The Use of the Computer in Serum Lipid and Lipoprotein Analysis

L. C. JENSEN, A. M. EWING, R. D. WILLS and F. T. LINDGREN,
Donner Laboratory, Lawrence Radiation Laboratory, University of California, Berkeley, California

Abstract

The usefulness of computers in data evaluation is generally recognized; however, the problem of utilizing a computer in the most intelligent manner deserves careful consideration.

Several programs are described which aid in serum lipid and lipoprotein analysis. Two programs requiring a minimum of manual measurements have been developed to analyze gas-liquid chromatograms. These programs perform many operations including corrections for baseline, linearity, Gaussian resolution, and variation in column conditions. The presentation in some detail of one of these programs for NCH elemental analysis illustrates the development and refinement of a program for a specific instrument.

Finally, a general purpose statistical analysis program has been developed which greatly aids in summarizing and correlating data from these programs, as well as other sources, such as ultracentrifugal data.

Introduction

TODAY THE UTILIZATION of computers in research work has become not only commonplace, but frequently a necessity. The increasing complexity of new and improved apparatus used in research often demands sophisticated and tedious calculations to proceed from experimental measurements to the final data for evaluation. These general considerations well apply to lipid and lipoprotein analysis, where considerable strides in technology and experimental techniques have been achieved over the past decade. Present computers are usually more than adequate to handle the specific problems to be calculated and the massive amounts of data to be processed and statistically evaluated.

It is our purpose here to outline some useful applications of computer technology we use in lipid and lipoprotein analysis. We shall consider several specific computer programs currently in use and illustrate in some detail how data of this or related kind ultimately may be processed effectively by a large capacity generalized statistical program. Also, we would like to emphasize by illustration in the development of a computer program, a most important factor in the application of computer techniques to analytical problems; this essential factor is the close working arrangement between the investigator and the computer programmer, frequently over prolonged periods of time.

Typical Programs in Lipid and Lipoprotein Analysis

Analytical Ultracentrifugation

The detailed study of serum-lipoprotein distributions with the analytical ultracentrifuge is a relatively difficult procedure (1). In particular, the manual analysis of the actual schlieren films is both tedious and time-consuming. Also, in order to correct for F [1] versus concentration effects as well as Johnston-Ogston (2) effects, rather extensive calculations are needed. Although Johnston-Ogston corrections have been rigorously treated and generalized for a two-component system (3), such corrections for a continuous distribution of lipoprotein macromolecules would appear not to have an exact mathematical solution. Among the difficulties are how to make precisely the F versus C corrections for a lipoprotein distribution, the dependence of F versus C on radial concentration and the time dependence of the concentration gradient within the analytical cell. This computer method (4) divides the low-density spectrum into 29 S_f rate intervals and the high-density spectrum into 15 F rate intervals. The required manual measurements include tracing a 5-fold enlargement of several schlieren patterns on a precision photo-offset template. The height of the schlieren curve at the midpoint of each interval is measured together with the distance of the peak position to the base-of-cell. These data, together with such additional information as rotor temperature, calibration factors and density (and viscosity) of the lipoprotein fraction, are transcribed onto a special key punch form. The program itself, illustrated schematically in Figure 1, calculates uncorrected lipoprotein concentrations, as well as corrected concentrations involving F versus C, Johnston-Ogston and the correction to standard reference conditions of temperature and density. If desired, all output data from each analysis may be directly punched by computer in proper format for use in the statistical program, described in detail later. Also, these output data may be used in a

[1] F rate is defined as Svedbergs of flotation at any given density. S_f rates are measured at 26C in a medium of 1.745 molal NaCl ($\rho_{26} = 1.0630$ g/ml). Flotation rates corrected for the effects associated with concentration dependence are indicated by the symbol S_f^0.

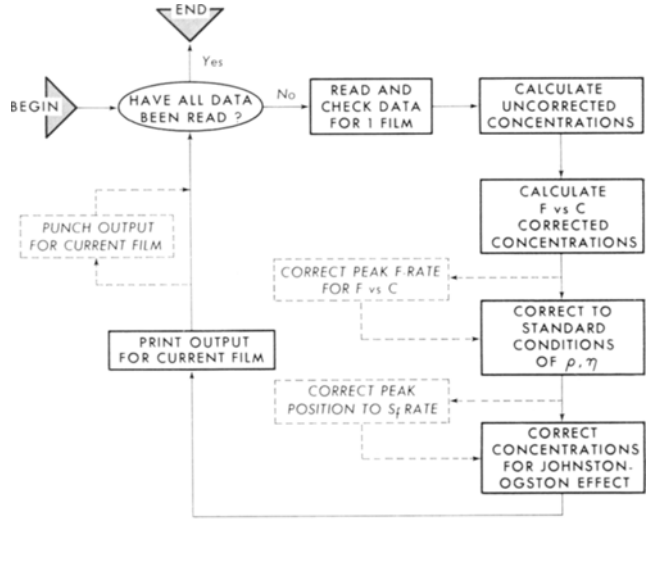


FIG. 1. Flow chart of ultracentrifuge schlieren pattern analysis program. Dashed lines indicate optional procedures.

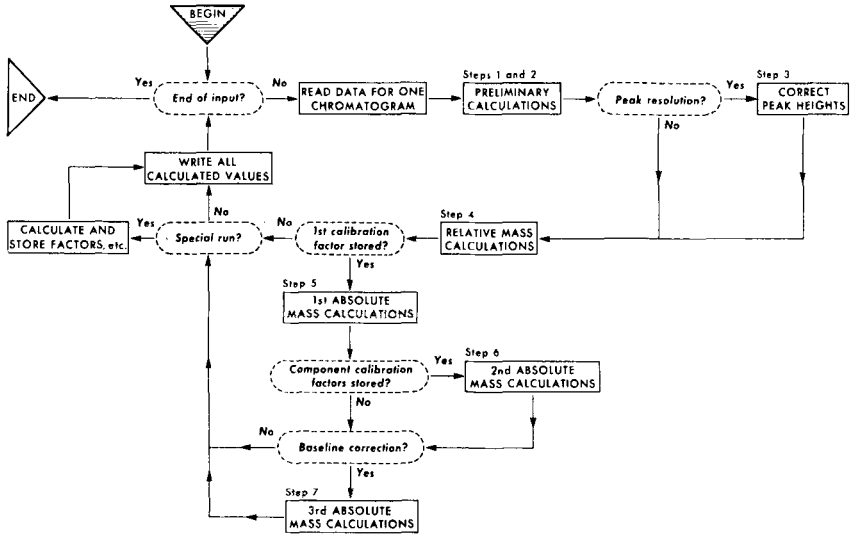


FIG. 2. Schematic diagram of general GLC program.

program which directly plots the corrected schlieren diagram using a standard Cal Comp plotter (California Computer Products, Inc., Anaheim, Cal.).

Another useful program (5) calculates classical moving boundary lipoprotein flotation rates (or sedimentation rates) by a best fit (least squares) straight line for the points ($\ln x_i,\omega^2 t_i$). The program also computes the deviation of each point from the best fit line and then recalculates the slope by successively omitting the most deviant point. Such an analysis frequently permits detection of film reading errors as well as subtle analytical cell leakage.

Other miscellaneous applications include calculations of estimated lipoprotein hydrated density by an ηF versus ρ program, molecular weight determinations, and calculations of $\int \omega^2 dt$ during analytic rotor acceleration to evaluate lipoprotein migration during this time period.

Preparative Ultracentrifugation

Using the dimensions of various swinging bucket rotors, it is possible to calculate lipoprotein subfractionation on a density gradient. This has been applied (6) to very low-density lipoproteins including chylomicrons by subdividing the nonlinear density (and viscosity) gradient into several (up to 17) regions. Ideal flotation was assumed within each region using the mean viscosity and density appropriate for that portion of the density gradient. Other applications include calculation of lipoprotein recovery in various types of preparative rotors.

Electron Microscopy

Using only the measurement of particle diameter as obtained from photographic prints, it is possible to calculate lipoprotein particle-size distributions and flotation rate distributions on the basis of particle number or particle mass. This program, which allows for a flexibility in the histogram intervals, has been applied to the chylomicron class of lipoproteins (6) where the simplifying assumptions of spherical shape and constant particle density were assumed.

Infrared Spectrometry

Using infrared absorbance (optical density) measurements at appropriate wave lengths, a computer program conveniently calculates content of total lipid and such constituent lipids as cholesterol, cholesteryl ester, glycerides, phospholipids and unesterified fatty acids. Fractions analyzed include whole serum lipid extracts as well as chromatographically separated lipid fractions.

Gas Chromatography

With the ever increasing application of gas-liquid chromatography (GLC) as an analytical tool, the calculation, tabulation, and analysis of gas liquid chromatograms can become a burdensome and time-consuming data-processing task. A variety of techniques is available for the calculation of areas under each chromatographic peak, such as planimetry, triangulation (7), the product of peak height and retention time (8–10) and mechanical and electronic integration (11). Also, a completely automatic procedure for computation of gas chromatographic data has been presented (12). This latter procedure, although elegant, requires conversion of detector voltage output to a frequency, storage of this information on a magnetic tape and, finally, presentation of this chromatographic information directly into a computer. Since such elaborate facilities are not generally available, a semiautomated computer analysis of gas chromatographic data has been developed, which required a minimum of manual data transcription from the actual chromatogram. This method, described in detail elsewhere (13), provides for absolute mass calibration, correction for nonlinearity of response of the apparatus, analysis of standard mixtures with correction of preexisting calibration factors, peak height correction for the contribution of interfering peaks in the same neighborhood (14) and subtraction from a sample chromatogram of contaminating components as obtained from an appropriate solvent blank run.

In all the above applications, it has been found advantageous to design individual coding sheets to facilitate data transcription. Further, the use of such coding forms tends to minimize errors both in the data transcription phase as well as in the key punching of the actual data cards. In nearly all our computation work, we have found standard 80 column IBM cards to be the most useful and flexible form initially in which to utilize and manipulate

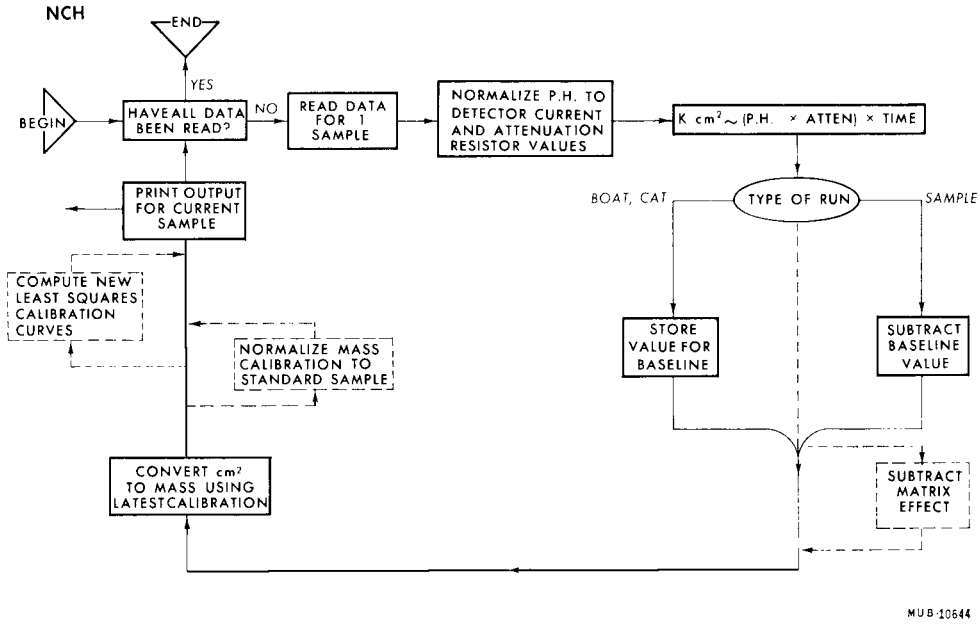


FIG. 3. Flow diagram of N-C-H elemental analysis program.

data. Of course, as card files of old data become excessively large, transcription to tape with considerable reduction in storage space is easily achieved. However, scrupulous labeling and inviolate storage of these data tapes is essential to prevent accidental loss or erasure.

C-H-N Elemental Analysis

In analytical and protein chemistry one of the most important quantitative procedures is for the elemental analysis of carbon, hydrogen and/or nitrogen. Such determinations, for instance, frequently are an essential basis for confirming the identity of a given molecular structure. Previously, these tedious analyses were done by the classical Pregl and Dumas methodology. With the introduction into the laboratory of one of the newly developed semiautomated CHN analyzers (F & M Scientific Model 185) the first problem was to calibrate the instrument and develop a systematic but flexible procedure for calculating the analytical results. Since this type of problem represents an excellent example of the manner in which a computer program is developed for a given instrument, including the close communication necessary between the investigator and the computer programmer, we are presenting some preliminary calibration studies. Also, we feel that because of the obvious interest in the performance characteristics of these new instruments, this program (15) although still under refinement, may help stimulate and hasten progress in the effective use of these elemental analyzers.

The final analytical step in this instrument is separation and quantitation by GLC of N_2, CO_2 and H_2O. Thus, the analysis of the resultant chromatograms could be achieved by several standard techniques such as peak height measurements (as recommended by the manufacturer) (16), triangulation, mechanical or electronic integration of peak area or by the product of peak height and retention time, a parameter in most GLC systems that is roughly proportional to peak area. Since we use this latter parameter in our general GLC program, described earlier, we decided initially on a similar approach for NCH analysis. However, instead of retention time, the time used was essentially the elution time as defined from the instant the gas valves switched to the furnace flow mode (following the timed combustion cycle of 20 sec) to the time corresponding to the appropriate peak maximum on the chromatogram. This empirical parameter was chosen not only because this time interval reflects chromatographic characteristics and performance of the instrument, but more importantly because it is the only relevant time interval that can be measured conveniently and accurately from the chromatogram.

During the initial phase of calibration, factors which either had to be controlled carefully or for which corrections were needed included the following: 1) Detector (bridge) current. 2) Recorder performance (including calibration). 3) Exact combustion time (20 sec). 4) Calcination of catalyst. 5) Background due to injection rod and sample boat contamination. 6) Catalyst mass control and measurement. 7) Contribution of sample matrix background. 8) Carrier gas flow-rate. 9) Combustion (oxidation) furnace temperature (line voltage regulator required). 10) Humidification of the carrier gas. 11) Two-boat system to prevent sample loss (one boat used as a press fit cover). 12) Avoidance of dust fallout on sample boats and injection rods during sample manipulation. 13) All usual precautions necessary for microanalytical work.

One goal in the use of the instrument was to be able to determine sample composition and absolute mass over the range of from 20–1000 μg. This would also be helpful in analytical work where the per cent content of NCH to total measured mass might be of importance. Also, we wished to be able to analyze elemental composition of lipid and lipoprotein fractions, the masses of which could not be easily determined in advance and which frequently were heavily contaminated by salts. For this application it was necessary to calibrate over total NCH mass ranges of from 20–1000 μg. The present version of the program incorporates nonlinear polynomial evaluation of elemental mass of NCH. It is our experience that any method for accurate calibration over this wide range of sample mass must consider and correct for nonlinear characteristics of the instrument.

Preliminary Calibration Studies

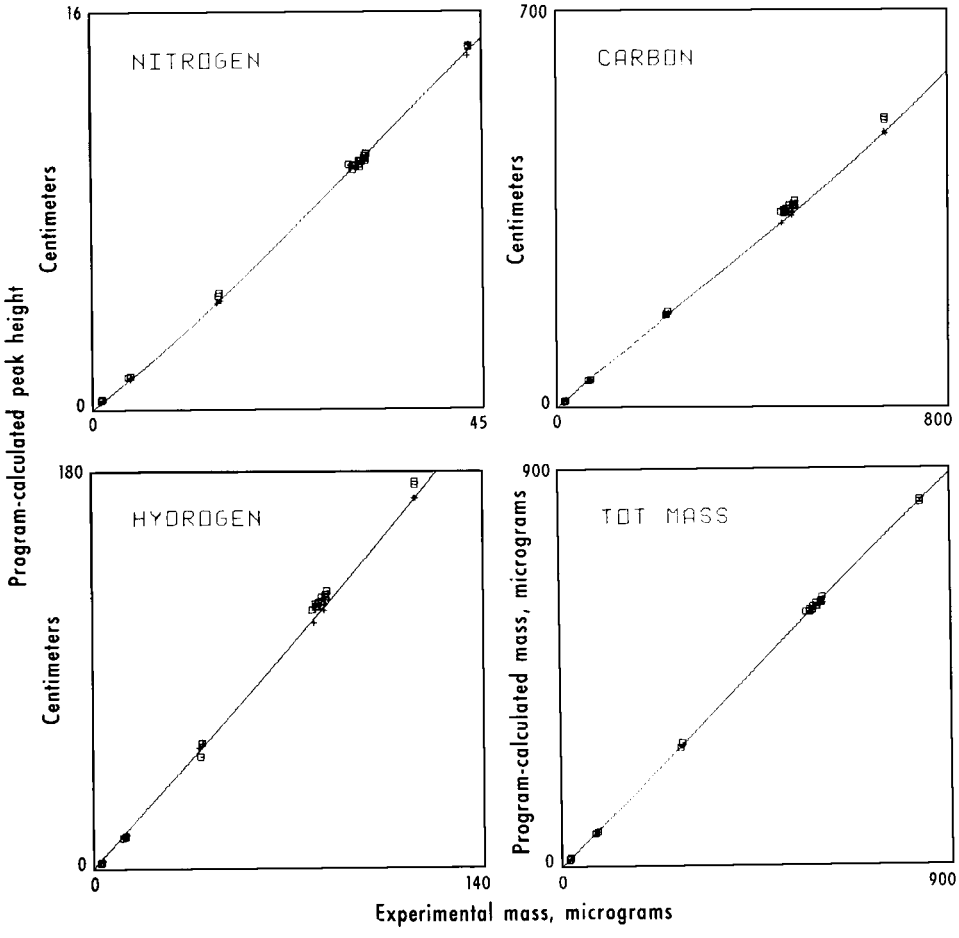
The first step in calibrating the instrument's output was a simple calculation of the parameter, elution time × peak height × recorder attenuation, for both catalyst blank run and samples. Blank run baseline values were subtracted from the sample values, and the resulting values for several standard sample runs of known weight and composition were graphed, yielding approximate linear regressions of the parameter on each elemental mass.

Even these calculations rapidly became too tedious on a desk calculator, so a short Fortran IV program was written to perform them. This program, contained on a deck of cards ¾ in. thick, was first run on the IBM 7094, and later modified to run on the larger CDC 6600. Like other programs, this one has grown substantially as more corrections and refinements have been added. The program deck is now over 6 in. thick, with the end not yet in sight. Subroutines soon were added to normalize peak height measurements to a standard detector bridge current, and to correct attenuation readings to calibrated values based on precision measurement of the resistor attenuator string.

The background baseline at this time merited further consideration. It was clear that it varied in magnitude for different amounts of catalyst. Subtracting an amount proportional to the mass of catalyst used was, however, an overcorrection, since a substantial part of the background was due to residual contamination on the injection rod and sample boats. At the present time we are storing background values of runs with rod and boats alone, one to three levels of catalyst mass, separately for each individual injection rod, and interpolating background values from these based on catalyst mass used in each sample run.

At this stage, the background calculations were sufficiently under control that it became possible to take a closer look at the actual mass calibration, which is slightly nonlinear, and the related question of the suitability of using our original parameter (peak height × time) or peak height alone. All the preceding corrections and refinements can be applied equally well to either parameter; the one to be used in the program may be specified by a single program card. By simply changing it and rerunning the program, the computer is assigned the task of recalculating the entire set of data, which results in two versions of output that can be compared. Such calculations obviously would be tedious and unreasonable to perform manually.

For each mode of calculation, polynomials were fitted to a series of calibration points for each element by a least-squares routine. Cubic polynomials provided a reasonably good fit of the data. A short auxilliary program utilizing the existing package of subroutines for producing graphical output on the Cal-Comp plotter yielded an invaluable graphical presentation of the curves. Figure 4 shows separate calibration curves for N, C and H as well as for



MUB-10646

FIG. 4, abc. Mass calibration curves for elemental N, C and H on the basis of program-corrected peak height. Cubic polynomials through origin are fitted by least-squares to 11 standard sample (stearamide) points (+). Calibration points plus several repeats run approximately 2 weeks later (□) line up along these curves but are displaced slightly, due to changing instrument characteristics.

d. Comparison of total NCH mass with program values calculated on the basis of the displaced calibration curves.

total NCH mass. The calibration for nitrogen generally exhibits a rising slope, in part the result of increased interpenetration of the N_2 peak position by CO_2 at high sample levels. In both the CO_2 and H_2O components there appears to be no significant interference from any other component. A secondary task of the program can be to write out data, such as elution time or differences in mass as calculated by alternate methods, in blocks corresponding to one day's sample runs for evaluation by the statistical program presented in detail in the next section.

A Generalized Computer Program for Statistical Analysis

In order to evaluate and correlate rapidly increasing amounts of experimental data from many sources, a group of programs has been written for statistical analysis. These programs were written in Fortran IV, originally for the IBM 7094 (17). The version presented here has been modified and expanded slightly to run on the CDC 6600 system. However, with minor modification these programs may be adapted to any large memory digital computer. They calculate the usual statistical parameters for a population and perform related tasks, such as listing the data conveniently or comparing a group of related populations in order to identify significant differences between them. A population, as the term is used here, means a set of related cases, for example, "all animals subjected to treatment A." Each case (here an individual animal) consists of a number of measurements, usually one for each variable being studied. A group of related populations is considered to be populations from the same experiment (or from similar experiments), for example, the three populations "all animals subjected to treatment A," "all animals subjected to treatment B" and "all control animals" might make up a group, assuming that the variables have been recorded in the same order and units of measurement for each of the three populations.

Data is entered into the system by means of punched cards, which allow a large degree of freedom in formating. The data can be coded directly onto coding forms for keypunching, or automatically punched by an earlier program, (such as the schlieren analysis program). If a particular data item is unavailable, it can be coded as "missing" simply by leaving its space blank. For flexibility of analysis, the group membership of each population and the population membership of each case can be defined by the arrangement of the input deck. This means that once data cards for all cases have been punched, a wide variety of breakdowns of the data can be analyzed merely by rearrangement of the input cards. This can generally be done with an IBM card sorter.

The first of these programs reads the input deck and stores the data for each population in a single record on the disc file, rapidly accessible to the remaining programs. The statistics program calculates several parameters (18,19). For each population it finds the mean (μ), a standard deviation (σ) and standard error of the mean, for each variable. The program then generates a six-part histogram for the intervals bounded by the points: $(\mu - 3\sigma)$, $(\mu - 2\sigma)$, $(\mu - \sigma)$, μ, $(\mu + \sigma)$, $(\mu + 2\sigma)$ and $(\mu + 3\sigma)$ for each variable. Any values less than $(\mu - 3\sigma)$ or greater than $(\mu + 3\sigma)$ are separately printed. This provides a means of checking the assumption of normal distribution, since the expected percentage of (normally-

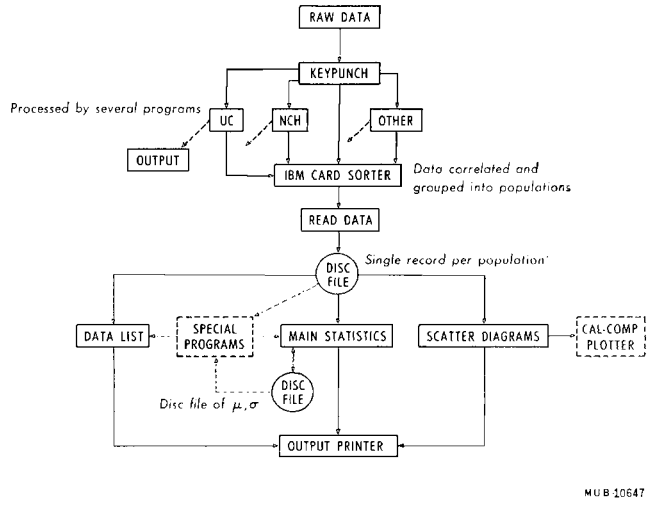


FIG. 5. Flow of data through a group of statistical programs. Data may be gathered from many sources, including some preprocessed by special instrument-oriented programs.

distributed) values within each such interval is well determined. If the normality assumption is untenable for a given variable, the subsequent program calculations involving this variable are probably invalid. Following the histogram calculations, the program computes the correlation coefficients (Pearson's r) for each pair of variables. The magnitude of r (which varies from -1 to $+1$) provides an indication of the degree of the linear relationship between the two variables; the sign of r is the sign of regression coefficient b_{xy} which can be computed from r_{xy} by $b_{xy} = r_{xy}\ \sigma_y/\sigma_x$. All r's significant at either the 5% or 1% level are so labeled. An option allows the calculation of multiple regression coefficients on selected variables. In all these calculations, provisions are made in the program so that missing values for any of the variables will cause no difficulty; the results will be adjusted appropriately and the printed output will state the number of measurements used in each calculation.

Population Comparisons

During these calculations, the results are being written onto a disc file or output tape for subsequent printing. In addition, the values of μ and σ are also written onto another disc file. At the end of a group of related populations these values are read back in, and, selecting all possible pairs of populations in turn, the program tests the two means for each variable in order to identify significant differences, using both the normal test, which is applicable for large values of n, and student's t-test, which is preferable for variables with small numbers of observations. In addition, the program applies the F-test to the pair of standard deviations because any significant differences in the two σ's will invalidate the t-test. Significant values at either the 5% or the 1% level are marked, and the results are printed in tabular form for each pair of populations.

At this stage of the program operation, the input data is still available on disc, and can be used as input to other programs. One lists the data for cross reference and detection of possible errors. Another program produces scatter diagrams of selected pairs of variables on the high-speed printer, and it can be modified to plot them on a Cal-Comp plotter.

In addition, using Fortran IV, it is a simple matter to write special programs to manipulate the data,

by merging populations or deriving new variables, and re-inputing it to the statistics program. An example would be to evaluate biological variation by studying the same variable in given populations at two different times.

Restrictions

The size of memory available to the program limits the amount of data that can be evaluated at one time. Thus upper limits must be established on the number of cases, variables, and total data items. The CDC 6600 version allows 20,000 data items, and up to either 400 variables or 500 cases. The IBM 7094 version allows 7,500 data items or 200 variables maximum, and also allows 500 cases.

In the event a population exceeds these limits, the read-in program will break the data into blocks of allowable size, and the statistics program will perform the calculations described above on these blocks. When the calculations for the last block in such a population are completed, the program will combine all data and calculate μ, σ, and the standard error of the mean for each variable in the population as a whole. In addition, no more than 30 populations may be compared with one another at a time. Should a group ever consist of more than 30 populations, the program will handle them in blocks of 30. Some mutual adjustment of these limits is possible with program modification, for example, a decrease in the number of permissible variables would permit an increase in the size of the data block. On the CDC 6600, it would also be possible to request additional memory, as several programs share the computer at once. It is anticipated that these restrictions will not affect the normal application of this program.

Output Identification

In order to identify correctly computer-produced analyses at some later time, two labeling options can be used. The first provides for a population description consisting of 78 characters (including spaces), which will be printed at the top of each page of output relating to the population. The second labeling procedure allows the use of one six-character (alphabetic or numeric) name for each variable; in the normal case, variable names are entered with the first population in each group and, from the point of view of the program, their entry serves to define the beginning of a new group. Thus, the careful use of the variable name option both enables the program to include the appropriate labels with the output and also provides a control over the comparisons that will be made.

Time required for the programs is quite variable, depending on the amount of data and the number of calculations required. A moderately large application will probably take no more than 2 min of IBM 7094 time or 45 sec of CDC 6600 central processor time, at a nominal cost.

ACKNOWLEDGMENTS

This work was supported by Research Grant 5-RO1-HE-01882-11 from the National Heart Institute, Public Health Service, Bethesda, Maryland, and by the United States Atomic Energy Commission.

REFERENCES

1. de Lalla, O., and J. W. Gofman, in "Methods of Biochemical Analysis, Vol. 1," edited by D. Glick, Interscience Publishers, New York, 1954, p 459.
2. Johnston, J. P., and A. G. Ogston, Trans. Faraday Soc. *42*, 789 (1946).
3. Trautman, R., V. N. Schumaker, W. F. Harrington and H. K. Schachman, J. Chem. Phys. *22*, 555 (1954).
4. Ewing, A. M., N. K. Freeman and F. T. Lindgren, in "Advances in Lipid Research, Vol. 3," edited by R. Paoletti and D. Kritchevsky, New York, Academic Press, Inc., 1965, Chap. 2, p 25.
5. Lindgren, F. T., N. K. Freeman, A. M. Ewing, and L. C. Jensen, JAOCS *43*, 281 (1966).
6. Bierman, E. L., T. L. Hayes, J. N. Hawkins, A. M. Ewing and F. T. Lindgren, J. Lipid Res. 7, 65 (1966).
7. James, A. T., and V. R. Wheatly, Biochem. J. *63*, 269 (1956).
8. Pecsok, R. L., "Principles and Practice of Gas Chromatography," New York, John Wiley and Sons, Inc., 1959, p 145.
9. Tandy, R. K., F. T. Lindgren, W. H. Martin and R. D. Wills, Anal. Chem. *33*, 665 (1961).
10. Carroll, K. K., Nature *191*, 377, (1961).
11. Johnson, R. D., D. D. Lawson and A. J. Havlik, J. Gas Chromatog. *3*, 303 (1965).
12. Johnson, H. W., Anal. Chem. *35*, 521 (1963).
13. Ewing, A. M., P. P. Walker, R. D. Wills and F. T. Lindgren, Semiannual Report, Lawrence Radiation Laboratory, UCRL-11184, Fall, 1963.
14. Bartlet, J. C., and D. M. Smith, Can. J. Chem *38*, 2057 (1960).
15. Jensen, L. C., F. T. Upham, G. S. Stevens and F. T. Lindgren, in preparation.
16. Hinsvark, O. N., and R. H. Waltz, Technical Paper No. 31, F&M Scientific Corp., Avondale, Pa., 1966.
17. Ewing, A. M., Semiannual Report, Lawrence Radiation Laboratory, UCRL-11833, Fall, 1964.
18. Hoel, P. G., "Introduction to Mathematical Statistics," New York, John Wiley & Sons, 1954.
19. Guilford, J. P., "Fundamental Statistics in Psychology and Education," New York, McGraw-Hill, 1956.