BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-87/31

FINAL REPORT ON BIPS GRAPHIC SUPPORT

by

Robert T. O'Reagan
Statistical Research Division
Bureau of the Census
Washington, D.C. 20233

Recommended by: Lawrence R. Ernst

Report completed: October 6, 1987

Report issued: October 6, 1987

FINAL REPORT ON BIPS GRAPHIC SUPPORT: SRD PROJECT #86-8

Background

BIPS (for Business Interactive Processing System) Graphic Support was originally called CAPS II Graphic Support. CAPS II stood for Census Automated Processing Systems, the II signifying second generation.

The earlier CAPS I had been the development and support of a set of files, programs, and network configurations for the Standard Statistical Establishment Listing (SSEL), in effect a master file of establishments in the economic areas including company affiliation, administrative data, and historical data, with capabilities for retrieving and updating this information.

The second generation or CAPS II effort -- with participation from Agriculture, Business, Construction, and Industry Divisions -- concerned itself with analytical tabulations of Economic Census data. The prototype systems for 1987 were being developed by the interdivisional Analytical Tabulation Review and Edit Referral Committee, initially chaired by Al Barten and later by Mark Wallace, both of Business Division. The ultimate configuration was to be used to produce an electronic listing of the 1987 Census tabulations, following the edit referral phase for individual records. These electronic listings aimed to eliminate the voluminous output of paper formerly required for analyst review, and to replace them with listings directly on display screens. From 600 microcomputers the analysts were to transmit corrections to hard disk for main frame processing, thus short circuiting a difficult and very time consuming cycle necessitated by the former paper corrections, keying, recycling, and the like. Much of the attention in 1985 was directed to record length, file size, screen content, up and downloading, and other feasibility factors.

Statistical Research Division proposed graphics as an aid for dynamic analysis of the tabulations. By early 1986, it became apparent that the different subject matter Divisions had divergent needs in terms of both hardware and software, so SRD focused on the Business Division application, or as it came to be called, BIPS.

Core of the BIPS graphics project

The goal remained to construct a paperless tabulation review. Working closely with Business Division analysts, SRD's graphics programmers familiarized themselves with the overall project development and created graphic outlier identification routines which are the basis of a graphics review system.

This new system, called PLOTLIER, uses a menu of chart choices that will plot the ratios of '87 to '82 data with the mean and two standard deviation limits shown for:

Establishments by KB
Sales by KB
Establishments by county
Sales by county

Examples of each chart format are included in attachments A through C. Demonstrations of the live system can be provided on request.

This system interactively prompts the user through a selection of menus to choose a particular chart type, whether for establishments or sales, by Kind of Business or county. After displaying a chart on the screen, the operator can elect to see a frequency plot of the same data, a table of the values that went into that chart or return to the higher level menu. The outliers are very noticeable in this chart layout since not only do they fall outside the standard deviation limits shown on the scatterplot but they are also flagged with the ID of the unit (county or KB code). While pinpointing of individual outlier records enables retrieval for detailed review, the bar-chart presentation permits easy identification of anomalous situations, bi-modal distributions, or suspicious skewness.

Though PLOTLIER was originally written for use on local micro-computers, a version was also created which permits portability to any mini-computer that uses standard FORTRAN.

Other Important Contributions

By the summer of 1986 two additional tasks had been identified. First, Mark Wallace requested SRD's assistance in providing some graphical insight into the appropriateness of symmetrical two sigma limits for outlier definition, that is, the assumption of a normal (or Gaussian) bell-shaped distribution. Second, he wanted some exploratory data analysis techniques applied to current-to-current ratios, such as sales to payroll, in addition to the current-to-prior period tests within a single field.

Since Texas had many observations (i.e. counties), SRD requested that state file for experimentation. Business Division provided it. New Jersey was also used as a subject for graphic exploration. The basic data files utilized correspond to Table 8 of the 1982 Census of Retail Trade publications, RC82-A-44 and RC82-A-31 respectively, as well as the corresponding 1977 data in some instances. These were therefore published or "clean" files by county and by KB (kind of business).

In addition to tabular summaries of statistical measures, such as attachments D1 and D2 which compute location measures (mean, median, mid-range) and dispersion measures (range, standard deviation, minimum and maximum, lower and upper quartile, etc.) and distributional measures (third and fourth moments, among others), SRD provided ten kinds of exploratory analytical graphic outputs:

1) scatterplots of values by county; attachments E and F are examples.
2) frequency histograms; attachments F and G are examples.
3) relative histograms; attachments H1 and H2 are examples.
4) frequency, or line plots, which are used in much the same way as histograms; attachments J1, J2.
5) percent point plots, which are interpreted in the same way as cumulative density plots; see attachment K for example.
6) draftsman plots, which show the graphic correlation between variables X and Y for each two variables at a time, for all variables of interest; see attachment L.
7) probability plots, which assume a certain theoretical underlying distribution and compare it to the observed distribution. A straight line at a 45 degree slope would indicate that the theoretical distribution type is a good fit. Attachment M1 for 1982 Annual Payroll/Establishments shows a fairly straight line but it does not pass through the origin nor have a 45 degree slope so it implies that a chi-square distribution was reasonable but the assumed mean and spread were not ideal. See attachments M2, M3, M4, M5, M6, M7 for examples with assumed Lognormal, Logistic, Extreme Value type 1, Normal, and

chi-square distributions being tested. Probability plots with linear subsegments and curved lines can also be interpreted.

8) PPCC or probability plot correlation coefficient plots show -- at their maximum against the Y axis -- which value of the distribution parameter (X axis) would produce the best match of theoretical and observed distributions. See attachments N1, N2, and N3 for examples assuming Pareto, Tukey Lambda, and Extreme Value type 2 distributions. The Tukey Lambda, for instance, has approximately a .98 correlation if -.4 is used as the parameter for the hypothetical distribution. Correlations that high indicate a strong correspondence between the observed and hypothesized distributions.

9) the starburst plot is a graphical technique for displaying multivariate data. The length of each ray of the star depicts the value of a different variable. In this case each star represents a county in the state, going across the page in alphabetical county order. Each ray of the star is one of several variables:

A Establishments
B Sales
C Annual Payroll
D First Quarter Payroll
E Employees

The first star is labelled with the above letter codes for orientation. The length of each ray (variable) is scaled against the minimum and maximum value observed within the state; a large star reflects a large county, etc. For New Jersey see attachment P1.

10) the normalized starburst plot is similar to the above but for scale. Small counties are shown too small for close inspection in the regular starburst pattern, but normalized, the shapes of the stars or multivariate relationships can be studied. Atypical shapes indicate atypical relationships of the variables (without regard for the county size). See Q1 and for Texas Q2.

We feel that the starburst plots show great potential for multivariate analysis.

Utility of the Project

PLOTLIER will probably be used by Business Division as part of their outlier identification review for the 1987 Census. To the best of our knowledge, this would be the first use the Bureau has ever made of analytical graphics for interactive data review.

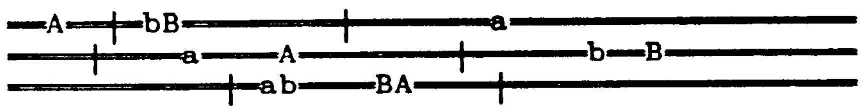
Business Division has not had an opportunity to thoroughly digest the graphical products which were provided as an aid in setting editing limits. It is fairly clear that the 1982 ÷ 1977 values cannot all be said to be normally distributed. Mark Wallace's staff will examine these outputs in much more detail when time permits.

The within-year ratios, such as sales to payroll, were run through the outlier identification programs (PLOTLIER) and also summarized in various of the ten graphic forms just described. We suggested that the data distributions might be better behaved if the ratios were computed as ratios of means rather than the mean of ratios, but in fact neither the mean nor the limits bore any obvious relation to the distribution when that was attempted. On the other hand, when ratios with a zero in the numerator (or denominator) were excluded from the computations, the mean and limits did appear to relate to the graphic distribution.

We look forward to the time when Business Division personnel can study the graphs more fully, so that we can interact with them.

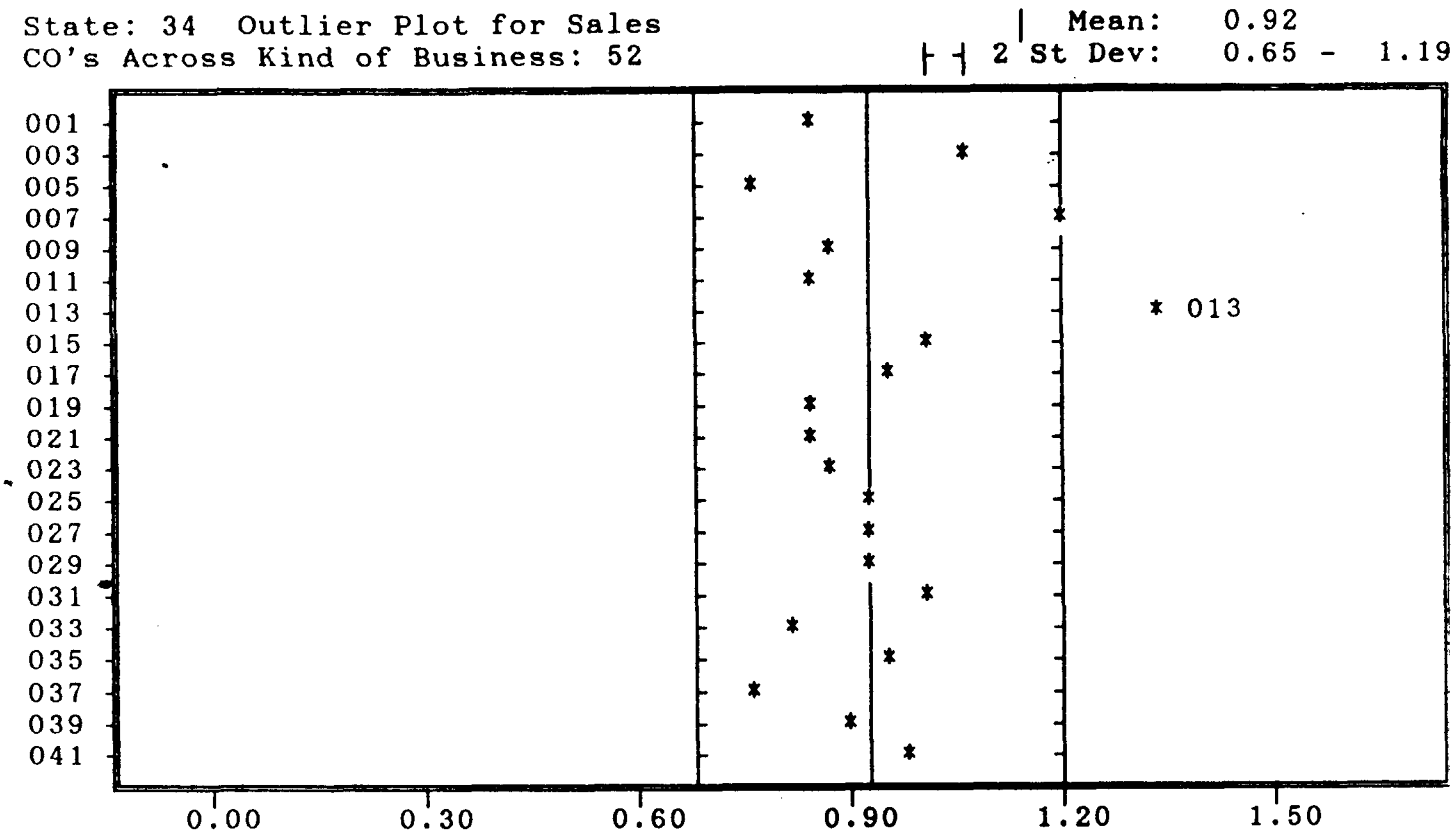
The following chart style could be used to analyze the outlier:

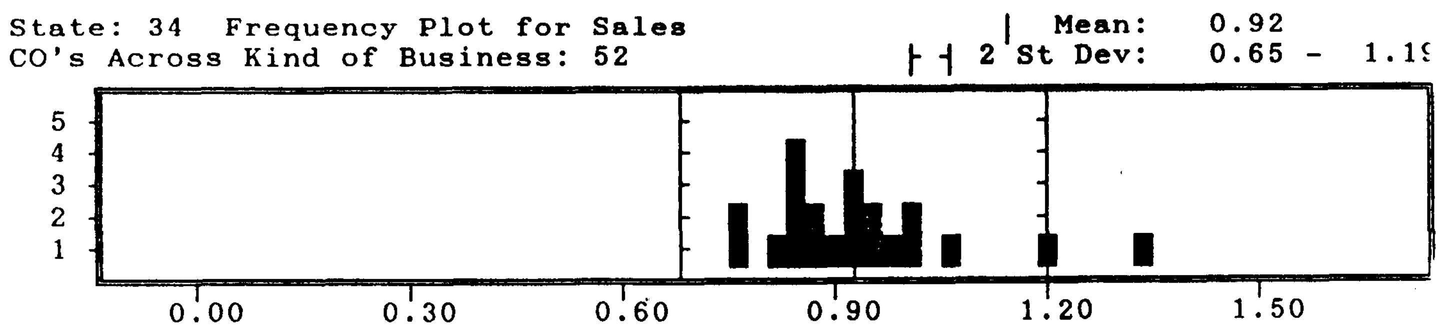
Sales
Annual Payroll
Employment



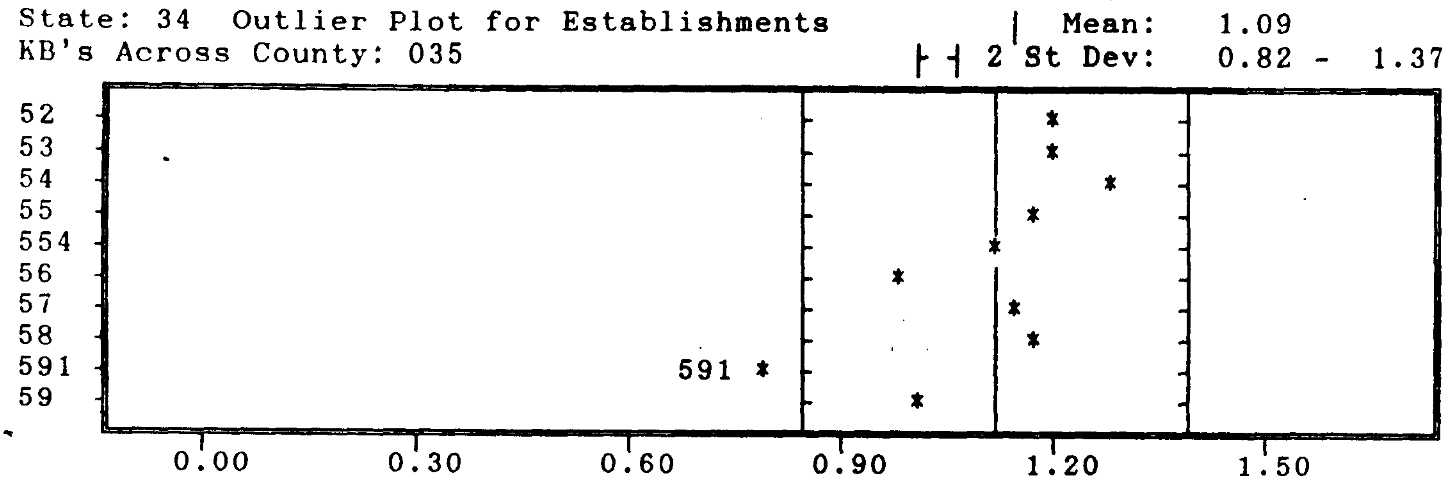
This chart uses one line for each of the three categories. The scale is not identified by values, but it extends the full range from its min to its max. The two tics projecting from the double line reflect the standard deviations from the mean. The lower case letters show the position of the establishment from the 1982 census and the upper case letters show their current status. If the establishment was not tabulated in the 1982 census then a lower case letter will not appear. Up to 26 outliers (A-Z) can be displayed on one chart.

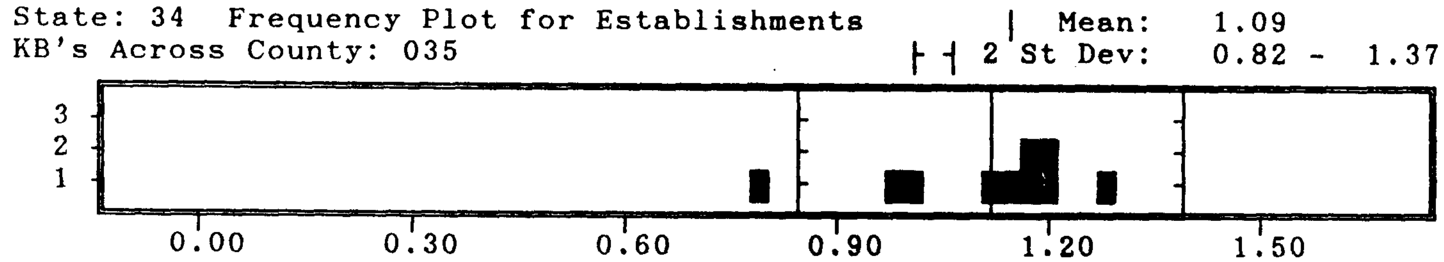
Attachment B





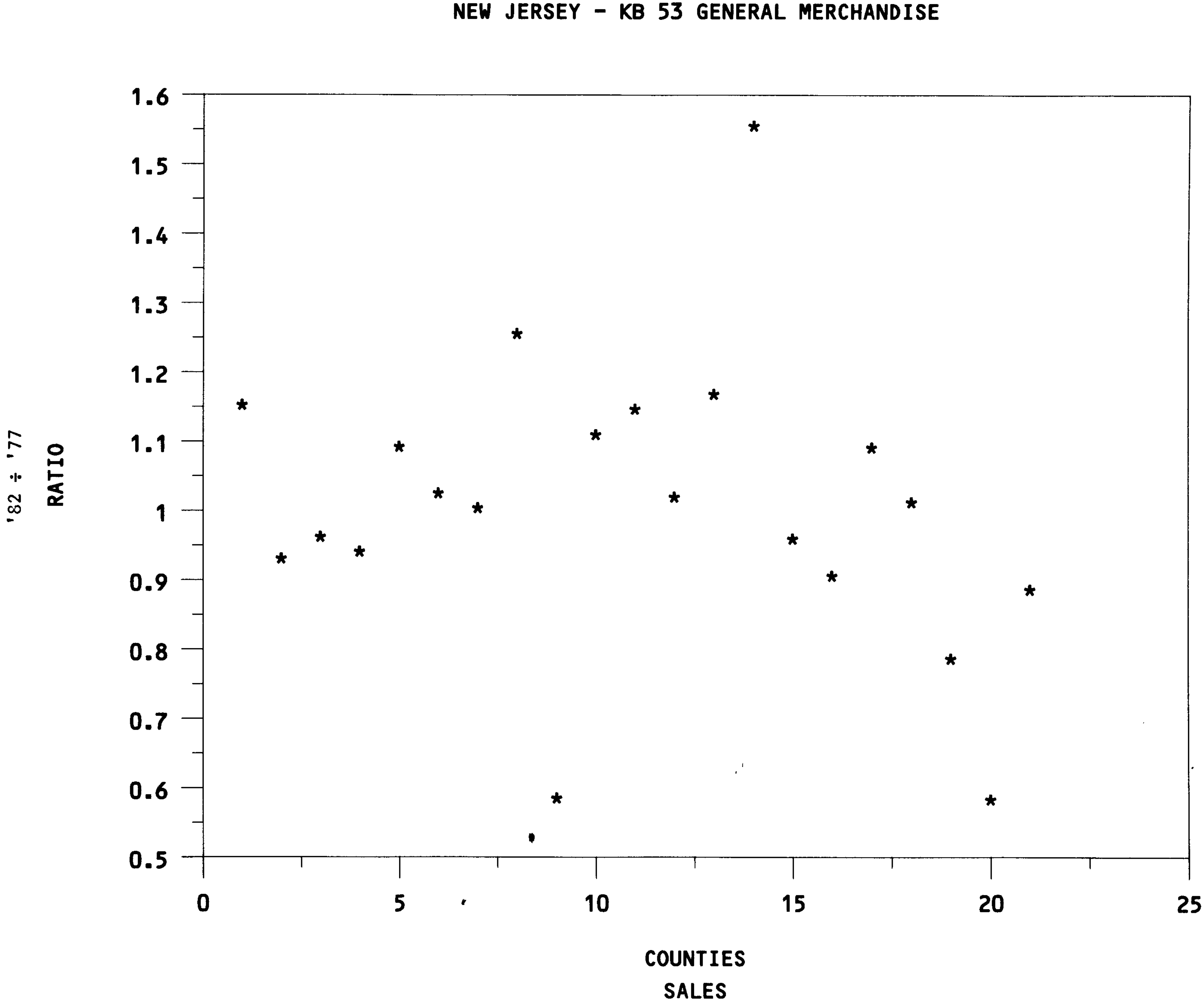
Attachment C



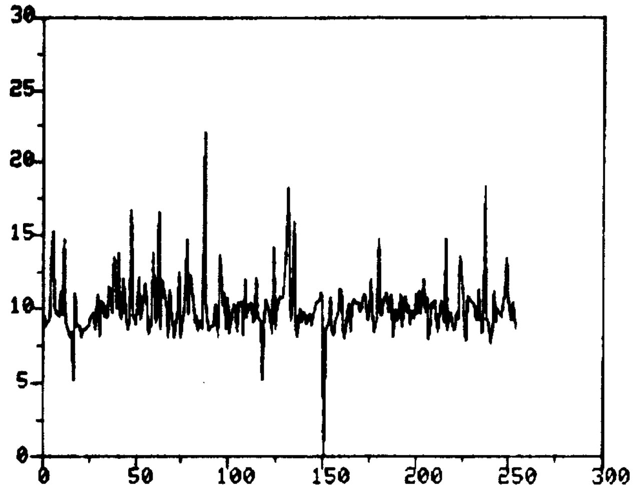


```
itle nj - kb 53 - sales ratio

THE TITLE HAS JUST BEEN SET TO
          NJ - KB 53 - SALES RATIO
>summary sr

                                   SUMMARY

                       NUMBER OF OBSERVATIONS =       21

*************************************************************************
*        LOCATION MEASURES          *       DISPERSION MEASURES        *
*************************************************************************
*  MIDRANGE      =    .1072945+001  *  RANGE         =    .9722500+000  *
*  MEAN          =    .1011954+001  *  STAND. DEV.   =    .2121337+000  *
*  MIDMEAN       =    .1070532+001  *  AV. AB. DEV.  =    .1462924+000  *
*  MEDIAN        =    .1015860+001  *  MINIMUM       =    .5868200+000  *
*                =                  *  LOWER QUART.  =    .9220700+000  *
*                =                  *  LOWER HINGE   =    .9346500+000  *
*                =                  *  UPPER HINGE   =    .1113400+001  *
*                =                  *  UPPER QUART.  =    .1131875+001  *
*                =                  *  MAXIMUM       =    .1559070+001  *
*************************************************************************
*       RANDOMNESS MEASURES         *     DISTRIBUTIONAL MEASURES      *
*************************************************************************
*  AUTOCO COEF   =    .7253829-001  *  ST. 3RD MOM.  =    .1234267+000  *
*                =    .0000000      *  ST. 4TH MOM.  =    .4012235+001  *
*                =    .0000000      *  ST. WILK-SHA  =   -.1157276+001  *
*                =                  *  UNIFORM PPCC  =    .9174360+000  *
*                =                  *  NORMAL  PPCC  =    .9587646+000  *
*                =                  *  TUK -.5 PPCC  =    .9802634+000  *
*                =                  *  CAUCHY  PPCC  =    .9521001+000  *
*************************************************************************
>
```

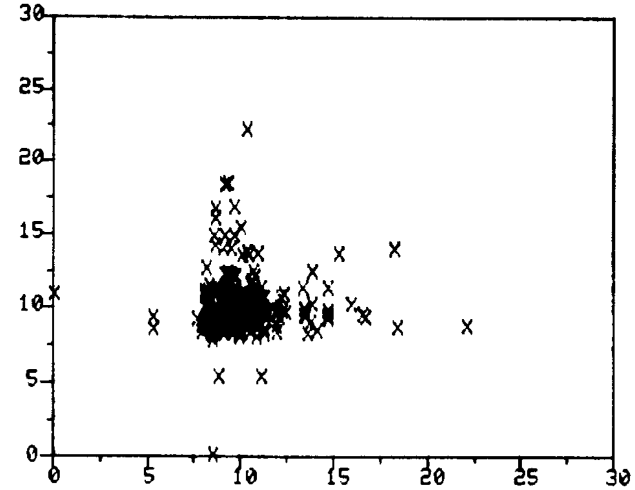
```
TITLE ANNUAL PAYROLL

THE TITLE HAS JUST BEEN SET TO
          ANNUAL PAYROLL
>SUMMARY X3

                                SUMMARY

                      NUMBER OF OBSERVATIONS =      254

***********************************************************************
*        LOCATION MEASURES          *        DISPERSION MEASURES      *
***********************************************************************
*  MIDRANGE     =    .1032044+007  *  RANGE         =   .2064089+007  *
*  MEAN         =    .3703960+005  *  STAND. DEV.   =   .1662460+006  *
*  MIDMEAN      =    .3666783+005  *  AV. AB. DEV.  =   .3479511+005  *
*  MEDIAN       =    .5672000+004  *  MINIMUM       =   .0000000      *
*               =                  *  LOWER QUART.  =   .1958500+004  *
*               =                  *  LOWER HINGE   =   .1996000+004  *
*               =                  *  UPPER HINGE   =   .1544300+005  *
*               =                  *  UPPER QUART.  =   .1598825+005  *
*               =                  *  MAXIMUM       =   .2064089+007  *
***********************************************************************
*       RANDOMNESS MEASURES         *     DISTRIBUTIONAL MEASURES     *
***********************************************************************
*  AUTOCO COEF  =   -.7435275-002  *  ST. 3RD MOM.  =   .9419373+001  *
*               =    .0000000      *  ST. 4TH MOM.  =   .1029869+003  *
*               =    .0000000      *  ST. WILK-SHA  =  -.2243571+003  *
*               =                  *  UNIFORM PPCC  =   .3272703+000  *
*               =                  *  NORMAL  PPCC  =   .4282459+000  *
*               =                  *  TUK -.5 PPCC  =   .6537895+000  *
*               =                  *  CAUCHY  PPCC  =   .7117193+000  *
***********************************************************************
>
```


NEW JERSEY - KB 53 GENERAL MERCHANDISE
1.6
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
'82 ÷ '77
RATIO
0
5
10
15
20
25
COUNTIES
SALES

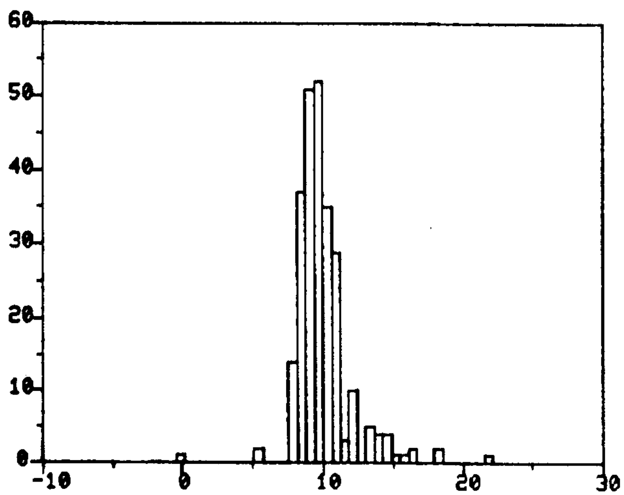
SALES/ANNUAL PAYROLL



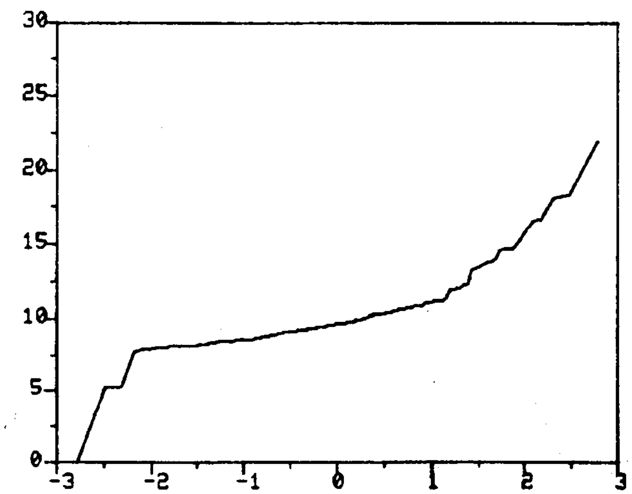
SALES/ANNUAL PAYROLL



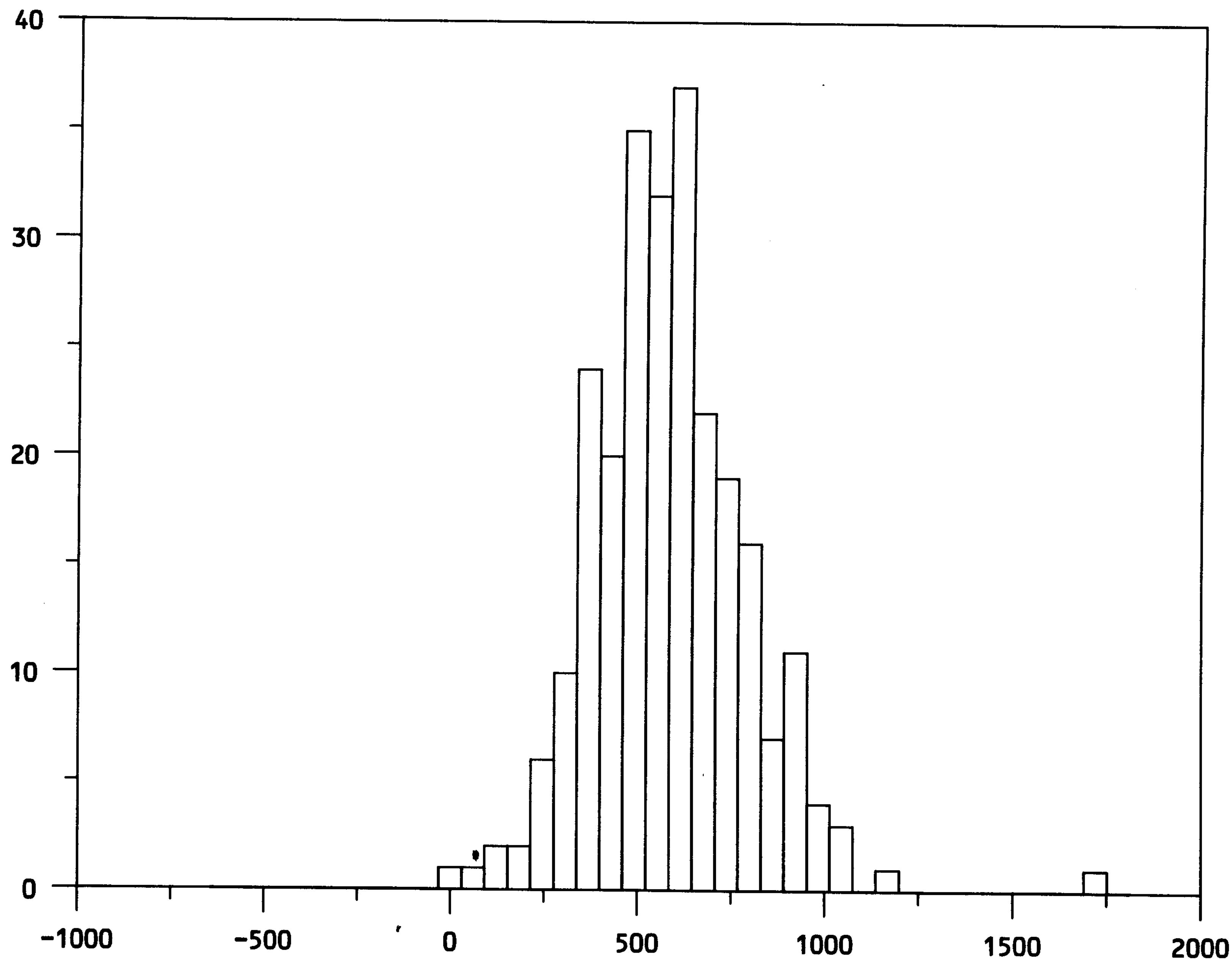
SALES/ANNUAL PAYROLL



SALES/ANNUAL PAYROLL

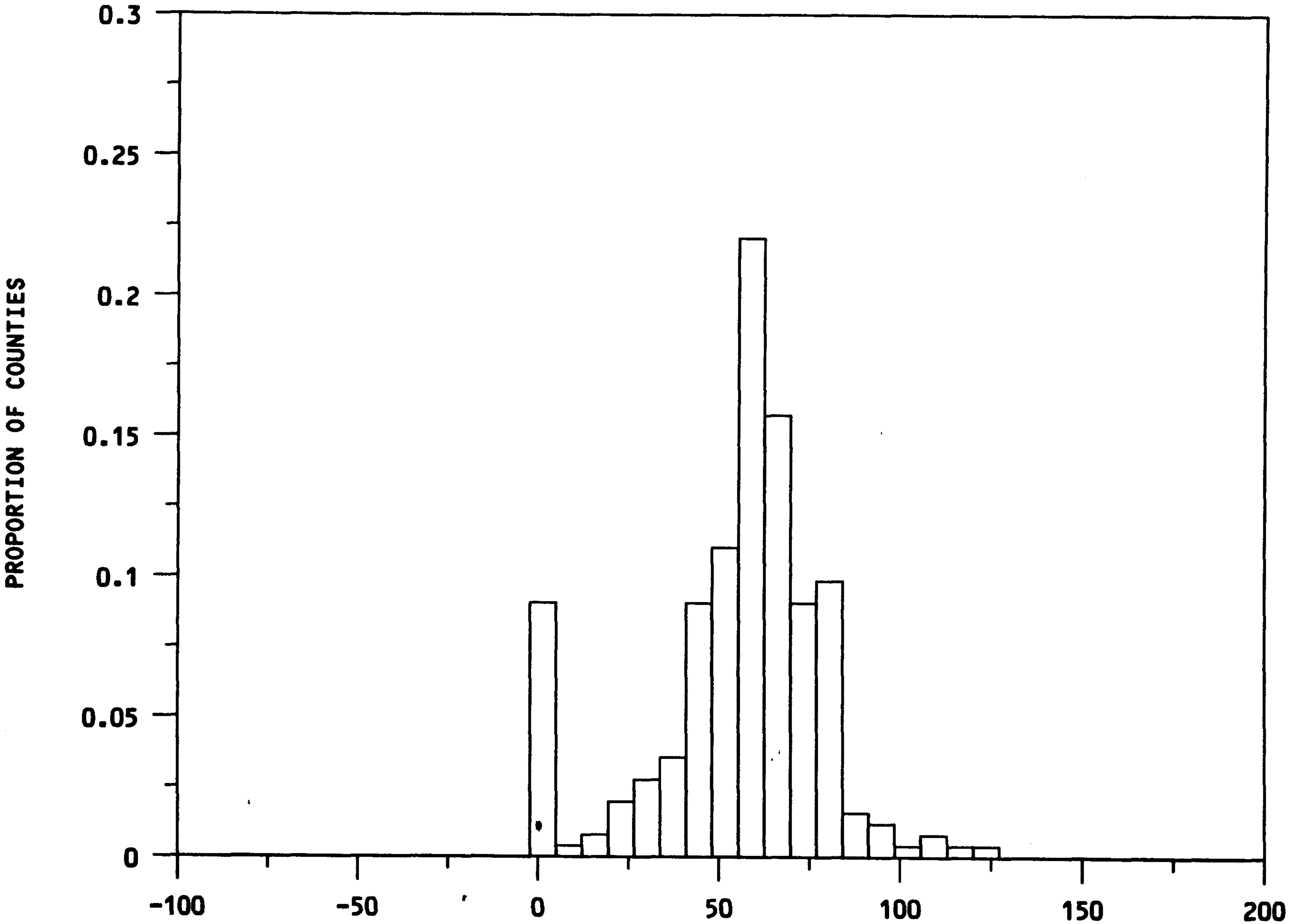


SALES/ESTABLISHMENTS

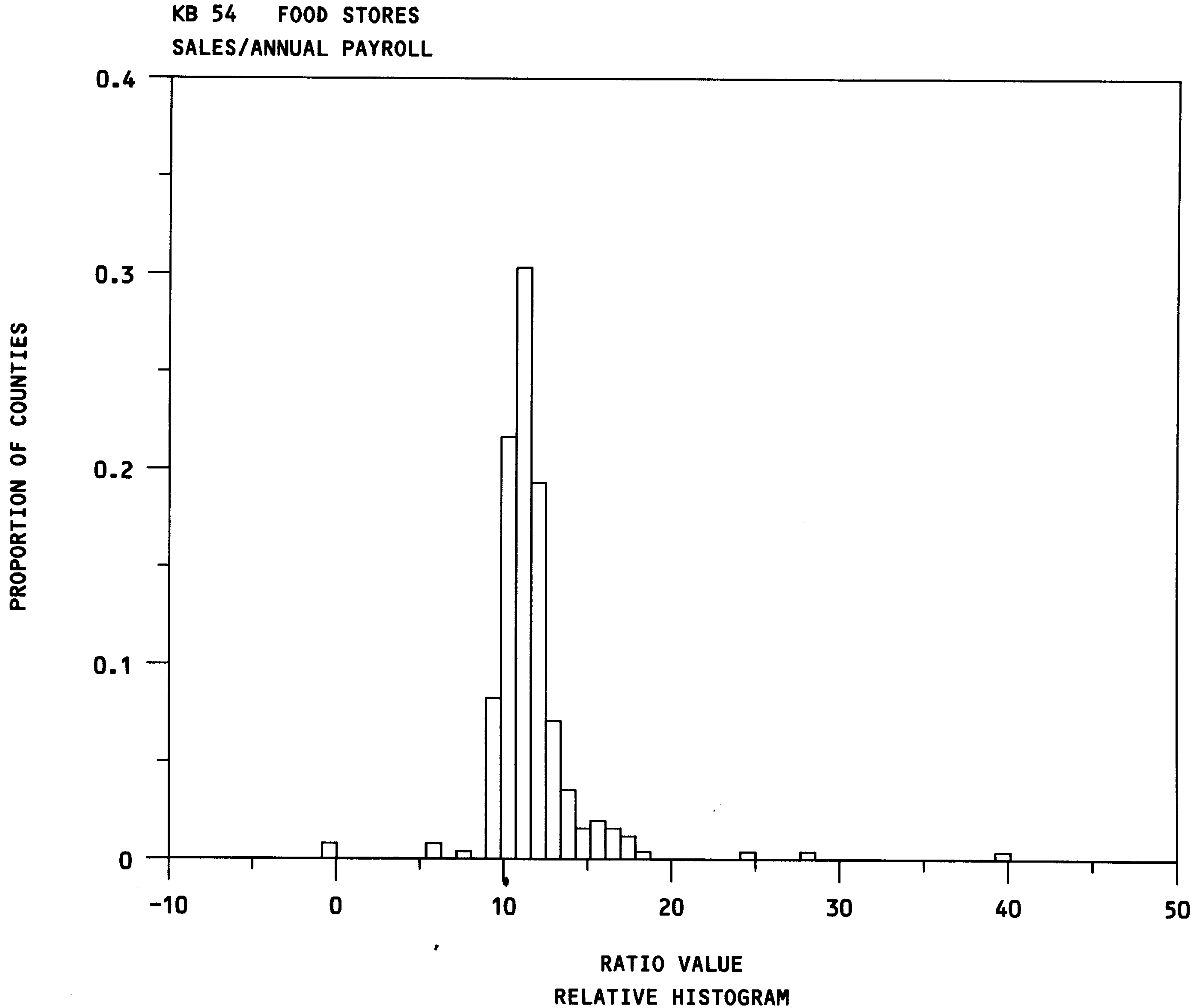


HISTOGRAM

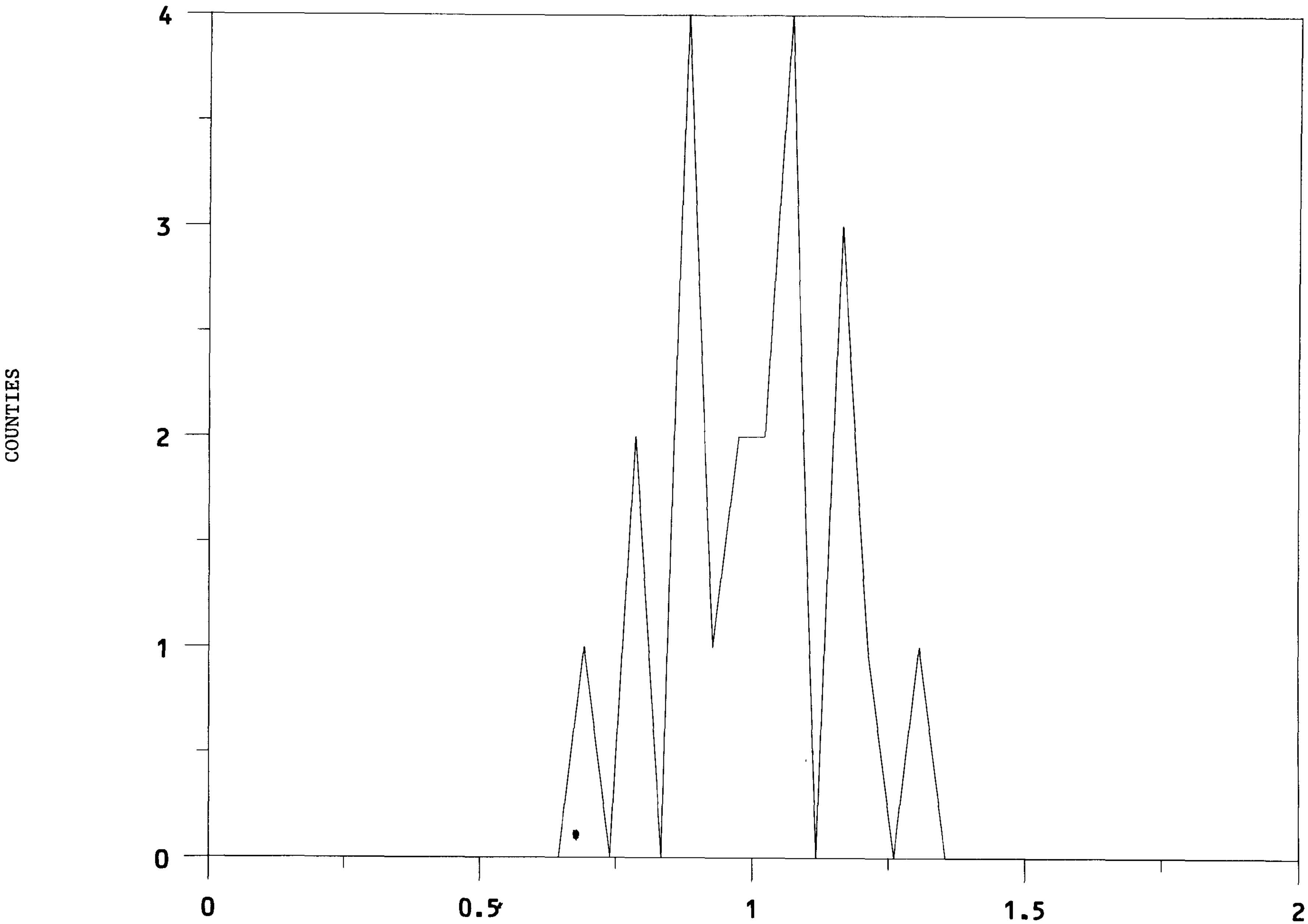
KB 53 GENERAL MERCHANDISE
SALES/EMPLOYEES



RATIO VALUE
RELATIVE HISTOGRAM


KB 54 FOOD STORES
SALES/ANNUAL PAYROLL
PROPORTION OF COUNTIES
0.4
0.3
0.2
0.1
0
-10
0
10
20
30
40
50
RATIO VALUE
RELATIVE HISTOGRAM

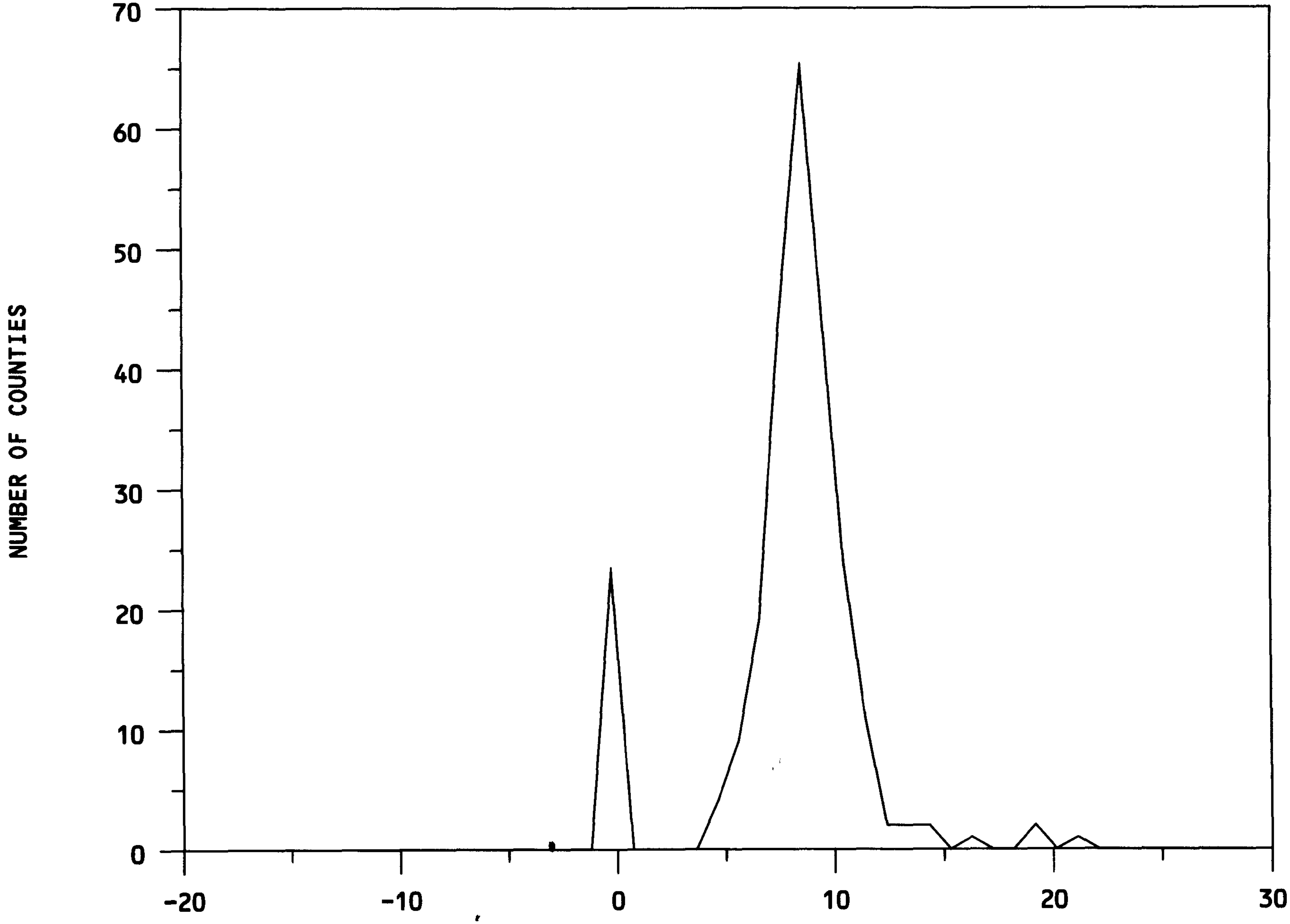
NEW JERSEY - KB 53 GENERAL MERCHANDISE



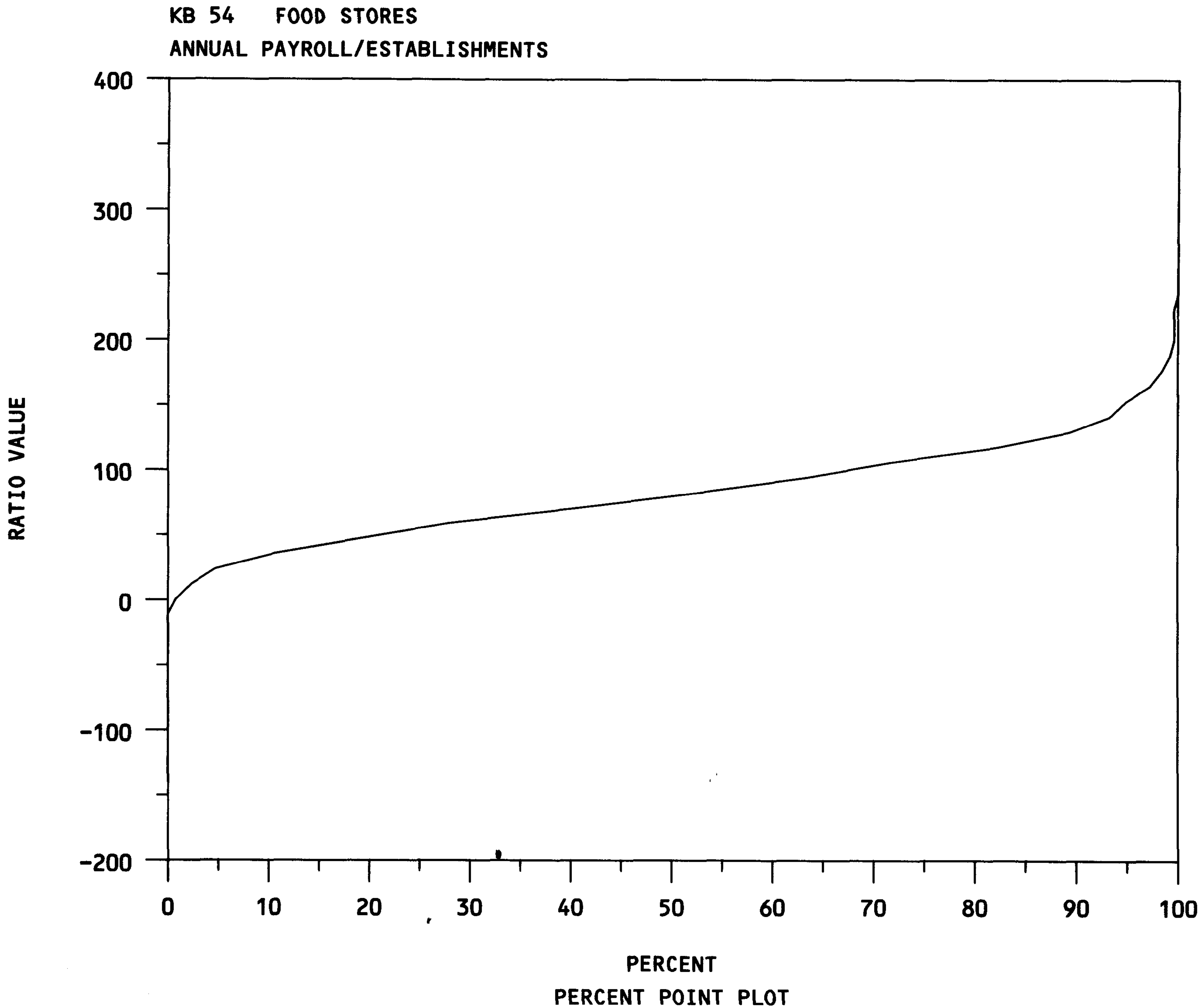
FREQUENCY PLOT

ESTABLISHMENTS RATIO '82 ÷ '77

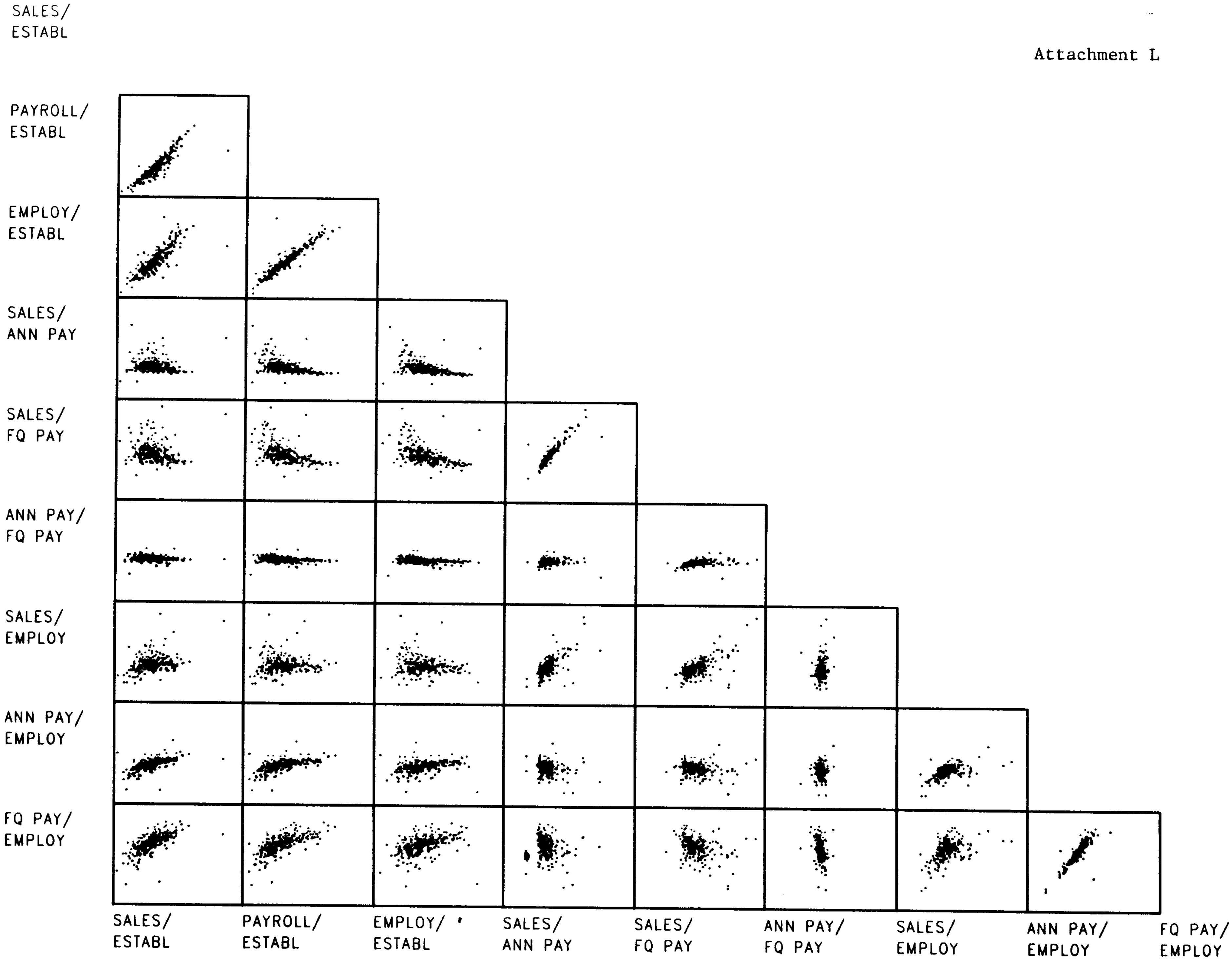
KB 53 GENERAL MERCHANDISE
SALES/ANNUAL PAYROLL



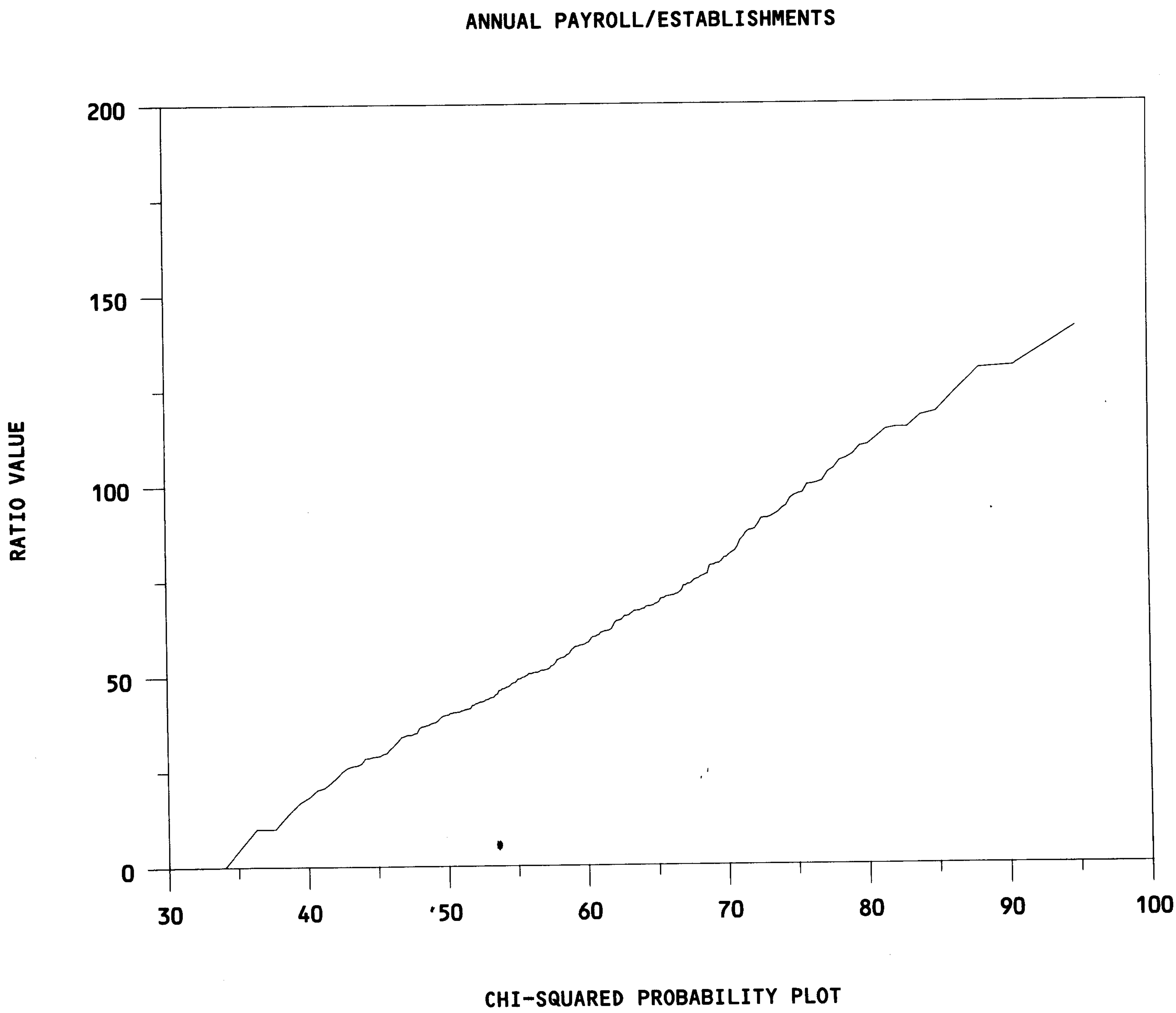
RATIO VALUE
FREQUENCY PLOT


KB 54 FOOD STORES
ANNUAL PAYROLL/ESTABLISHMENTS
400
300
200
100
0
-100
-200
RATIO VALUE
0
10
20
30
40
50
60
70
80
90
100
PERCENT
PERCENT POINT PLOT

Attachment L

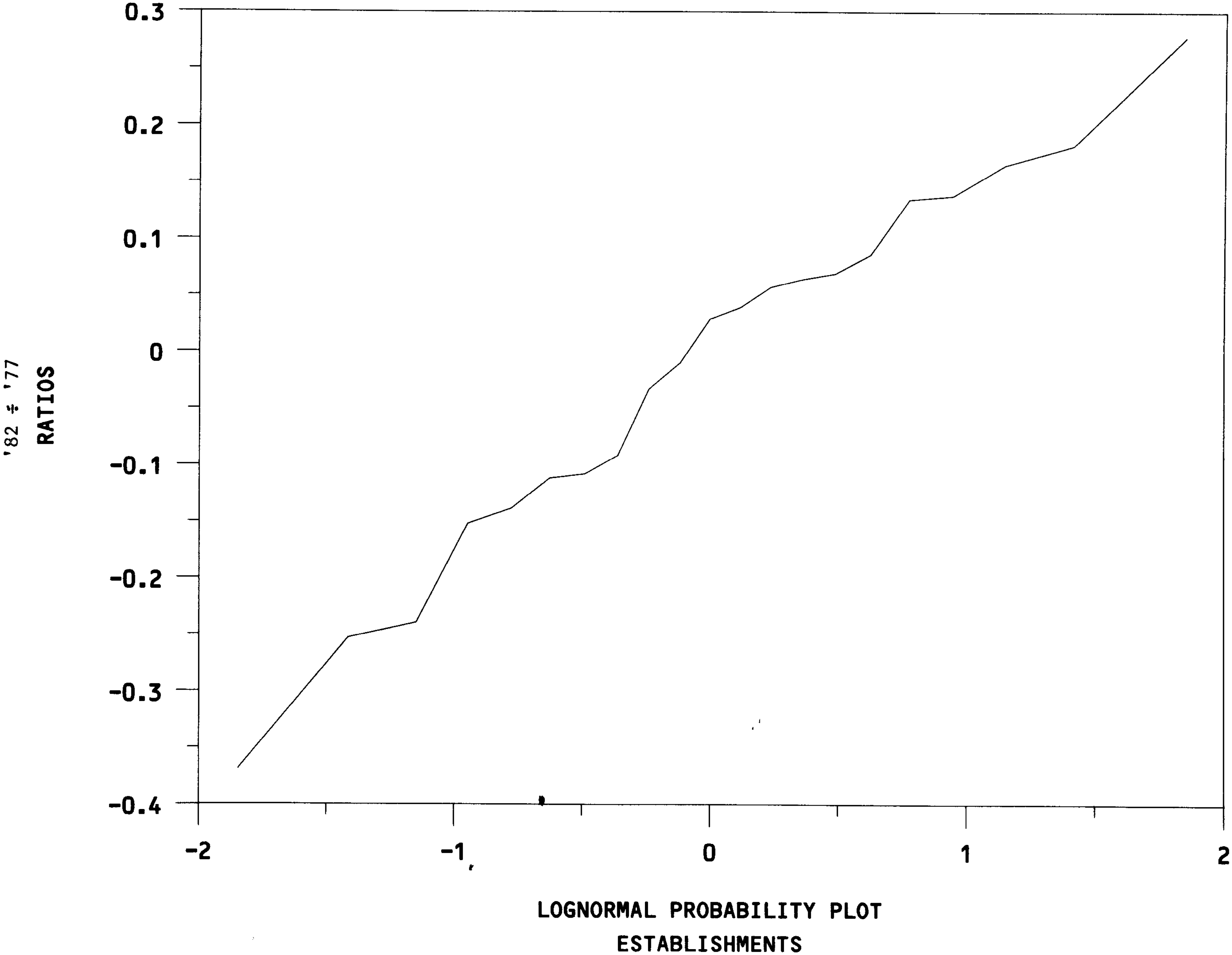


Attachment M1

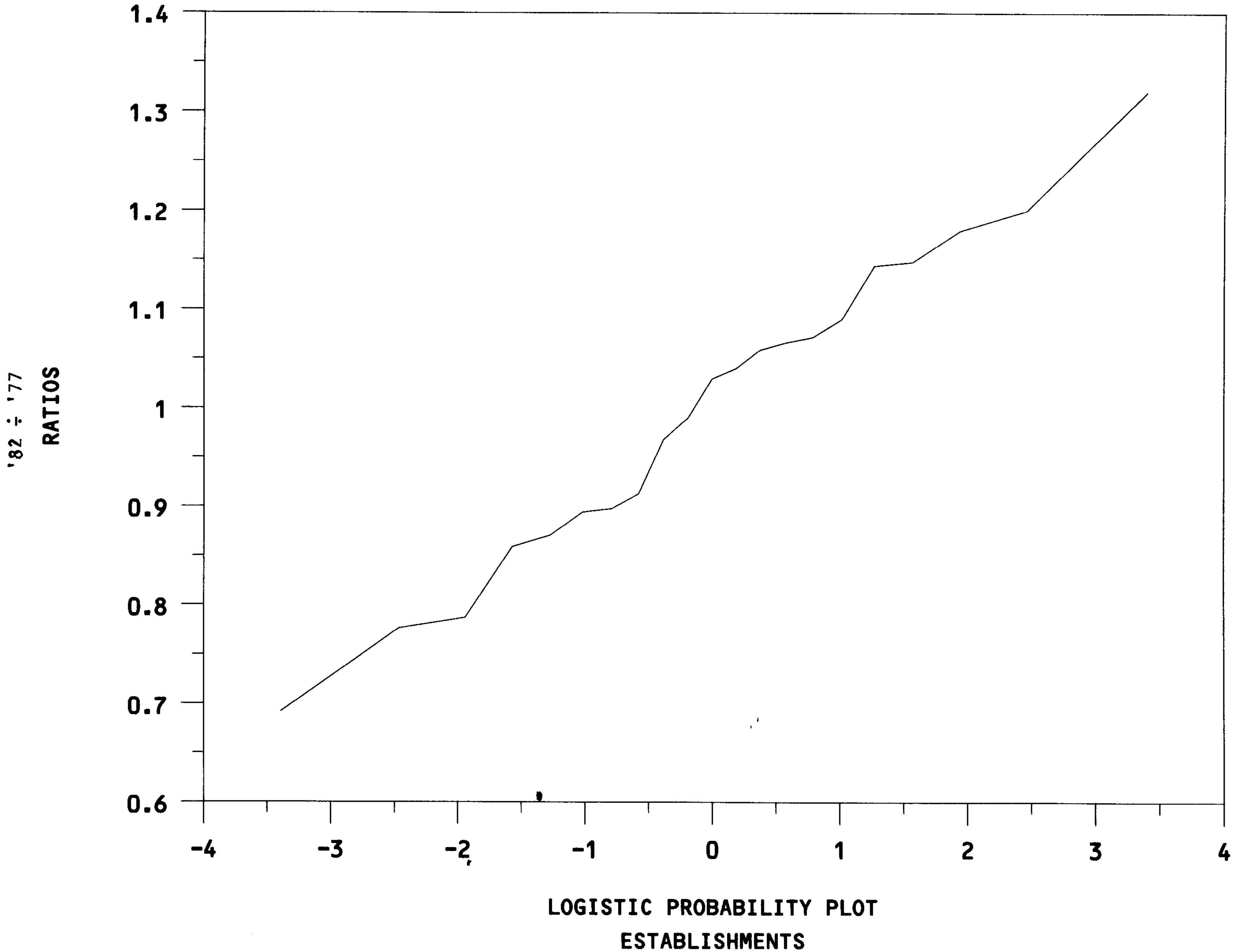


Attachment M2

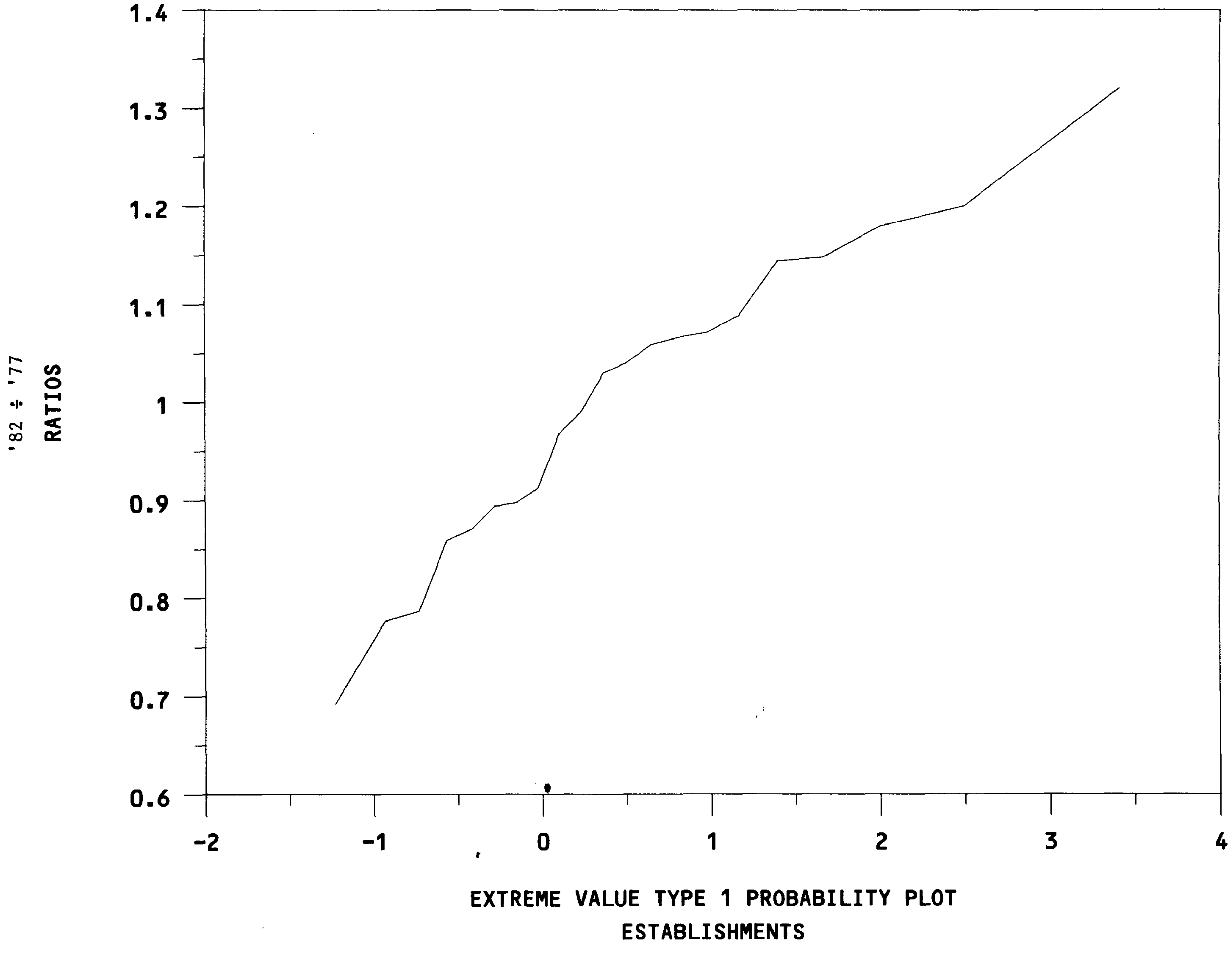
NEW JERSEY - KB 53 GENERAL MERCHANDISE



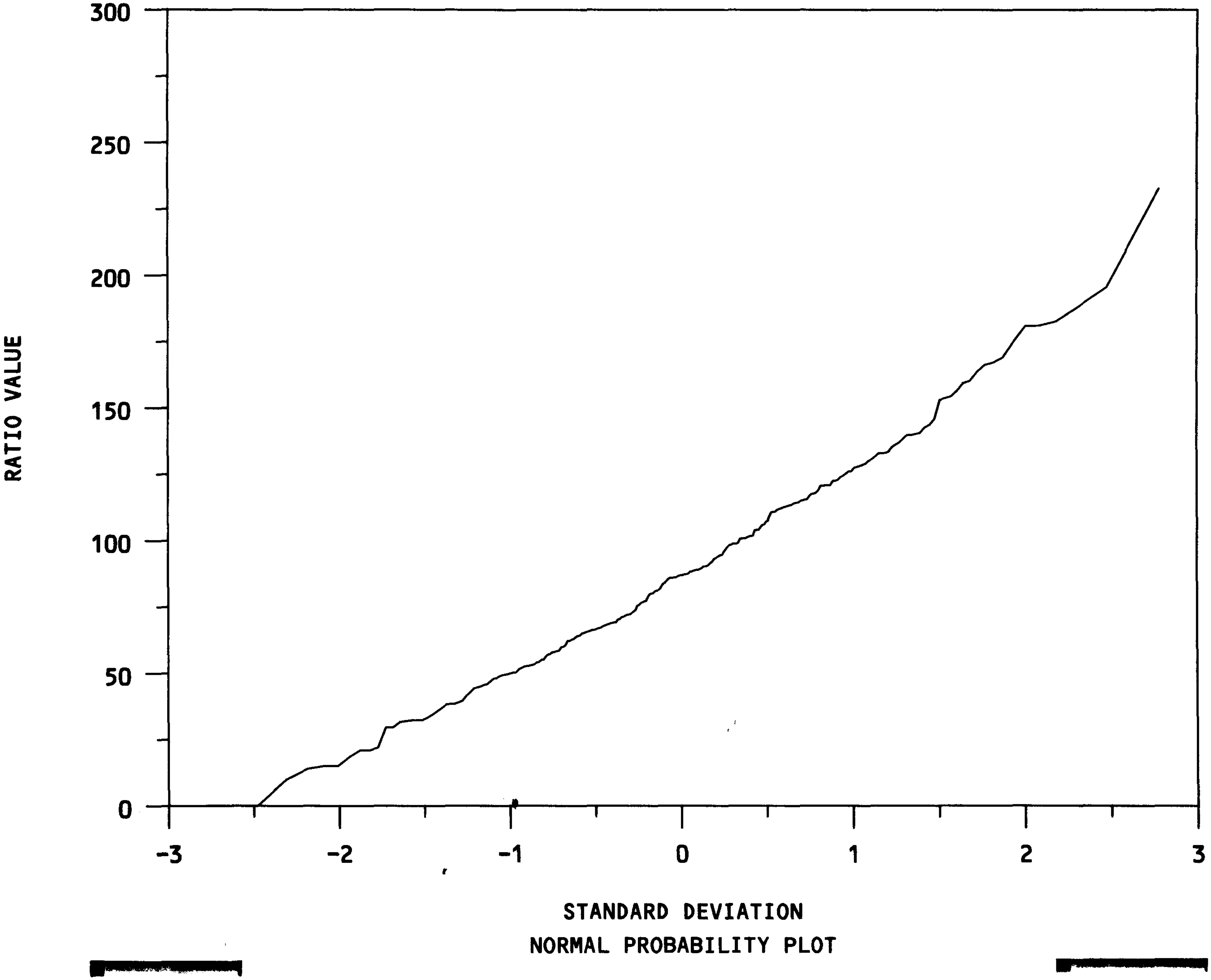
NEW JERSEY - KB 53 GENERAL MERCHANDISE



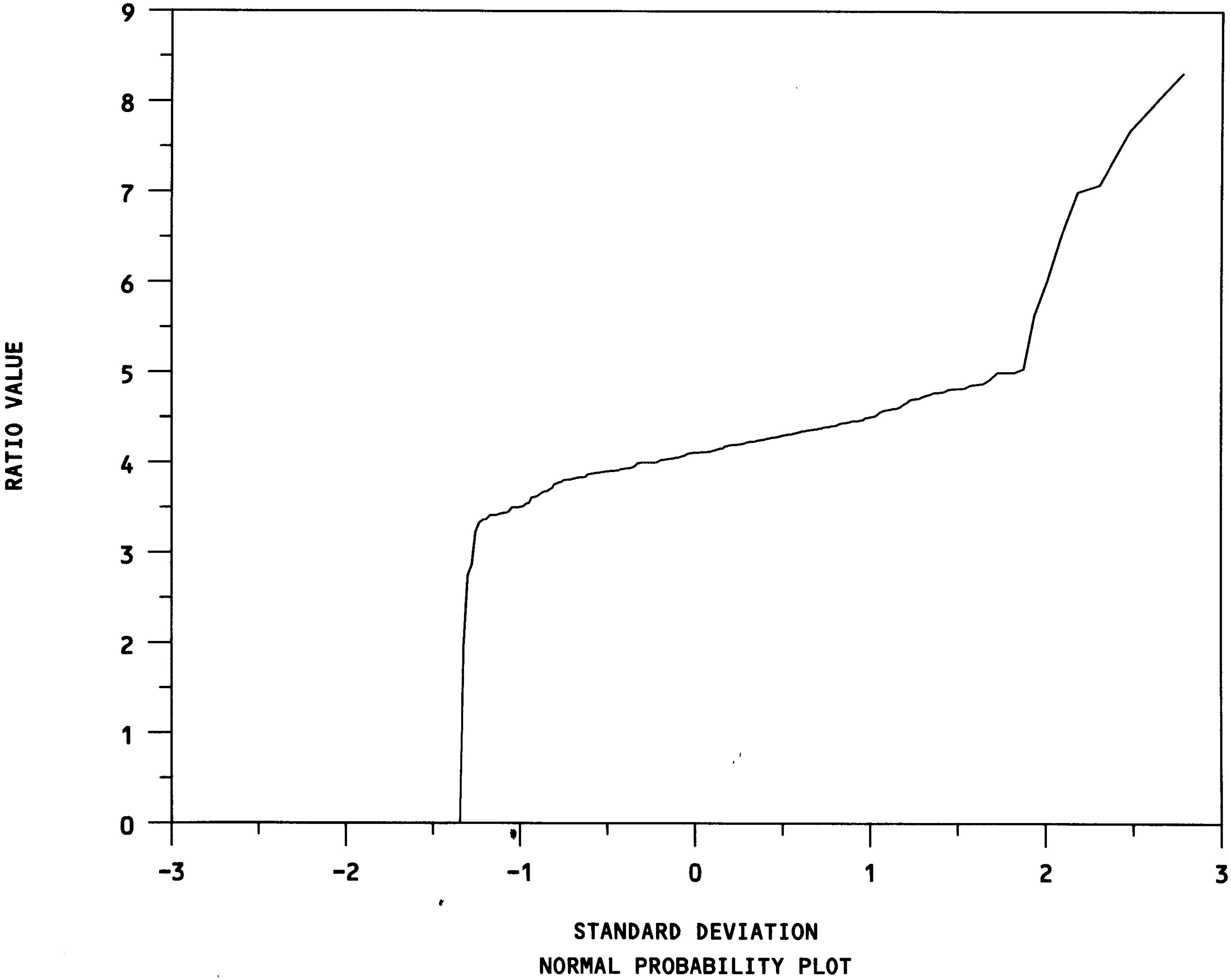
NEW JERSEY - KB 53 GENERAL MERCHANDISE



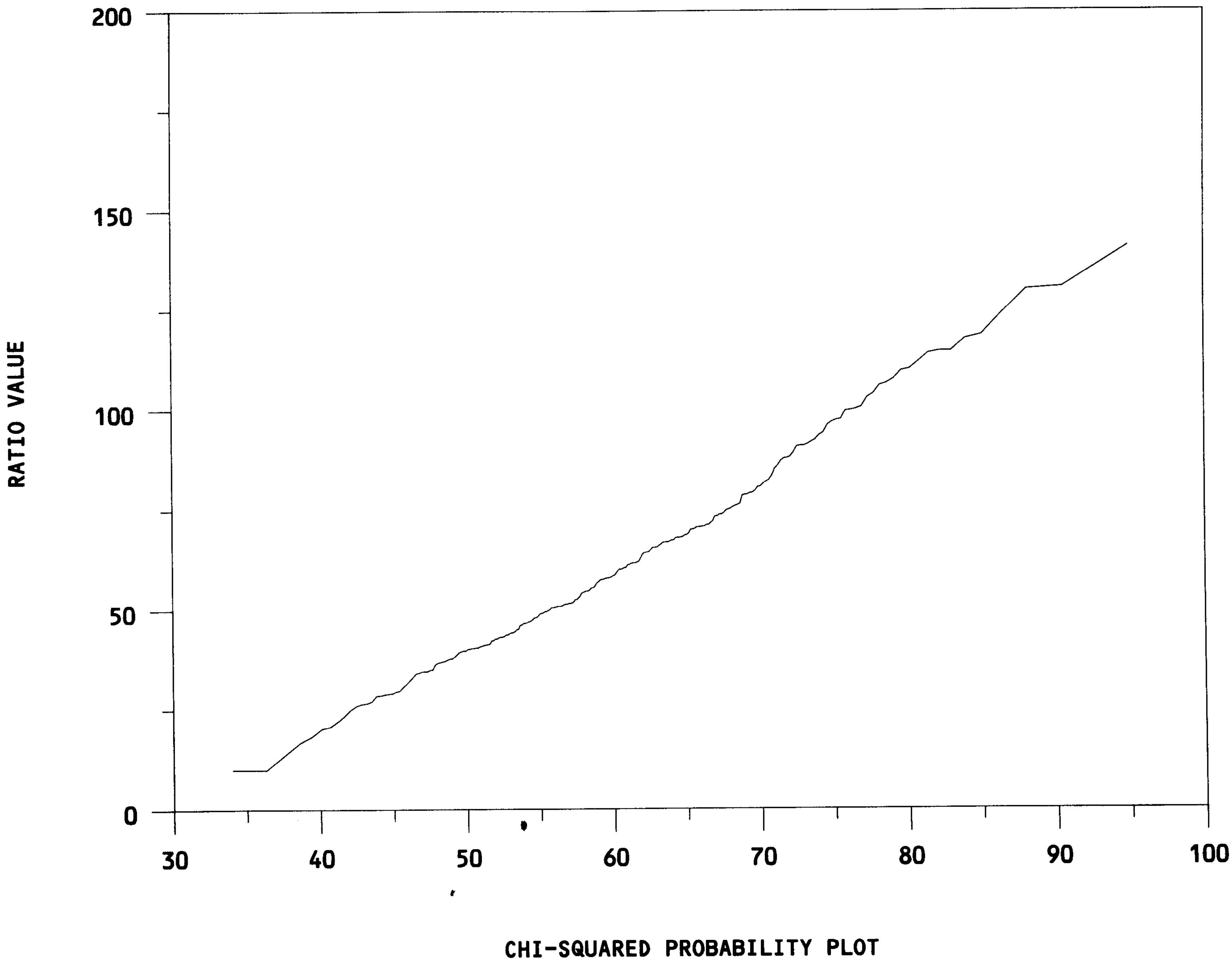
KB 54 FOOD STORES
ANNUAL PAYROLL/ESTABLISHMENTS



KB 53 GENERAL MERCHANDISE
ANNUAL PAYROLL/FIRST QUARTER PAYROLL

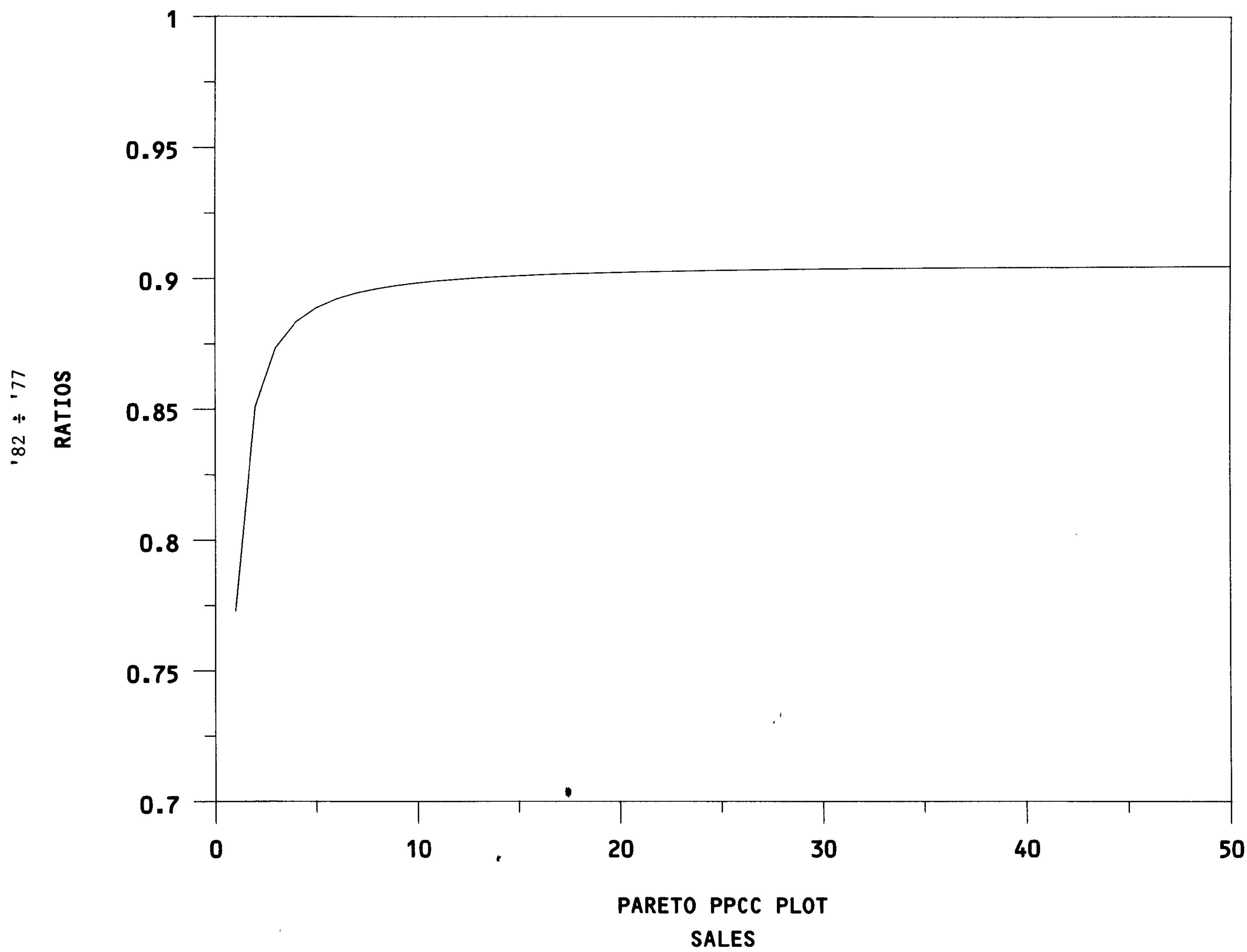


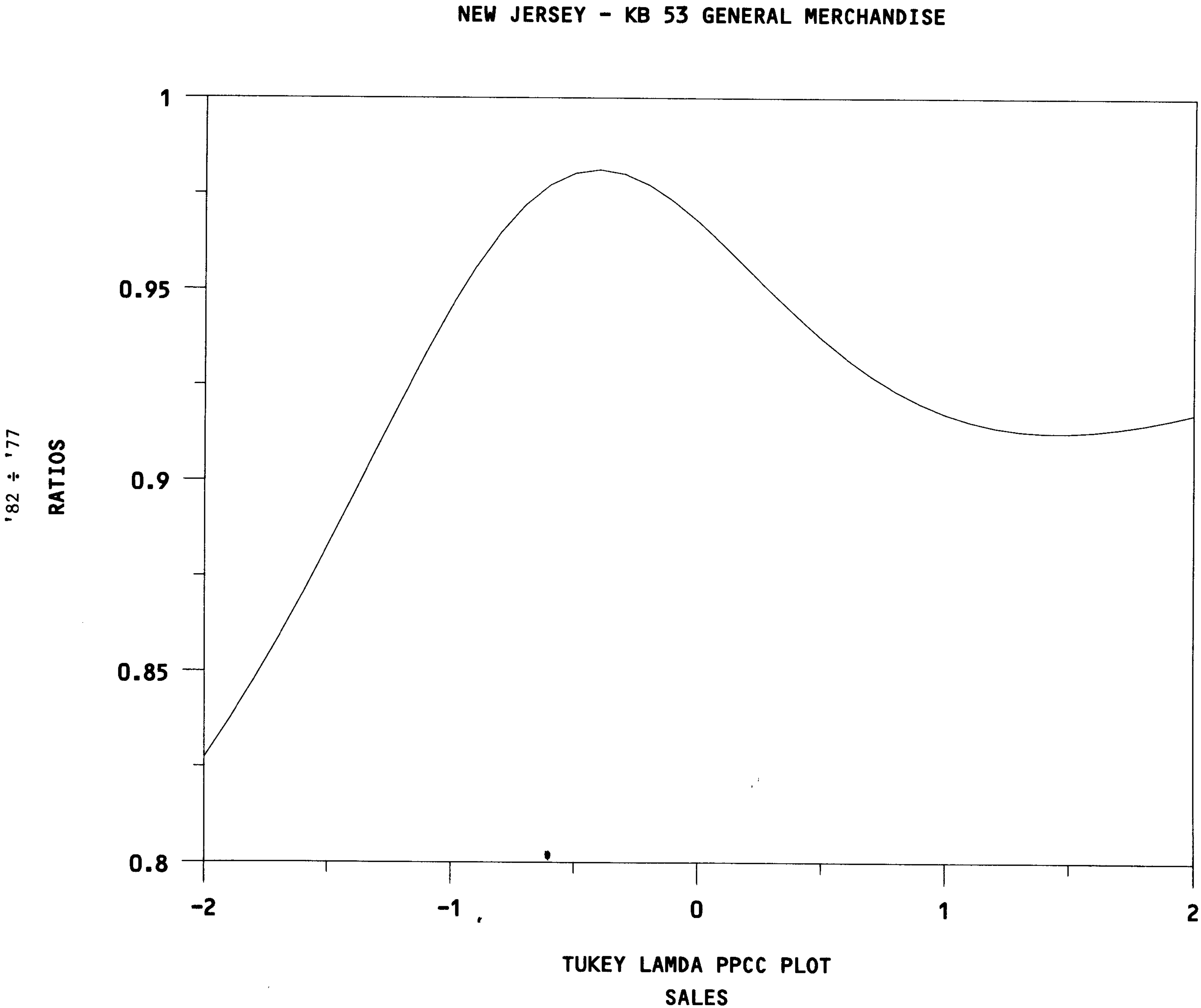
ANNUAL PAYROLL/ESTABLISHMENTS - NO LOVING



Attachment N1

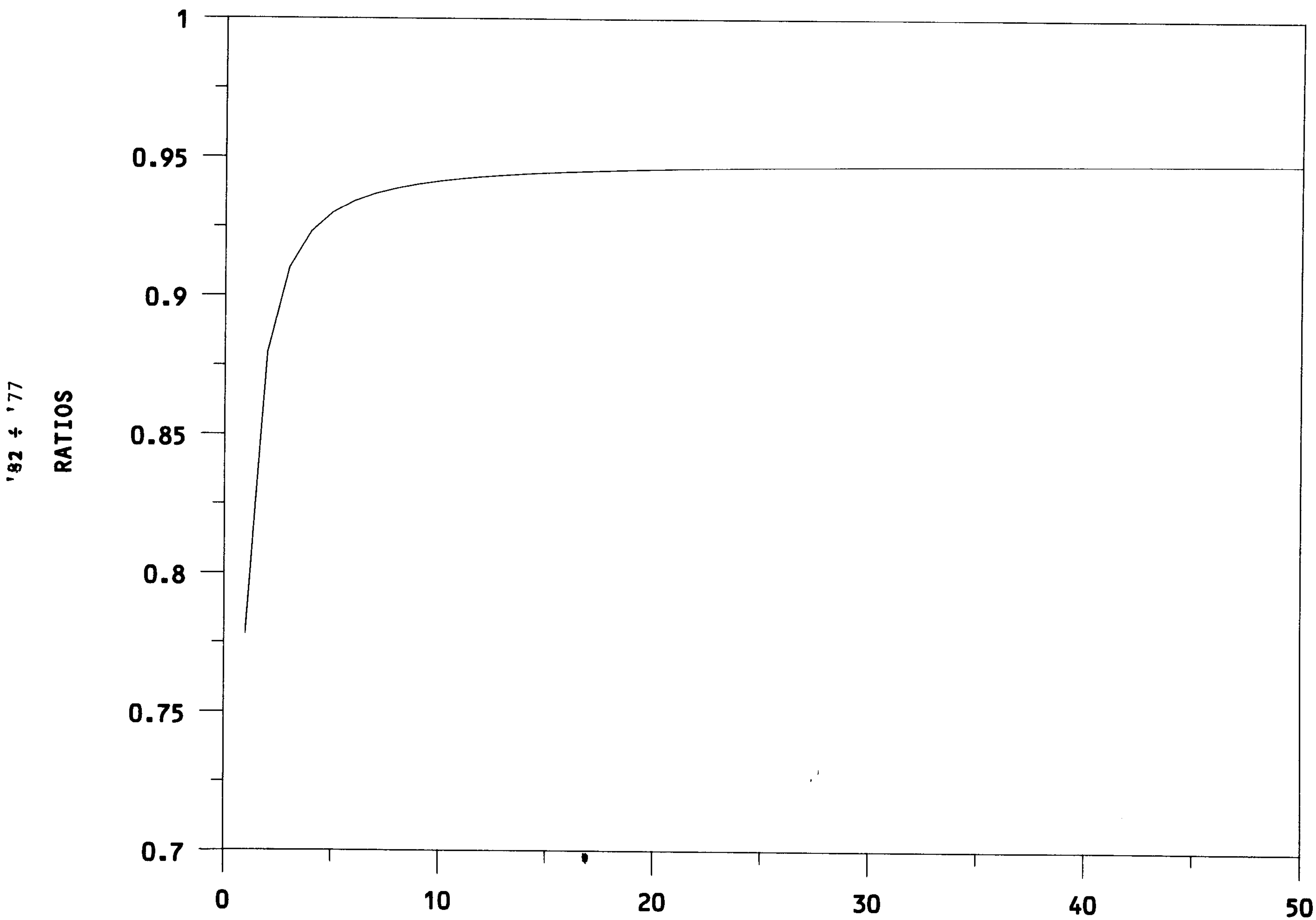
NEW JERSEY - KB 53 GENERAL MERCHANDISE





NEW JERSEY - KB 53 GENERAL MERCHANDISE
1
0.95
0.9
0.85
0.8
'82 ÷ '77
RATIOS
-2
-1
0
1
2
TUKEY LAMDA PPCC PLOT
SALES

NEW JERSEY - KB 53 GENERAL MERCHANDISE

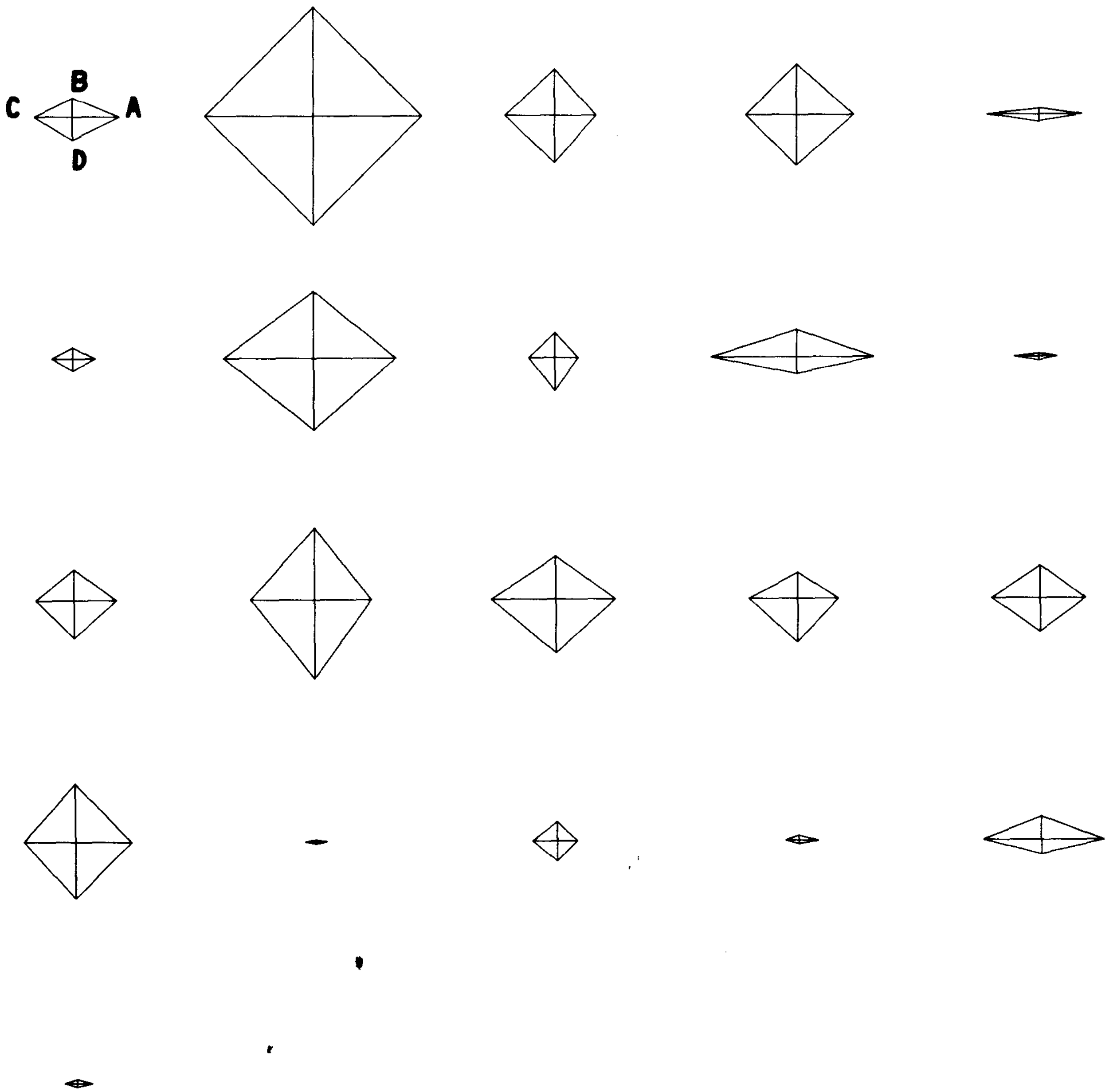


EXTREME VALUE TYPE 2 PPCC PLOT

SALES

NJ KB 53 O-1 ONLY

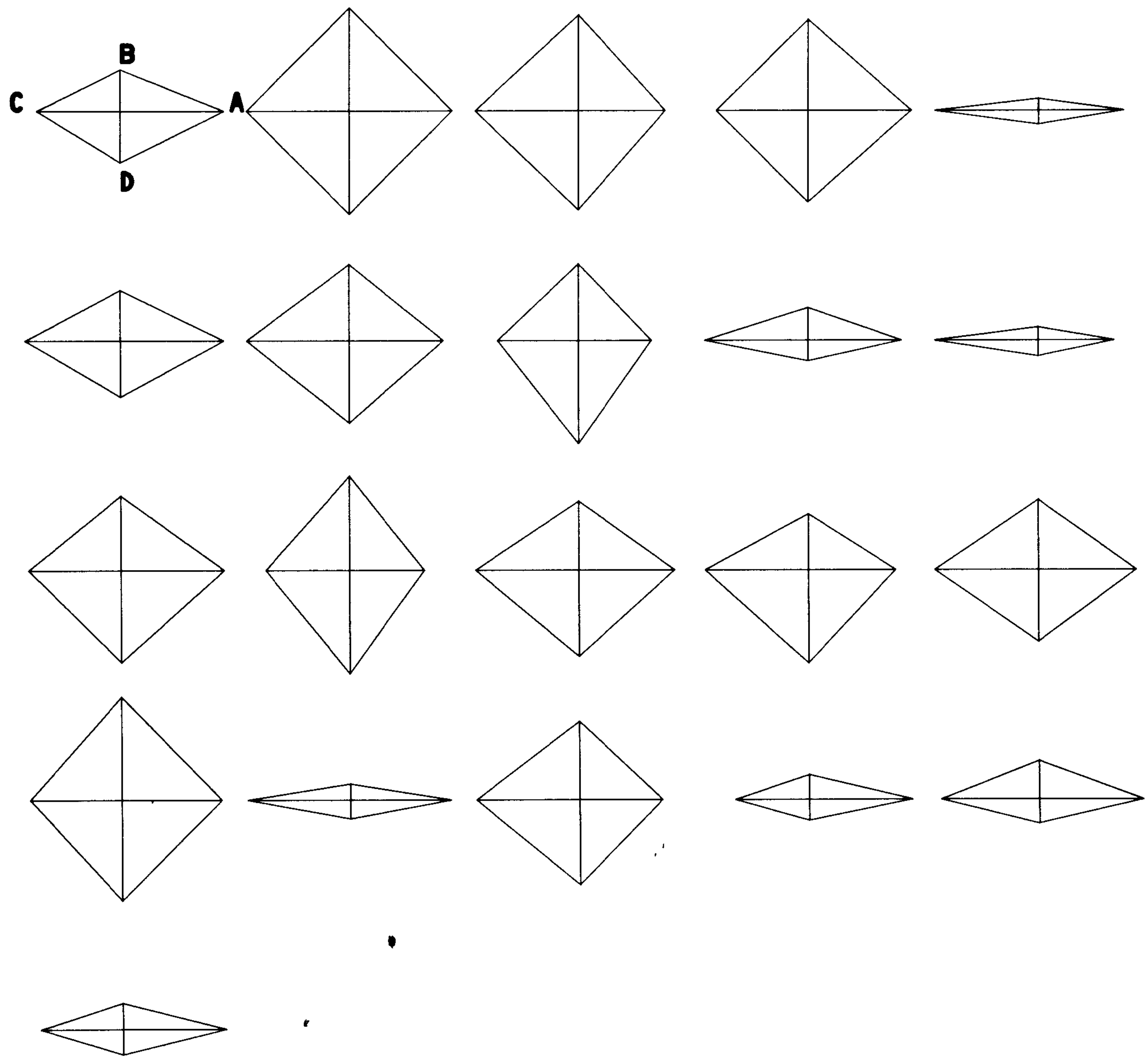
STARBURST PLOT



Attachment Q1

NEW JERSEY KB 53 NORMALIZED

STARBURST PLOT



NORMALIZED

STARBURST PLOT