Extensions of Fractile Graphical Analysis

P. C. MAHALANOBIS
Indian Statistical Institute
INDIA

1. INTRODUCTION

The technique of fractile graphical analysis (F.G.A.), in the form of bivariate graphs, was developed and applied in the analysis of economic data in a series of papers (Mahalanobis, 1958a, 1958b, 1960, 1962). The wide applicability of this method was further demonstrated in recent papers by Linder (1963), Rhea Das (1960a, b, 1964) and others. Some conjectures made by the author were studied in a number of theoretical papers by Kawada (1961), Kitagawa (1960), Mitrofanova (1961), Sethuraman (1961, 1963) and Takeuchi (1961). The computational aspects through the use of unit record machines were examined by Roy and Kalyanasundaram (1963). The object of the present paper is to provide further theoretical foundations of extensions of F.G.A.

Genesis of the problem. Let us suppose that we wish to study the differences in the distribution of per capita consumption of cereals between two different regions or over time in a single region. It is a common practice in such cases to compare the mean values, standard deviations and other measures characterizing the distributions. Such overall comparisons may not be meaningful or completely informative especially when differences in the consumption of cereals are not the same for all 'comparable sub-sets' of the populations under comparison (in two regions or at different points of time). Thus it may be pertinent to ask whether there is any *differential* increase in the consumption of cereals between the 'poor' and 'rich' sections of the populations. Such a problem leads us to define comparable subsets of populations such as poor, rich and so on. For this purpose we use a suitable concomitant variable such as per capita income of an individual as an indication of economic status. We note that values of income in real terms, that is, making adjustments for changes in prices, at different times or in different regions are not, however, comparable and therefore comparable subsets cannot be defined as groups of individuals having the same per capita income. But groups of individuals having the same relative economic status in the two populations, as defined by ranks with respect to income within a population, may be amenable to meaningful comparisons. In other situations there may be other ways of defining comparable subsets of the populations. After choosing a number of comparable subsets on the basis of a concomitant variable, we examine the difference in the distributions of a main variable for every pair of comparable subsets. Fractile graphical analysis is a convenient technique by which the desired comparisons can be made through appropriate graphs drawn on the basis of sample data. The actual computations involved when there are one main variable and one concomitant variable are briefly explained in Section 2, and certain generalizations to cases inclusive of mean values, standard deviations etc. are discussed in other sections.

2. BIVARIATE FRACTILE GRAPH

A fractile graph in the case of two dimensional data is defined as follows. Let (Y, X) denote two variables and $(Y_1, X_1), \ldots, (Y_N, X_N)$, be N independent observations. One particular variate, say X, is selected for ranking the observations in ascending order. Replacing the X values by ranks and arranging the observations in ascending order of X, we obtain

$$(2.1) \qquad (Y_{(1)}, 1), (Y_{(2)}, 2), \ldots, (Y_{(N)}, N)$$

where $Y_{(i)}$ is the Y value associated with the X value of rank i. Now divide the observations (2.1) into a chosen number, g, of groups such that each group consists of $h = N/g$ consecutive observations. These are called fractile groups. The ith fractile group represented by $[i]$ consists of the observations

$$(2.2) \quad (Y_{(ih-h+1)}, ih-h+1), (Y_{(ih-h+2)}, ih-h+2), \ldots, (Y_{(ih)}, ih)$$

which are replaced by the pair

$$(2.3) \qquad (Y_{[i]}, i)$$

where i represents the ith fractile group and $Y_{[i]}$ is any statistic (such as the mean, median, maximum, etc.) based on the Y values of the observations in $[i]$,

$$(2.4) \qquad Y_{(ih-h+1)}, \ldots, Y_{(ih)}.$$

The g pairs

$$(2.5) \qquad (Y_{[1]}, 1), \ldots, (Y_{[g]}, g)$$

provide the fractile graph, by plotting $Y_{[i]}$ against i, $i = 1, \ldots, g$ and joining the successive points by straight lines.

The graph so obtained is represented by G_g. The fractile graph for the entire population, using the same procedure for all members of the population, may be designated Γ_g. As the sample size increases, G_g will provide a consistent estimator of Γ_g.

Separation between graphs. To compare two fractile graphs based on independent samples from two different populations it is necessary to consider the difference between graphs of parallel samples from the same population. For this purpose we divide each sample into two independent halves. The graphs of the two half samples from the first population are denoted by G_{g1} and G_{g2} and that of the entire sample by G_g. Similarly we have the corresponding graphs G'_{g1}, G'_{g2} and G'_g for a sample from the second population. We then choose a measure of separation $\|A-B\|$ between any two graphs A and B. To test the significance of the observed difference between G_g and G'_g we may use a statistic of the type

$$(2.6) \qquad M = 4\sqrt{\frac{n\,n'}{n+n'}}\,\frac{\|G_g - G'_g\|}{(\sqrt{n}\|G_{g1}-G_{g2}\| + \sqrt{n'}\|G'_{g1}-G'_{g2}\|)}$$

The overall measure of separation, proposed in the earlier papers, is the area between graphs. The exact distribution of M is unknown.

The purpose of drawing the fractile graphs is not, however, only to make an overall comparison over the whole range as provided by a statistic of the type (2.6). It also seems possible to compare the graphs at each fractile point or at sets of consecutive fractile points and draw inferences. If necessary, a test of

the type (2.6) may be used for particular sections of the graphs to examine the significance of the observed differences.

In practice a significance test is hardly necessary when there is a clear separation of the graphs, over the whole range or over any portion of the graphs, indicated by the fractile points for the two halves of one sample being completely above or below those for the other sample.

The F.G.A. can be used in any situation in which parallel or interpenetrating network of samples (I.P.N.S.) can be drawn, for example, in the study of consumption of cereals in India (Mahalanobis, 1962) or in testing the normality of frequency distributions by Linder (1963). The F.G.A. can also be used to test the significance of differences in concentration curves (Mahalanobis, 1960)[1].

[1] It may be noted that the F.G.A. is based on concepts which have no connection with concentration curves. The innovation in F.G.A. is the introduction of the concept of a graphical error. It is therefore possible to carry out tests of significance with F.G.A. which is not possible with a concentration curve. This crucial point of using the graphical error for the tests of significance was missed by Swamy (1963).

Surmises. The exact distribution of the separation in FGA is not known. The author made some surmises (Mahalanobis 1958b) which have been later approximately verified by model sampling experiments.

The error area e between two fractile graphs (with sample sizes N_1 and N_2, and g the fixed number of fractile groups for both graphs) would tend to decrease as $\sqrt{(N_1+N_2)/N_1N_2}$, and to increase proportionately to g. Also, as each fractile group consists of $h = N/g$ observations, it follows that, when sample sizes are kept constant, then the error area e would tend to vary approximately as $g^{3/2}$. Also, if g is changed to gk, then changes in the error area would vary proportionately to $k^{3/2}$, which is a most useful property in testing the significance of the separation by changing values of g.

3. EXTENSION OF F.G.A TO METRICISED CLASSIFICATORY VARIATE.

The method of F.G.A. has been used so far by superposing the fractile groups for the two sets of samples from two populations, and drawing all fractile graphs in such superposed positions. In this procedure, as the ranking order of x alone is taken into consideration, the values of each Y_i and Y'_i etc can be plotted against any set of i points on the x-scale, for example, at equal or at random intervals. The surmises mentioned above would still hold good.

The object of the extension of F.G.A., discussed in the present paper, is to include in the comparisons the two mean values and the two standard deviations of the x-variate for the two combined samples. If it is assumed that the distribution of x is normal, then the appropriate (but to some extent approximate) procedure would be to plot values of Y_i and Y'_i for the two sub-samples and the combined sample from both Lot 1 and Lot 2 on the normal-probit points on the x-scale, that is, against 1.75, 0.68 etc.

Tests of Normality. As a prior step, it is possible to use F.G.A. to test the normality of the x-distribution, following the procedure of Linder (1963). It is desired to make a bivariate (x, y) comparison for two samples, called Lot 1 and Lot 2, from two "populations", processed under different conditions of treatment. Two sub-samples, of 100 each, were drawn from Lot 1 and Lot 2, and measurements were taken of x and y on a percentage scale. Variate x (represented on the x-scale in all accompanying charts) was chosen as the ranking variate, that is, as the 'independent' variate, to study changes in y with variations of x.

Each of two sub-samples drawn from Lot 1 was separately ranked in ascending order of the value of x, and was divided into 10 decile groups, each consisting of 10 observations; the two sub-samples were pooled, again ranked, and divided into 10 decile groups. The basic data are given in the form of average values of x and y for each decile, separately, for sub-sample 1 (s.s.1), sub-sample 2 (s.s.2), and the combined sample in Table (0.1) for Lot 1. Similar data for Lot 2 are given in Table (0.2).

Average values of x and y
in fractile groups based on x-ranking

serial number	s.s.1		s.s.2		combined	
	average values		average values		average values	
	x	y	x	y	x	y
	Table (0.1) Lot 1					
1	42.0	7.45	39.9	6.12	40.9	6.75
2	48.8	8.00	46.7	6.30	47.7	7.11
3	53.0	7.05	51.9	6.82	52.6	7.01
4	55.0	6.65	57.5	8.60	55.7	7.53
5	58.1	7.58	62.1	8.15	60.2	8.18
6	63.0	7.45	65.1	7.70	64.3	7.66
7	67.0	8.18	68.9	8.18	67.7	8.44
8	68.9	8.70	72.5	8.30	70.8	7.70
9	73.7	7.28	79.3	8.85	76.2	8.44
10	82.4	9.30	90.4	9.30	86.9	9.15
	Table (0.2) Lot 2					
1	36.5	8.20	33.3	8.62	34.8	8.22
2	40.6	6.95	41.7	8.40	41.2	7.80
3	45.8	8.48	47.0	8.52	46.4	8.56
4	51.1	10.52	50.9	9.20	51.0	9.86
5	53.4	8.20	54.9	10.38	53.9	9.46
6	56.4	10.45	58.8	9.42	57.8	9.96
7	60.0	9.55	63.1	11.48	61.1	10.46
8	64.2	11.08	66.8	10.80	65.8	10.55
9	71.4	11.18	72.6	10.50	72.0	10.97
10	82.7	15.15	85.8	15.25	84.2	15.30

The mean value and the standard deviation of the combined sample were found to be 62.31 and 13.34 respectively for Lot 1. The corresponding normal-probit points, on the above basis, were drawn on the x-scale, and are shown as 1.75, 1.05, 0.68, 0.39, 0.13, minus and plus, in Chart (1). These x-values are given for each decile in col. (2) of Table (1.1). The corresponding observed mean values of x for each decile are given, in cols. (3), (4) and (5) of the same Table (1.1).

Table (1.1) Lot 1 : Test of normality of x-distribution

serial no. of decile fractile groups	mean values of decile fractile groups			
	x-scale for normal distribution	y-scale for observed mean values of y		
		s.s.1	s.s.2	combined
(1)	(2)	(3)	(4)	(5)
1	38.9	42.0	39.9	40.9
2	48.4	48.8	46.7	47.7
3	53.3	53.0	51.9	52.6
4	57.1	55.0	57.5	55.7
5	60.6	58.1	62.1	60.2
6	64.0	63.0	65.1	64.3
7	67.5	67.0	68.9	67.7
8	71.3	68.9	72.5	70.8
9	76.2	73.7	79.3	76.2
10	85.7	82.4	90.4	86.9

The x-scale gives the location of mean values of decile fractile groups for a normal distribution with observed mean value = 62.31 and standard deviation = 13.34

The observed mean values of x for s.s.1 in col. (3) are plotted on the y-scale on corresponding normal-probit points given in col. (2); and adjoining points of y are joined by a straight line to give the fractile graph for s.s.1 for Lot 1, as shown in the lower part of Chart 1.1. The fractile graphs for s.s.2 and for the combined sample are plotted

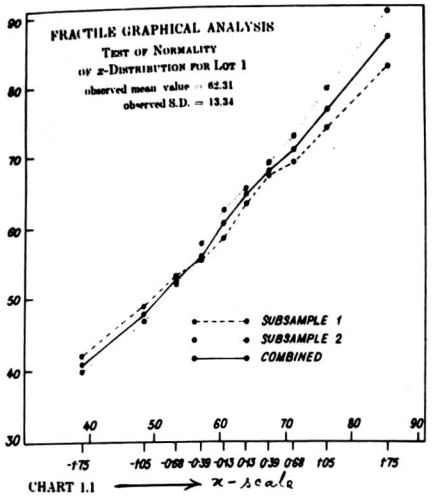

CHART 1.1

in the same way. The area between the two sub-sample graphs gives the error-area associated with the combined fractile graph for Lot 1. From Chart 1, it can be seen, that a straight line can be drawn lying within the error-area over practically the whole range with the exception of the bottom decile which deviates a little upwards. The distribution of x may therefore be considered approximately normal with some deviation at the lower end. It is also seen that the error-area increases indicating wider dispersion towards the top end.

Data for Lot 2 were processed in the same way. The observed mean value was 56.83, and the standard deviation 14.31. Relevant figures for the normal-probit points on the x-scale, and corresponding mean values of x for each decile, are shown in Table (1.2). The fractile graphs are shown in Chart

Table (1.2) Lot 2 : Test of normality of x-distribution

serial no. of decile fractile groups	mean values of decile fractile groups			
	x-scale for normal distribution	y-scale for observed mean values of y		
		s.s.1	s.s.2	combined
(1)	(2)	(3)	(4)	(5)
1	31.7	36.5	33.3	34.3
2	41.9	40.6	41.7	41.2
3	47.1	45.3	47.0	46.4
4	51.3	51.1	50.9	51.0
5	55.0	53.4	54.9	53.9
6	58.6	56.4	58.3	57.8
7	62.4	60.0	63.1	61.1
8	66.5	64.2	66.3	65.8
9	71.8	71.4	72.6	72.0
10	81.9	82.7	85.3	84.2

The x-scale gives the location of mean values of decile fractile groups for a normal distribution with observed mean value = 56.83 and standard deviation = 14.31

(1.2), from which it can be seen that a straight line can be drawn within the error-area, over the whole range except again for the bottom decile.

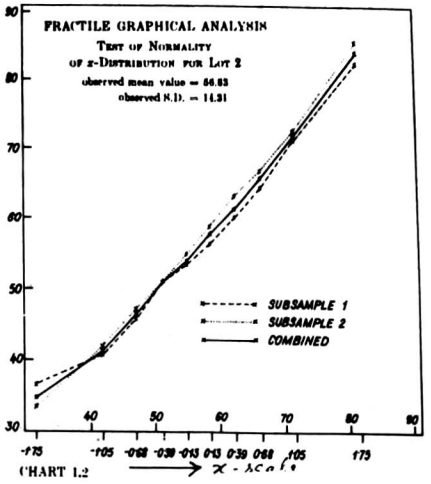

CHART 1.2

Generalised F.G.A. method for testing distributions. It is possible to generalise the F.G.A. method for testing the normality of a distribution. Fractile groups can be always formed for samples of observations whatever be the distribution of the variate. Also, for any distribution ϕ of a variate x, there is a theoretical mean value of x for each fractile group which can be located on the x-scale. These points may be called the ϕ-probit points on the x-scale. Also, for any distribution ϕ of x, if the theoretical mean values for different fractile groups are plotted on the y-scale on the corresponding ϕ-probit points on the x-scale, then all the y-points must lie on a straight line. This property makes it possible to use F.G.A. for testing the goodness of fit of a given ϕ-distribution to any set of observation. It may be noted that I have used the word 'probit' in a much more general sense than is the usual practice.

Bivariate metricised F.G.A. of y on x. As the distribution of x is approximately normal for both Lot 1 and Lot 2, it is possible to plot the mean value of y for each fractile group on the corresponding normal-probit point on the x-scale, and to obtain the fractile graph by joining adjoining y-points. The normal-probit point on the x-scale for each decile group of Lot 1

Table (2.1) Lot 1 : Bivariate Fractile Graphical Analysis of y on x

serial no. of decile fractile groups	decile fractile groups based on x-ranking			
	x-scale for normal probits	y-scale for observed mean values of y		
		s.s.1	s.s.2	combined
(1)	(2)	(3)	(4)	(5)
1	40.4	7.45	6.12	6.75
2	48.5	6.00	6.30	7.11
3	53.3	7.05	6.82	7.01
4	57.2	6.65	8.60	7.53
5	60.6	7.58	6.15	8.18
6	64.0	7.45	7.70	7.66
7	67.4	8.18	8.18	8.44
8	71.3	8.70	8.30	7.70
9	76.1	7.28	8.85	8.44
10	84.3	9.30	9.30	9.15

is given in col (2) of Table (2.1), and the corresponding observed mean values of y are given for s.s. 1, s.s. 2, and the combined sample respectively in col. (3), (4) and (5) of Table (2.1). The fractile graphs are obtained by plotting y-values on the corresponding normal probit-points on the x-scale, and joining adjacent y points. The three graphs for Lot 1 are shown in the lower part of Chart 2.

Similar data for Lot 2 are given in Table (2.2), and the fractile graphs are shown in the upper part of Chart 2. Because the mean value and standard deviation of Lot 1 are appreciably different from the mean value of standard deviation of Lot 2, the end points Y_1 and Y'_1, in the first decile, do not lie on the same point on the x-scale. It is possible, at this stage, to introduce a simple rule of construction, namely, to join the end-points by a horizontal and a vertical line further away from the end-points. In this case a horizontal line is drawn from Y_1 to the point y_1 on which a vertical line can be dropped from Y'_1. At the top end, a vertical line is drawn from Y'_{10} to the point y_{10} which can be joined by a horizontal line with Y_{10}. The 'separation' between Lot 1 and Lot 2 can be defined as the area bounded by the two combined graphs and by two horizontal and two vertical lines drawn under the above rule of construction, that is, the area, $y_1\ Y'_1\ Y'_{10}\ y_{10}\ Y_{10}\ Y_1 y_1$. The two error-areas for Lot 1 and 2 do not overlap except for a small portion at the bottom end. The 'separation' between Lot 1 and Lot 2, in respect of y on x, may be, therefore, considered to be significant. It is also seen that y increases with x at an appreciably greater rate for Lot 2 in comparison with Lot 1, showing that y is more strongly associated with x for Lot 2.

Table (2.2) Lot 2 : Bivariate Fractile Graphical Analysis of y on x

serial no. of decile fractile groups	decile fractile groups based on x-ranking			
	x-scale for normal probits	y-scale for observed mean values of y		
		s.s.1	s.s.2	combined
(1)	(2)	(3)	(4)	(5)
1	33.3	8.20	8.02	8.22
2	42.0	6.95	8.40	7.60
3	47.2	8.48	6.52	8.56
4	51.3	10.52	9.20	9.66
5	55.0	8.20	10.38	9.46
6	58.6	10.45	9.42	9.96
7	62.3	9.55	11.48	10.46
8	66.5	11.08	10.80	10.55
9	71.7	11.18	10.50	10.97
10	80.4	15.15	15.25	15.30

It would be noted that in Tables (2.1) and (2.2) and in Chart 2, the normal-probit points on the x-scale, for the combined samples, have been used for plotting not only the values of y for the combined samples but also for plotting the values of y for the two sub-samples for both Lot 1 and Lot 2. The approximation introduced by this procedure can be easily removed. The normal-probit points on the x-scale for sub-sample 1 can be used for plotting the values of y for sub-sample 1, and a similar procedure can be adopted for plotting values of y for sub-sample 2. The error-area between the fractile graphs for two sub-samples can be then formed by using the same construction as has been already described for forming the area of the separation between the fractile graphs for the combined samples for Lot 1 and Lot 2. A consistent method can be, therefore, developed for error-areas and separation by using the above procedure.

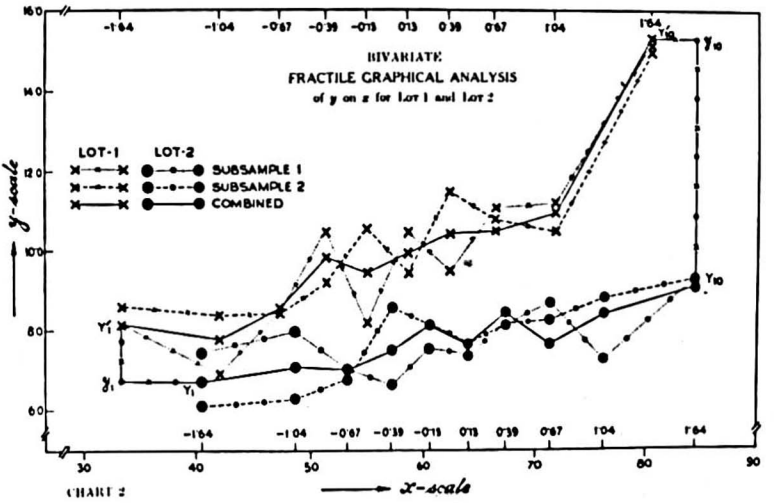

CHART 2

References

Das, Rhea S. (1960a). "Applications of Fractile Graphical Analysis to Psychometry : I– Item Analysis," *Psychological Studies*, 5, 11–18.

Das, Rhea S. and Sharma, K. N. (1960b). "Applications of Fractile Graphical Analysis to Psychometry : II– Reliability," *Psychological Studies*, 5, 71–77.

Das, Rhea S. (1964). "Item Analysis by Probit and Fractile Graphical Methods," *Brit. J. Statist. Psych.*, 17, 51–64.

Kawada, Y. (1961). "Some Remarks Concerning the Expectation of the Error Area in Fractile Analysis," *Sankhyā, Ser. A*, 23, 155–160.

Kitagawa, T. (1960). "Sampling Distributions of Statistics Associated with a Fractile Graphic Method," *Bull. Math. Statist.*, Fukuoka, Japan, 9, 10–42.

Linder, A. (1963). Convocation address. Indian Statistical Institute.

Mahalanobis, P. C. (1958a). "A Method for Fractile Graphical Analysis with Some Surmises of Results," *Transactions of the Bose Research Institute*, Calcutta, 22, 223–230.

Mahalanobis, P. C. (1958b). Lectures in Japan : Fractile Graphical Analysis, Indian Statistical Institute.

Mahalanobis, P. C. (1960). "A Method of Fractile Graphical Analysis," *Econometrica*, 28, 325–351.

Mahalanobis, P. C. (1962). "A Preliminary Note on the Consumption of Cereals in India," *Bull. Inst. Internat. Statist.*, 39, 53–76.

Mitrofanova, N. M. (1961). "On Some Problems of Fractile Graphical Analysis," *Sankhyā, Ser. A*, 23, 145–154.

Roy, J. and Kalyanasundaram, G. (1963). "Punched Card Processing of Sample Survey Data for Fractile Graphical Analysis," *Contributions to Statistics*, (Volume presented to Professor P. C. Mahalanobis on the occasion of his 70th birthday), Statistical Publishing Society, Calcutta, 411–418.

Sethuraman, J. (1961). "Some Limit Distributions Connected with Fractile Graphical Analysis," *Sankhyā, Ser. A*, 23, 79–90.

Sethuraman, J. (1963). "Fixed Interval Analysis and Fractile Analysis," *Contributions to Statistics*, (Volume presented to Professor P. C. Mahalanobis on the occasion of his 70th birthday) Statistical Publishing Society, Calcutta, 449–470.

Takeuchi, K. (1961). "On Some Properties of Error Area in the Fractile Graph Method," *Sankhyā, Ser. A*, 23, 65–78.