COMPARISONS OF DATA PROCESSING PROCEDURES FOR THE STATISTICAL ANALYSIS OF VISUAL FIELDS

Katherine Freeman, Montefiore Medical Center

OBJECTIVE

Practical experience with regard to analyzing visual field data is rare. This paper intends to illustrate practical and efficient methods of handling and analyzing data that are represented as coordinates of a matrix.

STATEMENT OF PROBLEM

Medical: The visual field is defined by 74 points in ones's field of perception, depicting a somewhat octagonal shaped matrix. The measurement taken at each point in the matrix is the threshold or the lowest intensity of light(measured in decibels) perceived by the eye at that point in the visual field. The technique of static automated perimetry has been used successfully in glaucomatous eyes to detect visual field loss over time by examining changes in corresponding threshold points from visual fields obtained at two points in time. However, a problem still exists with regard to defining those changes that represent measurement errors which are due to such factors as different medications, alertness, pupil size and relative place in the visual field, etc., in contrast with those changes that represent true clinically notable loss. Due to variablity in these measurements, ranges have not as yet been established that define whether changes in threshold that occur over time are within the 'normal' range(i.e. are due to random components of change) vs changes that represent some systematic non-random component representing true clinically notable changes(those changes that lie outside the 'normal range'). Equivalently, the problem can be stated as how does one distinguish normal/expected fluctuations in response from true clinical progression of disease. The problem becomes less academic when one considers that clinical decisions to treat a patient in a particular way are based on the magnitude of these changes. In addition, although information regarding variability in normal eyes is available, fluctuation due to measurement error in glaucomatous eyes has not as yet been studied, and anecdotal evidence indicates that variability in glaucomatous eyes is generally larger compared with normal eyes.

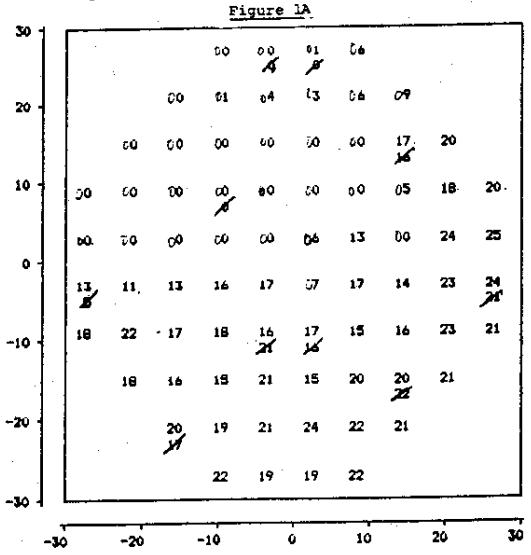
Statistical: The problem becomes finding a measure to represent fluctuations over time that would be valid statistically and intuitively palatable. Since the threshold values were considered to be normally distributed (although this assumption is somewhat difficult to validate given the relatively small sample size in the current study), the measure of fluctuation or dispersion was chosen to be the variance of the threshold measurement evaluated at each corresponding point in the six visual fields. The 'normal' range within which random fluctuations produced by measurement error could occur would be that range marking two standard deviations about the mean threshold at the point; if the data were normally distributed, this interval would encompass approximately the central 95.5% of all observations. A further problem arises when one considers all 74 points in the visual field and the variability about each of these points. Thus, a single representative measure of fluctuation for the visual field is not readily apparent; one estimate can be derived by pooling estimates of variability obtained for each of 74 points in the visual field across the six visual fields.

Data Management Problem: Data for 74 threshold measurements (printed in an octagonal matrix pattern) represented one evaluation(visual field) for one eye of a patient with glaucoma. There were five patients, each of whom had two glaucomatous eyes. Evaluations were performed at baseline, 1, 2 and 3 weeks after baseline, and 2 and 3 months after baseline, twice a day--once in the morning and once in the afternoon. Several problems existed with regard to managing the data. Firstly, the data were not submitted in a format that could be keyed onto tape readily. Secondly, in order to streamline procedures for working with this volume of variables, it was considered an advantage to name each variable by the variable's 'location' or (x,y) coordinate point within the visual field, and in addition that the name be short and adaptable for performing SAS procedures and functions. Thirdly, there is large amount of data--10 eyes x 2 times daily x 6 serial evaluations with 74 data points per evaluation.

MATERIALS AND METHODS

Serial visual fields were obtained using program 32 on an Octopus 2000R perimeter. An example of the original data sheets is shown in Figure 1A.

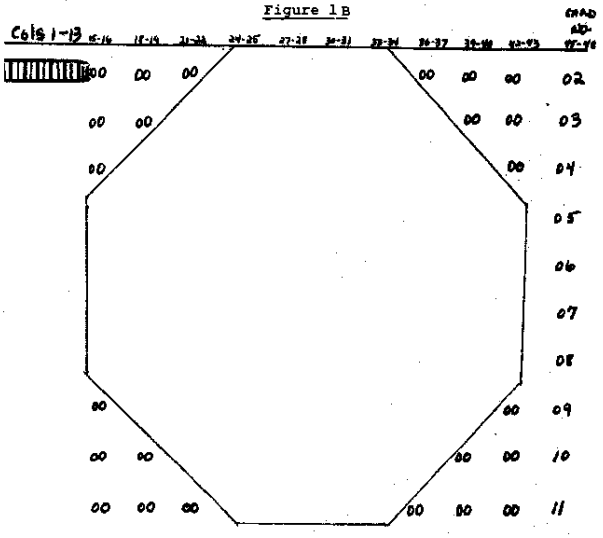
Figure 1A



The typed numbers are those generated by the OCTOPUS 2000R; if a threshold measure was a single digit, a zero was written to the left of the digit so that each threshold would have two digits for data entry. Ten points in the visual field have two determinations(called double determinants); the points were chosen randomly and the average of the two, rounded up, was used as the threshold measure for the point.

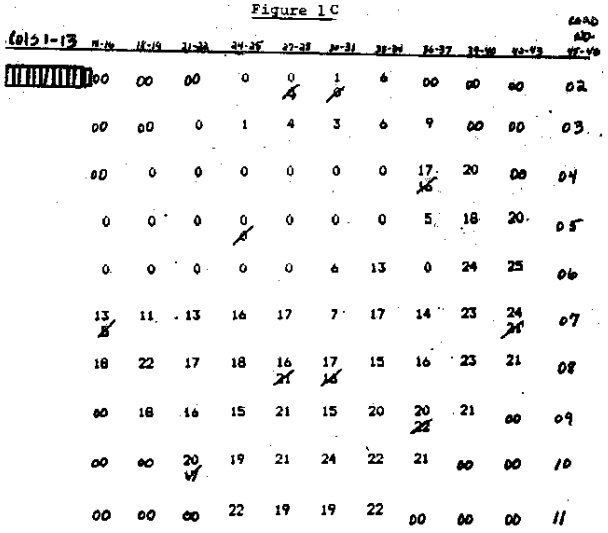
Since data were not in a form that could be keyed readily, a cover sheet or template was devised to specify columns for data values; the template(Figure 1B) had a cut-out window so that data from the visual field beneath the template window would be visible (Figure 1C); columns 1-13 on each 'card' are reserved for patient identification information. The clinicians were instructed to code all visual field measurements as two digit numbers, so that data would be keyed

Figure 1B



correctly within the columns specified. Data entry instructions were printed on the template; columns for data entry were specified so that there would be a blank column between each pair of variable fields. This would facilitate reading the data using a combination of list input and grouped format lists as part of SAS input routines.

Figure 1C



Data were keyed onto tape and transferred to disk attached to Montefiore Medical Center's IBM 3081 mainframe computer running VM/MVS/TSO with ISPF. At the time these data were processed, SAS version 5.08 was in use. The code used to read and store the data as a permanent SAS dataset are shown in Figure 2A. The variables ID1 through ID10 represent the same unique identification number for a given patient. DF1 through DF10 represent the date of evaluation for the visual field, entered in MMDDYY6. format. TIME1 through TIME10 represent the time of day(AM vs PM) at which the visual field was evaluated, and R_L1 through R_L10 represent whether the visual field represented the right or left eye. Of these identifier variables, the variables ID1, DF1, TIME1, and R_L1 were kept and were subsequently renamed ID, DF, TIME, and R_L, respectively. Thus the variables ID, DF, TIME and R_L represent the unique identification of a specific visual field for a patient. Points in the visual field are represented by the variable name Vxy, where 'V' is the 'root' of the variable name and x and y are numbers such that x represents the row of the visual field matrix numbered sequentially from top(row=1) to bottom(row=10), and y represents the column of the visual field matrix numbered sequentially from left(column=0) to right (column=9).

FIGURE 2A: INITIAL INPUT STATEMENT--ALL DATA;

```
DATA OPHTHAM2; INFILE RAWD2;
BATCH="&SYSDATE"; **IDENTIFIES DATE OF BATCH;
INPUT
ID1 DF1 MMDDYY6. TIME1 R_L1 (V10-V19) (3.)
CARD2/
ID2 DF2 MMDDYY6. TIME2 R_L2 (V20-V29) (3.)
CARD3/
ID3 DF3 MMDDYY6. TIME3 R_L3 (V30-V39) (3.)
CARD4/
ID4 DF4 MMDDYY6. TIME4 R_L4 (V40-V49) (3.)
CARD5/
ID5 DF5 MMDDYY6. TIME5 R_L5 (V50-V59) (3.)
CARD6/
ID6 DF6 MMDDYY6. TIME6 R_L6 (V60-V69) (3.)
CARD7/
ID7 DF7 MMDDYY6. TIME7 R_L7 (V70-V79) (3.)
CARD8/
ID8 DF8 MMDDYY6. TIME8 R_L8 (V80-V89) (3.)
CARD9/
ID9 DF9 MMDDYY6. TIME9 R_L9 (V90-V99) (3.)
CARD10/
ID10 DF10 MMDDYY6. TIME10 R_L10 (V100-V109)
(3.) CARD11;
PROC PRINT;
TITLE1 'INITIAL INPUT STATEMENT--ALL DATA';
```

An alternative approach using SAS MACRO language to perform iterative processing for reading the raw data is shown in Figure 2B.

FIGURE 2B: MACRO INPUT STATEMENT--
SUBSET OF VISUAL FIELD VARIABLES;

```
          DATA OPHMAC2; INFILE RAWD2;
BATCH="&SYSDATE"; **IDENTIFIES DATE OF BATCH;
%MACRO INMAC(FIRST=1,LAST=9);
     %LOCAL I;
     %LET ZERO=0;
     %LET NINE=9;
INPUT
     ID1 DF1 MMDDYY6. TIME R_L1
     %DO  I=&FIRST  %TO  &LAST;
     @15 (V&I&ZERO-V&I&NINE)  (3.) /
     %END;
     @15 (V100-V109)  (3.);
DROP V10--V12  V17--V21  V28--V30  V39 V80
     V89--V91 V98--V102 V107--V109;
     RUN;
PROC PRINT;
TITLE1 'MACRO INPUT STATEMENT';
%MEND INMAC;
%INMAC(FIRST=1,LAST=9);
```

The SAS MACRO solution avoids reading in redundant information (i.e identification variables other than ID1,DF1,TIME1,R_L1) that were useful only in conjunction with data entry.

Once the data were validated, a temporary SAS dataset was created to find the variability associated with each corresponding point across the six visual fields. It was assumed that the variability in light intensity perceived(threshold) at a single point was constant over time(regardless of the length of intervals between evaluations), and the true variability across visual fields measured at a single point is expected to be the same across all 74 points in the visual field. The estimates of variability for the point V(x,y,z) are expressed mathematically for the (q)th patient in Figure 3.

FIGURE 3: VARIANCES ACROSS VISUAL FIELDS AND POOLED ACROSS POINTS

$$\text{Estimate of Variance for Point V(x,y) Across z=6 fields} = \sum_{z=1}^{6} \frac{(V(x,y,z) - \bar{V}(x,y))^2}{5}$$

$$= \text{VAR}(V(x,y))$$

$$\text{Estimate of Variance Across Time Pooled Over All 74 Points} = \sum_{x=1 \text{ to } 10;\ y=0 \text{ to } 9} \frac{\text{VAR}(V(x,y))}{74}$$

where x represents the xth row of the visual field matrix
y represents the yth column, and
z represents the zth visual field such that z=1 for baseline and z=6 for 3 months.

$$\bar{V}(x,y) = \sum_{z=1}^{6} V(x,y,z)/6 \ .$$

APPROACH 1: PROC SUMMARY WITH SAS MEANS FUNCTION

Estimates of variability were derived using SAS PROC SUMMARY in conjunction with the SAS means function. The code is provided in Figure 3A. Firstly, PROC SUMMARY was used with a class statement containing ID and R_L so that variances could be computed individually for right and left eyes of a patient. Note, for this study, although evaluations of right and left eyes of the same patient are usually not considered to be statistically independent, the clinicians felt that progression of disease in each of the two eyes are independent and should thus be treated statistically as such. The SUMMARY procedure was preferred to the MEANS procedure because it allowed for variables to be specified in the VAR statement in an abbreviated variable list without the necessity of specifying individual variable names of the newly created variances in the OUTPUT dataset; this is not an option in the MEANS procedure. This is particularly useful when statistics are required for each of many variables, as we have for the visual field data matrix. The SAS means function was then applied to the newly created variables (V13--V106) in the OUTPUT dataset VARDS1, each of which represents the variance of the (x,y) threshold point across the six fields.

FIGURE 3A: SUMMARY PROCEDURE WITH MEANS FUNCTION

```
10          DATA OPHTHAMA; SET OPHTHAM;
11               IF ID<10; *GLAUCOMATOUS PATIENTS ONLY;
12               IF NVISIT<6; *FIRST 4 WEEKS,2,3MOS;
13

NOTE: DATA SET WORK.OPHTHAMA HAS 109 OBSERVATIONS AND 98 VARIABLES. 58 OBS/TRK.
NOTE: THE DATA STATEMENT USED 0.18 SECONDS.

13            PROC SORT; BY ID R_L;
14

NOTE: 4 CYLINDERS DYNAMICALLY ALLOCATED ON SYSDA FOR EACH OF 3 SORT WORK DATA SETS.
NOTE: DATA SET WORK.OPHTHAMA HAS 109 OBSERVATIONS AND 98 VARIABLES. 58 OBS/TRK.
NOTE: THE PROCEDURE SORT USED 0.42 SECONDS.

14            PROC SUMMARY; CLASS ID R_L; VAR V13--V106;
15               OUTPUT OUT=VARDS1 VAR= ;
NOTE: THE DATA SET WORK.VARDS1 HAS 18 OBSERVATIONS AND 80 VARIABLES. 72 OBS/TRK.
NOTE: THE PROCEDURE SUMMARY USED 0.72 SECONDS.

16          DATA VARDS1; SET VARDS1;
17               MVARDS1=MEAN(OF V13--V106);
18          TITLE1 'POOLED-VARIANCES FOR ALL POINTS ACR VIS--WKS1-4,AND 2,3 MOS';
19

NOTE: DATA SET WORK.VARDS1 HAS 18 OBSERVATIONS AND 81 VARIABLES. 72 OBS/TRK.
NOTE: THE DATA STATEMENT USED 0.13 SECONDS.

19            PROC PRINT DATA=VARDS1;
20            VAR ID R_L _TYPE_ _FREQ_ MVARDS1;
21          TITLE1 'POOLED VARIANCES FOR ALL PTS ACROSS VIS 1-4 WKS,2,3MOS--SUMM';
22          TITLE2 'VISUAL FIELD STUDY--VISITS 1ST 4 WKS,2,3MOS  ';
```

APPROACH 2: USE OF ARRAYS WITH MEANS PROCEDURE(FIGURE 3B)

An array was created of all the points in the visual field. The threshold corresponding with each point was then output as a single observation. The data were then sorted by ID and R_L, and the MEANS procedure was used with a BY statement with the variables in the SORT procedure to compute the variances of each of the points across visual fields and create an OUTPUT dataset(STATS0) of variances corresponding with each point (x,y) in the visual field. Another MEANS procedure was used to pool(take tbe mean of) the variances for each point across all points in the visual field.

FIGURE 3B: ARRAY STATEMENT WITH MEANS PROCEDURES

```
27          DATA ALLV; SET OPHTHAMA;

NOTE: DATA SET WORK.ALLV HAS 109 OBSERVATIONS AND 98 VARIABLES. 58 OBS/TRK.
NOTE: THE DATA STATEMENT USED 0.15 SECONDS.

28          DATA ALLV; SET ALLV;
29          ARRAY V V13--V106;
30          DO OVER V;
31          ALLV=V; INDEX=_I_ +12;
32          IF V^=. THEN OUTPUT;
33          END;
34

NOTE: DATA SET WORK.ALLV HAS 8066 OBSERVATIONS AND 100 VARIABLES. 58 OBS/TRK.
NOTE: THE DATA STATEMENT USED 0.99 SECONDS.

35          DATA ALLV; SET ALLV;
36          KEEP ALLV R_L NVISIT ID INDEX;

NOTE: DATA SET WORK.ALLV HAS 8066 OBSERVATIONS AND 5 VARIABLES. 1066 OBS/TRK.
NOTE: THE DATA STATEMENT USED 1.20 SECONDS.

37          PROC SORT; BY INDEX ID R_L;

NOTE: DATA SET WORK.ALLV HAS 8066 OBSERVATIONS AND 5 VARIABLES. 1066 OBS/TRK.
NOTE: THE PROCEDURE SORT USED 0.99 SECONDS.

38          PROC PRINT DATA=ALLV(OBS=10);
39          TITLE1 'PRINT OF 10 OBS FOR ALLV DATASET';
40
NOTE: THE PROCEDURE PRINT USED 0.12 SECONDS AND PRINTED PAGE 2.

41          DATA ALVONE; SET ALLV;

NOTE: DATA SET WORK.ALVONE HAS 8066 OBSERVATIONS AND 5 VARIABLES. 1066 OBS/TRK.
NOTE: THE DATA STATEMENT USED 0.58 SECONDS.

42          PROC SORT; BY INDEX ID R_L ;

NOTE: DATA SET WORK.ALVONE HAS 8066 OBSERVATIONS AND 5 VARIABLES. 1066 OBS/TRK.
NOTE: THE PROCEDURE SORT USED 0.91 SECONDS.

43          PROC MEANS NOPRINT; BY INDEX ID R_L;
44           VAR ALLV; OUTPUT OUT=STATS0 VAR=VARALLV;
NOTE: THE DATA SET WORK.STATS0 HAS 740 OBSERVATIONS AND 4 VARIABLES. 1304 OBS/TRK.
NOTE: THE PROCEDURE MEANS USED 1.63 SECONDS.

45          PROC SORT; BY ID R_L;

NOTE: DATA SET WORK.STATS0 HAS 740 OBSERVATIONS AND 4 VARIABLES. 1304 OBS/TRK.
NOTE: THE PROCEDURE SORT USED 0.23 SECONDS.

46          PROC MEANS NOPRINT; BY ID R_L;
47          VAR VARALLV;
48          OUTPUT OUT=STATS1 MEAN=MVARALLV;
NOTE: THE DATA SET WORK.STATS1 HAS 10 OBSERVATIONS AND 3 VARIABLES. 1676 OBS/TRK.

NOTE: THE PROCEDURE MEANS USED 0.19 SECONDS.

49          PROC PRINT DATA=STATS1;
50          TITLE1 'POOLED VARIANCES FOR ALL PTS ACROSS VIS 1-4 WKS,2,3MOS--MEAN';
51.
```

FIGURE 4: SUMMARY PROCEDURE WITH MEANS FUNCTION(OUTPUT)

POOLED VARIANCES FOR ALL PTS ACROSS VIS 1-4 WKS,2,3MOS--SUMM
VISUAL FIELD STUDY--VISITS 1ST-4 WKS,2,3MOS

OBS	ID	R_L	_TYPE_	_FREQ_	MVARDS1
1	.	.	0	109	12.9524
2	.	0	1	56	11.2967
3	.	1	1	53	15.2152
4	1	.	2	24	19.2728
5	2	.	2	23	6.1068
6	3	.	2	22	16.4242
7	4	.	2	24	16.7283
8	5	.	2	16	6.1588
9	1	0	3	12	22.8990
10	1	1	3	12	17.9217
11	2	0	3	12	5.5006
12	2	1	3	11	7.5133
13	3	0	3	12	14.8140
14	3	1	3	10	21.0414
15	4	0	3	12	8.4802
16	4	1	3	12	26.6278
17	5	0	3	8	8.1554
18	5	1	3	8	5.2346

FIGURE 5: ARRAY STATEMENT WITH MEANS PROCEDURES(OUTPUT)

POOLED VARIANCES FOR ALL PTS ACROSS VIS 1-4 WKS,2,3MOS--MEAN

OBS	ID	R_L	MVARALLV
1	1	0	22.8990
2	1	1	17.9217
3	2	0	5.5006
4	2	1	7.5133
5	3	0	14.8140
6	3	1	21.0414
7	4	0	8.4802
8	4	1	26.6278
9	5	0	8.1554
10	5	1	5.2346

DISCUSSION

Two ways of reading raw data were illustrated; the first presented was efficient in the sense that it utilized grouped format lists and abbreviated variable lists. It was also efficient in the sense that it could be easily interpretable by relatively inexperienced SAS programmers. In contrast, the second approach required an iterative MACRO %DO loop and other local or global MACRO variables. This approach was advantageous because only the necessary variables were read into the permanent SAS database.

In performing the required statistical operations, however, the SUMMARY procedure in conjunction with the SAS means function did an effective job of computing the required variances using less computer time than did the approach using arrays with the means procedures. The summary procedure and means function(including data statements and print) used 1.44 seconds of CPU time with PROC SUMMARY using .71 seconds(almost half) of this time. The array approach using means procedures used 6.82 seconds of CPU time, almost 5 times the amount used for the summary procedure approach. For the array approach, the means procedure alone used 1.82 seconds of CPU time. Although the print of the OUTPUT file from the SUMMARY procedure is not as readily interpretable as that of the output for the MEANS procedure, the clarity and conciseness of SAS statements as well as savings in CPU time make it the preferred approach in this example.