ACHARYA JAGADISH CHANDRA BOSE
BIRTH CENTENARY

Vol. XXII, 1958, of the

TRANSACTIONS OF THE
BOSE RESEARCH INSTITUTE
CALCUTTA

XXI. A Method of Fractile Graphical Analysis with some Surmises of Results

By P. C. Mahalanobis

XXI. A METHOD OF FRACTILE GRAPHICAL ANALYSIS WITH SOME SURMISES OF RESULTS

By P. C. MAHALANOBIS

Indian Statistical Institute, Calcutta

(*Received for publication on November 18, 1958*)

1.1. Interpenetrating samples. Consider any bivarite 'population' or 'universe' of two random variates, x and y, with or without correlation. Draw two interpenetrating samples from this population in accordance with any 'probability design'[*] and with replacement.

1.2. These two interpenetrating samples would be statistically equivalent and would supply equally valid estimates. These two samples may also be called 'sub-samples' to emphasize the fact that they can be pooled together to give one complete sample.

2.1. Fractile groups. Suppose each sub-sample consists of n elementary units, each unit being a pair of values of the two variates x and y. Consider the first sub-sample of, say, n sample units. Rank them in order of ascending values of x.

2.2. It is now possible to divide the n units into g groups $(1, 2, \ldots i, \ldots g)$, each of equal number, say, n'; so that $n = g \,.\, n'$. These may be called fractile groups.

3.1. Fractile graphs. Next calculate the mean value (or median) of n' values of y in each group to give in the first sub-sample the values $y'_1, y'_2, y'_3, \ldots, y'_g$ corresponding to the serial number of the group $1, 2, 3, \ldots i, \ldots g$.

3.2. Take g equi-distant points $(1, 2, 3, \ldots i, \ldots g)$ on the x-axis to represent the g groups; and plot the corresponding values of $y'_1, y'_2, \ldots, y'_g$. Finally, draw straight lines to join each pair of adjoining points y'_1 and y'_2; y'_2 and $y'_3, \ldots y'_{i-1}$ and $y'_i, \ldots, y'_{g-1}$ and y'_g. This connected chain of lines will be called the fractile graph $G(1)$.

3.3. Now consider the second sub-sample. It is possible to go through a similar process of ranking the sample units in order of ascending values of x; dividing them after ranking into g groups each of equal number n'; calculating the values of $y''_1, y''_2, \ldots, y''_g$; and drawing a second fractile graph $G(2)$. We can then have two sub-sample graphs $G(1)$ and $G(2)$, which have equal statistical validity.

3.4. It is also possible to mix together the two sub-samples to form a combined sample, and rank the sample units again; divide into g groups of equal number $(2n')$; calculate the new values of $y_1, y_2, \ldots, y_g$; and draw the fractile graph $G(1, 2)$ for the two sub-samples taken together, that is, for the combined sample.

3.5. It should be observed that each of the g groups would consist of the same proportion of the total sample units in all three cases (that is, in the first sub-sample, in the second sub-sample, and in the combined sample); but the end-points of values of x for any particular group would be different, in general, in the three cases.

4.1. Error area. It is possible to measure on paper the area bounded by the two sub-sample fractile graphs $G(1)$ and $G(2)$. We shall (semi-intuitively) call this area $a(1, 2)$ as the 'error' to be associated with the combined fractile graph $G(1, 2)$.

[*] Which would enable valid estimates being made regarding the population characteristics in accordance with the theory of probability.

4.2. The combined fractile graph $G(1, 2)$ will lie, in general, partly within and partly outside the area $a(1, 2)$; and can lie wholly inside or wholly outside this area in particular cases. It is possible also to measure the area bounded by the combined fractile graph $G(1, 2)$ which lies outside the area $a(1, 2)$; and this area may be called b_{12}.

4.3. It would be observed that the two sub-sample fractile graphs $G(1)$ and $G(2)$ will either not intersect at all, or will intersect in $1, 2, \ldots,$ or $(g-1)$ points. For each pair of sub-sample graphs there would be a definite number of points of intersection $(0, 1, 2, \ldots, g-1)$.

5.1. Separation. It is further possible to consider a second population from which a pair of interpenetrating sub-samples are drawn, and a second set of fractile graphs, say $G'(1)$, $G'(2)$ and $G'(1, 2)$, are constructed in exactly the same way. The area bounded by $G'(1)$ and $G'(2)$ would give the second error area $a'(1, 2)$ to be associated with the second pooled graph $G'(1, 2)$.

5.2. It is now possible to measure the area lying between the two combined fractile graphs $G(1, 2)$ and $G'(1, 2)$; and this area may be called the 'separation' between $G(1, 2)$ and $G'(1, 2)$ and may be written as, say, $S(1, 2)$; or simply, S, when samples from only two populations are under consideration.

5.3. It may be observed that it is always possible to draw the fractile graphs $G(1)$, $G(2)$, $G(1, 2)$, $G'(1)$, $G'(2)$ and $G'(1, 2)$; and to measure the error areas $a(1, 2)$ and $a'(1, 2)$ and also to measure the separation area $S(1, 2)$. All these quantities are thus operationally defined; and can be studied by experimental sampling. (*See* the charts J.F.G. 1, 2, 3, 4 at the end of the paper.)

6.1. Surmises. I shall now give a number of semi-intuitive surmises * some of which are being experimentally studied. I may mention that these experimental results, although still very meagre, do not contradict the surmises.

6.2. The area $a(1, 2)$ which has been called the error associated with $G(1, 2)$ would decrease statistically in proportion to $\frac{1}{\sqrt{n'}}$ with increasing size of the sample n' of each group (when g is constant); and would increase in proportion to g (when n' is kept constant); as a first approximation.

6.3. The combined fractile graph $G(1, 2)$ would tend statistically to lie more and more within the area $a(1, 2)$ with increasing values of n' (with g constant).

6.4. The number of points of intersection of $G(1)$ and $G(2)$ would tend statistically to be distributed like changes in 'runs' of heads and tails in g throws of an unbiased coin.

6.5. The error to be associated with the 'Separation', $S(1, 2)$, to be called, say, E, can be found in the usual way from the two error areas, $a(1, 2)$ and $a'(1, 2)$, associated respectively with the two combined fractile graphs, $G(1, 2)$ and $G'(1, 2)$, representing respectively the two populations from which the two pairs of sub-samples are drawn. That is, it is possible to take $E = \sqrt{[a^2(1, 2)+a'^2(1, 2)]}$.

6.6. To test the significance of the observed separation, it is possible to use the criterion S^2/E^2 which would tend to be distributed, as a first approximation, like Chi-square.

* Which were first stated in a lecture at the Indian Statistical Institute, Calcutta, in April 1958; and were repeated at Berkeley, Chicago, East Lansing, New York and other places in the U.S.A. in May and June 1958.

6.7. All the above results would remain valid even when the two sets of sub-sample values are subjected to any combination of linear and non-linear transformations which do not disturb the ranking of the variates.

7.1. Further observations. Although I have so far assumed that values of x and y were given in the form of measurements, this is not necessary. For the x-variate, it is sufficient that the observations are capable of being ranked as a whole. Also, let us assume that the y-variate (also cannot be measured but) can be ranked for the whole set of $2n$ sample units. After dividing the sample units into g groups (in all three cases, namely, in the first sub-sample, in the second sub-sample and in the combined sample), each group would consist of n' units in the first and second sub-samples and $2n'$ units in the combined sample, and each of these sample units will have a rank-number for y. It would be then possible to calculate the average of the rank-numbers of y (of the median rank-number) in each group; to plot these values; and draw the graphs as already explained. The above results would, therefore, hold also for variates which can be ranked but cannot be measured.

7.2. Secondly, the results given above would hold for all functional relations between x and y, and also when x and y are statistically independent.

7.3. The results can be extended when y is given in the form of observations of a 'time-series' with x as co-ordinates of time. The only modification would be that the values of y (in the first and the second sub-samples) would not refer to the same time-interval (which would constitute the group) but to a system of alternate time-intervals. The procedure in this case would be extremely simple. The series of observations would be automatically ranked in order of x (representing time). Either single values of y, or the average of values of y occurring in a time-interval, would have to be plotted. Each sub-sample would then consist of one alternate set of observations (that is, of observations occurring in one of two alternate sets of intervals of time); and the graphs can be constructed by joining two adjoining points of observation in each sub-sample or in the combined sample.

8. The results given in this note are of great generality and also have invariance of a high order. It would be of interest either to prove or to disprove all or some of the results. If any of the surmises given here is correct as a first approximation, it would be also of interest to find more accurate results either generally or under different sets of specified conditions regarding population, design of sampling, type of transformation, type of functional relation between the variates; and in the case of time-series, the magnitude and pattern of the system of time-intervals used to record the observations.

9.1. Model sampling experiments. The surmises given in paragraph 6 are being now studied by model sampling experiments in the Indian Statistical Institute. One series of experiments consisted of drawing a pair of sub-samples from a (theoretical) bivariate normal population with—

mean value of x	$= 0$
mean value of y	$= 0$
standard deviation of x	$= 1$
standard deviation of y	$= \sqrt{2}$
coefficient of correlation	$= 1/\sqrt{2}$.

9.2. The results of 18 experiments are given in the following Table (J.F.G. 1).

The number of groups, g, is shown in each case in col. (2); the number of each group, n', in col. (3); and the size of each sub-sample, $n = n' . g$, in col. (4). The next col. (5)

TABLE (J.F.G. 1)

Fractile Graphical Analysis : Results of Model Sampling Experiments : Calcutta 1958

Sampling from bivariate normal population : $m_x = m_y = 0$;

$$\sigma_x = 1,\ \sigma_y = \sqrt{2} = 1{\cdot}41\ ;\ \rho_{xy} = \frac{1}{\sqrt{2}} = 0{\cdot}71$$

Sl. No.	Number				Area between two sub-sample graphs		Function of area between sub-sample graphs	
	of groups (g)	in each group (n')	in each sub-sample ($n=n'.g$)	of model experiments (m)	average ($\bar{a}$)	standard deviation $s(a)$	$\sqrt{n'}\,[\bar{a} \pm s(\bar{a})]$	$\frac{\sqrt{n'}}{g}\,[\bar{a} \pm s(\bar{a})]$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	2	24	48	48	0·254	0·173	1·24 ± 0·12	0·62 ± 0·06
2	2	48	96	48	0·219	0·140	1·52 ± 0·14	0·76 ± 0·07
3	4	12	48	48	1·09	0·569	3·78 ± 0·28	0·94 ± 0·07
4	4	24	96	48	0·767	0·347	3·76 ± 0·25	0·94 ± 0·06
5	4	48	192	25	0·629	0·147	4·36 ± 0·20	1·09 ± 0·05
6	8	6	48	48	3·25	1·10	7·96 ± 0·39	1·00 ± 0·05
7	8	12	96	48	2·35	0·70	8·14 ± 0·35	1·02 ± 0·04
8	8	24	192	25	1·74	0·63	8·51 ± 0·62	1·06 ± 0·08
9	16	1	16	102	14·5	2·84	14·48 ± 0·28	0·90 ± 0·02
10	16	3	48	148	9·39	2·10	16·26 ± 0·30	1·02 ± 0·02
11	16	6	96	48	6·75	1·99	16·54 ± 0·70	1·03 ± 0·04
12	16	12	192	37	4·76	1·15	16·48 ± 0·66	1·03 ± 0·04
13	16	24	384	18	3·28	0·65	16·08 ± 0·75	1·00 ± 0·05
14	16	48	768	6	2·43	0·40	16·81 ± 1·13	1·05 ± 0·07
15	32	3	96	50	19·0	3·37	32·92 ± 0·83	1·03 ± 0·03
16	32	6	192	24	13·8	1·81	33·79 ± 0·90	1·06 ± 0·03
17	48	1	48	48	42·2	5·17	42·15 ± 0·75	0·88 ± 0·02
18	64	3	192	24	40·4	5·57	69·90 ± 1·97	1·09 ± 0·03

gives the number, m, of model sampling experiments. In each experiment the actual area, $a(1,\ 2)$, between the two sub-sample fractile graphs (based on the averages) was measured on squared paper; and the average value of these areas, called $\bar{a}$, is shown in col. (6). The standard deviation of these areas was directly calculated from the observations and is called $s(a)$ and is shown in col. (7). The standard deviation of the average value, $\bar{a}$, can be obtained by dividing $s(a)$ by $\sqrt{m}$, where m is the number of experiments as shown in col. (5).

9.3. The calculated values of $\sqrt{n'}[a \pm s(\bar{a})]$ are shown in col. (8); and of $\frac{\sqrt{n}}{g}[\bar{a} \pm s(\bar{a})]$ in col. (9). It will be noticed that, for the same value of g, $\sqrt{n'}[\bar{a} \pm s(\bar{a})]$ tends to be constant showing that $\bar{a}$ tends to vary as $\frac{1}{\sqrt{n'}}$, as predicted in paragraph 6.2.

9.4. It will be further noticed that $\sqrt{n'}[\bar{a} \pm s(\bar{a})]/g$ given in col. (9), also tends, as a first approximation, to be constant showing that $\bar{a}$ tends to increase in proportion to

g as predicted in paragraph 6.2. It would be noticed that the observed values of $\sqrt{n'}[\bar{a} \pm s(\bar{a})]/g$ are appreciably lower for $g = 2$ than for higher values of g which indicates that present results are only approximate.

10.1. Data from the National Sample Survey. Some observational data based on an enquiry into household budgets for consumer expenditure in the 7th Round of the National Sample Survey of India (October 1953–March 1954) have also been used for model sampling experiments. For this purpose 1,412 household schedules covering the rural area of the whole of India were selected; and the two variates chosen were 'total expenditure in rupees per person per 30 days' as x, and the corresponding 'expenditure on food items per person per 30 days as y. In each experiment a pair of sub-samples was drawn with replacement giving equal probability to each household and the percentile graphs were drawn in the usual way; and the area, $a(1, 2)$, between the two sub-sample graphs (based on medians) was measured directly on squared paper.

10.2. Results of a series of 14 experiments are shown in the following Table (J.F.G. 2) in which the different columns are arranged in exactly the same way as in Table (J.F.G. 1).

TABLE (J.F.G. 2)

Fractile Graphical Analysis : Results of Model Sampling Experiments : Calcutta 1958

Sampling from consumer expenditure data,* National Sample Survey, 7th Round, October 1953–March 1954 : All India Rural : 1,412 households

Sl. No.	Number				Area between two sub-sample graphs		Function of area between sub-sample graphs	
	of groups (g)	in each group (n')	in each sub-sample ($n=n'.g$)	of model experi-ments (m)	average ($\bar{a}$)	standard deviation $s(a)$	$\sqrt{n'}\,[\bar{a} \pm s(\bar{a})]$	$\frac{\sqrt{n'}}{g}\,[\bar{a} \pm s(\bar{a})]$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	2	24	48	50	1·12	0·829	5·51 ± 0·575	2·76 ± 0·29
2	4	12	48	50	3·88	2·08	13·4 ± 1·02	3·35 ± 0·26
3	4	24	96	50	2 40	0·969	11·8 ± 0·67	2·95 ± 0·17
4	4	48	192	50	1·85	1·07	12·8 ± 1·05	3·20 ± 0·26
5	8	6	48	50	11·39	4·84	27·9 ± 1·68	3·49 ± 0·21
6	8	12	96	50	7·16	2·78	24·8 ± 1·36	3·10 ± 0·17
7	8	24	192	50	5·80	2·44	28·4 ± 1·69	3·55 ± 0·21
8	16	1	16	99	48·32	15·65	48·3 ± 1·57	3·02 ± 0·10
9	16	3	48	200	29·57	9·29	51·2 ± 1·14	3·20 ± 0·07
10	16	6	96	100	20·81	6·85	51·0 ± 1·68	3·19 ± 0·10
11	16	12	192	50	15·69	5·72	54·3 ± 2·80	3·39 ± 0·18
12	16	24	384	25	11·38	3·21	55·7 ± 3·14	3·48 ± 0·20
13	32	3	96	50	58·98	14·10	102·1 ± 3·45	3·19 ± 0·11
14	48	1	48	50	135·83	25·61	135·8 ± 3·62	2·83 ± 0·08

* x = total expenditure in rupees per person per 30 days.
y = expenditure in rupees on food items per person per 30 days.

10.3. It will be noticed that $\sqrt{n'}[\bar{a} \pm s(\bar{a})]/g$ tends to remain constant but the values for $g = 2$ is again appreciably lower showing the need of a more rigorous expression for small values of g.

10.4. The results of the model sampling experiments, as far as they go, are in accordance with the surmises as a first approximation. This is encouraging and shows that systematic studies on both experimental and theoretical lines are likely to lead to fruitful results.

10.5. I have much pleasure in acknowledging the help given by my young colleague, Nikhilesh Bhattacharjee, under whose supervision the model sampling experiments were done.

FRACTILE GRAPHICAL ANALYSIS: MODEL SAMPLING EXPERIMENTS: CALCUTTA 1958

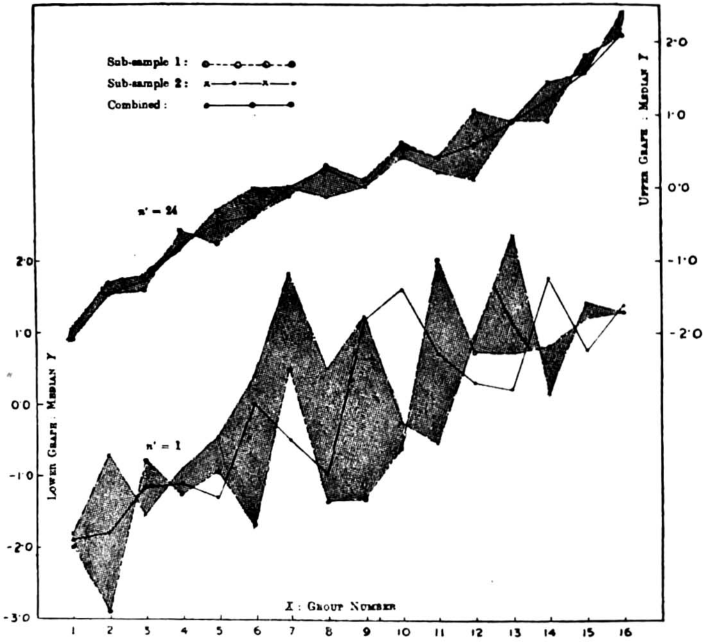

CHART (JFG.1): ACTUAL SAMPLE GRAPHS SHOWING AREA (a) BETWEEN SUB-SAMPLES FOR DIFFERENT VALUES OF n' : $g = 16$

[Bivariate normal population : $m_x = m_y = 0$: $\sigma_x = 1$, $\sigma_y = \sqrt{2}$, $\rho_{xy} = \frac{1}{\sqrt{2}}$]

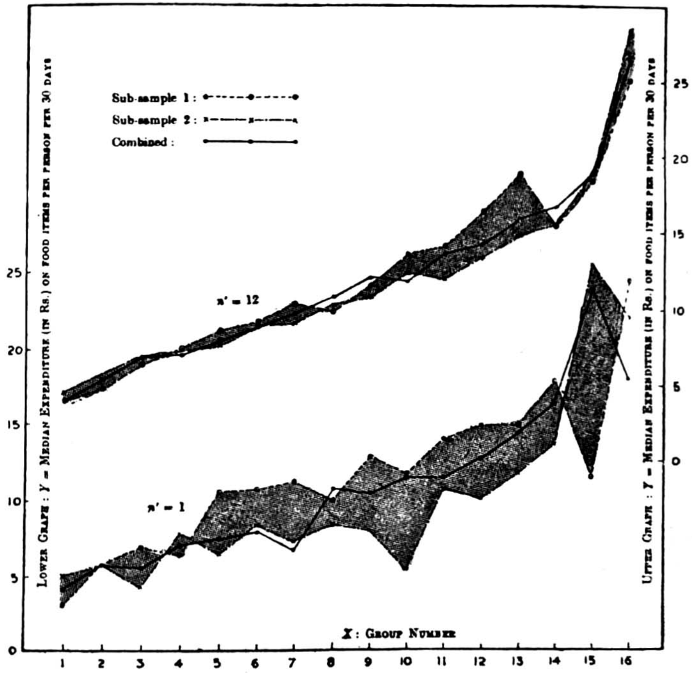

CHART (JFG.2): ACTUAL SAMPLE GRAPHS SHOWING AREA (a) BETWEEN SUB-SAMPLES FOR DIFFERENT VALUES OF n' : $g = 16$

[Consumer expenditure data, NSS, 7th Round, October 1953–March 1954 : All-India Rural, 1412 households]

FRACTILE GRAPHICAL ANALYSIS : MODEL SAMPLING EXPERIMENTS : CALCUTTA 1958

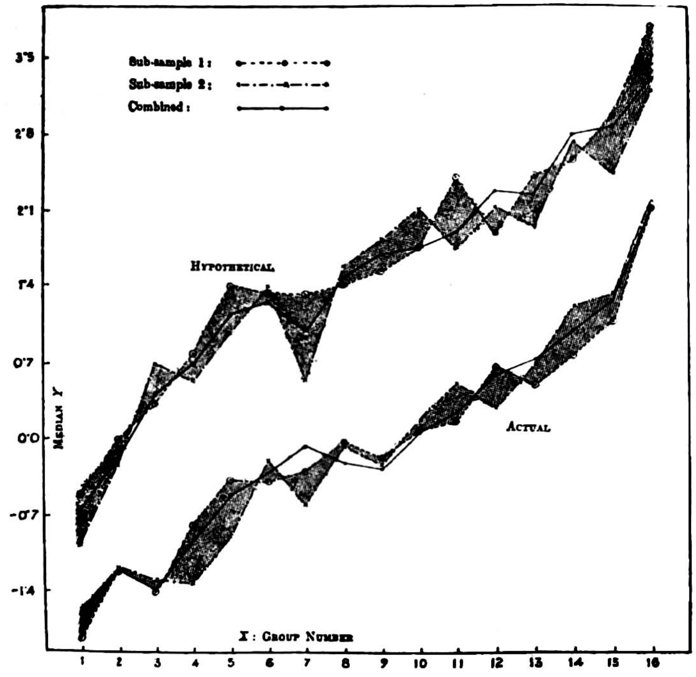

CHART (JFG.3): ILLUSTRATING SEPARATION BETWEEN (SAMPLE) GRAPHS FOR AN ACTUAL AND A HYPOTHETICAL POPULATIONS : $g = 16$, $n' = 24$

["Actual" : Bivariate normal population : $m_x = m_y = 0$; $\sigma_x = 1$, $\sigma_y = \sqrt{2}$, $\rho_{xy} = \frac{1}{\sqrt{2}}$]

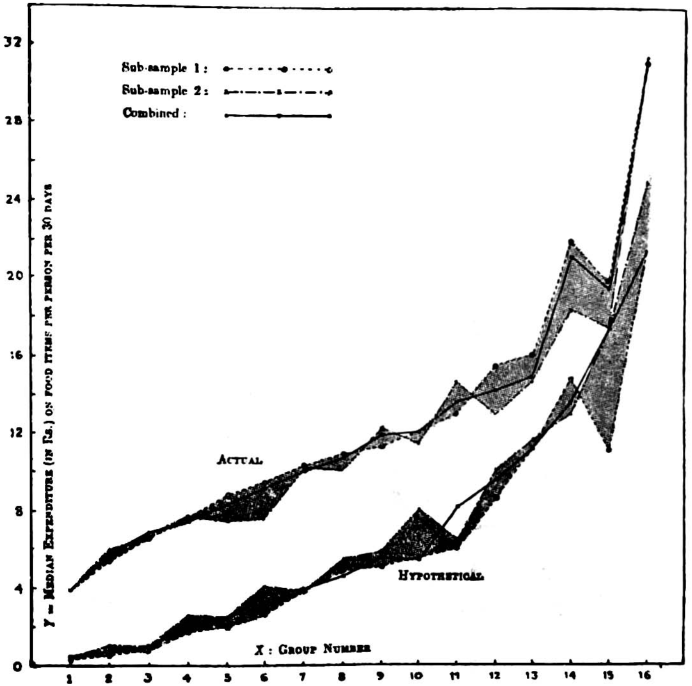

CHART (JFG.4): ILLUSTRATING SEPARATION BETWEEN (SAMPLE) GRAPHS FOR AN ACTUAL AND A HYPOTHETICAL POPULATIONS : $g = 16$, $n' = 12$

["Actual" : Consumer expenditure data, NSS, 7th Round, October 1953–March 1954 : All-India Rural, 1412 households]

Printed at the Baptist Mission Press, 41a Lower Circular Road, Calcutta 16, India