M. N. Murthy

An Approach to Classification of Frequency Distributions with an Application to Classification of Age Distributions

1, Introduction

IN practice there is considerable interest in comparing frequency distributions of specified characteristics over time and space. Mainly owing to the lack of a suitable methodology for such comparisons, the users confine themselves to comparing some specific statistical measures such as measures of location, which include mean, median and mode. The concentration curve (Lorenz curve) is commonly used for comparing the concentrations of distributions as measured from the line of equal distribution and the measure used for this purpose is the coefficient of concentration. But this measure does not effectively discriminate between concentration curves with a bulge near the lower values and those having a bulge near the higher values and does not bring out fully the differences in distribution patterns.

In this note, an attempt is made to develop a methodology for structurally comparing the frequency distributions unaffected by the levels of the frequencies and the units of measure in which the variables are expressed and, on the basis of this comparison to classify meaningfully the frequency distributions which can be considered to be similar structurally. For this purpose the concept of distance as developed by Mahalanobis is used.

The methodology developed in this note will have considerable application in classifying countries, and regions within countries, on the basis of the struc-

tares of their frequency distributions of size of land holding, income, consumer expenditure, establishments by number of workers, or other variables.

2. The Approach

Any frequency distribution has three basic characteristics, which bring out the structure or the pattern of the distribution. These three characteristics are dispersion measured by the *coefficient of variation* (ratio of the standard deviation (a) to the mean), *skewness* measured by Y_1 (ratio of the third central moment to o^3) and *kurtosis* measured by Y_s (ratio of the fourth central moment to a^4, minus 3). These three measures are particularly suitable for oar purpose as they are pure numbers not affected by the units of measurement hi which the variables are expressed.

Representing each distribution by its values of coefficient of variation, skewness and kurtosis as a point in a three dimensional space, it is possible to measure the distance between any two distributions *A* and *B* by the distance measure defined by

$$D^2 = (C_A - C_B)^2 + (\gamma_{1A} - \gamma_{1B})^2 + (\gamma_{2A} - \gamma_{2B})^2,$$

where C_t Yi and Ya stand for the coefficient of variation and measures of skewness and kurtosis, respectively.

Thus if there are *k* frequency distributions, say, the same characteristic for " *k* countries or regions within countries or some other groups, or distributions of *k* characteristics, they can be represented as *k* points in a three dimensional space. On the basis of their coefficient of variation, skewness and kurtosis, ${}^kC_2D^2$ values can be computed. Using these values, the distributions can be classified into a number of groups such that the distance between any two distributions in any group is less than a prespecified value.

To facilitate the formation of the groups, it is desirable to arrange the distributions in ascending or descending order of their distances from a farthest point or from a point which can be considered to be the representation of an ideal, desirable or target distribution. The ${}^kC_2\ D^2$ values are computed and given in the form of a symmetric matrix and the groups are formed by forming squares progressively such that any two distributions in a square has a distance less than the prespecified value.

Another possible way of forming groups is to arrange the frequency distributions according to their distance from an ideal, desirable or target distribution as measured by the D^2 measure and form the groups linearly classified on the basis of the Devalues.

3. An Application

To illustrate the method, it is applied to the classification of the countries in the ESCAP (Economic and Social Commission for Asia and the Pacific) region oa the basis of the structures of the age distributions of their populations. There are 34 countries in the ESCAP region, but the data on age distribution by five-year age groups were available for only 27 of them. For these 27 countries, the values of the coefficient of variation,-skewncss and kurtosis (Table I) and-the $^{27}C_2.D^2$ values were computed. Thus the 27 countries can be represented as 27 points in three dimensional space and these points can be classified on the basis of the D^2 values.

Arranging the countries according to the distances of their age distributions from that of Australia, and considering a D^2 value of less than or equal to 0.25 between any two countries as qualifying them to be classified in the same group, it can be seen from Table 2 that the countries get classified by this method into four groups such that the distance between any two countries from the same group is less than or equal to 0,25. The composition of the four groups is *as* follows':

Group 1	*Group 2*	*Group 3*	*Group 4*
Australia	Burma	Malaysia	Western Samoa
New Zealand	Papua New Guinea	Nepal	
Japan	Singapore	Brunei	
Nauru	Korea	Thailand	
Hong Kong	India	Philippines	
	Khmer	Trust Territory	
	Sri Lanka	Fiji	
	Indonesia	Gilbert Ellice Isl.	
	Pakistan	Iran	
		British Solomon Islands	
		Tonga	
		Cook Islands	

The aim of this note is not to enter into a discussion of why certain countries "are in one group and not in the other. However, it may be mentioned that , the countries in Group 1 have a favourable age distribution in the sense of "nearness to the ideal, desirable or target age distribution referred to in the next

TABLE 1—SHOWING THE VALUES OF COEFFICIENT OF VARIATION, SKEWNESS AND KURTOSIS FOR THE AGE DISTRIBUTIONS OF THE COUNTRIES IN THE ESCAP REGION TOGETHER WITH THE YEAR TO WHICH THE DATA RELATE

Country	*Year*	c. v.	*Skewness* (γ_1)	*Kurtosis* (γ_2)
Australia	1972	0.6856	0.5111	−0.6992
British Solomon Isl.	1972	0.8239	1.0448	0.5494
Brunei	1973	0.7895	0.9707	0.3264
Burma	1970	0.7714	0.7596	−0.2946
Cook Islands	1971	0.8811	1.1757	0.6063
Fiji	1972	0.7627	0.9815	0.4872
Gilbert and Ellice Isl.	1973	0.8328	1.0596	0.4870
Hong Kong	1973	0.7150	0.6784	−0.5031
India	1973	0.7719	0.8574	0.0200
Indonesia	1971	0.7914	0.8828	0.0998
Iran	1971	0.8440	1.0704	0.4919
Japan	1972	0.6364	0.4441	—0.5889
Khmer	1962	0.7880	0.8959	0.0239
Korea	1972	0.7561	0.8592	−0.0227
Malaysia	1970	0.8154	1.0034	0.2590
Nauru	1966	0.7119	0.5147	−0.5222
Nepal	1971	0.7762	0.9163	0.3112
New Zealand	1971	0.7185	0.5921	-0.6311
Pakistan	1968	0.8131	0.9353	0.1398
Papua New Guinea	1971	0.7955	0.8093	−0.1451
Philippines	1972	0.8097	0.9986	0.3826
Singapore	1972	0.7301	0.8571	−0.0313
Sri Lanka	1970	0.7915	0.8955	0.0694
Thailand	1970	0.8151	1.0083	0.3485
Tonga	1966	0.8362	1.0656	0.5733
Trust Territory of P.I.	1972	0.8097	1.0492	0.4135
Western Samoa	1970	0.8811	1.2461	1.0623
ESCAP* excl. *1* countries	1962-1973	0.7688	0.8303	−0.0585

* Excluding Afghanistan, Bangladesh, Bhutan, China, Laos, Mongolia and Vietnam.

(Sources of Basic Data: United Nations (1973) Demographic Yearbook, New York, and ESCAP (1973) Statistical Yearbook for Asia and the Far East, Bangkok.)

TABLE 2—SHOWING THE FORMATION OF THE GROUPS SUCH THAT THE DISTANCE BETWEEN THE STRUCTURES OF THE AGE DISTRIBUTIONS OF ANY TWO COUNTRIES IN EACH GROUP AS MEASURED BY D^2 MEASURE IS LESS THAN OR EQUAL TO 0.25

country	AUS	N.Z	JAP	NAU	HON	BUR	PNG	SIN	KOR	IND	KHM	S.L	I'A	PAK	MAL	NEP	BRU	THA	PHI	TTP	FIJ	GEI	IRA	BSI	TON	C.I	W.S
Australia	x																										
New Zealand	x	x																									
Japan	x	x	x																								
Nauru	x	x	x	x																							
Hong Kong	x	x	x	x	x																						
Burma	x	x	x	x	x	x																					
Papua New Guinea				x	x	x	x																				
Singapore						x	x	x																			
Korea						x	x	x	x																		
India						x	x	x	x	x																	
Khmer						x	x	x	x	x	x																
Sri Lanka						x	x	x	x	x	x	x															
Indonesia						x	x	x	x	x	x	x	x														
Pakistan						x	x	x	x	x	x	x	x	x													
Malaysia							x	x	x	x	x	x	x	x	x												
Nepal							x	x	x	x	x	x	x	x	x	x											
Brunei							x	x	x	x	x	x	x	x	x	x	x										
Thailand								x	x	x	x	x	x	x	x	x	x	x									
Philippines								x	x	x	x	x	x	x	x	x	x	x	x								
Trust Territory								x	x	x	x	x	x	x	x	x	x	x	x	x							
Fiji										x	x	x	x	x	x	x	x	x	x	x	x						
Gilbert Ellice Isls.											x	x	x	x	x	x	x	x	x	x	x	x					
Iran												x	x	x	x	x	x	x	x	x	x	x	x				
British Solomon Isls.													x	x	x	x	x	x	x	x	x	x	x	x			
Tonga														x	x	x	x	x	x	x	x	x	x	x	x		
Cook Islands															x	x	x	x	x	x	x	x	x	x	x	x	
Western Samoa																										x	x

x + value of D^2 measure is less than or equal to 0.25.

paragraph., whereas Group 2 countries are possibly in the process of improving the structure-of their age distribution followed by countries in Group 3 and the one country in Group 4 has possibly the most unfavourable age distribution.

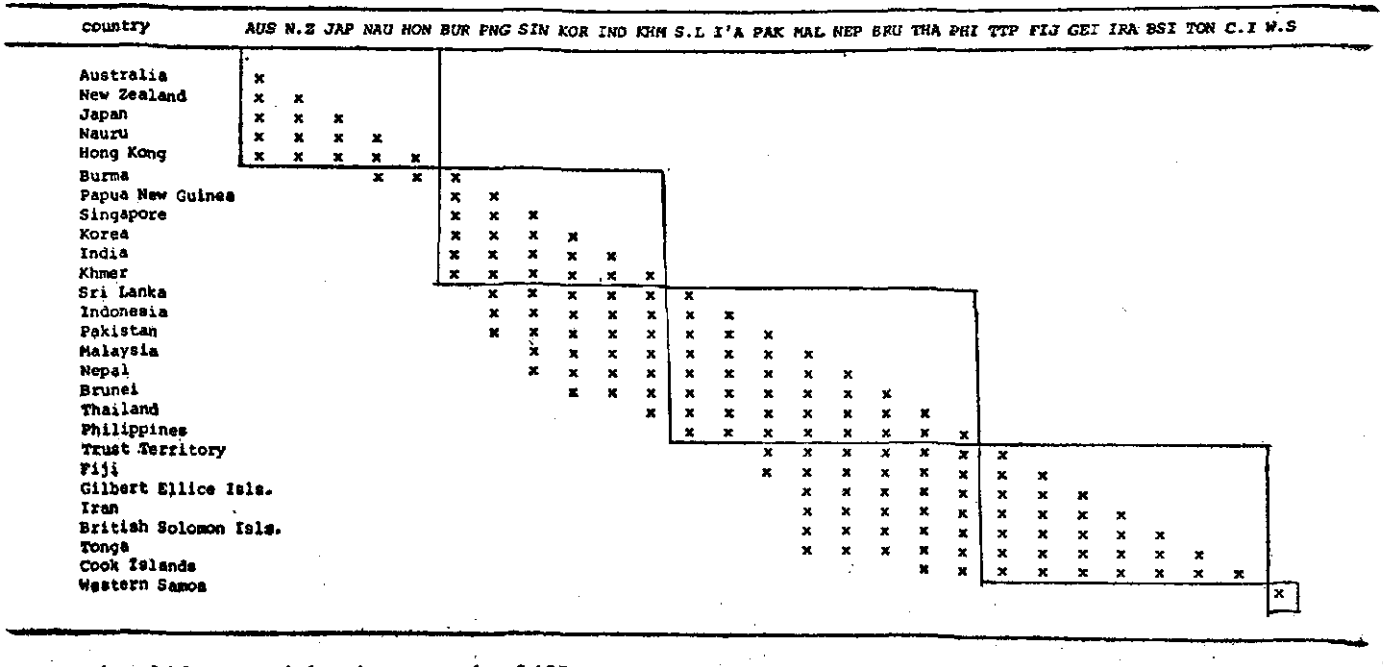
Instead'of arranging the countries in ascending order of the distance of their age distributions from that of Australia, one may consider an ideal, desirable or target age distribution as a rectangular one with a life span of 100 years (say), for .which the coefficient of variation turns out to be 0.5774, skewness is 0 and kurtosis is.— 1,200. Arranging the countries in increasing order of the distances of their age distributions from the target age distribution mentioned here and proceeding as above, it is found that the same four groups as obtained earlier are again obtained, showing the stability of the method, particularly since Australia is having an age distribution structurally nearest to the target distribution in the ESCAP region.

Arranging the countries in increasing order of the distance of their age distributions from that of all the countries taken together for formation of groups is not advantageous, as equality of distance does not ensure similarity in the distribution pattern. Hence this approach has not been used in this note. It may be noted that even if this arrangement is used, the same four groups would be obtained, though the formation of groups becomes a bit cumbersome.

If more groups are required with a view to emphasizing greater similarity of countries within groups, than the maximum distance requirement in a group can be reduced to any desired quantity before forming the groups. Similarly, if broader groups are required, then the maximum distance requirement in a group can be increased to any desired quantity prior to the formation of the groups. For instance, if the maximum D^2 value requirement in a group is reduced from 0.25 considered earlier to 0.125, then the countries get grouped into the following six groups, as can be seen from Table 3.

Group 1	*Group 2*	*Group 3*	*Group 4*	*Group 5*
Australia	Burma	Sri Lanka	Trust Territory	Western
New Zealand	Papua New	Indonesia	Fiji	Samoa
Japan	Guinea	Pakistan	Gilberg Ellice	
Nauru	Singapore	Malaysia	Islands	
Hong Kong	Korea	Nepal ,	Iran	
	India	Brunei	British Solomon	
	Khmer	Thailand	Islands	
		Philippines	Tonga	
			Cook Islands	

TABLE 3—SHOWING THE FORMATION OF THE SUCH GROUPS THAT THE DISTANCE BETWEEN THE STRUCTURES OF THE AGE DISTRIBUTIONS OF ANY TWO COUNTRIES IN EACH GROUP AS MEASURED BY D^2 MEASURE IS LESS THAN OR EQUAL TO 0.125



x—value of D^2 measure is less than or equal to 0.125.

Tables 4 gives the countries arranged in ascending order of their distances or their age distributions from the target age distribution as measured by the value of the D^2 measure. Groups of countries with similar age distributions can also be formed by classifying them linearly by the value of this distance measure from the target distribution. If the class intervals are taken as 0-1, 1-2, 2-3, 3-4, 4-5 and > 5, we get the following six groups :

Group 1	*Group 2*	*Group 3*	*Group 4*	*Group 5*	*Group 6*
Australia	Burma	Singapore	Nepal	Gilbert Ellice	Western
Japan	Papua New	Korea	Malaysia	Islands	Samoa
New Zealand	Guinea	India	Brunei	Iran	
Nauru		Khmer	Thailand	British	
Hong Kong		Sri Lanka	Philippines	Solomon	
		Indonesia	Trust	Islands	
		Pakistan	Territory	Tonga	
			Fiji	Cook Islands	

TABLE 4--SHOWING COUNTRIES ARRANGED IN ASCENDING ORDER OF THE DISTANCE OF THEIR AGE DISTRIBUTION AS MEASURED BY D^2-STATISTIC

country	*D^2-statistic from target distribution*	*country*	*D^2-statistic from target distribution*
Australia	0.5237	Nepal	3.1629
Japan	0.5741	Malaysia	3.1921
New Zealand	0.6941	Brunei	3.3171
Nauru	0.7424	Thailand	3.4710
Hong Kong	0.9648	Philippines	3.5558
Burma	1.4344	Trust Territory	3.7582
Papua New Guinea	1.8153	Fiji	3.8443
Singapore	2.1238	Gilbert Ellice Islands	4.0340
Korea	2.1562	Iran	4.0794
India	2.2613	British Solomon Islands	4.2130
Khmer	2.3449	Tonga	4.3471
Sri Lanka	2.4591	Cook Islands	4.7372
Indonesia	2.5146	Western Samoa	6.7630
Pakistan	2.7254		

4. Concluding Remarks

This method of classifying frequency distributions on the basis of their structures or patterns is more meaningful than classifying them on the basis of point parameters. The emphasis in comparison is shifting away from aggregates, means, ratios and differences to patterns of distributions such as the distribution of land, income and other characteristics among the population with a view to focussing attention on the disequilibrium or social injustice reflected in the distributions. Comparison of countries on the basis of the patterns of their income distributions, land holding distributions, household size distributions, etc. is more meaningful than classifying them on the basis of their per capita income, average land holding size, average household size, etc. The method given in this note is proposed as a tentative one with the hope that the method will get developed and improved with its increasing application to practical problems.

Acknowledgement

The author wishes to thank Mr. K. Sato and other staff of the Computer Section of the Institute of Developing Economies for getting all the computations for this note done through the electronic computer.