*This work was supported by the U.S. National Heart, Lung, and Blood Institute Contracts NIH-NHLBI-71-2245 and 5-732-H607005-03 from National Institutes of Health.

THE USE OF MULTIPLE MEASUREMENTS TO MONITOR PROTOCOL ADHERENCE IN EPIDEMIOLOGICAL STUDIES

by

H.M. Schey and C.E. Davis

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1099

December 1976

ABSTRACT

We describe a method for detecting forged data that arise in multiple readings of a single variable when some of the reported values are not true readings but rather copies of one or two genuine values. The method exploits the fact that zero differences between pairs of readings will occur with large frequency in such forged data. The identification of those technicians in whose data zero differences occur often enough to raise suspicions of forgery is accomplished by eye from a frequency plot of zero differences, aided by a simple application of cluster analysis. The method is illustrated using the multiple blood pressure measurements made by seventy-eight clinicians in the Lipids Research Prevention Trial.

Introduction

In many epidemiological studies multiple measurements of certain variables are taken on a subject at one sitting because the mean of several readings provides a measure with less variability than that of a single reading. It sometimes happens, however, that technicians will only make one or two genuine readings and substitute copies of them in place of the remaining readings of the series. Obviously some procedure must be set up to detect forgeries of this kind.

One detection scheme is suggested by the fact that in any series of numbers, some of which have been manufactured by copying, an anomalously large number of values will be identical, and in a frequency plot one value will occur too often. Because the value which appears too frequently will no doubt differ from case to case, it is advantageous to standardize this procedure by examining differences among the reported readings. A frequency plot of such differences will then expose the forgery by showing an abnormally large number of zero differences.

If this scheme of examining the frequency of zero differences is adopted, one then must decide what constitutes too large a number of zero differences. It is often the case that field data are not especially uniform, particularly if they are taken by different groups of technicians, using different instruments, working in different places on different populations, and under different conditions. When this is the case it is difficult or impossible to establish any general underlying distribution which can be utilized in determining how large a fraction of zero difference frequencies is to be taken as evidence of forgery. Under these circumstances it is possible that a separation of those zero differences which occur often enough to raise suspicions of forgery from those which occur with acceptably small frequencies can be done by eye from a frequency plot of zero differences. Simple cluster analysis may provide some

additional guidance for separating the data into acceptable and unacceptable groups. It is important to recognize that any criteria developed in this way will be crude and must not be applied stringently. They are to be regarded as warning signals which facilitate the detection of protocol breaks without the need to review manually a large number of data forms.

In the following we illustrate the method using the multiple blood pressure readings taken as part of the Lipids Research Clinics (LRC) Prevention Trial.[1] The analysis will lead to a pair of criteria useful in detecting forged (copied) blood pressure data. The first of these criteria follows the pattern outlined above while the second is a simple and direct extension of it also based on frequencies of differences.

Zero Frequency Criterion

The protocol of the Lipids Research Clinics Prevention Trial stipulates that four blood pressure measurements be made at one sitting according to the procedure followed in the Hypertension Detection and Follow-Up Program (HDFP).[2] The first and third measurements are made using a standard sphygmomanomter, while the second and fourth are taken on a random-zero sphygmomanomter.[3] Since every set of readings is accompanied by the identification number of the blood pressure technician who took them, we can analyze the data for each technician separately.

In analyzing these data, we found that for a number of the technicians, the first and second readings were identical in a large fraction of the cases. Similarly, the third and fourth readings were identical with high frequency. Evidently these technicians were not making four distinct blood pressure measurements; instead they were taking the first and third readings correctly, but the second reading was merely a copy of the first, and the fourth, a copy of the third. Thus we were led to seek a statistical procedure of the kind outlined above which can be applied to the data of each technician and allow us to detect

such infractions of the protocol. To accomplish this we denote by ω the difference between the first and second systolic blood pressure readings made on a subject in any visit: $\omega = S_1 - S_2$. If two readings of blood pressure are identical then $\omega = 0$. Next let us define f_0 as the fraction of a technician's blood pressure measurements in which $\omega = 0$. If the technician is copying some of his/her data as described above, then the value of f_0 for that technician will be "too large." Our task is to find a criterion for distinguishing between "large" and "small" values of f_0.

The frequency plot of f_0 shown in Figure 1 is based on data taken from a group of 78 LRC technicians, each of whom has made 50 or more sets of blood pressure measurements. The bulk of the readings lie between $f_0 \sim .04$ and $f_0 \sim .33$ with a maximum of 11 readings around $f_0 = .12$. In addition, there are scattered readings up to $f_0 \sim .96$. What Figure 1 seems to indicate is that the readings in the range from .04 to .33 come from those clinicians whose second readings are done correctly. The remaining readings are abnormally large and presumably stem from faulty technique, intentional or otherwise.

If one accepts this interpretation, a means must be found to determine the cut-off point beyond which a value of f_0 will be judged "too large." To accomplish this we have made use of a hierarchical cluster analysis based on an algorithm due to Johnson[4] as implemented by Barr, et al.[5] The algorithm uses a nearest neighbor criterion and a Euclidean metric. The result of this analysis is to split the data into two clusters, one for which $f_0 \leq .322$, and the other for which $f_0 \geq .580$. This is probably the clustering one would make by eye using the distribution given in Figure 1. The upper cluster is made up of readings we shall regard as "too large" and shall therefore take as incorrect. Thus, we have established our "zero difference criterion": If $f_0 \leq .322$, accept; otherwise, reject.

Before moving on, several comments about this criterion may be in order.

First, we note that it is based only on systolic readings. This is the case because the diastolic values provide no additional information; essentially the same pattern of zero difference readings is found in systolic and diastolic blood pressure measurements. Second, no use has been made of the third and fourth readings because, again, they provide no additional information; the distribution of zero differences between the third and fourth readings is very much the same as that for the first and second. Moreover, we could detect no evidence of protocol infractions which involved the first pair of readings on the one hand with the second pair on the other. Finally, it is worth repeating a point made earlier: This criterion will never be used as the final word in deciding whether a technician is or is not doing a satisfactory job of reading blood pressure. Rather, it is simply a warning to us to check the relevant forms, and any action we take will be based on what we find there.

The Two Dimensional Criterion

The criterion established above is based exclusively on the quantity f_0, the fraction of measurements in which the variable ω assumes the value 0. In effect we have divided each technician's readings into two groups, those for which $\omega = 0$, and those for which $\omega \neq 0$. We now subdivide this second group according to the sign of ω so that a technician's readings can be described by three numbers, f_0, f_+, and f_-, which are, respectively, the fraction of readings in which $\omega = 0$, $\omega > 0$, and $\omega < 0$. We may then speak in geometric terms and associate with each technician a point in a three dimensional "fraction space" with axes f_-, f_+, and f_0 as illustrated in Figure 2a. Fortunately, however, it is not necessary for us to use a three dimensional representation, because, for any technician,

$$f_- + f_+ + f_0 = 1$$

which is the equation of a plane in fraction space as shown in Figure 2b. The

point representing any technician's readings lies in this plane, indeed, lies in that part of the plane in the positive octant -- the propped up triangle in Figure 2b. If we lay this triangle down flat, it provides a convenient way to represent the data in two dimensions, as illustrated in Figure 3. Note that we have introduced two coordinates, ξ and η, which can be used to specify any point in this "fraction plane."[6] A little analytic geometry shows that

$$\xi = \sqrt{1/2}\ (f_+ - f_-)$$

and

$$\eta = \sqrt{3/2}\ f_0.$$

We shall now develop a two dimensional criterion, a generalization of the zero frequency result obtained above, based on the two variables ξ and η. As before, we shall reject readings for which f_0 (or equivalently, η) is "too large." These are the cases treated above in which $\omega = 0$ occurs too often. But now we shall be able to examine more carefully the readings which correspond to small values of η and distinguish among them by the distribution of ξ values.

Some notion of how data in the acceptable range of η might split up can be formed rather easily from the definition of the companion variable ξ. We see that, apart from the factor $\sqrt{1/2}$ which is geometric in origin, ξ is the fraction of readings for which $\omega > 0$, less the fraction for which $\omega < 0$. But as a subject acclimatizes to his surroundings and to the technician, he relaxes, and as a consequence his blood pressure usually decreases. We should therefore expect $\omega = S_1 - S_2$ to be positive more frequently than negative, and this means that, as a rule, $f_+ > f_-$. Hence we expect the observed values of ξ to be predominately positive.

This expectation is born out when our data are plotted on the fraction plane (See Figure 4). We note that some data points are situated at large values of η ($\eta > .70$), but the bulk are at low values ($\eta < .40$). This is

merely the observation that led to the zero frequency criterion established in the previous section ($f_0 = .322$ corresponds to $\eta = .394$). What is new in Figure 4 is the fact, suggested above, that for most of the measurements, ξ is positive. In fact, expecting f_+ to be larger than f_- leads us to be suspicious of the measurements for which $\xi < 0$, but as before we turn to cluster analysis to discern patterns in the data.

Figure 5a is Figure 4 redrawn but now showing clustering into two groups. (The boundaries of the regions should not be taken literally. Here and in what follows we enclose points belonging to the same cluster by a polygon just for the sake of simplicity). These are precisely the two clusters obtained above in establishing the zero-difference criterion. At the three cluster level we get the grouping shown in Figure 5b. The data points at small values of η now cluster roughly according to the sign of ξ, and we expect that all or most of the lower left cluster will be rejected since points there correspond to blood pressure which increases from the first to the second reading. In fact we formulate a ξ criterion <u>almost</u> as indicated in Figure 5b. Specifically, we shall accept a measurement if $\xi \geq 0$ and reject it if $\xi < 0$. Note that in the interests of obtaining a criterion that is easy to implement, we have overridden the cluster analysis results to the extent that one small region of the $\xi - \eta$ plane has been re-allocated from rejection to acceptance. As it happens, with our data only one point is re-categorized as a result.

We now summarize our two dimensional criterion, phrased in terms of the three f's:

$$\text{If } f_0 \leq .322 \text{ and } f_+ \geq f_- \text{ , accept; otherwise, reject.}$$

As in the case of the earlier zero-frequency criterion of which this is a generalization, only the first two systolic readings were used in establishing this result, and our earlier comments on this fact apply here as well. Figure 6 shows the boundaries of the acceptance and rejection regions in the fraction plane.

Verification

Although we have not yet had the opportunity to apply our criteria in situations other than the one in which they were developed, we can adduce some evidence that indicates our results are sensible. The first and most important evidence is the response of those technicians whose blood pressure measurements were unacceptable according to one or both of the criteria, and who were confronted with this fact. Often "reasons" were given to justify the faulty readings, but no technician denied that he or she was indeed forging data.

Another piece of evidence goes back to a preliminary analysis we made just on the data of 48 clinicians and restricted to the first visit of the series each subject makes to his clinic. That analysis was very similar to the more extensive one discussed here, although the data were grouped solely by the eye with no use made of cluster theory. It led to essentially the same criteria we have obtained here for the larger sample of 78 clinicians and five visits.

There is another measure of the similarity between our 48 and 78 clinician analyses, and that is in the estimates each provides for the fraction of cases in which $\omega < 0$, $\omega = 0$, and $\omega > 0$.

Calling these $\hat{p}_-$, $\hat{p}_o$, and $\hat{p}_+$, respectively, we show their values for each of the two studies in Table 1. They are seen to be similar. Note that these values are calculated from the data points which remain after those failing one or both criteria have been discarded.

	$\hat{p}_-$	$\hat{p}_o$	$\hat{p}_+$
Preliminary Study	.289	.133	.578
Present Study	.310	.134	.555
HDFP	.295	.116	.589

(Table 1)

The last piece of evidence we have to offer in support of our criteria is based on the estimated of the three p's coming from the Hypertension

Detection and Follow-Up Program.[7] Using only that part of the HDFP data pertaining to men in the 35-59 year age group and in the lower diastolic stratum, we obtain the numbers given in Table 1 under the heading "HDFP." Again we see a close similarity to the values obtained from that portion of our own data which is acceptable according to the two criteria established here.

Summary

In this paper we have presented a method for detecting a type of forged data that can arise in epidemiological studies when a series of readings is made of a single variable. If the data are not all genuine readings, but rather copies of one or two bona fide values, we can exploit the fact that zero differences between readings will occur more frequently than they would in clean data. The criterion for distinguishing real from forged data requires a decision as to how great a zero difference frequency must be to indicate forgery, and this decision can be made by eye from a frequency plot, or by a simple application of cluster analysis. We have presented an example of this scheme which uses the repeated blood pressure measurements made in the LRC Prevention Trial and have shown how it can result in criteria by means of which forged data can be detected without an expensive, tedious examination of a large number of data forms.

In our example we have, in fact, developed two criteria only one of which is based on zero difference frequencies. The other represents an extension of this idea to non-zero differences. Admittedly this second criterion is rather special to our example, resting as it does on certain characteristics of blood pressure. Its inclusion here emphasizes the fact that in any given situation our basic ideas may be extended to obtain criteria tailored to the data under analysis.

The methods given in this paper lead to a rather crude test which must

therefore not be applied blindly -- it merely indicates which parts of the data require closer inspection. In the face of data (typical of epidemiological studies) which do not conform to some well defined distribution, such procedures, crude as they are, may be the best we can obtain. Nevertheless, we believe they can provide effective and easily implemented methods for uncovering forged data.

Footnotes

[1]Lipids Research Clinics: Protocol for the Lipids Research Clinics Type II Coronary Primary Prevention Trial. Unpublished protocol. Central Patient Registry and Coordinating Center, University of North Carolina, 1973.

[2]Remington, R. D., Hypertension Detection and Follow-Up Program, Inserm, 8 Sept., 1973, Vol. 21, pp. 185-194.

[3]Wright, B. M. and Dore, C. F., A random-zero sphygmomanometer, The Lancet, February 14, 1970, pp. 337-338.

[4]Johnson, S. D., "Hierarchical Clustering Schemes," Psychometrika, XXXII (1967), pp. 241-254.

[5]Barr, A. J., Goodnight, J. H., Sall, J. P., and Helwig, J. T., A User's Guide to SAS 76 (1976), SAS Institute, Inc., pp. 72-79.

[6]This method of representing data in two dimensions has been used in the past, although usually in conjunction with a different system of coordinates. See, for example, Mosteller, F. and Tukey, J. W. (1968), Data analysis, including statistics. Handbook of Social Psychology, Eds. G. Lindzey and E. Aronson, pp. 80-203, Reading, Mass.: Addison Wesley. We wish to express our thanks to Steven J. Samuels for pointing out this reference to us.

[7]We are grateful to the staff of HDFP for providing us with the data from which the three values of p were calculated.

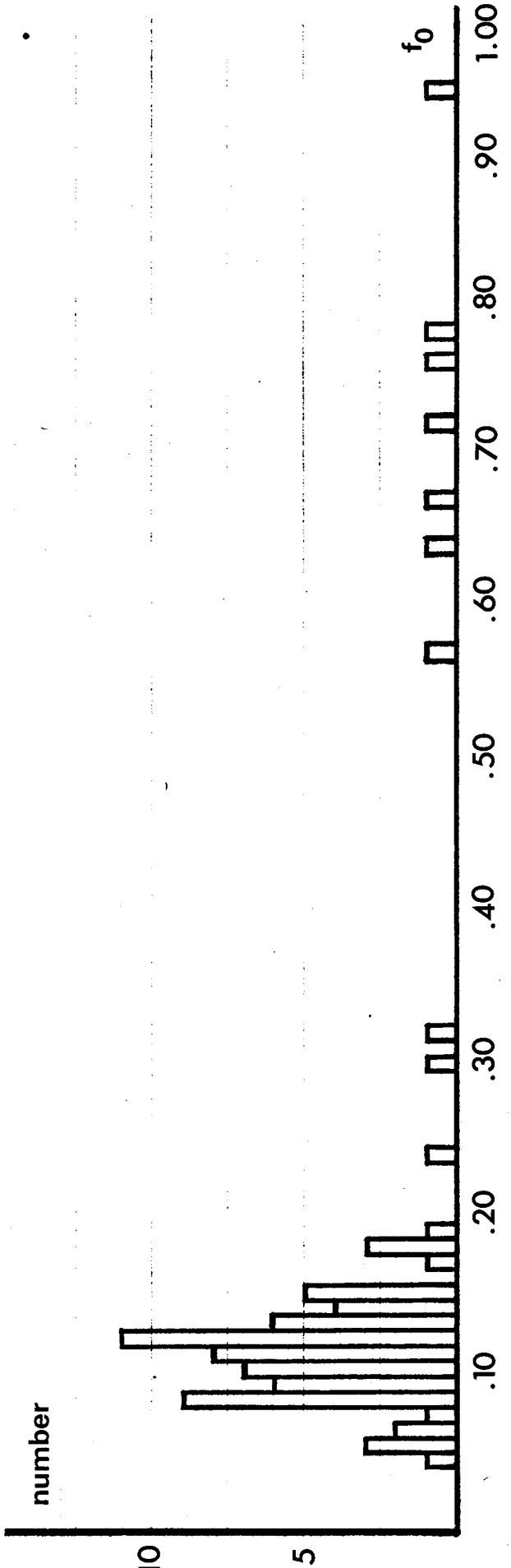

Figure 1. The frequency distribution of f_0

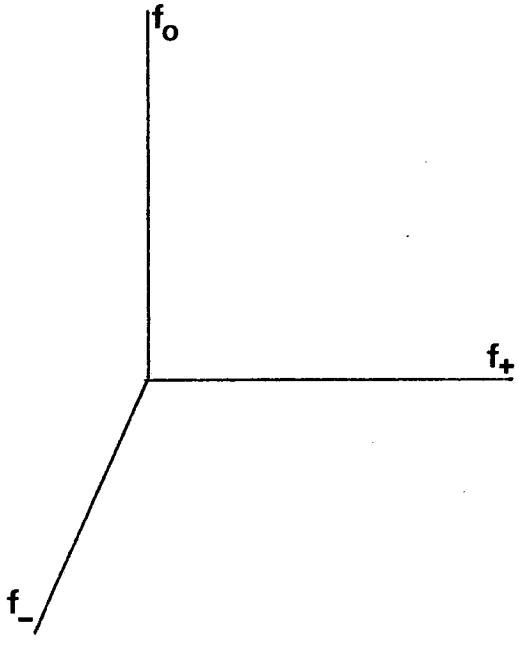

Figure 2a. Fraction space

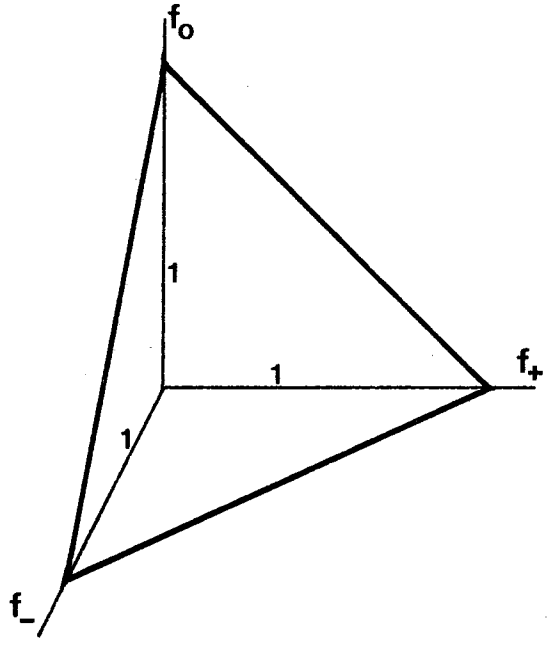

Figure 2b. The fraction plane in fraction space

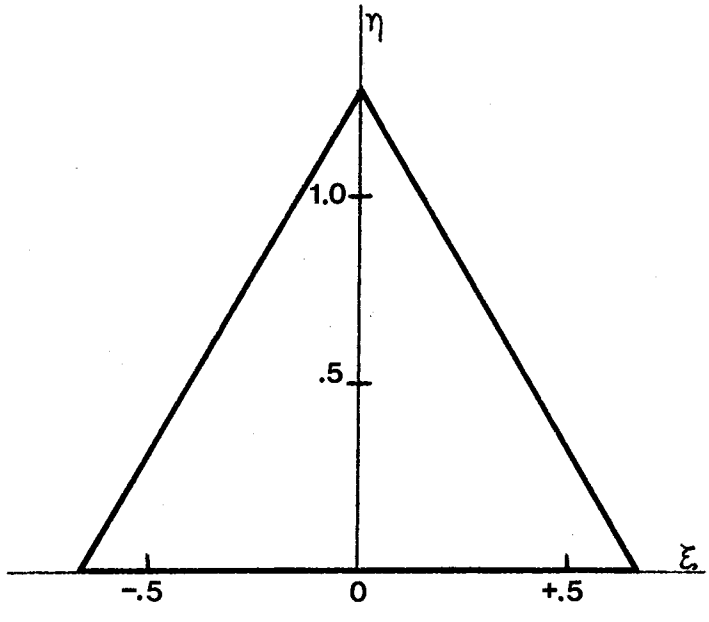

Figure 3. The fraction plane

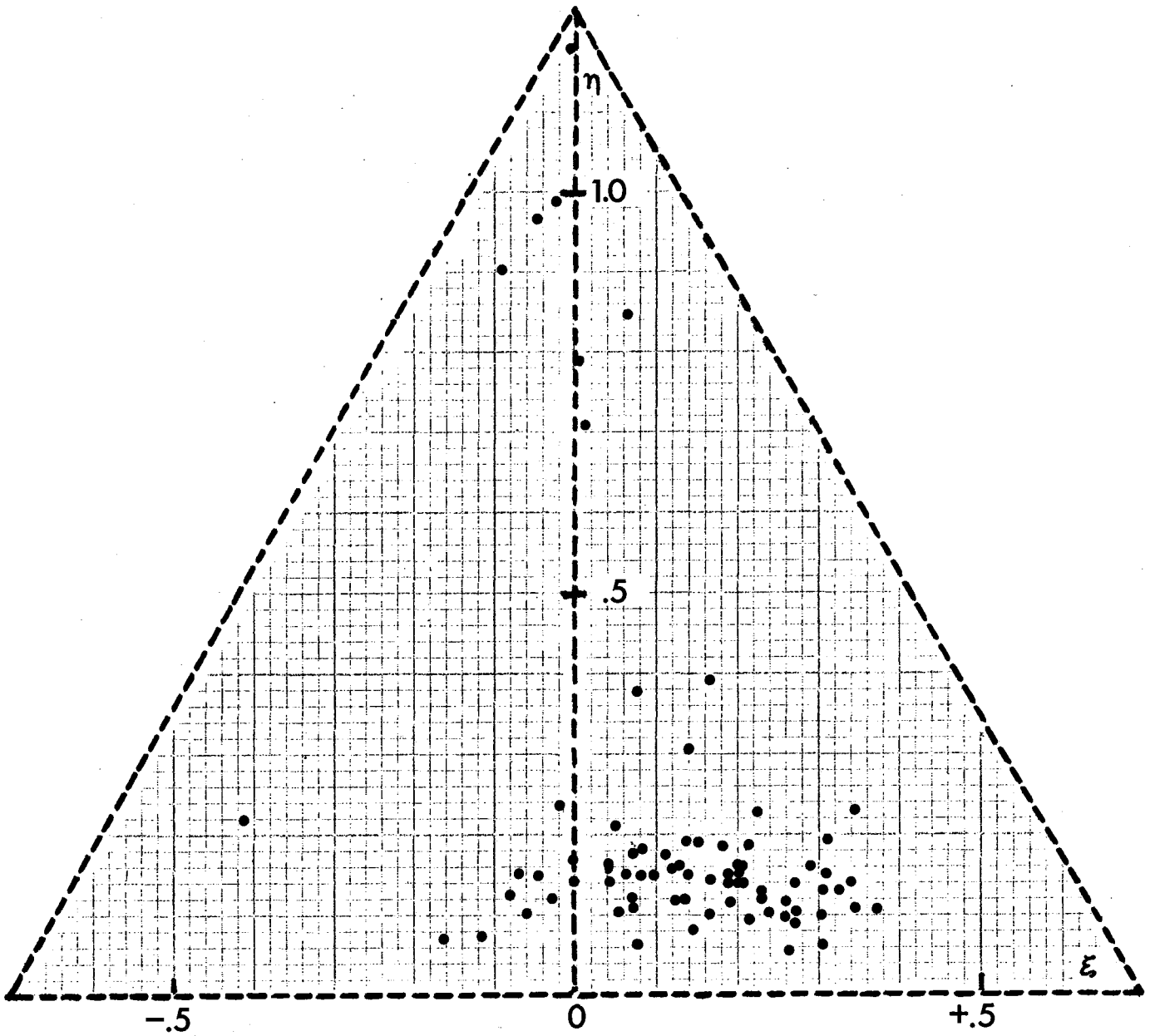

Figure 4. The (ξ,η) values for 78 technicians

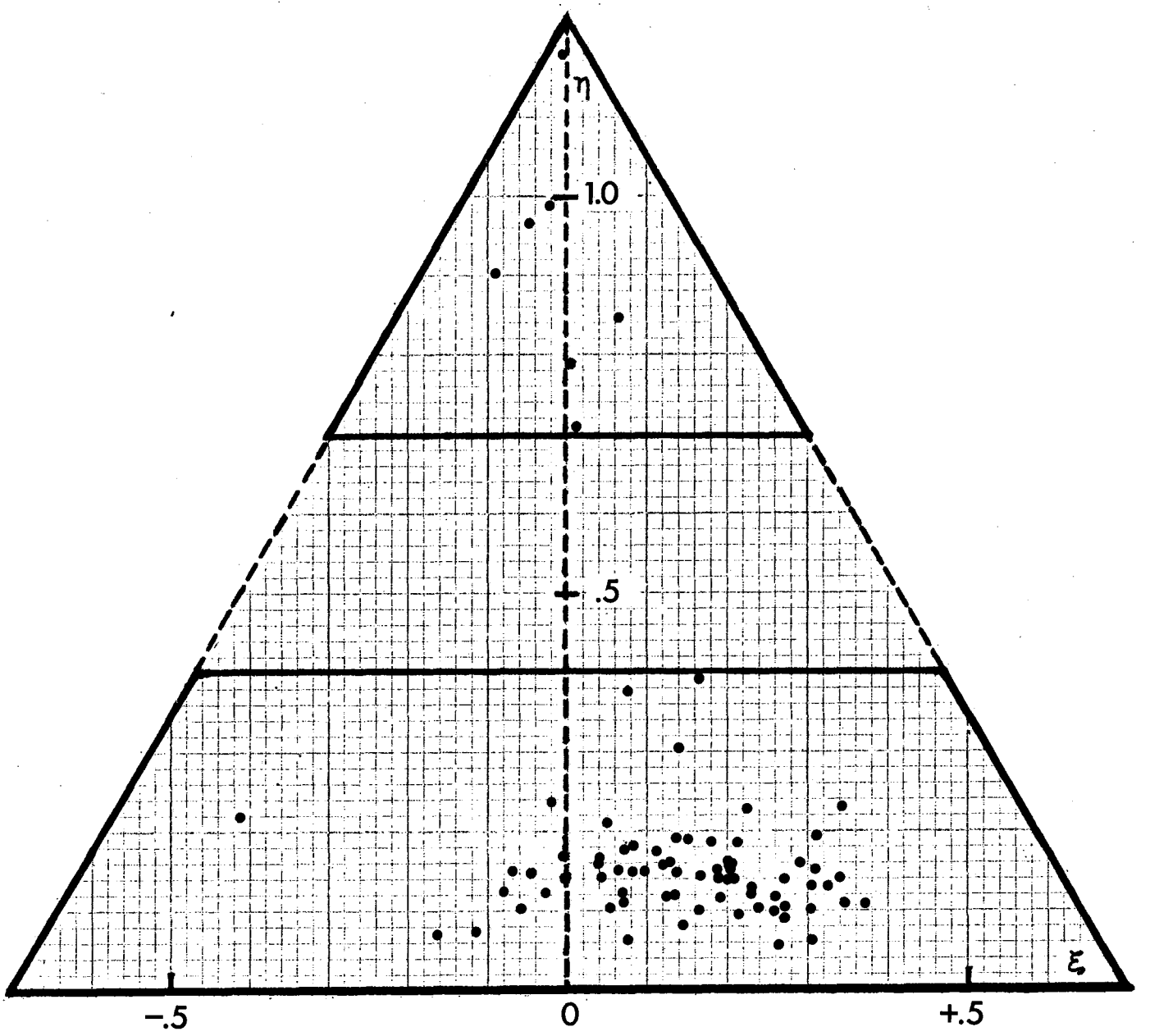

Figure 5a. Cluster analysis results: two clusters

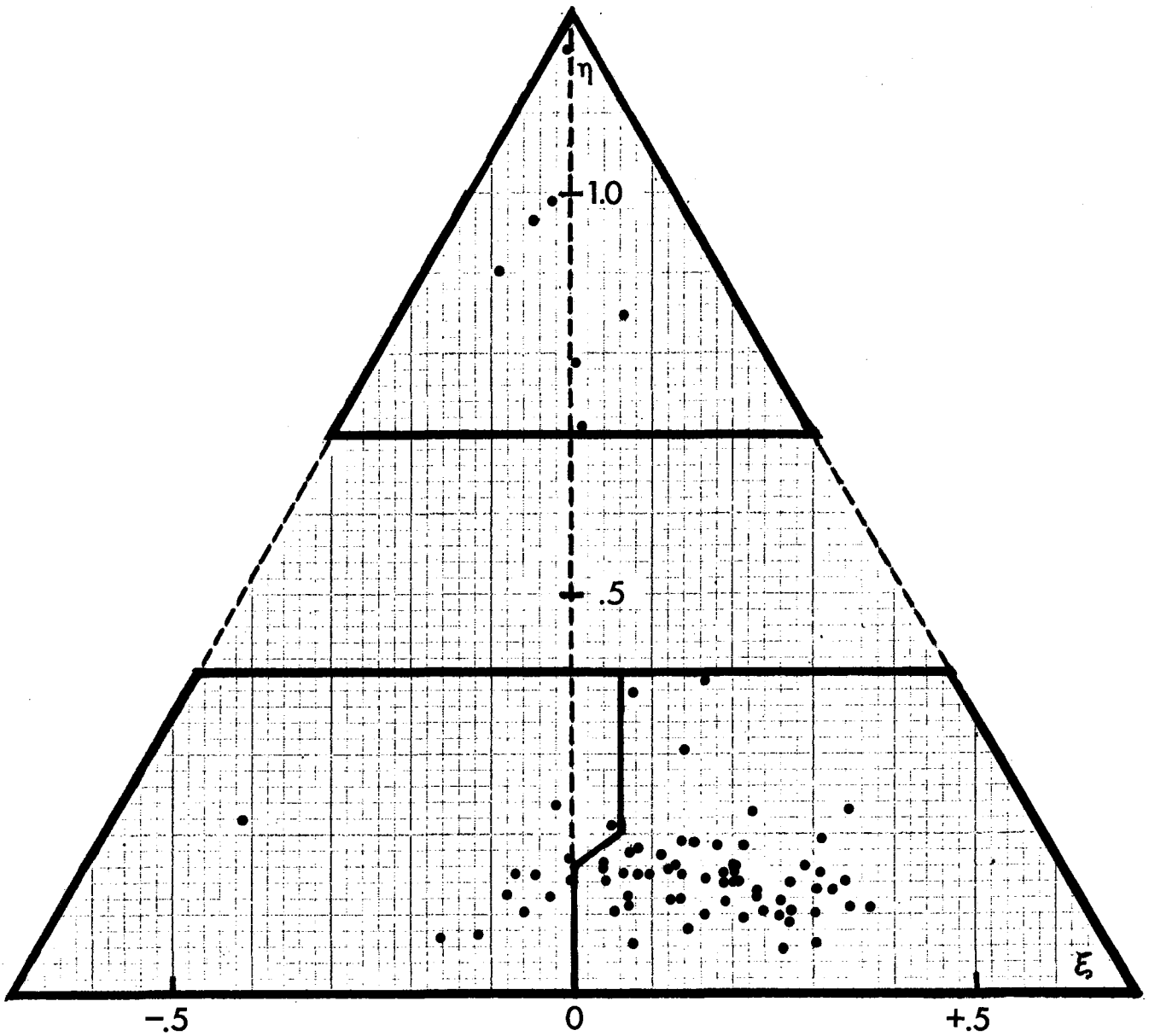

Figure 5b. Cluster analysis results: three clusters

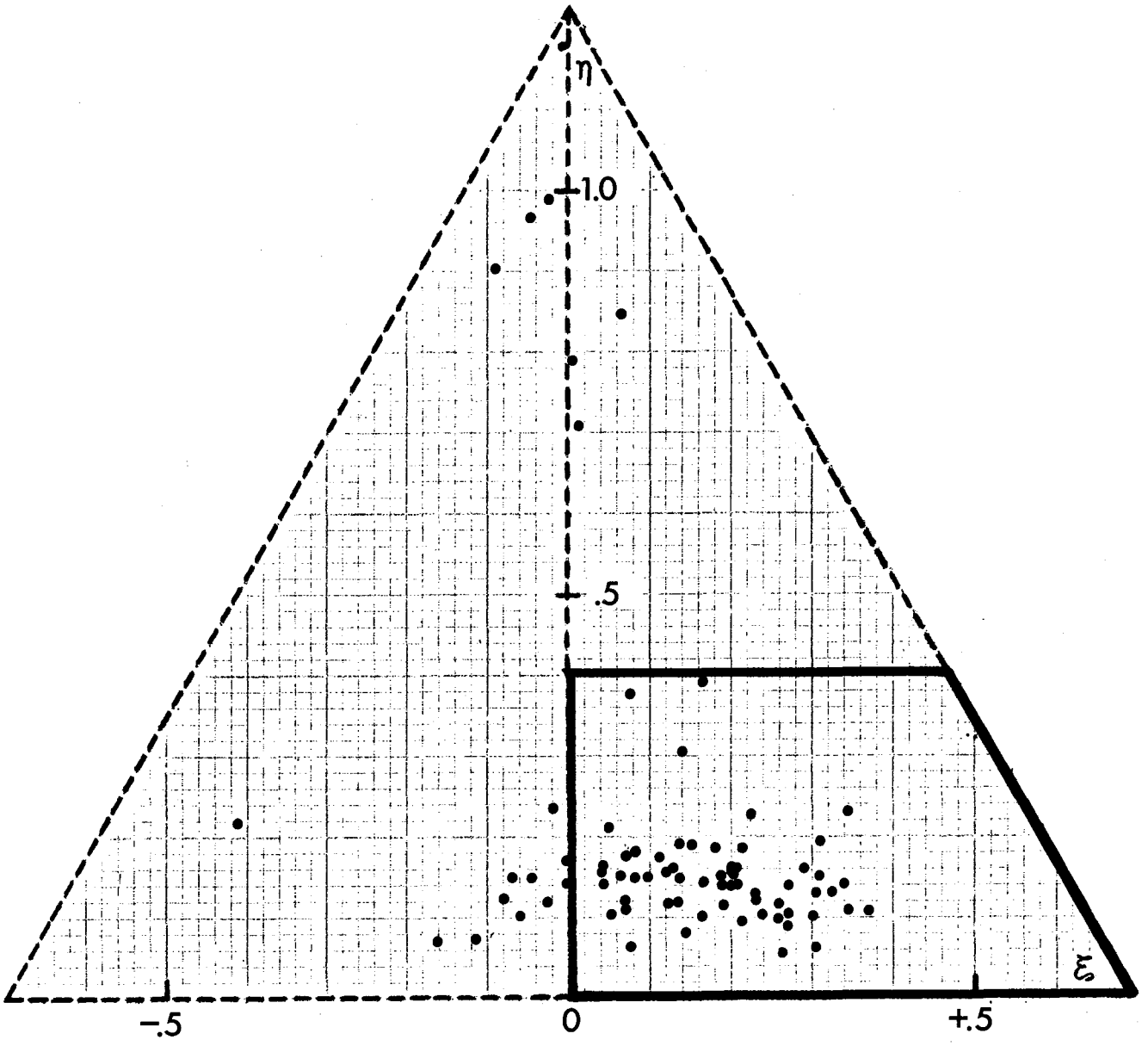

Figure 6. The acceptance region in the ξ-η plane