficients; analysis of variance table of SS attributable to regression and deviations from regression; multiple correlation coefficient; standard error of estimate; standard deviations of regression coefficients; ratio of regression coefficients to standard error; partial correlation coefficients; variance of components of each additional regression coefficient in order computed; residuals; ratio of range of residuals to standard error of estimate.

The other programs have the multiple problem capacity and specifications are supplied in a similar manner.

The capacity of the regression program designated BIMED 06 is 30 variables and 5000 observations for each problem. BIMED 07 contains all the features of BIMED 06 but also includes the feature of specifying the observations as belonging to various subgroups (maximum 28 subgroups). Problems and subproblems as specified in BIMED 06 may be performed on any collection of these sub-

samples without limit. The largest total sample size in any selection of subgroups is 5000; however, the total sample size of all subsamples may be 32000.

The polynomial regression program BIMED 08 computes polynomial regressions for each degree successively up to the degree (max. 10) specified by a problem card. An analysis of variance table is provided at each step for determining the goodness-of-fit.

BIMED 09 is an adaptation of the stepwise regression program written by M. A. Efroymson [1] of Esso Research and Engineering Company. Data are scaled and transformed as described for BIMED 06 and a succession of intermediate regression equations are obtained adding one variable at a time. At each stage the variable which makes the greatest improvement in goodness-of-fit is added to the regression. After each addition to the regression function the previously added variates are examined for deletion. Both addition and deletion of variates are determined by comparison with F ratios which are specified in the problem card.

This program will handle 59 independent variables and will operate satisfactorily even though variates are included which are constant or which are linearly dependent on other variates. Readout available on call includes descriptive statistics such as sums, means, variances, correlations, residuals, and so on. Also in this program the constant term may be specified to be zero.

The first analysis of variance program, BIMED 11, is written for a complete factorial design with as many as eight factors with the number of categories or levels of each, L_i limited by $\prod L_i \leqq 20{,}000$ and max $L_i \leqq 1000$. The number of replications is essentially unlimited.

This program will provide a breakdown of sums of squares into orthogonal components (linear, quadratic, cubic, and so on) for four variates and provides, for these variables, tables of interactions, that is, tables of residuals. The number of levels for these variables is limited to 9.

The program description contains a discussion of the modification of the analysis of variance tables which are appropriate in nested designs.

BIMED 12 is an analysis of variance program permitting 14 variables with no limitation on the number of replicates and with the same limitations on the L_i given above. The number of interaction sum of squares to be computed can be designated in the control card.

Both analysis of variance programs permit transformation on call and a series of problems may be analyzed at one time.

The analysis of covariance program BIMED 13 will analyze 2 to 6 analysis of variance variables with as many as eight covariates. The number of replicates is limited to 1000 and the number of levels L_i of each variable must satisfy $\prod L_i \leqq 2000$. Transformation may be made on any or all covariates as well as on the variate. The notation used in this description follows that given by Scheffé [2].

In addition to the analysis of variance table for the model

$$E(y_\alpha) = \eta_\alpha + \gamma_1 x_{1\alpha} + \gamma_2 x_{2\alpha} + \cdots + \gamma_p x_{p\alpha}, \tag{1}$$

where η_α is the full factorial design vector, for example,

$$\eta_\alpha = \mu + \alpha_i + \beta_i + \delta_{ij}, \tag{2}$$

the program computes a full factorial analysis of variance table, estimates $\hat{\gamma}_i$, and the variance and covariances of these estimates. Also computed are ratios of regression coefficients to their standard errors and an F ratio for the hypothesis

$$H: \gamma_i = \mathring{\gamma}_i, \tag{3}$$

where $\mathring{\gamma}_i$ are specified in the problem card.

Since the appropriate analysis for missing values may be obtained by introducing a dummy covariate (1 corresponding to missing value and zero otherwise) this program may be used for factorial experiments with a small number of missing values.

The component analysis program BIMED 02 computes for a maximum of 25 variates on a maximum of 150 observations, the correlation coefficients, the eigenvalues including the cumulative proportion of total variance for each component, the eigenvectors, and the rank order of each standardized case ordered by size of each principal component separately. The definitions here follow Kendall [3].

BIMED 03 includes the computation given in BIMED 02 and further computes coefficients of regression for as many as 20 different dependent variates on the orthogonal components treated as independent variables. Also computed are the reductions in sum of squares of residuals due to these orthogonal components and coefficients of the regression equation when the first one, first two, and first three components are used as independent variables (each component is expressed in terms of the standardized data).

These two programs will be extended in capacity by use of an iterative matrix inversion procedure and a factor analysis rotation program will be based on the extended BIMED 02.

The two discriminant analysis programs are BIMED 04 and 05. One program computes for a maximum of five groups 25 variables and 150 observations, the mean scores, the covariance matrix and its inverse, the coefficients of the discriminant or classification functions, and the evaluation of these classification functions for each case. Each individual is classified into groups according to the largest computed value for the classification functions. The matrix of these classifications is tabulated.

The other program computes the discriminant function for separating two groups and has additional features of various selections of variates to be used in the discrimination and a readout of the discriminant function evaluated for each

case, showing the amount and type of overlap remaining for the two groups using this discriminant function.

A stepwise discriminant function for two groups can, of course, be computed by using BIMED 09 with the dependent variable 0, 1 for the two groups.

The general linear hypothesis program BIMED 14 could be classified either as a regression or analysis of variance program. This program was written for assistance in the solution of incomplete factorial designs, factorial designs with unequal numbers of observations per cell, and factorial experiments with missing values including the use of covariates in these cases.

The model may include $p \leqq 60$ analysis of variance variables and $q \leqq 60$ covariates with $p + q \leqq 60$. Nine hypotheses may be stated for test. A proper set of stated hypotheses will also produce sums of squares for other F-tests than the F-tests automatically computed. As in the other cases, transformation of the data may be specified.

Design cards are prepared to specify the parameters to be estimated. For example, to specify a factor A with 3 categories whose effects may be parameterized by a_1, a_2, a_3 with $\sum a_i = 0$, we would designate for three groups the design

	a_1	a_2
Group 1	1	0
2	0	1
3	−1	−1

to indicate a_1 for the first group, a_2 for the second group, and $a_3 = -a_1, -a_2$ for the third group. Other crossed or nested classifications and/or interaction are specified in a similar manner.

Hypothesis cards specify which parameters of the model are to be estimated and which parameters are to be set equal to zero in the process of estimation by least squares. The minimum sum of squares is computed and the least squares estimates of the included parameters are listed. These minimum sums of squares may be used in F-tests. The program computes the standard tests of this type.

BIMED 15 is called data screening No. 1. For $n \leqq 2000$ observations on p variates $p \leqq 50$ and $np \leqq 20{,}000$, this program, for each variate, transforms, standardizes $z = (x - \bar{x})/s$ and cross-tabulates this variate with any other specified variates (specified in problem card). The tabulation of each variate is made into ten classes according to the deciles of a standardized normally distributed variable, that is, $z < -1.282$; $-1.282 < z < -.842$, and so on. Each of these ten categories can be expected to contain $N/10$ observations. Significant departures from this expectation indicate nonnormality. A χ^2 is computed to test departure. From the observed character of the departure we may be able to conclude the form of the nonnormality, for example, skewness, high tails, bimodal form, extreme observations.

From the cross-tabulation one can note independence or dependence (a χ^2 is computed for the 100 cells of the 10×10 table under expectation $N/100$), the form of dependence, if any, bimodality, outliers, and so on.

The output includes means and standard deviations and a list of all items whose standardized value exceeds in absolute value a quantity stated in problem card. Also on call, one may obtain a listing of the transformed and untransformed data. Here, again, problems may be in series and one may compare tabulation of transformed data with tabulation of untransformed data.

Data screening No. 2 (BIMED 16) provides a description of strata of multivariate data. For $n \leqq 650$ observations on $p \leqq 30$ variates this program transforms the data as desired and stratifies the observations on any variate according to levels ($\leqq 20$) specified in the problem card. Within each stratum the means, variances, standard deviations, standard error of means, and correlation coefficients of all variates are computed.

There are additional programs in preparation: orthogonal factor rotation; covariance with a single variable of classification; cross-tabulation; chi square analysis, based on the methods given by Cochran [4]; screening 3; time series analysis of variance.

5. Problems of the biomedical sciences

To give an idea of the kinds of problems which arise in the biomedical sciences to which these programs and other special ones have been applied, four examples will be described.

EXAMPLE 5.1. A clinic, studying various types of heart disease, records for several types of patients various lipid determinations from blood drawings and various excretion hormones. This involves two illustrations of data screening programs. The lipids are cholesterol, phospholipids, and α- and β-lipoproteins. The hormones are estrogen, 17 keto-steroids, and Porter-Silber corticoids. There were 163 cases with blood determinations only and 174 cases with urine determination only. In 107 of these cases both blood and urine variables were observed. BIMED 16 was used to describe these data for all three groupings of the data; within each grouping, age was used as a stratifying variable. Therefore, 18 tabulations (six age decades for each of three groupings) were prepared by this program of the type shown in table II.

One of the preliminary steps in the preparation of these data involved the use of BIMED 15. An example of the output is shown in table III.

Various other programs including the discriminant analysis BIMED 5 were used for comparing patients with myocardial infarct with normal patients, using various selections of variables. A briefer example of the output of this program is given next.

EXAMPLE 5.2. A method for predicting those patients with hypertension due to unilateral renal or renal arterial disease who will be cured by appropriate surgery (C. C. Winter et al.).

Various authors in the past have suggested many different criteria. In this study these criteria were all evaluated on a group of patients. A discriminant analysis will aid in evaluating individual criteria and combination of these

TABLE II

DESCRIPTION OF DATA FOR INTERVAL 3
Age equal to or greater than 50.00 but less than 60.00

Conditioning Variable	Frequency	Mean	Variance	Std. Dev.	Std. Error
1 Age	29	54.62069	8.52958	2.92054	0.54233
Conditioned Variables					
2 Cholesterol	29	2.40677	0.00644	0.08027	0.01491
3 Phospholipids	29	2.36860	0.00524	0.07237	0.01344
4 α-lipoprotein	29	1.31004	0.02232	0.14941	0.02774
5 β-lipoprotein	29	1.89435	0.00162	0.04027	0.00748
6 Urinary estrogen	29	−0.38612	0.06643	0.25774	0.04786
7 17 keto-steroids	29	0.76818	0.06134	0.24766	0.04599
8 Porter-Silber corticoids	29	0.71101	0.05379	0.23193	0.04307

Correlation Coefficients

Row 1							
1.00000	−0.20005	−0.04447	−0.03481	0.03345	−0.03295	−0.14824	0.02759
Row 2							
−0.20005	1.00000	0.49785	−0.23602	0.20094	0.10192	0.11940	0.26772
Row 3							
−0.04447	0.49785	1.00000	−0.15323	0.07695	0.01558	0.05854	0.15973
Row 4							
−0.03481	−0.23602	−0.15323	1.00000	−0.95594	0.14066	−0.12405	−0.16136
Row 5							
0.03345	0.20094	0.07695	−0.95594	1.00000	−0.04004	0.14076	0.16815
Row 6							
−0.03295	0.10192	0.01558	0.14066	−0.04004	1.00000	−0.00895	−0.04737
Row 7							
−0.14824	0.11940	0.05854	−0.12405	0.14076	−0.00895	1.00000	0.28413
Row 8							
0.02759	0.26772	0.15973	−0.16136	0.16815	−0.04737	0.28413	1.00000

criteria which are, of course, not independent. The criteria used are (1) age-sex; (2) duration of hypertension; (3) renal function tests; (4) ratio of individual renal urine volume; (5) serum creatinine; (6) radioisotope renogram; (7) aortography; (8) excretory urography.

Fourteen cured and twelve uncured patients were analyzed. A stepwise regression BIMED 9 was performed first on the (0, 1) variate for cured and uncured with F-level to enter 2.00. Variables 1, 8, and 4 were selected. As an illustration of the discriminant analysis program BIMED 5 these three selected variates were used. In addition to showing the statistic D^2 we can see the amount of separation obtained with these three tests.

The machine output is given in table IV, including the evaluation of the discriminant function for each case. This is printed out in two columns so that one can easily observe the amount of overlap remaining between the two groups.

EXAMPLE 5.3. This example illustrates use of analysis of variance to investigate sources of laboratory error in the Department of Medicine, which assays quantities of infectious agents in urine.

The experiment was designed covering: (1) two days; (2) three urines (nested within day); (3) each urine diluted by four methods to obtain 1/1000 dilution

TABLE III

Example of Output of BIMED 15

Problem No. 1
Selection 2 — 1 Male MI Patients
Sample Size 163
Variable 2 (Column) Phospholipids
Transformation 3 (log)
Mean 2.41068
Std. Dev. 0.08121
Chi Square 4.30061
Extremes
Item No. Value
72 3.0

Joint Chi Square 205.71

Variable 3 (Row) α-Lipoprotein
Transformation 3 (log)
Mean 2.36764
Std. Dev. 0.08125
Chi Square 3.31902
Extremes
Item No. Value

												N	Diff.
+Inf	*	1	0	0	2	3	1	0	2	0	8	17	0.7
1.282	*	0	0	1	0	1	2	2	5	3	3	17	0.7
0.842	*	0	0	0	0	1	1	5	2	4	2	15	−1.3
0.524	*	1	0	1	2	1	0	1	3	2	0	11	−5.3
0.253	*	0	0	2	2	4	1	3	0	4	1	17	0.7
0.	*	0	1	4	2	1	4	6	1	1	0	20	3.7
−0.253	*	0	1	2	4	2	2	1	3	1	0	16	−0.3
−0.524	*	2	1	3	5	2	1	0	0	0	1	15	−1.3
−0.842	*	3	4	7	0	3	2	0	0	0	0	19	2.7
−1.282	*	8	5	2	1	0	0	0	0	0	0	16	−0.3
−Inf	*												
	*	*	*	*	*	*	*	*	*	*	*		
	−Inf	−1.282	−.842	−.524	−.253	0.0	.253	.524	.842	1.282	+Inf		
N		15	12	22	18	18	14	18	16	15	15	163	
Diff.		−1.3	−4.3	5.7	1.7	1.7	−2.3	1.7	−0.3	−1.3	−1.3		

TABLE IV

MACHINE OUTPUT OF DISCRIMINANT ANALYSIS, PROBLEM NO. 5

Variable Means by Group and Difference in Means

Variable	Mean 1	Mean 2	Difference
1	0.52142856E 01	0.83333333E-01	0.51309523E 01
2	0.36428571E 01	0.28333333E 01	0.80952382E 00
3	0.57142857E	0.50000000E 01	0.71428567E 00

Sum of Products of Dev. from Means

0.20927381E 03	−0.77618992E 01	−0.23142852E 02
−0.77619031E 01	0.18880954E 02	0.35714321E 01
−0.23142853E 02	0.35714340E 01	0.22857147E 02

Inverse of Sum of Products of Dev. from Means

0.54082225E-02	0.12236881E-02	0.52846218E-02
0.12236898E-02	0.54853339E-01	−0.73318554E-02
0.52846215E-02	−0.73318618E-02	0.50246273E-01

Discriminant Function Coefficients

0.32514665E-01	0.45446739E-01	0.57070016E-01

Mahalanobis D Square = 0.58652569E 01
$F(3, 22)$ = 0.11580122E 02

Pop. No.	Sample Size	Mean Z	Variance Z	Std. Dev. Z
1	14	0.66121109E 00	0.13998279E-01	0.11831432E-00
2	12	0.41682538E-00	0.57865334E 00	0.76069266E 00

Rank	First Group Values	Second Group Values	First Group Item No.	Second Group Item No.
1	0.8739		2	
2	0.7843		5	
3	0.7713		12	
4	0.7272		14	
5	0.7192		1	
6	0.7192		7	
7	0.7143		10	
8	0.6456		11	
9	0.6363		6	
10	0.5812		8	
11	0.5812		13	
12		0.5812		8
13	0.5567		4	
14	0.4787		9	
15	0.4671		3	
16		0.4671		2
17		0.4671		4
18		0.4671		6
19		0.4216		1
20		0.4216		7
21		0.3971		10
22		0.3878		12
23		0.3762		9
24		0.3762		11
25		0.3646		5
26		0.2737		3

as follows: $(1/10)^3$, 1/1000, (1/10)(1/100), $(1/\sqrt{1000})^2$; (4) two samples from each; (5) two plates for each sample.

The design is a $2 \times 3 \times 4 \times 2 \times 2$ with nesting as noted above.

The counts from each plate are approximately Poisson so the analysis was performed using the square root transformation. It was also desired to perform separately four $2 \times 3 \times 2 \times 2$ analyses for each dilution. BIMED 11 was used to obtain the analysis. The results will not be given in detail. It may be of interest, however, to note the residual error for the four dilution methods. These variances are .18, .15, .10, .08 respectively for the four methods indicated above.

EXAMPLE 5.4. This example is taken from the Cancer Detection Clinic in Los Angeles. About 10,000 women are screened every year for cancer. Most women are not found to have cancer, but some are classed by Papanicolou smear as showing dysplasia, cancer in situ, or more advanced stages of cancer of the cervix, and some have uterine corpus cancer, breast cancer, and so on.

Information is gathered on about thirty variables in the hope of discovering something about the etiology of cancer and perhaps differential etiology for different types or different sites of cancer. The variables studied are: religion, race, circumcision status of husband, type of contraceptive used, age, age at marriage, total duration of marriages, number of marital events, currently married, age at menarche, age at menopause, years since menopause, number of pregnancies, number of children, estrogen index slide, occupation, weight, and ten variables covering history of cancer in the family.

The cases studied include 212 showing dysplasia, 234 with cancer of the cervix, 48 with uterine corpus cancer, and 138 with breast cancer.

As an example of one small step in the analysis a stepwise discrimination was carried out between corpus and breast cancer cases using BIMED 9. In this example very little reduction in variance was obtained from even the best variates, the variance being reduced from .19 to .17. The four variables selected first were age, age at menopause, weight, and age at menarche. The relation is negative for the age variables, that is, younger in favor of breast cancer. A description of this work will appear in the journal *Cancer*.

REFERENCES

[1] M. A. EFROYMSON, "Multiple regression analysis," *Mathematical Methods for Digital Computers*, New York, Wiley, 1960.

[2] HENRY SCHEFFÉ, *The Analysis of Variance*, New York, Wiley, 1959.

[3] M. G. KENDALL, *A Course in Multivariate Analysis*, New York, Hafner, 1957.

[4] W. G. COCHRAN, "Some methods for strengthening the common χ^2 tests," *Biometrics*, Vol. 10 (1954), pp. 417–451.