/NONPARAMETRIC DENSITY ESTIMATION FOR UNIVARIATE AND BIVARIATE DISTRIBUTIONS WITH APPLICATIONS IN DISCRIMINANT ANALYSIS FOR THE BIVARIATE CASE/

by

MARK HAUG

B.S., Kansas State University, 1984

MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1986

Approved by:

James J. Higgins

Major Professor

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. James J. Higgins, for his help and guidance throughout my report.

INTRODUCTION

Density estimation dates back to A.D. 1538, when Henry VIII ordered a crude census be taken. By 1625, perodic lists of births and deaths were printed, and in 1661 John Graunt presented a paper to the Royal Society on the "bills of mortality". The paper attempted to assign probabilities that an individual (in London) would die within a particular age interval. Graunt's histogram had set a precedent for exhibiting an empirical probability density function. (1)

It is the purpose of this report to look at an alternative form of density estimation known as the kernel estimate, an approach to nonparametric density estimation. The kernel estimate gives us a smooth probability density estimate rather than a step-function constructed by the method of histograms. It is this smoothness that makes the kernel estimator appealing.

In 1962 Parzen defined and investigated properties of kernel estimate of probability density function. They have been widely investigated since that time (2). Other nonparametric methods have also been given (1,3). In spite of their popularity, the SAS, SPSS, and BMDP programs do not have the procedure for constructing kernel estimates. Moreover, software for microcomputers to construct kernel estimates does not appear to be readily available.

As a major part of this investigation, microcomputer software was developed for the purpose of constructing univariate and bivariate nonparametric density estimates of the kernel type. The microcomputer is an ideal environment for implementing nonparametric density estimation due to the interactive nature of the estimation procedure. This newly developed software will allow researchers

(1) to fit univariate densities to data

(2) obtain estimated probabilities

(3) obtain plots of density

(4) fit bivariate distributions to data

(5) obtain plots of bivariate densities

(6) obtain nonparametric discrimination for bivariate discriminant analysis

A detailed set of instructions is given in the report to enable the nonstatistician to implement the density estimation procedure. Examples are given to illustrate the output. The procedures are applied to three examples:

(1) data from a known distribution

(2) actual waiting time data

(3) actual bivariate sociological data.

The kernel estimate can be applied to any of the applications that a histogram estimate may be applied to, and the kernel estimate is more appropriate for continuous data. The software created will be practical for the researcher and instructive for the more traditional histogram followers.

II. DENSITY ESTIMATION

2.1 METHOD OF HISTOGRAMS

Relative frequency for an interval is the frequency of that interval divided by the total number of measurements. For Graunt's paper, "bills of mortality", he took an historical count of deaths in London. Then, after determining the age intervals, he would simply count the number of deaths (frequency) for each age interval. The count for each age interval divided by the entire count is Graunt's probability of deaths (relative frequency). See Figure 2.1.1. (1).

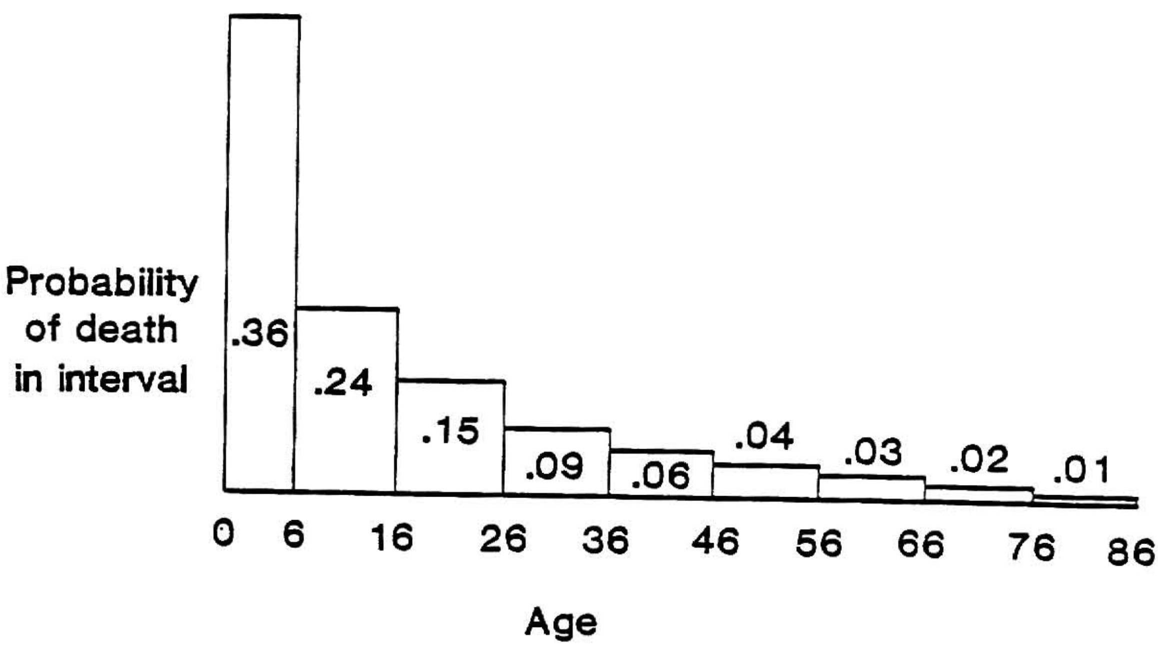


Figure 2.1.1

One problem with the histogram is in selecting the number of intervals and inverval widths in order to obtain an accurate representation of the population distribution. Too few intervals suppresses fine details in the appearance of the density. Too many intervals cause too rough a representation of the distribution. Huntsberger (4) has proposed an equation determining an adequate number of intervals (k) as a function of the sample size (n). The equation is

$$k = 1 + 1.4 \log_e(n). \tag{2.1.1}$$

With Huntsberger's k, the computed range of a data set, and the frequencies of each interval (determined by the range divided by K), an adequate histogram can be determined. This, then, is one estimate of the data's probability density.

2.2 METHOD OF KERNELS

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population with continuous density f(x). By the method of kernels, f(x) can be estimated by $\hat{f}(x)$, where

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \tag{2.2.1}$$

Here $K(\cdot)$ is a bounded symmetric probability density function, called the kernel (3). Throughout this report the kernel will be the Normal (Gaussian) probability density function. The constant h is called the "window" size.

The choice of functional form $K(\cdot)$ is not critical as long as $K(\cdot)$ is a smooth function (3). However, the window size is critical. The window size is analogous to the histogram's interval size in that the same problems surface with an undersized window or an oversized window.

The kernel estimate is a weighted average of data near x. The window size and the function $K(\cdot)$ determine the amount of weight on each data point, X_i. The constant h must be chosen such that it is not too large, thereby not assigning too great a weight to the more distant observations in determining $\hat{f}(x)$ at x. Also h must not be chosen so small as to ignore the more influential observations near x.

An empirical method of determining window size was developed and used in the computer program as follows.

Using Huntsberger's equation (2.1.1), k is determined. Dividing k into the range of a data set, the histogram interval size is determined. This quantity is used as the window size, h, for samples of size less than or equal to 40. Based on experience with actual data sets, the following modifications were made for larger samples. The empirical h was multiplied by 3/4 for $40 < h \leq 75$, by 1/2 for $75 < n \leq 200$, and by 1/4 for $n > 200$. These window sizes are merely suggested values. The user of the program, working interactively with the computer, may select any window size for his/her data that seems appropriate.

III. ESTIMATING THE UNIVARIATE DISTRIBUTION

3.1 METHOD OF KERNELS

Equation 2.2.2 becomes

$$\hat{f}(x) = \frac{(2\pi)^{-\frac{1}{2}}}{nh} \sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{x-X_i}{h}\right)^2} . \qquad (3.1.1)$$

Let y_1 and y_n be the smallest and largest values of the sample data set, respectively. Then, a suggested initial value for h, based on empirical studies, is

$$h = \frac{(y_n - y_1)}{1 + 1.4 \log_e(h)} \cdot c \qquad (3.1.2)$$

where $c = 1$ for $n \le 40$, $c = 3/4$ for $n > 40$, $c = 1/2$ for $n > 75$, and $c = 1/4$ for $n > 200$.

3.2 A COMPUTER PROGRAM TO COMPUTE THE KERNAL ESTIMATE OF AN UNIVARIATE DISTRIBUTION WITH APPLICATIONS FOR PROBABILITY ESTIMATION

Using 3.1.1 and 3.1.2, $\hat{f}(x)$ can be computed for any univariate data set. Computer software is available from the Department of Statistics, Kansas State University. Using an IBM compatible arrangement, the following steps must be taken in order for the user to achieve results (graphical representation of the density estimate and probability estimation). The program is written in GW-Basic.

1. Load BASIC.
2. Insert diskette into slot A (available from the Department of Statistics).
3. Type LOAD"A:UDKE, press "Return"
4. Type RUN, press "Return"
5. Proceed to type in data (each entry followed by "Return")
6. At the end of the data, press "Return" as instructed in #5. Then type -9999 and press "Return". (The program sorts the data and displays the suggested window size)
7. The user is asked what window size he/she wishes to use for the density estimate. Type user's window size and push "Return".
8. There is a waiting period.
9. The density estimate is displayed on the screen. For a printed copy push the print screen function.
10. After viewing the density estimate, push "Return" and the monitor prompts the user for a lower integral limit. (There may be a pause before the user is prompted.)

11. Type the lower integral limit and press "Return".

Then, type the up integral limit and press "Return"

(These are a and b respectively for $\int_a^b \hat{f}(x)\,dx$.

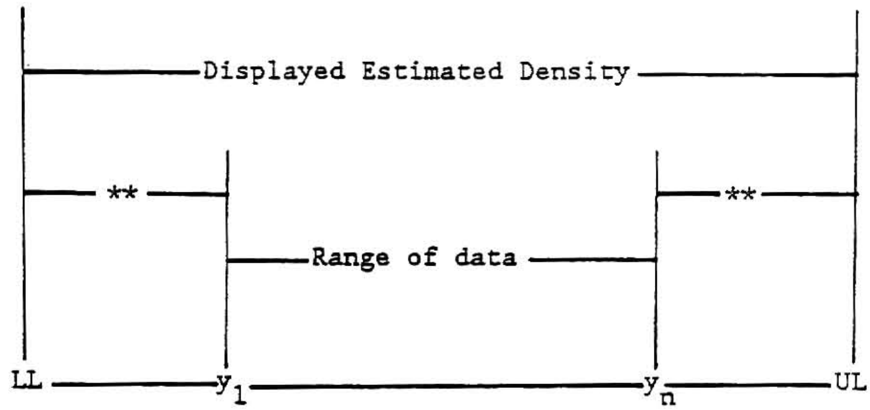
The trapezoidal rule is employed for computation.)

12. The estimate is computed and the user is prompted as to whether he/she desires another set of probability limits. Type YES and push "Return" if this is desired (See #11). Type NO and push "Return", and the user is asked whether he/she desires to use any other window size.

13. Type YES and press "Return" if this is desired (See #7). If NO, the program is terminated.

The limits a and b for $\int_a^b \hat{f}(x)\,dx$ must be chosen such that $LL \le a < b \le UL$, where $LL = y_1 - \frac{(y_n - y_1)}{4}$ and

$UL = y_n + \frac{y_n - y_1}{4}$. See Figure 3.2.1



** one-quarter range of data

Figure 3.2.1

Two examples follow. The first example is a random sample of fifty observations from a normal density with known mean and standard deviation of 2 and 1, respectively. (Figure 3.2.2) The results of the probability estimation are recorded in Figure 3.2.3. Graphical displays of the density estimate are represented in Figure 3.2.4 to Figure 3.2.6.

The second example is one hundred thirty two observations that are the lengths of time (in minutes) from the first customer-salesperson contact to the close of the sale or the customer's departure (5) (Figure 3.2.7.). After estimating the density in the second example, it is fairly evident that the waiting times appear to be Gamma distributed. Using method-of-moments estimators, a gamma distribution was fitted to the data. The results of the estimated parameter density and several kernel estimates for probability estimation are recorded in Figure 3.2.8. Graphical displays are represented in Figure 3.2.9 to Figure 3.2.11. This is an example of nonparametric density estimation used as a first step in model identification.

.045
.644
.717
.719
.884
.933
.978
1.146
1.221
1.253
1.257
1.263
1.35
1.443
1.444
1.481
1.53
1.553
1.564
1.631
1.654
1.662
1.705
1.727
1.816
1.915
1.935
2.072
2.103
2.122
2.15
2.166
2.187
2.251
2.284
2.359
2.373
2.427
2.458
2.521
2.897
3.02
3.056
3.331
3.499
3.643
3.726
3.801
4.183
4.635

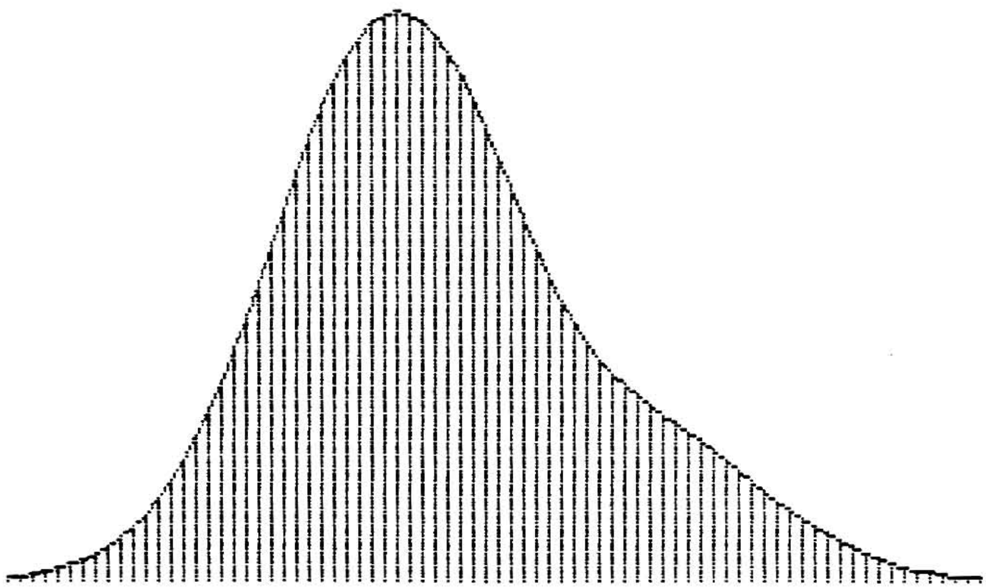
Data from Normal(2,1)

Figure 3.2.2

	-1 to 0	0 to 1	1 to 2	2 to 3	3 to 4	4 to 5
Actual	.022	.126	.341	.341	.126	.022
Histogram	.000	.140	.400	.280	.140	.040
1/2 S.W	.011	.159	.426	.257	.128	.036
SW	.022	.183	.373	.264	.128	.043
2 SW	.064	.190	.270	.241	.152	.059

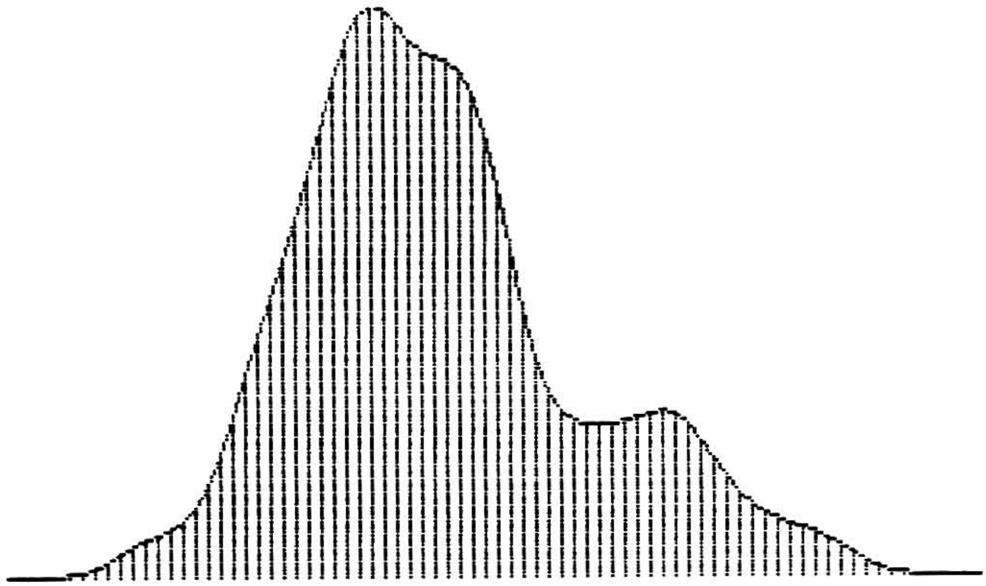
Probability Estimation for example one.

Figure 3.2.3



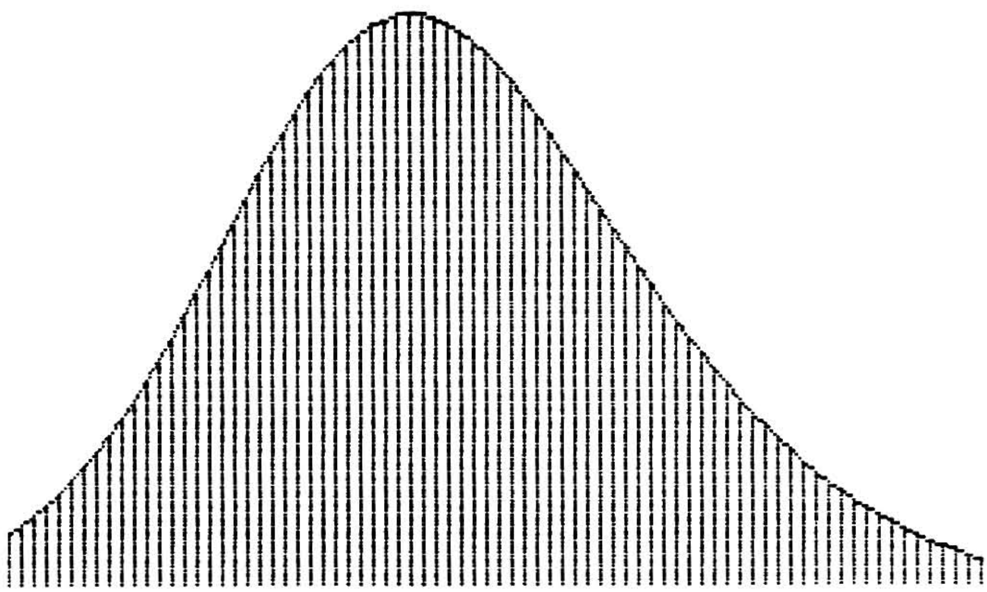
Density with suggested window size.

Figure 3.2.4



Density with one-half the suggested window size.

Figure 3.2.5



Density with two-times the suggested window size.

Figure 3.2.6

.4
.7
.8
.9
1
1.1
1.5
1.7
1.9
3
3.2
3.3
4.1
4.5
5.1
5.4
6
6
6.1
7
7
7.4
7.4
7.6
7.7
7.8
7.9
7.9
8
8.100001
8.100001
8.399999
8.399999
8.8
8.899999
9
9.7
10
10.1
10.3
10.5
10.9
10.9
11
11
11.1
11.1
12
12.2
12.3
12.4
12.5
12.8
13
13.3
13.5
14
14.2
14.2
14.9
14.9
15
15
15
15.1
15.5
15.9
15.9
16.2
16.9
17.4
17.6
17.7
18
18.4
18.7
19.2
20.1
20.1
20.6
21.9
22.3
23
24.8
25.1
25.4
25.6
26
26.2
26.4
27.2
27.7
27.7
28.7
29.2
30
30
30
30.5
30.5
31.8
32.2
33.3
34.6
35
35.1
35.4
37
38.1
38.2
39.2
39.9
40
40.1
41.3
41.6
42.1
42.3
43.1
44.4
47.6
48.6
49.1
50.1
60.2
66.1
69.1
77.1
81
98.2
105.2
118.4

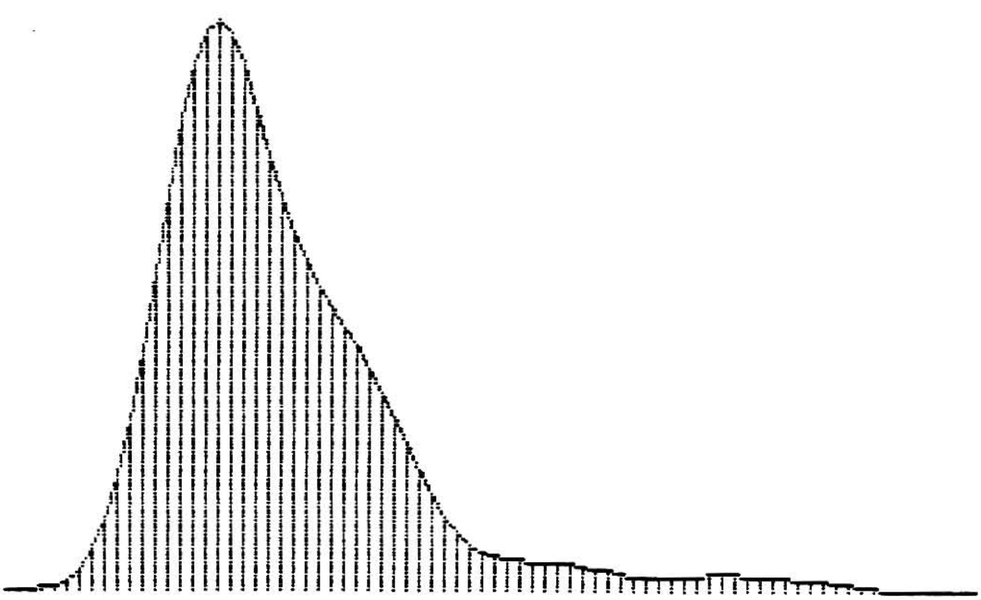
Waiting Time Data

Figure 3.2.7

	0 - 10	0 - 20	0 - 30	30 - 60
M. of M.(gamma)	.321	.570	.733	.207
Histogram	.288	.583	.742	.197
1/2 SW	.274	.553	.689	.190
SW	.221	.490	.634	.199
2 SW	.156	.369	.514	.229

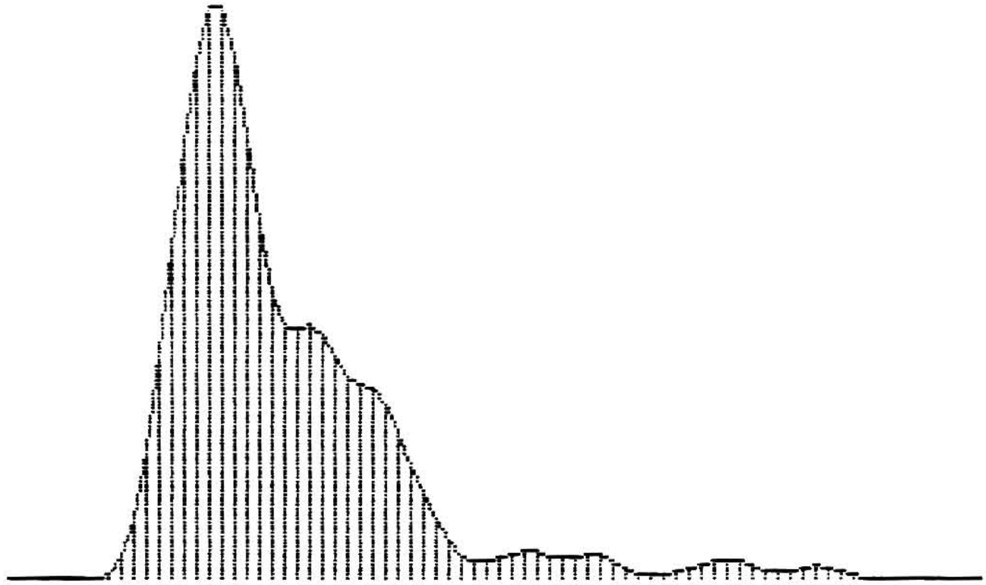
Probability Estimation for example two.

Figure 3.2.8



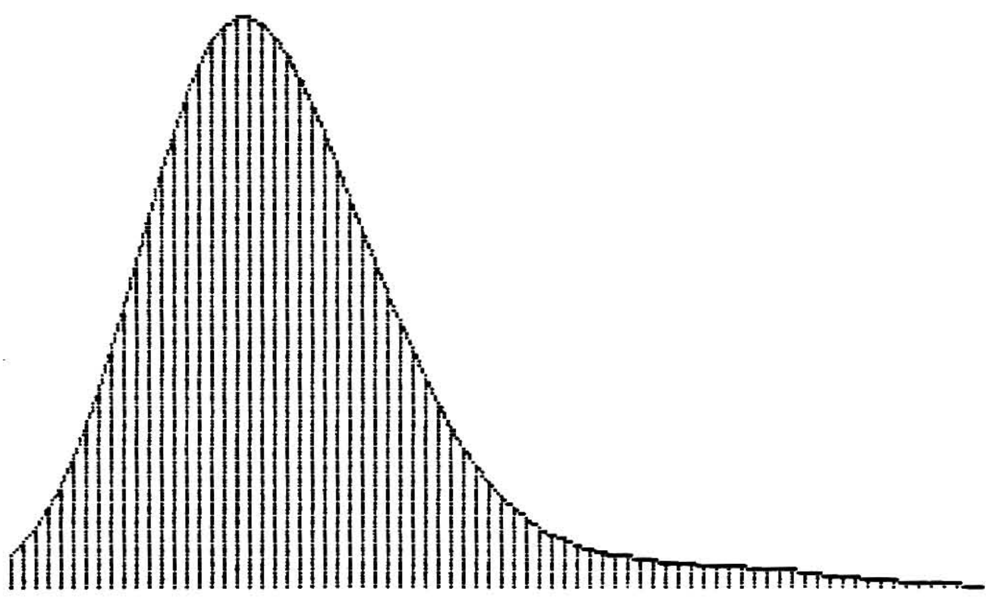
Density with suggested window size.

Figure 3.2.9



Density with one-half suggested window size.

Figure 3.2.10



Density with two times suggested window size.

Figure 3.2.11

IV ESTIMATING THE BIVARIATE DISTRIBUTION

4.1 METHOD OF KERNELS

Let (X_{1i}, X_{2i}), i=1...n be a random sample from bivariate population with continuous density $f(x_1, x_2)$. From Prakasa Rao (3), the bivariate density estimator for $f(x_1, x_2)$ is

$$\hat{f}(x_1, x_2) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} K_1\left(\frac{x_1 - X_{1i}}{h_1}\right) \sum_{i=1}^{n} K_2\left(\frac{x_2 - X_{2i}}{h_2}\right). \quad (4.1.1)$$

Different window sizes h_1 and h_2 may be used for each variate and different kernels may also be used. For our purposes $K_j(\cdot)$ will be the Normal kernel for $j = 1, 2$. The window size will be computed the same as it was in the univariate case. (See Eq. 3.1.2).

Thus, the density estimate is:

$$\hat{f}(x_1, x_2) = \frac{1}{2\pi nh_1h_2} \sum_{i=1}^{n} e^{-1/2\left(\frac{x_1 - X_{1i}}{h_1}\right)^2} \sum_{i=1}^{n} e^{-1/2\left(\frac{x_2 - x_{2i}}{h_2}\right)^2} \quad (4.1.2)$$

4.2 THE BIVARIATE DISCRIMINANT RULE

Let $\hat{f}_1$ and $\hat{f}_2$ be the estimated densities for populations 1 and 2 respectively. For discriminant analysis, an observation (x_1, x_2) is identified as being from population one if $\hat{f}_1(x_1, x_2) / \hat{f}_2(x_1, x_2) \geq 1$. Otherwise the observation is identified as being from population 2.

4.3 A COMPUTER PROGRAM TO COMPUTE THE KERNEL ESTIMATE OF A BIVARIATE DISTRIBUTION WITH APPLICATIONS TO DISCRIMINANT ANALYSIS

Using 3.1.1 and 4.1.2, $\hat{f}(x_1, x_2)$ can be computed for any bivariate data set. Computer software is available from the Department of Statistics, Kansas State University. Using an IBM compatible computer arrangement, the following steps must be taken in order to achieve results. The program is designed to operate on two samples of two variates. The program is written in GW-Basic.

The software's possiblities are graphical representation of density and discriminant analysis. To begin:

1. Load BASIC
2. Insert diskette into slot A (available from the Department of Statistics)
3. Type LOAD"A:BDKE press "Return"
4. Type RUN press "Return"
5. Proceed to type in data (for Sample 1). (Each entry followed by press "Return")
6. At the end of the data, press "Return" as instructed in #5. Then type -9999 and press "Return" twice. (The program prints the data set and the number of observations in the set).
7. Follow instructions #5 and #6 for the second sample of data (In the case where there is only one sample for the software, and the user wishes for only graphics of that

sample's density, then the user should repeat instructions #5 and #6 with several "dummy" observations. For example, four "one"s followed by -9999 and two "Returns").

8. The program suggests a window size for each variate. Then it prompts the user to choose a window size for each variate. This step is repeated for the second sample. DO NOT CHOOSE A WINDOW SIZE OF ZERO.
9. After the four window sizes have been assigned values, the user is asked whether he/she wishes to view sample one's density. If yes, continue with instruction #10. If no, type NO and push "Return", continue with instruction #14.
10. Type YES and push "Return". There will be a waiting period. Eventually "KONS = 1500" will appear. The user is asked to choose a "KONS" value. This is merely a graphics scaling factor. Choose KONS to be between 1000 and 10,000.
11. After choosing KONS, press "Return" and the density is drawn. If you wish to have a printed copy of the display's density, press "Shift" and "Prt Scr" simultaneously.
12. After viewing the density, press "Return". The user is asked if he/she wishes to choose a new KONS. If so, continue with instruction #11.
13. If the user does not wish to have a new KONS, he/she may type NO and press "Return" and be asked if he/she wishes to have new windows. If yes, the user will continue at instruction #8. (The user will have two other

opportunities to return to instruction #8, and thus new windows).

14. If the user does not wish to have new windows, he/she is asked whether he/she would like to view the second sample's density. If yes, type YES and press "Return", continue at instruction #10.
15. If the user decides not to view the second sample's density, then the user will be asked whether he/she would like to use discriminant analysis.
16. If no, the program is terminated. If yes, all observations from both samples are classified. After a period of time, the user is shown the proportion of observations from the sample data set correctly classified. The user is asked if he/she would like new windows. If yes, continue at instruction #8.
17. If the user is pleased with the current windows type NO "Return". Then, the user may type the value of each variate of an observation to be classified. This procedure may be repeated as often as the user likes. When finished, the user types NO "Return" in response to the question of whether the user has any other observations to be classified. NO "Return" completes the program.

In reference to #16, smaller window sizes yield better discriminating results for the observations from each sample. However, smaller windows also exaggerate features of the density.

(See Figure 4.2.1). One must be concerned with the reality of the density picture before he/she can classify future observations.

One example follows. Two samples of data were collected on women who had a history of physical abuse and women who did not have this history. Two variates were measured. The first variate was a measurement of the testing anxiety. The second variate was a measure of identification with negative female attributes. The data consisted of twenty-six females who were physically abused and thirty females who were not(6).(See figure 4.2.1)

For the fifty-six observations, there are graphical representations of both densities for several windows (Figure 4.2.2 - 4.2.5) and a summary of the discriminant analysis procedure performed on each observation of the sample data set (Figure 4.2.6).

Figure 4.2.5 is merely a grid because of the large window size. Most all features have been removed by the estimation procedure.

Figure 4.2.5 indicates that most sample observations are being classified to Population 2, the women who have not experienced or have not reported abuse. The reason for this is the compactness of the data for sample 2 giving larger estimates of f(x) over a smaller range as opposed to the relatively large ranges for both variates of sample one.

56	13
50	8
48	11
21	7
31	11
39	7
29	11
41	9
45	9
41	12
36	12
27	9
40	10
26	3
38	12
22	8
35	4
24	11
29	3
34	7
36	4
40	5
27	9
53	12
35	4
61	5
42	12
23	6
22	8
33	6

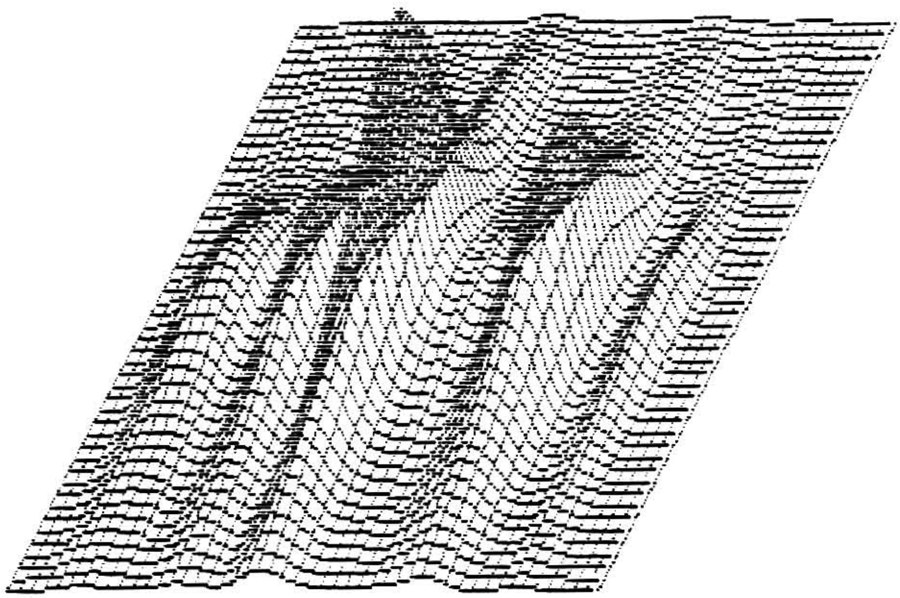
NUMBER OF OBS FROM POPULATION TWO: 30

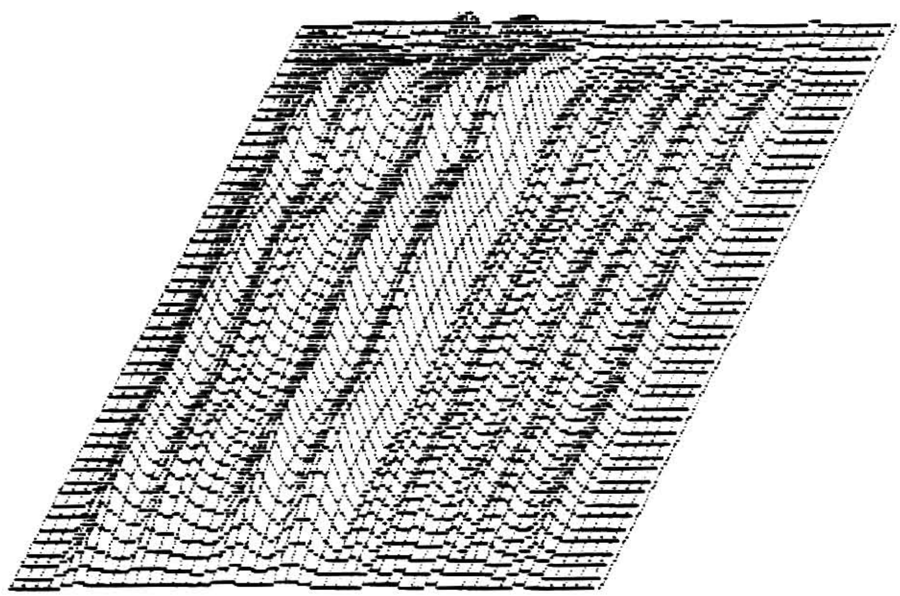
70	4
23	5
50	9
46	7
58	7
29	1
67	8
37	16
67	14
51	9
47	8
40	7
63	10
46	4
47	6
80	9
49	7
65	8
35	8
49	10
68	8
26	4
46	9
38	1
54	12
71	10

NUMBER OF OBS FROM POPULATION ONE: 26

Sociological Data

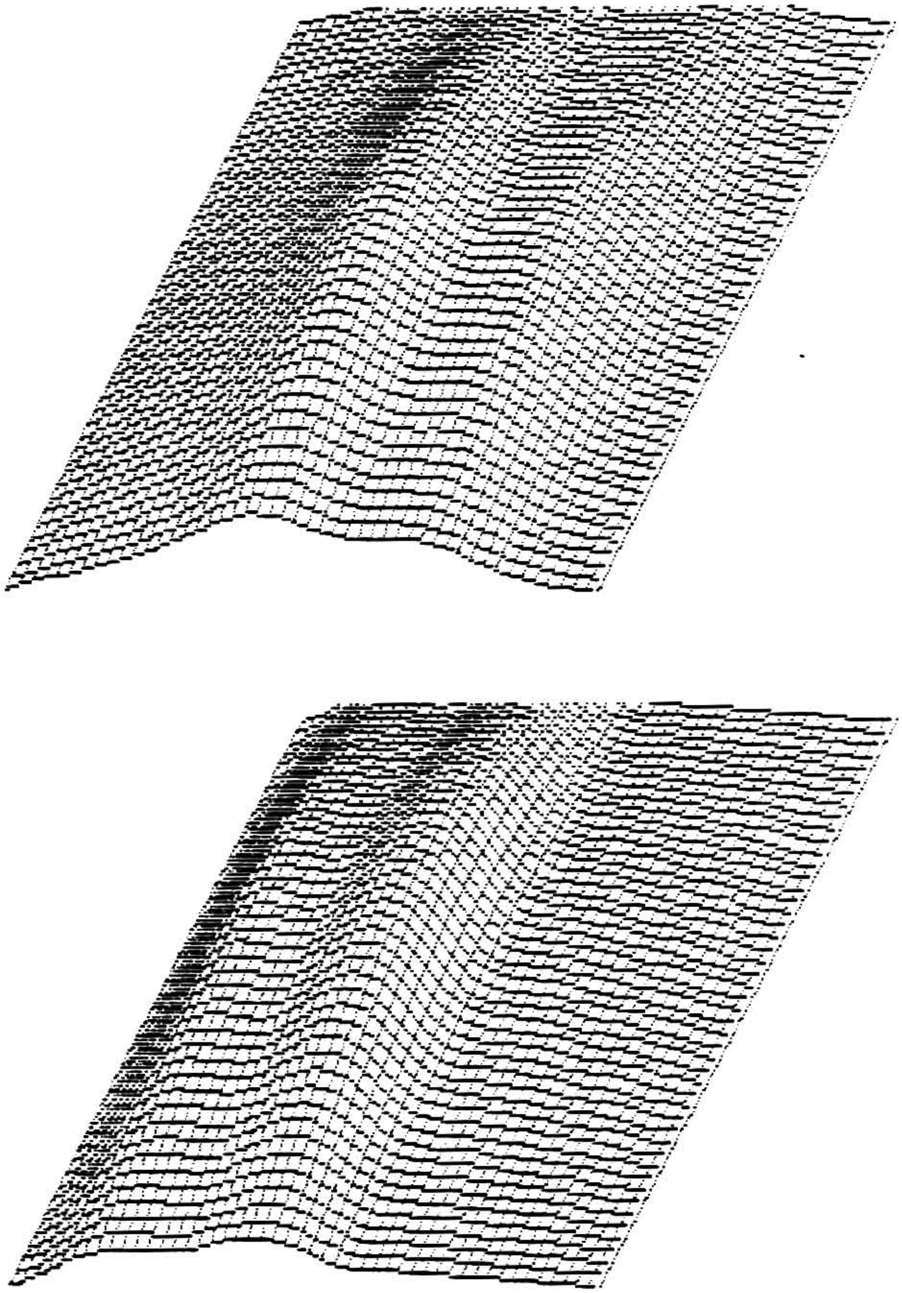
Figure 4.2.1





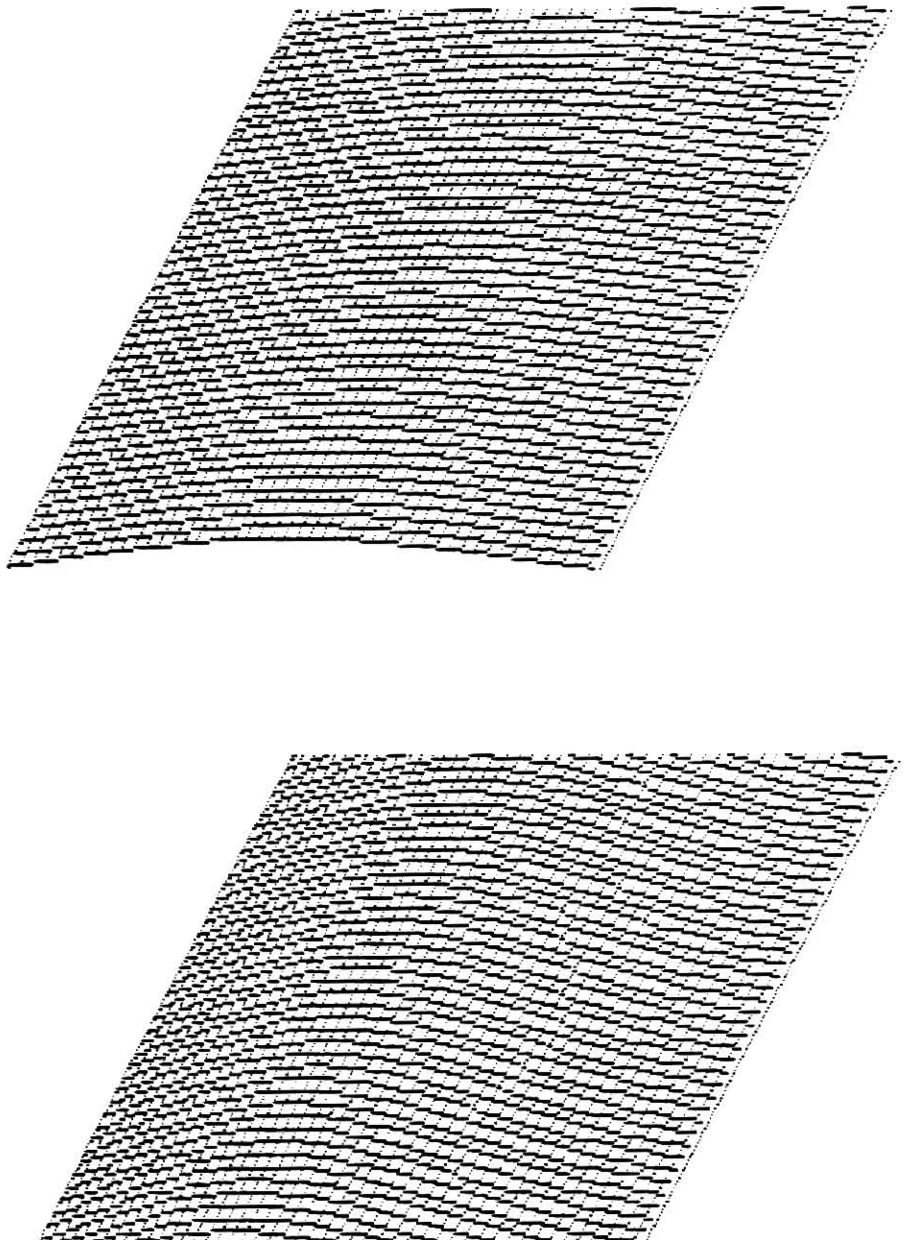
One-quarter suggested window sizes.
(Abused women, top)

Figure 4.2.2



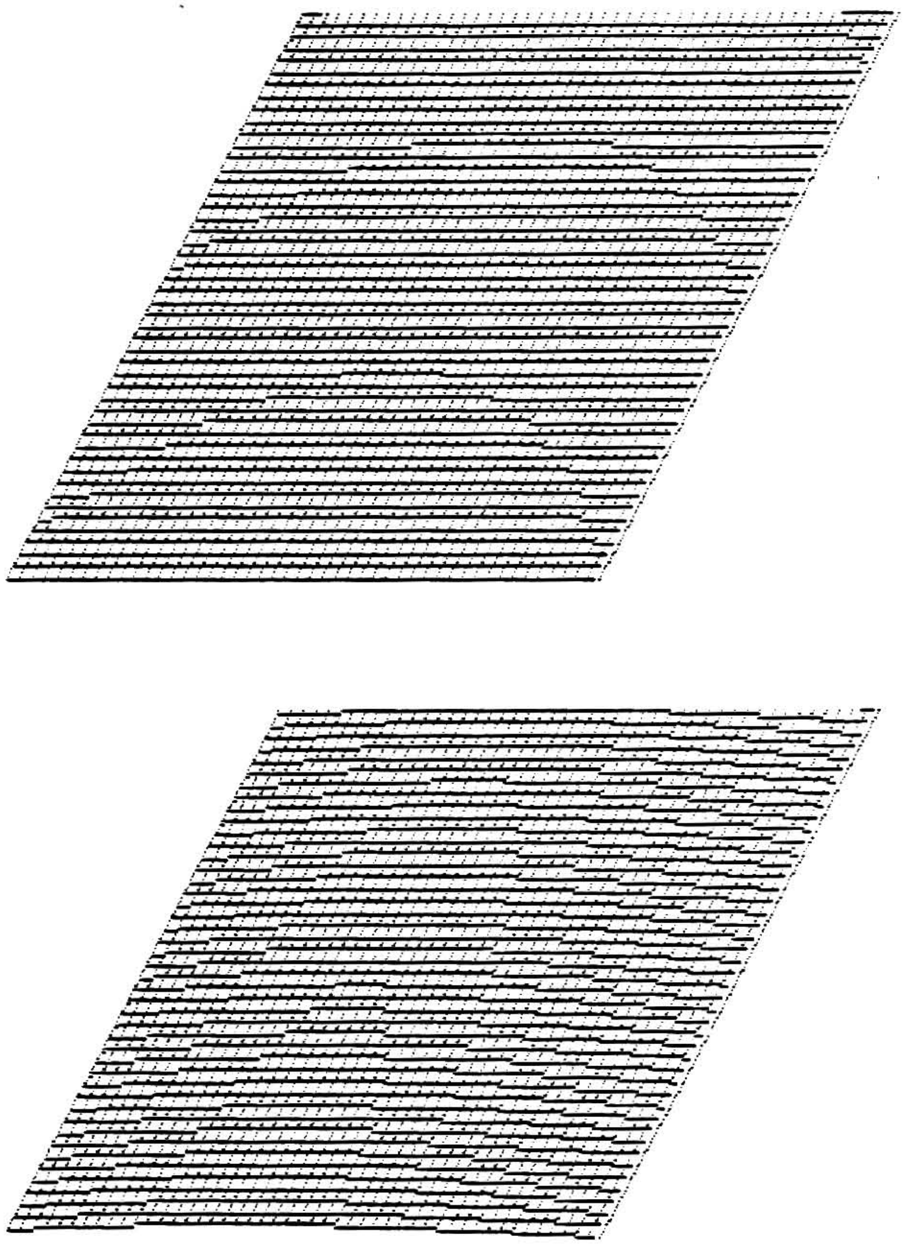
One-half suggested window sizes.
(Abused women, top)

Figure 4.2.3



Suggested window sizes.
(Abused women, top)

Figure 4.2.4



Two times suggested window sizes.
(Abused women, top)

Figure 4.2.5

	(1/4)SW	(1/2)SW	SW	(2)SW
Proportion correctly classified to Population one.	.808	.615	.269	0
Proportion correctly classified to Population two.	1	.833	1	1

Summary of Discrimination procedure.

Figure 4.2.6

V. APPENDICES

A. The Computer Program for Univariate Density Estimation

B. The Computer Program for Bivariate Density Estimation

Appendix A

```
10 DIM S(900),M(80),HT(80),HTT(80),ORD(80)
20 N=0
30 CLS
40 FOR I=1 TO 900
50 INPUT "ENTER VALUE OF DATA "; S(I)
60 N=N+1
70 IF S(I)=-9999 THEN GOTO 100
80 NEXT I
90 LPRINT "N= "N
100 FOR I=1 TO N-1
110 FOR J=1 TO N-I
120 X=S(J)
130 W=S(J+1)
140 IF X<=W THEN GOTO 170
150 S(J)=W
160 S(J+1)=X
170 NEXT J
180 NEXT I
190 FOR I= 1 TO N-1
200 S(I)=S(I+1)
210 LPRINT S(I)
220 NEXT I
230 N=N-1
240 LPRINT "SAMPLE SIZE= "N
250 C=S(N)-S(1)
260 D=C/(1+(1.4*LOG(N)))
270 DR=1
280 IF N>40 THEN DR=.75
290 IF N>75 THEN DR=.5
300 IF N>200 THEN DR=.25
310 DREV=DR*D
320 LPRINT "SUGGESTED WINDOW= "DREV
330 PRINT "SUGGESTED WINDOW= "DREV
340 INPUT "YOUR WINDOW ";DD
350 LPRINT ""
360 LPRINT "YOUR WINDOW= "DD
370 FOR J=1 TO 78
380 M(J)=S(1)-(C/4)+(J*C/52)
390 NEXT J
400 FOR J=1 TO 78
410 R=0
420 FOR I=1 TO N
430 Z=(M(J)-S(I))/DD
440 Y=EXP(-Z*Z/2)/(DD*2.51)
450 R=R+Y
460 NEXT I
470 HT(J)=R/N
480 ORD(J)=HT(J)
490 NEXT J
```

```
500 FOR I=1 TO 77
510 FOR J=1 TO 78-I
520 X=ORD(J)
530 W=ORD(J+1)
540 IF X<=W THEN GOTO 570
550 ORD(J)=W
560 ORD(J+1)=X
570 NEXT J
580 NEXT I
590 KONS=165/ORD(78)
600 SCREEN 1
610 CLS
620 FOR J=2 TO 79
630 HTT(J)=INT(170-(KONS*HT(J-1))+.5)
640 P=4*J-4
650 LINE (P,170)-(P,HTT(J))
660 NEXT J
670 FOR J=2 TO 78
680 P=4*J-4
690 Q=4*J
700 LINE(P,HTT(J))-(Q,HTT(J+1))
710 NEXT J
720 FOR J=1 TO 78
730 HT(J)=(170-HTT(J+1))/KONS
740 NEXT J
750 X$=INKEY$: IF X$="" THEN GOTO 750
760 SCREEN 2
770 CLS
780 INPUT "INTEGRAL LOWER LIMIT, U ";U
790 INPUT "INTEGRAL UPPER LIMIT, V ";V
800 PRINT "LOWER LIMIT= "U
810 PRINT "UPPER LIMIT= "V
820 UT=INT(((U-S(1))*52/C)+14.5)
830 VT=INT(((V-S(1))*52/C)+14.5)
840 F=HT(UT)+HT(VT)
850 FOR J=UT+1 TO VT-1
860 E=2*HT(J)
870 F=E+F
880 NEXT J
890 F=F*C/102
900 PRINT "ESTIMATED U<=PROBABILITY<=V = "F
910 INPUT "OTHER INTEGRAL LIMITS? YES OR NO ";D$
920 IF D$="YES" THEN GOTO 780
930 INPUT "TRY OTHER WINDOWS? YES OR NO ";E$
940 IF E$="YES" THEN GOTO 320
950 SAVE "UDKE
```

```
10 DIM S1(500),S2(500),SS1(500),SS2(500),MS(48),MT(48),KF(48)
20 DIM T1(500),T2(500),TT1(500),TT2(500),HSS(48),HTT(48)
30 N1=0
40 N2=0
50 FOR I=1 TO 500
60 INPUT "ENTER FIRST VALUE OF OBS OF POPULATION ONE, X1";S1(I)
70 INPUT "ENTER SECOND VALUE OF OBS OF POPULATION ONE, X2";T1(I)
80 N1=N1+1
90 IF S1(I)=-9999 THEN GOTO 110
100 NEXT I
110 FOR I=1 TO N1-1
120 LPRINT S1(I),T1(I)
130 NEXT I
140 N1=N1-1
150 LPRINT "NUMBER OF OBS FROM POPULATION ONE: "N1
160 FOR I=1 TO 500
170 INPUT "ENTER FIRST VALUE OF OBS OF POPULATION TWO, Y1";S2(I)
180 INPUT "ENTER SECOND VALUE OF OBS OF POPULATION TWO, Y2";T2(I)
190 N2=N2+1
200 IF S2(I)=-9999 THEN GOTO 220
210 NEXT I
220 FOR I=1 TO N2-1
230 LPRINT S2(I),T2(I)
240 NEXT I
250 N2=N2-1
260 LPRINT "NUMBER OF OBS FROM POPULATION TWO: "N2
270 FOR I=1 TO N1
280 FOR J=1 TO N1+1-I
290 X=S1(J)
300 W=S1(J+1)
310 IF X<=W THEN GOTO 340
320 S1(J)=W
330 S1(J+1)=X
340 NEXT J
350 NEXT I
360 FOR I=1 TO N1
370 FOR J=1 TO N1+1-I
380 X=T1(J)
390 W=T1(J+1)
400 IF X<=W THEN GOTO 430
410 T1(J)=W
420 T1(J+1)=X
430 NEXT J
440 NEXT I
450 FOR I=1 TO N2
460 FOR J=1 TO N2+1-I
470 X=S2(J)
480 W=S2(J+1)
490 IF X<=W THEN GOTO 520
```

```
500 S2(J)=W
510 S2(J+1)=X
520 NEXT J
530 NEXT I
540 FOR I=1 TO N2
550 FOR J=1 TO N2+1-I
560 X=T2(J)
570 W=T2(J+1)
580 IF X<=W THEN GOTO 610
590 T2(J)=W
600 T2(J+1)=X
610 NEXT J
620 NEXT I
630 FOR I=1 TO N1
640 S1(I)=S1(I+1)
650 T1(I)=T1(I+1)
660 NEXT I
670 FOR I=1 TO N2
680 S2(I)=S2(I+1)
690 T2(I)=T2(I+1)
700 NEXT I
710 FOR I= 1 TO N1
720 SS1(I)=S1(I)
730 TT1(I)=T1(I)
740 NEXT I
750 FOR I= 1 TO N2
760 SS2(I)=S2(I)
770 TT2(I)=T2(I)
780 NEXT I
790 R1ONE=S1(N1)-S1(1)
800 R2ONE=T1(N1)-T1(1)
810 R1TWO=S2(N2)-S2(1)
820 R2TWO=T2(N2)-T2(1)
830 D1=(1.4*LOG(N1))+1
840 REV=1
850 IF N1>40 THEN REV=.75
860 IF N1>75 THEN REV=.5
870 IF N1>200 THEN REV=.25
880 DSONE=REV*R1ONE/D1
890 DTONE=REV*R2ONE/D1
900 LPRINT "SUGGESTED WINDOW SIZE FOR VARIATE ONE OF POPULATION ON
910 PRINT "SUGGESTED WINDOW SIZE FOR VARIATE ONE OF POPULATION ONE
920 LPRINT "SUGGESTED WINDOW SIZE FOR VARIATE TWO OF POPULATION ON
930 PRINT "SUGGESTED WINDOW SIZE FOR VARIATE TWO OF POPULATION ONE
940 INPUT "YOUR WINDOW SIZE FOR VARIATE ONE OF POPULATION ONE";DS1
950 LPRINT "YOUR WINDOW SIZE FOR VARIATE ONE OF POPULATION ONE"DS1
960 INPUT "YOUR WINDOW SIZE FOR VARIATE TWO OF POPULATION ONE";DT1
970 LPRINT "YOUR WINDOW SIZE FOR VARIATE TWO OF POPULATION ONE"DT1
980 D2=(1.4*LOG(N2))+1
990 REV=1
1000 IF N2>40 THEN REV=.75
1010 IF N2>75 THEN REV=.5
1020 IF N2>200 THEN REV=.25
1030 DSTWO=REV*R1TWO/D2
```

```
1040 DTTWO=REV*R2TWO/D2
1050 LPRINT "SUGGESTED WINDOW SIZE FOR VARIATE ONE OF POPULATION T
1060 PRINT "SUGGESTED WINDOW SIZE FOR VARIATE ONE OF POPULATION TW
1070 LPRINT "SUGGESTED WINDOW SIZE FOR VARIATE TWO OF POPULATION T
1080 PRINT "SUGGESTED WINDOW SIZE FOR VARIATE TWO OF POPULATION TW
1090 INPUT "YOUR WINDOW SIZE FOR VARIATE ONE OF POPULATION TWO";DS
1100 LPRINT "YOUR WINDOW SIZE FOR VARIATE ONE OF POPULATION TWO"DS
1110 INPUT "YOUR WINDOW SIZE FOR VARIATE TWO OF POPULATION TWO";DT
1120 LPRINT "YOUR WINDOW SIZE FOR VARIATE TWO OF POPULATION TWO"DT
1130 INPUT "DO YOU WISH TO VIEW SAMPLE ONE'S DENSITY? YES OR NO";A
1140 IF AAAA$="YES" THEN GOTO 3450
1150 INPUT "DO YOU WISH TO VIEW SAMPLE TWO'S DENSITY? YES OR NO";B
1160 IF BBBB$="YES" THEN GOTO 4540
1170 INPUT "DO YOU WISH TO USE DISCRIMINANT ANALYSIS? YES OR NO";C
1180 IF CCCC$="YES" THEN GOTO 2380
1190 GOTO 5430
1200 INPUT "ENTER VALUE OF VARIABLE ONE OF OBSERVATION";S
1210 INPUT "ENTER VALUE OF VARIABLE TWO OF OBSERVATION";T
1220 LPRINT "VARIABLE ONE= "S
1230 LPRINT "VARIABLE TWO= "T
1240 RS=0
1250 FOR I=1 TO N1
1260 ZS=(S-S1(I))/DD1
1270 YS=EXP(-ZS*ZS/2)/(DD1*2.51)
1280 RS=RS+YS
1290 NEXT I
1300 HHS=RS/N1
1310 RT=0
1320 FOR I=1 TO N1
1330 ZT=(T-T1(I))/DD1
1340 YT=EXP(-ZT*ZT/2)/(DD1*2.51)
1350 RT=RT+YT
1360 NEXT I
1370 HHT=RT/N1
1380 ORD1=HHS*HHT
1390 RS=0
1400 FOR I=1 TO N2
1410 ZS=(S-S2(I))/DD2
1420 YS=EXP(-ZS*ZS/2)/(DD2*2.51)
1430 RS=RS+YS
1440 NEXT I
1450 HHS=RS/N2
1460 RT=0
1470 FOR I=1 TO N2
1480 ZT=(T-T2(I))/DD2
1490 YT=EXP(-ZT*ZT/2)/(DD2*2.51)
1500 RT=RT+YT
1510 NEXT I
1520 HHT=RT/N2
1530 ORD2=HHS*HHT
1540 ORD=ORD1+ORD2
1550 PROP1=ORD1/ORD
1560 PROP2=ORD2/ORD
```

```
1570 PRINT "LIKELIHOOD OF OBS FROM POP ONE: "PROP1
1580 LPRINT "LIKELIHOOD OF OBSERVATION FROM POPULATION ONE: "PROP1
1590 LPRINT "LIKELIHOOD OF OBSERVATION FROM POPULATION TWO: "PROP2
1600 PRINT "LIKELIHOOD OF OBS FROM POP TWO: "PROP2
1610 INPUT "LIKE TO TRY OTHER OBSERVATIONS WITH THESE WINDOWS ? YE
1620 IF Z$="YES" THEN GOTO 1200
1630 INPUT "LIKE TO TRY OTHER WINDOWS ? YES OR NO ";Y$
1640 IF Y$="YES" THEN GOTO 830
1650 GOTO 5430
1660 FOR I=1 TO N1
1670 SS1(I)=S1(I)
1680 TT1(I)=T1(I)
1690 NEXT I
1700 FOR I=1 TO N2
1710 SS2(I)=S2(I)
1720 TT2(I)=T2(I)
1730 NEXT I
1740 GOTO 2380
1750 LPRINT "NUMBER OF OBS FROM POPULATION TWO: "N2
1760 FOR I=1 TO N1
1770 FOR J=1 TO N1+1-I
1780 X=S1(J)
1790 W=S1(J+1)
1800 IF X<=W THEN GOTO 1830
1810 S1(J)=W
1820 S1(J+1)=X
1830 NEXT J
1840 NEXT I
1850 FOR I=1 TO N1
1860 FOR J=1 TO N1+1-I
1870 X=T1(J)
1880 W=T1(J+1)
1890 IF X<=W THEN GOTO 1920
1900 T1(J)=W
1910 T1(J+1)=X
1920 NEXT J
1930 NEXT I
1940 FOR I=1 TO N2
1950 FOR J=1 TO N2+1-I
1960 X=S2(J)
1970 W=S2(J+1)
1980 IF X<=W THEN GOTO 2010
1990 S2(J)=W
2000 S2(J+1)=X
2010 NEXT J
2020 NEXT I
2030 FOR I=1 TO N2
2040 FOR J=1 TO N2+1-I
2050 X=T2(J)
2060 W=T2(J+1)
2070 IF X<=W THEN GOTO 2100
2080 T2(J)=W
2090 T2(J+1)=X
2100 NEXT J
2110 NEXT I
```

```
2120 FOR I=1 TO N1
2130 S1(I)=S1(I+1)
2140 T1(I)=T1(I+1)
2150 NEXT I
2160 FOR I=1 TO N2
2170 S2(I)=S2(I+1)
2180 T2(I)=T2(I+1)
2190 NEXT I
2200 R1ONE=S1(N1)-S1(1)
2210 R2ONE=T1(N1)-T1(1)
2220 D1=(5.6*LOG(N1))+4
2230 DONE=R1ONE*R2ONE/(D1*D1)
2240 D1REV=((N1+100)/(N1+2))*DONE
2250 LPRINT "SUGGESTED WINDOW FOR SAMPLE ONE= "D1REV
2260 PRINT "SUGGESTED WINDOW FOR SAMPLE ONE= "D1REV
2270 INPUT "YOUR WINDOW";DD1
2280 LPRINT "YOUR WINDOW= "DD1
2290 R1TWO=S2(N2)-S2(1)
2300 R2TWO=T2(N2)-T2(1)
2310 D2=(5.6*LOG(N2))+4
2320 DTWO=R1TWO*R2TWO/(D2*D2)
2330 D2REV=((N2+100)/(N2+2))*DTWO
2340 LPRINT "SUGGESTED WINDOW FOR SAMPLE TWO= "D2REV
2350 PRINT "SUGGESTED WINDOW FOR SAMPLE TWO= "D2REV
2360 INPUT "YOUR WINDOW";DD2
2370 LPRINT "YOUR WINDOW= "DD2
2380 CTR1=0
2390 CTR2=0
2400 DD1=DS1*DT1
2410 DD2=DS2*DT2
2420 FOR K= 1 TO N1+N2
2430 IF K>N1 GOTO 2480
2440 HGH=1
2450 S=SS1(K)
2460 T=TT1(K)
2470 GOTO 2520
2480 L=K-N1
2490 HGH=2
2500 S=SS2(L)
2510 T=TT2(L)
2520 RS=0
2530 FOR I=1 TO N1
2540 ZS=(S-S1(I))/DD1
2550 YS=EXP(-ZS*ZS/2)/(DD1*2.51)
2560 RS=RS+YS
2570 NEXT I
2580 HTS=RS/N1
2590 RT=0
2600 FOR I=1 TO N1
2610 ZT=(T-T1(I))/DD1
2620 YT=EXP(-ZT*ZT/2)/(DD1*2.51)
2630 RT=RT+YT
2640 NEXT I
2650 HTT=RT/N1
```

```
2660 ORD1=HTS*HTT
2670 RS=0
2680 FOR I=1 TO N2
2690 ZS=(S-S2(I))/DD2
2700 YS=EXP(-ZS*ZS/2)/(DD2*2.51)
2710 RS=RS+YS
2720 NEXT I
2730 HTS=RS/N2
2740 RT=0
2750 FOR I=1 TO N2
2760 ZT=(T-T2(I))/DD2
2770 YT=EXP(-ZT*ZT/2)/(DD2*2.51)
2780 RT=RT+YT
2790 NEXT I
2800 HTT=RT/N2
2810 ORD2=HTS*HTT
2820 ORD=ORD1+ORD2
2830 PROP1=ORD1/ORD
2840 PROP2=ORD2/ORD
2850 HG=2
2860 IF PROP1>.5 THEN HG=1
2870 IF HG>HGH THEN GOTO 2910
2880 IF HG<HGH THEN GOTO 2910
2890 IF HGH=2 THEN CTR2=1+CTR2
2900 IF HGH=1 THEN CTR1=1+CTR1
2910 NEXT K
2920 PERC1=CTR1/N1
2930 PERC2=CTR2/N2
2940 LPRINT "PROPORTION CORRECTLY DISCRIMINATED TO POPULATION ONE
2950 PRINT "PROPORTION CORRECTLY DISCRIMINATED TO POPULATION ONE"
2960 LPRINT "PROPORTION CORRECTLY DISCRIMINATED TO POPULATION TWO
2970 PRINT "PROPORTION CORRECTLY DISCRIMINATED TO POPULATION TWO"
2980 INPUT "LIKE TO TRY OTHER WINDOWS ? YES OR NO ";Y$
2990 IF Y$="YES" THEN GOTO 830
3000 GOTO 1200
3010 GOTO 3450
3020 N=0
3030 CLS
3040 FOR I=1 TO 300
3050 INPUT "VALUE OF X1 OF OBSERVATION ";SS(I)
3060 INPUT "VALUE OF X2 OF OBSERVATION ";TT(I)
3070 N=N+1
3080 IF SS(I)= -9999 THEN GOTO 3100
3090 NEXT I
3100 FOR I=1 TO N-1
3110 LPRINT SS(I) ,TT(I)
3120 NEXT I
3130 FOR I=1 TO N-1
3140 FOR J=1 TO N-I
3150 XS=SS(J)
3160 WS=SS(J+1)
3170 IF XS<=WS THEN GOTO 3200
3180 SS(J)=WS
3190 SS(J+1)=XS
3200 NEXT I
```

```
3210 FOR I=1 TO N-1
3220 FOR J=1 TO N-I
3230 XT=TT(J)
3240 WT=TT(J+1)
3250 IF XT<=WT THEN GOTO 3280
3260 TT(J)=WT
3270 TT(J+1)=XT
3280 NEXT J
3290 NEXT I
3300 FOR I=1 TO N-1
3310 SS(I)=SS(I+1)
3320 TT(I)=TT(I+1)
3330 NEXT I
3340 N=N-1
3350 LPRINT "NUMBER OF OBSERVATIONS= "N
3360 CS=SS(N)-SS(1)
3370 CT=TT(N)-TT(1)
3380 DST=(5.6*LOG(N))+4
3390 D=CS*CT/(DST*DST)
3400 DREV=((N+100)/(N+2))*D
3410 PRINT "SUGGESTED WINDOW= "DREV
3420 LPRINT "SUGGESTED WINDOW= "DREV
3430 INPUT "YOUR WINDOW";DD
3440 LPRINT "YOUR WINDOW= "DD
3450 CS=R1ONE
3460 CT=R2ONE
3470 DD=DS1*DT1
3480 FOR J=1 TO 48
3490 MS(J)=SS1(1)-(CS/6)+(J*CS/36)
3500 MT(J)=TT1(1)-(CT/6)+(J*CT/36)
3510 NEXT J
3520 FOR J=1 TO 48
3530 RS=0
3540 FOR I=1 TO N1
3550 ZS=(MS(J)-SS1(I))/DD
3560 YS=EXP(-ZS*ZS/2)/(DD*2.51)
3570 RS=RS+YS
3580 NEXT I
3590 HSS(J)=RS/N1
3600 NEXT J
3610 FOR J=1 TO 48
3620 RT=0
3630 FOR I=1 TO N1
3640 ZT=(MT(J)-TT1(I))/DD
3650 YT=EXP(-ZT*ZT/2)/(DD*2.51)
3660 RT=RT+YT
3670 NEXT I
3680 HTT(J)=RT/N1
3690 NEXT J
3700 KONS=1500
3710 PRINT "SUGGESTED SCALING FACTOR= "KONS
3720 INPUT "YOUR SCALING FACTOR ";KONS
3730 CLS
```

```
3740 SCREEN 2
3750 FOR J=1 TO 48
3760 P=8*J+12
3770 FOR I=1 TO 48
3780 KF(I)=INT(170-(KONS*HSS(J)*HTT(I))-(3.476*(I-1)+.5))
3790 NEXT I
3800 KP=P+4
3810 LINE(P,KF(1))-(KP,KF(2))
3820 FOR I=2 TO 47
3830 KP=P+(4*I)
3840 LINE-(KP,KF(I+1))
3850 NEXT I
3860 NEXT J
3870 FOR J=1 TO 48
3880 FOR I=1 TO 48
3890 KF(I)=INT(170-(KONS*HSS(I)*HTT(J))-(3.476*(J-1)+.5))
3900 NEXT I
3910 P=20+(4*J)
3920 LINE(P-4,KF(1))-(P,KF(2))
3930 FOR I=3 TO 48
3940 P=P+8
3950 LINE-(P,KF(I))
3960 NEXT I
3970 NEXT J
3980 X$=INKEY$: IF X$="" THEN GOTO 3980
3990 CLS
4000 PRINT "PREVIOUS SCALING FACTOR: "KONS
4010 INPUT "TRY NEW SCALING FACTOR? YES OR NO ";B$
4020 IF B$="YES" THEN GOTO 3710
4030 INPUT "TRY OTHER WINDOWS? YES OR NO ";A$
4040 IF A$="YES" THEN GOTO 830
4050 GOTO 1150
4060 INPUT "DO YOU WISH TO ESTIMATE PROBABILITIES? YES OR NO";DDD
4070 MGH=1
4080 IF DDDD$="YES" THEN GOTO 5190
4090 GOTO 1150
4100 DIM SS(300), TT(300), MS(48), MT(48), HSS(48), HTT(48), KF(4
4110 N=0
4120 CLS
4130 FOR I=1 TO 300
4140 INPUT "VALUE OF X1 OF OBSERVATION ";SS(I)
4150 INPUT "VALUE OF X2 OF OBSERVATION ";TT(I)
4160 N=N+1
4170 IF SS(I)= -9999 THEN GOTO 4190
4180 NEXT I
4190 FOR I=1 TO N-1
4200 LPRINT SS(I) ,TT(I)
4210 NEXT I
4220 FOR I=1 TO N-1
4230 FOR J=1 TO N-I
4240 XS=SS(J)
4250 WS=SS(J+1)
4260 IF XS<=WS THEN GOTO 4290
```

```
4270 SS(J)=WS
4280 SS(J+1)=XS
4290 NEXT I
4300 FOR I=1 TO N-1
4310 FOR J=1 TO N-I
4320 XT=TT(J)
4330 WT=TT(J+1)
4340 IF XT<=WT THEN GOTO 4370
4350 TT(J)=WT
4360 TT(J+1)=XT
4370 NEXT J
4380 NEXT I
4390 FOR I=1 TO N-1
4400 SS(I)=SS(I+1)
4410 TT(I)=TT(I+1)
4420 NEXT I
4430 N=N-1
4440 LPRINT "NUMBER OF OBSERVATIONS= "N
4450 CS=SS(N)-SS(1)
4460 CT=TT(N)-TT(1)
4470 DST=(5.6*LOG(N))+4
4480 D=CS*CT/(DST*DST)
4490 DREV=((N+100)/(N+2))*D
4500 PRINT "SUGGESTED WINDOW= "DREV
4510 LPRINT "SUGGESTED WINDOW= "DREV
4520 INPUT "YOUR WINDOW";DD
4530 LPRINT "YOUR WINDOW= "DD
4540 CS=R1TWO
4550 CT=R2TWO
4560 DD=DS2*DT2
4570 FOR J=1 TO 48
4580 MS(J)=SS2(1)-(CS/6)+(J*CS/36)
4590 MT(J)=TT2(1)-(CT/6)+(J*CT/36)
4600 NEXT J
4610 FOR J=1 TO 48
4620 RS=0
4630 FOR I=1 TO N2
4640 ZS=(MS(J)-SS2(I))/DD
4650 YS=EXP(-ZS*ZS/2)/(DD*2.51)
4660 RS=RS+YS
4670 NEXT I
4680 HSS(J)=RS/N2
4690 NEXT J
4700 FOR J=1 TO 48
4710 RT=0
4720 FOR I=1 TO N2
4730 ZT=(MT(J)-TT2(I))/DD
4740 YT=EXP(-ZT*ZT/2)/(DD*2.51)
4750 RT=RT+YT
4760 NEXT I
4770 HTT(J)=RT/N2
4780 NEXT J
4790 KONS=1500
4800 PRINT "SUGGESTED SCALING FACTOR= "KONS
4810 INPUT "YOUR SCALING FACTOR ";KONS
4820 CLS
4830 SCREEN 2
```

```
4840 FOR J=1 TO 48
4850 P=8*J+12
4860 FOR I=1 TO 48
4870 KF(I)=INT(170-(KONS*HSS(J)*HTT(I))-(3.476*(I-1)+.5))
4880 NEXT I
4890 KP=P+4
4900 LINE(P,KF(1))-(KP,KF(2))
4910 FOR I=2 TO 47
4920 KP=P+(4*I)
4930 LINE-(KP,KF(I+1))
4940 NEXT I
4950 NEXT J
4960 FOR J=1 TO 48
4970 FOR I=1 TO 48
4980 KF(I)=INT(170-(KONS*HSS(I)*HTT(J))-(3.476*(J-1)+.5))
4990 NEXT I
5000 P=20+(4*J)
5010 LINE(P-4,KF(1))-(P,KF(2))
5020 FOR I=3 TO 48
5030 P=P+8
5040 LINE-(P,KF(I))
5050 NEXT I
5060 NEXT J
5070 X$=INKEY$: IF X$="" THEN GOTO 5070
5080 CLS
5090 PRINT "PREVIOUS SCALING FACTOR: "KONS
5100 INPUT "TRY NEW SCALING FACTOR? YES OR NO ";B$
5110 IF B$="YES" THEN GOTO 4800
5120 INPUT "TRY OTHER WINDOWS? YES OR NO ";A$
5130 IF A$="YES" THEN GOTO 830
5140 GOTO 1170
5150 INPUT "DO YOU WISH TO ESTIMATE PROBABILITIES? YES OR NO";EEE
5160 MGH=2
5170 IF EEEE$="YES" THEN GOTO 5190
5180 GOTO 1170
5190 INPUT "LOWER INTEGRAL LIMIT OF VARIATE ONE";AA
5200 INPUT "UPPER INTEGRAL LIMIT OF VARIATE ONE";BB
5210 INPUT "LOWER INTEGRAL LIMIT OF VARIATE TWO";CC
5220 INPUT "UPPER INTEGRAL LIMIT OF VARIATE TWO";DD
5230 LPRINT "LIMITS FOR VARIATE ONE"AA,BB
5240 LPRINT "LIMITS FOR VARIATE TWO"CC,DD
5250 FF=0
5260 AAS=INT(((AA-SS1(1))*36/CS)+6.5)
5270 BBS=INT(((BB-SS1(1))*36/CS)+6.5)
5280 CCT=INT(((CC-TT1(1))*36/CT)+6.5)
5290 DDT=INT(((DD-TT1(1))*36/CT)+6.5)
5300 FOR J=CCT TO DDT
5310 EE=0
5320 FOR I=AAS TO BBS
5330 EE=EE+(HSS(I)*HTT(J))
5340 NEXT I
5350 FF=FF+EE
5360 NEXT J
5370 LPRINT "PROBABILITY OF REGION: "FF
5380 PRINT "PROBABILITY OF REGION: "FF
5390 INPUT "DO YOU WISH TO COMPUTE OTHER PROBABILITIES? YES OR NO
5400 IF FFFF$="YES" THEN GOTO 5190
5410 IF MGH=2 THEN GOTO 1170
5420 IF MGH=1 THEN GOTO 1150
5430 SAVE"BDKE
```

VI REFERENCES

(1) Tapia, R. A. and Thompson, J.R. (1978), Nonparametric Density Estimation, The Johns Hopkins University Press, Baltimore and London.

(2) Parzen, E. (1962), "On Estimation of a Probability Density Function and Mode", Annals of Mathematical Statistics, 33:1065-1076.

(3) Prakasa Rao, B.L.S. (1983), Nonparametric Functional Estimation, Academic Press, Orlando, San Diego, San Francisco, New York, London, Toronto, Montreal, Sydney, Tokyo and Sao Paulo.

(4) Huntsberger, D. V. (1967), Elements of Statistical Inference, Allyn and Bacon, Boston.

(5) McClave, J. T. and Benson, P. G. (1985), Statistics for Business and Economics, Dellen Publishing Company, San Fransisco.

(6) Davis, Al (1985), College of Human Ecology, Kansas State University, Manhattan, Kansas.

NONPARAMETRIC DENSITY ESTIMATION FOR UNIVARIATE AND BIVARIATE DISTRIBUTIONS WITH APPLICATIONS IN DISCRIMINANT ANALYSIS FOR THE BIVARIATE CASE

by

MARK HAUG

B.S., Kansas State University, 1984

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1986

ABSTRACT

An alternative to the histogram density estimate has been proposed in the literature. The alternative is the kernel type nonparametric density estimate. The kernel type estimator requires the same information that the histogram estimator requires to construct the estimate. However, the kernel type estimator produces a continuous estimated density.