Statistical pattern analysis and its procedure

by Mitsuo Fujioka[1] and Hiroshi Iwai[2]

1 Purpose

When analysing statistical data, the integrated use of several kinds of data often facilitates a comprehensive understanding. To take an example, for an analysis of data concerning workers' health, it is necessary to use several indices based on statistical data for occupational injury, occupational mortality, general mortality and working conditions. Nevertheless, it is not easy to analyse several indices simultaneously for integrated use of them.

On the other hand, the remarkable development of computer systems, in terms of not only hardware but also software, makes possible new and varied methods of statistical analyses. That is to say, by applying developed computer systems to data processing, very complicated statistical methods can now be processed by computer, something which was almost impossible using classical statistical methods manually developed in the past.

In fact, easy access to computers in daily life has played a major role in improving overall statistical methods concerning data collection, recording, control, calculation and analysis. Moreover, the popularization of personal computers, including software, has accelerated this improvement. However, it is also true that conducting statistical processes with the help of computers very often needs a considerable degree of specialist knowledge.

The "Statistical Pattern Analysis" (SPA) using a computer system is perhaps one of the simplest and most useful methods for statistical analysis in the case of integrated use of several indices. In this paper, two types of SPA are presented: firstly, an integrated observation method for several kinds of statistical data in the form of cross-table; and secondly, an analysis method for large scale data in the form of data matrices.

---

[1] Professor, Faculty of Humanities and Social Sciences, Shizuoka University, Japan.

[2] Professor, Faculty of Economics, Kansai University, Japan.

2 Implications of the Statistical Pattern Analysis

2-1 Outline of SPA

The first step in statistical observation is data classification. Although the work of classification is applied generally in many academic fields as a basic task, it does demand a considerable degree of specialist and technical knowledge when we analyse a complicated natural or social phenomenon. For example, in the field of labour statistics, where industrial and occupational classifications have been modified according to economic developments, the

overall picture is rather complex. In the case of Japan, there are no less than 463 industrial sub-classifications and 376 occupational sub-classifications.

However, even if an original classification of the statistics is complicated, for general observation a simple classification or a dichotomy of the data can often be useful with particular threshold values. There might be above or below an average level, increasing or decreasing tendency, large-medium-small, productive workers or un-productive workers, and so on. Incidentally, a comprehensive observation would be easier with these categories combined into a set of patterns.

Research into statistical comparison of infant mortality conducted by Dr. Hiroshi Maruyama (former professor of Osaka University) could be quoted as an example. In his research (from 1930s to 1950s), simple classification is used from a comprehensive view point for analysing quantitative indices (infant mortality rate, neonatal death rate, and stillbirth rate), as well as qualitative indices ("-index: infant mortality rate ÷ neonatal death rate [this index shows the degree of socio-economic causes of infant mortality]). In other words, firstly, he classified each district according to one of three categories (a , b , c) in relation to infant mortality rate (= R) and "-index (= A), i.e., type-a: R > 10(A-1), type-b: R < 10(A-1), and type-c: R…10(A-1). In the same way, regarding stillbirth rate (= S) and neonatal death rate (= N), he classified them into three categories (X, Y, Z), i.e., type-X: S > N, type-Y: S < N, type-Z: S…N. He then combined these three categories into a set of nine pattern groups for easy comparison (see Table 1).[1]

With the assistance of a computer system, Iwai and Fujioka attempted to apply this method to the research into fluctuations in employment structure. In parallel with this attempt, Fujioka tried to use the method in practice for their research based upon the advice of Dr. Maruyama.[2] After a detailed discussion between Iwai and Fujioka on the project and through trial-and-error in their own concrete application, the implication, variety and procedure for the Statistical Pattern Analysis have gradually became clearer.[3] Finally, Iwai applied it in his research on cross-analysis of employment structure according to industry and occupation.[4] Fujioka used this method for an analysis of employment structure according to industry, occupation, sex, age and prefecture (the largest administrative unit in Japan).[5]

Table 1 Category and pattern classification on infant mortality

S & N	R & A	Type-a: R > 10(A-	Type-c: R…10(A-1)	Type-b: R < 10(A-	Number of cases
Type-X: (Number	S > N of cases)	aX (0)	cX (0)	bX (6)	(6)
Type-Z: (Number	S…N of cases)	aZ (0)	cZ (1)	bZ (5)	(6)
Type-Y: (Number	S < N of cases)	aY (0)	cY (5)	bY (12)	(17)
Total		(0)	(6)	(23)	(29)

Notes: S = Still birth rate, N = Neonatal death rate, R = Infant mortality rate, A = " index (R ÷ N), 10(A-1): Scale adjusted to infant mortality rate for comparison

Source: Hiroshi Maruyama, *Research in Socio-Medical Science I: Infant Mortality*, 1976, Iryotosho Publisher, p.406

There are two types of SPA: firstly, an integrated observation method for several kinds of indices; and secondly, an analysis method for large scale data in the form of data matrices.

The first type of SPA is designed to promote efficiency in the most fundamental approach for the analysis of general statistical tables, and it can be used simply by PC (personal computer) spread-sheet. This method is useful for structural analysis using cross-tables and or analysis of fluctuations. The second type is designed to compare cases according to pattern group, or to analyse the tendency, regularity and correlation of many kinds of indices in the form of data matrices, using either spread-sheet or data-base software. This method of processing a large quantity of accumulated data can be utilized thanks to the revolutionary developments that have taken place in the computer industry.

The Statistical Pattern Analysis which we use can be summarized in six steps: (1) classifying data into categories; (2) combining these categories into a set of patterns; (3) indicating selected patterns; (4) re-arranging the cases according to pattern group; (5) counting the number of these cases; and (6) summarizing this case-counting in tabular form. Although steps (1) and (2) are common for both type of SPA, steps from (3) to (6) are varied by the type.

In fact, the idea of data-classification is not new; it has already been used for data processing in the fields of ecology, biology, behavioral science etc.. It is also used for statistical data processing such as cluster analysis, categorical data analysis, or statistical pattern recognition.[6] Nevertheless, SPA is characterized by the fact that : (1) a set of combined patterns indicates several kinds of summarized information simultaneously; (2) it is exceedingly simple, having no complicated numerical formulae for classifying/combining categories, thus making it readily accessible to non-specialists; (3) it is a standardized data processing method, applying commonly used procedures, with the assistance of a computer system, i.e. classification and combination with repetitions, sorting, selecting, counting and comparing.

2-2 Classification into categories

The first step of SPA is classifying data into categories. When classifying data, the classification of indices into categories should be simplified as far as possible. The simplest classification is a dichotomy, such as above-average level and below-average level, or increasing and decreasing tendency.[7] If there is a large number of categories, then the number of pattern groups will also be too enormous to analyse.

For example, in the case of 3 kinds of indices and 2 categories for each set of indices, the number of pattern groups combined with these categories is 2 cubed: 2^3, which gives us eight pattern groups. However, if the number of categories is increased from 2 to 10, the number of pattern groups will be changed to 10 cubed: 10^3, that is 1000. As a result, SPA for purposes of simplifying and summarizing the data will become impracticable because of the great number of patterns.

From the viewpoint of classification of variables according to the size of the range set, a familiar classification method for variables involves the following three types: (1) a continuous

variable; (2) a discrete variable; and (3) a binary or dichotomous variable. Additionally, according to their scale of measurement, classification of variables is as follows:

(1) "A nominal scale", which "merely distinguishes between classes";
(2) "An ordinal scale", which "induces an ordering of the objects";
(3) "An interval scale", which "assigns a meaningful measure of the difference between two objects";
(4) "A ratio scale", which "is an interval scale with a meaningful zero point".[8]

There are two types of statistical classification: qualitative and quantitative classification. Frequently, "variables on nominal and ordinal scales are referred to as categorical variables or qualitative variables", and "variables on interval or ratio scales are then referred to as quantitative variables." These scale definitions are put in hierarchical order from "nominal" up to "ratio", and "by giving up information one may reduce a scale to any lower order scale".[9] When analysing statistical data, it is often necessary to use mixed scales of measurement. For the integrated use of several kinds of data including descriptive records, it must be useful to transform variables into either a nominal or an ordinal scale.

Although this demotion of variables involves an information loss for each piece of data, it widens the scope of the information provided because several kinds of data are being used simultaneously. For comparison, a simple scale is sometimes useful when using similar but heterogeneous data, such as statistics for occupational injuries. Furthermore, the loss of information can be significant for statistical analyses which safeguard privacy when using individual case records.

Classifying data into several categories means transforming the variables into a nominal or an ordinal scale. Each category classified should then be stated in figures such as 1, 2, 3. For example, below-average level can be termed category-1, average level category-2, and above-average level category-3, or decreasing tendency category-1 and increasing tendency category-2. It is better to avoid using 0 because of its special characteristics in calculation.

When classifying data, the number of categories and the criteria/threshold values for classification of each type of data have to be distinct, such as below-average level and average level or over in the case of the two categories. Instead of an average for a criterion, certain threshold values can be used. For example, in the field of demography, 1.0 of NRR(net reproduction rate, the average number of live daughters that would be born to a hypothetical female birth cohort) is used as a theoretical criterion of population replacement. When the threshold values and classification into categories are significant, then a set of patterns becomes discernible. On the contrary, if the threshold values and categories are more vague, the analysis of pattern groups could be meaningless. Therefore the first problem for SPA is to determine criteria for classification into categories and the number of these categories.

2-3 Combination of categories into pattern groups

The second step of the SPA is combining each category into a set of patterns. For instance, with the first index at average level or over, we have category-2, and with the second index at below-average level, category-1. Then the combined pattern of these two categories is 21. Additionally, if the third index shows a rising tendency, we also have category 2, in which case the pattern is 212.

Although the implication of a pattern such as 212 is simple, it is useful for analysing a great deal of data because it includes several kinds of information. However, the process of classifying data into various categories and then combining them into patterns is exceedingly intricate especially if the quantity of data is very large. SPA simplifies the process through utilization of a computer system.

Some attentive consideration is necessary when combining categories into patterns. Firstly, the kinds and number of indices for combination should be carefully checked. Secondly, the order in which we combine these categories is also of importance. If we increase the number of indices to be combined, complicated analysis may well become possible, but in cases where the implication of the order for category-combination is not distinct, the combined patterns might be of very little significance. Therefore the number of indices has to be kept within the limits of minimal necessity, and the order of indices should be exactly arranged for analysis.

2-4 Indication of the results

There are two methods of observation for analysing patterns.

The first method is indicating the patterns in the form of a cross-table. After combining the categories into a set of patterns, each pattern is indicated on the cross-table, and selected patterns can be shown individually on the table if desired. This is particularly useful in the case of a large-scale cross-table; any special patterns which the computer system has automatically identified can be easily observed.

Another method is to summarize the case-counting in tabular form according to pattern group. For analysis of large scale data with fields and records in a computer data-base, it is too complicated to observe patterns which are placed irregularly. Therefore, the following three steps are necessary for easy observation and analysis: firstly, re-arranging the records according to the pattern group; secondly, counting the number of cases (= records) in each pattern group; and thirdly, summarizing this case-counting in tabular form.

3 Procedure for the Statistical Pattern Analysis

3-1 An integrated observation method for various indices

The first type of SPA: i.e. an integrated observation method for various indices in the form of regular statistical tables, consists of three steps: firstly, classifying each index into categories; secondly, combining these categories into patterns; and thirdly, indicating the patterns selected.

There are three indices in table 2 as an example: (1), (2), SMR = standardized mortality ratio of males aged 35-54 (total employed = 100) according to occupation and causes of death in 1970 and 1985 respectively; and (3), comparison of SDR (standardized death rate, 1970 = 100) for males aged 35-54 for 1970 and 1985. Standardized mortality ratios (index A and index B) are indicated from B3 to F8 and from B12 to F17 on a work-sheet of Microsoft-Excel. The index of standardized death rate (index C) is to be found between B21 to F26.

For example, when classifying index A into categories, the number of categories and criteria/threshold values for classification have to be decided. If the number of categories is 3 and the threshold values are 80 and 120, each index A will be classified into 3 categories, i.e. category 1 for below 80 (low level), category 2 for 80 - 119 (medium level), and category-3 for 120 or over (high level). In the same way, index B and index C will be classified as follows:

Index A: 1 = below 80 [low], 2 = 80 - 119 [medium], 3 = 120 or over [high]
(SMR 1970: Total employed = 100)
Index B: 1 = below 80 [low], 2 = 80 - 119 [medium], 3 = 120 or over [high]
(SMR 1985: Total employed = 100)
Index C: 1 = below 66 (66: Comparative index for SDR [all causes]
for category "Total employed",
2 = 66 - 99 (higher than 66 but declining), 3 = 100 or over (rising)
(Comparison of SDR: index number for 1985 based upon 100 for 1970)

Concrete operation on the work-sheet of MS-Excel is as follows:

Step 1: Classifying each index into categories

In the case of each index A from B3 to F8 being below 80, 1 will be indicated from I3 to M8 on the work-sheet, and when this is above 80 and below 120, 2 will be indicated, as well as 3 for 120 or over.

Move the cursor to I3.
Type a functional formula : =IF(B3<80,1,IF(B3<120),2,3))
Copy the formula from I3 to M8.

Table 2 Standardized mortality ratio (SMR, 1970, 1985) and comparison of standardized death rate (SDR, 1970-1985) of male 35-54 by occupation and death causes (Japan): Work-sheet on MS-EXCEL

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Index A	SMR	1970 (Emp	=	)			Ctg		Ix	A				Pt		A	B	C	
2	Causes	All	Cance	Heart	Cereb	Accid			Al	Cn	Hr	C	Ac			Al	Cn	Hr	Cb	Ac
3	Employed	100	100	100	100	100		Em	2	2	2	2	2		Em	222	222	222	221	221
4	Mang&Ad	45	62	58	37	26		Mn	1	1	1	1	1		Mn	113	123	113	112	113
5	Clerical	106	138	123	93	60		Clr	2	3	3	2	1		Clr	221	332	322	221	111
6	Product	97	84	90	101	120		Prd	2	2	2	2	3		Prd	211	212	222	221	321
7	Sales	124	125	137	131	78		Sal	3	3	3	3	1		Sal	321	322	322	321	111
8	Service	118	108	122	127	93		Srv	2	2	3	3	2		Srv	232	233	333	331	221
9																				
10	Index B	SMR	1985 (Emp	=	)			Ctg		Ix	B				Pt	111	121	211	221	
11	Causes	All	Cance	Heart	Cereb	Accid			Al	Cn	Hr	C	Ac			Al	Cn	Hr	Cb	Ac
12	Employed	100	100	100	100	100		Em	2	2	2	2	2		Em				221	221
13	Mang&Ad	71	82	66	72	67		Mn	1	2	1	1	1		Mn					
14	Clerical	99	128	104	89	60		Clr	2	3	2	2	1		Clr	221			221	111
15	Product	79	73	80	82	93		Prd	1	1	2	2	2		Prd	211			221	
16	Sales	102	105	101	108	66		Sls	2	2	2	2	1		Sls					111

17	Service	152	135	166	173	115	Srv	3	3	3	3	2	Srv					221
18	Index C																	
19	Change:	SDR	1985		(1970	=100	Ctg		Ix	C			Pt	333	233	133	123	113
20	Causes	All	Cance	Heart	Cereb	Accid		Al	Cn	Hr	C	Ac		Al	Cn	Hr	Cb	Ac
21	Employed	66	89	89	43	44	Em	2	2	2	1	1	Em					
22	Mang&Ad	104	118	100	81	114	Mn	3	3	3	2	3	Mn	113	123	113		113
23	Clerical	61	82	75	41	44	Clr	1	2	2	1	1	Clr					
24	Product	53	78	79	35	34	Prd	1	2	2	1	1	Prd					
25	Sales	54	75	66	35	37	Sls	1	2	2	1	1	Sls					
26	Service	85	111	121	58	54	Srv	2	3	3	1	1	Srv		233	333		

Codes: Causes = death causes, SMR = standardized mortality ratio, SDR = standardized death rate, Ctg = category, Pt = pattern, Ix = index
Death causes; Cance = Cancer, malignant neoplasms, Heart = Heart disease, Cereb = Cerebrovascular disease, Accid = Accidents and adverse effects
Occupation; Emp, Employed = Total employed, Mang&Ofi = Managers and officials, Clerical = Clerical and related workers, Product = Craftsmen, production process workers and labourers, Sales = Sales workers, Service = Service workers

In the same way, for index B and index C three categories, such as 1, 2 or 3 will be indicated using the following formulae:

Move the cursor to I12.
Type: =IF(B12<80,1,IF(B12<120),2,3))
Copy the formula from I12 to M17.

Move the cursor to I21.
Type: =IF(B21<$B$21,1,IF(B21<100,2,3))
Copy the formula from I21 to M26.

Step 2: Combining the categories into patterns

When combining the three categories into patterns, a 3-digit figure will be indicated from P3 to T8 on the work-sheet. For example, when the index A is 2, index-B is 2, and index C is 2, then the pattern will be indicated as 222 at P3. The command and the formula in this case is as follows:

Move the cursor to P3.
Type: =(I3*100)+(I12*10)+(I21)
Copy the formula from P3 to T8.

Step 3: Indicating selected patterns

When it is necessary to observe specified patterns such as 111, 121, 211, 221, which were at average or below average levels of SMR in 1970/1985 and showed decrease tendencies in SDR, these will be indicated from P12 to T17 if the pattern is typed at P10, Q10, R10 and S10.

Move the cursor to P10.
Type: 111
Type 121, 211, and 221 at Q10, R10 and S10.
Move the cursor to P12.
Type: =IF(OR(P3=$P$10,P3=$Q$10,P3=$R$10,P3=$S$10),P3, "")
Copy the formula from P12 to T17.

In the same way, 333, 233, 133, 123 and 113, which were both high and low levels of SMR in 1970/1985 but showed increase tendencies in SDR, typed at P19 and T19 will be indicated from P21 to T26.

Steps (4), (5) and (6) described in section 2-1: i.e. re-arranging the cases according to pattern group, counting the number of the cases, and summarizing this case-counting in tabular form, are not usually necessary in this type of SPA.

As a result of these analyses and indications, we have found nine patterns which were at below and or average levels of SMR and showed decrease-tendencies in SDR. These include, for example, the cerebrovascular diseases of clerical and production workers. When observing the fluctuations in detail for these two groups of occupation, we note that there has been an obvious decline in the index for SDR, i.e. from 100 in 1970 to 41 and 35 respectively in 1985.

In contrast, the table indicates the patterns which show increase-tendencies in SDR, e.g. the cancer and heart diseases of managers and service workers. The worst pattern 333, showing high levels of SMR (both in 1970 and 1985) as well as an increase-tendency in SDR, was related to heart disease among service workers. Moreover, SDR in relation to all causes of death had only increased for managers, although levels of SMR for this occupational group remained low. Generally, variations in SMR between the group “managers” and the group “production workers” are significant; these differentials only seem to be slight in Japan, i.e. 71 and 79 for all causes of death and 66 and 80 for death caused by heart disease.

This analysis method would be exceedingly effective when applied to large-scale cross tables, which include several hundred rows and columns for sub-classifications such as those for types occupation and industry.

3-2 Data analysis method for large scale data file

The second type of SPA, i.e. a data analysis method in the form of data matrices, aims at analysing a large scale data file with fields and records, such as population indices (= fields) according to countries (= records). This method has wide applicability for data analysis, because some multi-dimensional cross data can be transformed into a data format of two dimensions with fields and records. Although this method is also useful for analysing individual case records, analysts should proceed with caution in order to safeguard privacy. It consists of six steps: (1) classifying each index into categories; (2) combining the categories into patterns; (3) indicating selected patterns; (4) re-arranging each record with pattern into pattern groups; (5) counting the number of records (= cases) belonging to each pattern group; and (6) summarizing this case-counting in tabular form.

Table 3, indicated on the Excel-work-sheet as an example, is a table on mortality among children under 5 years of age for various countries in the world. There are 129 records

(= countries), and three fields (= indices) for each record, i.e. infant mortality rates in 1989 (index A), the mortality ratio for children under five years of age in 1989 (index B), and an " 5 index (index C, showing the degree of social causes of death for children under five years of age, [see note in Table 4]). Names of countries are shown from A2 to A130. Index A, B and C are at B2 - B130, C2 - C130 and D2 - D130 respectively.

For example, when threshold values provided by UNICEF are used for classification, index A (infant mortality rate) has been classified into four categories, i.e. 1 = low level for below 25; 2 = medium level for 25 - 54; 3 = high level for 55 - 99, 4 = very high level for 100 or over. There are further four categories for index B (mortality ratio for under-5s) , i.e. 1 = low level for below 31; 2 = medium level for 31 - 94; 3 = high level for 95 - 169; 4 = very high level for 170 or over. For index C (" 5 index), when a dichotomy of below-median or above-median is used for simple classification into two categories, each index has to be classified into either 1 (= below-median) or 2 (= above-median). These categories are then combined into patterns such as 111 or 222. However, observation of a large quantity of data placed irregularly is too complicated, even if it is indicated by summarized patterns. Therefore, after indicating patterns, it is necessary to summarize the data again for easy recognition. Each record with its pattern should first be re-arranged according to its pattern group, and then the number of records belonging to each pattern group can be counted.

Table 3 Infant mortality rate and U5MR in the world (1989)

	A	B	C	D	E	F	G	H	I
1	Countries Indices	IM	U5M	" 5idx		CtgA	CtgB	CtgC	Ptn
2	Afghanistan	169	296	1.75		4	4	2	442
3	Albania	25	30	1.20		2	1	1	211
4	Algeria	70	102	1.46		3	3	2	332
5	Angola	173	292	1.69		4	4	2	442
6	Argentina	31	36	1.16		2	2	1	221
7	Australia	8	9	1.13		1	1	1	111
8	Austria	8	10	1.25		1	1	1	111
9	Bangladesh	116	184	1.59		4	4	2	442
10	Belgium	10	12	1.20		1	1	1	111
11	Benin	89	150	1.69		3	3	2	332
-	*****	-	-	-		-	-	-	-
-	*****	-	-	-		-	-	-	-
128	Zaire	81	132	1.63		3	3	2	332
129	Zambia	78	125	1.60		3	3	2	332
130	Zimbabwe	63	90	1.43		3	2	2	322

Codes: IMR = Infant mortality rate (deaths before 1 year of age ÷ number of births, 1989)
U5MR = Mortality ratio of children under 5 years of age (deaths before 5 years of age ÷ number of births, 1989), the units both of IMR and U5MR; per 1,000 births
" 5 idx = " 5 index (U5MR ÷ IMR), Ctg = category, Ptn = pattern
Categories: IMR; 1 = below 25 [low level], 2 = 25 - 54 [medium level], 3 = 55 - 99_@[high level], 4 = 100 or over [very high level]
U5MR; 1 = below 30 [low level], 2 = 30 - 94 [medium level], 3 = 95 - 169 [high level], 4 = 170 or over [very high level]

" 5 idx; 1 = below 1.37 (median) [low level], 2 = 1.37 or over [average level or over]
(The threshold values of IMR and U5MR are provided by UNICEF.)
Notes: The " 5 index is a ratio of U5MR to IMR, and it has been designed by Dr. Hiroshi Maruyama. It is a qualitative index which indicates the degree of social causes of death for children under 5 years of age, such as accidents and effects of adverse living environment. This is because causes related to social factors account for the vast majority of mortalities in age group 1 - 4 years .
" 5 index = U5MR ÷ IMR = [mortality for children under 1 year of age + mortality for those 1 - 4 years of age] ÷ mortality for children under 1 year of age $ 1
Source: UNICEF

Concrete operation is as follows:

Step 1: Classifying each index into categories

In the case of each index A from B2 to B130 being below 25, 1 will be indicated from F2 to F130, and when it is at 25 - 54, 2 will be indicated, with 3 for 55 - 99, 4 for 100 or over. For index B (C2 - C130), four categories (1,2,3 or 4), and for index C (D2 - D130), two categories (1 for below 1.37 [median] or 2 for 1.37 or over), will be indicated in the same way.

Move the cursor to F2.
Type a functional formula : =IF(B2<25,1,IF(B2<55,2,IF(B2<100,3,4)))
Copy the formula from F3 to F130

Move the cursor to G2.
Type: =IF(C2<30,1,IF(C2<95,2,IF(C2<170,3,4)))
Copy the formula from G3 to G130.

Move the cursor to H2.
Type: =IF(D2<1.37,1,2)
Copy the formula from H3 to H130.

Step 2: Combining the categories into patterns

When combining the three categories into patterns, a 3-digit figure will be indicated from I2 to I130 using the following command and formula:

Move the cursor to I2.
Type: =(F2*100)+(G2*10)+H2
Copy the formula from I3 to I130

Step 3: Indicating selected patterns

Omitted because usually unnecessary.

Step 4: Re-arranging the records according to pattern group

After indicating a pattern for each record, the re-arrangement of the records according to pattern group should be carried out by using the sorting function. When using this function, it is better to change the formulae for indicating categories and patterns into values. The result of sorting is shown in table 4. Commands are as follows:

Change the formula of each cell into values;
Range: F2:I130
/EC
/ESV

Sort each record according to its pattern groups (column I);
Range: A2:I130
/DS
S: (column) I
A

Step 5: Counting the number of cases according to pattern group

When all the records have been re-arranged according to pattern groups, it is easy to count the number of cases (= records) belonging to each one. The idea of functional formulae and a retrieving/extracting function is extremely useful when summarizing case-counting in tabular form.

For counting the number, and indicating the total number of cases in each pattern group, the following command and formulae are used:

Type No at K1 and L1.
Move the cursor to J2 and type 1.

Move the cursor to J3.
Type the following formula: =IF(I3=I2,J2+1,1)
Copy the formula from J4 to J130.

Move the cursor to K2, and type: =IF(J2>=J3,J2,"")
Copy the formula from K3 to K130.

Step 6: Summarizing case-counting in tabular form

To tabulate the case-counting according to pattern group, the following commands for retrieving/extracting each total number of records are used:

Move the cursor to L2, and type: >=1
Type Ptn at M1 and No at N1.

Retrieve and extract the figures from K2 to K130 which fulfill the required condition: i.e. above 1 or over (result of case-counting).

/DFA
O
L: $I$1:$K$130
C:$L$1:$L$2
T:$M$1:$N$130

Finally, a summary table of case-counting according to pattern group can be indicated as in table 4, showing 37 cases for the pattern 111, and 27 cases for the pattern 442 etc.. The component ratio of 111, i.e. the pattern of low levels of both infant mortality rate and U5MR, and a below-average level of the " 5 index, is 28.7 % of all the countries in this table. It can be established that this pattern mainly consists of developed countries.

Table 4 A summing-up table of mortality patterns

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Countries Indices	IM	U5M	" 5idx		CtgA	CtgB	CtgC	Ptn		No	No	Ptn	No	Cp.Rt.
2	Australia	8	9	1.13		1	1	1	111	1		>	111	37	28.7
3	Austria	8	10	1.25		1	1	1	111	2			112	2	1.6
4	Belgium	10	12	1.20		1	1	1	111	3			121	2	1.6
5	Bulgaria	14	17	1.21		1	1	1	111	4			211	1	0.8
6	Canada	7	9	1.29		1	1	1	111	5			221	16	12.4
7	Chile	20	27	1.35		1	1	1	111	6			222	3	2.3
8	Costa Rica	18	22	1.22		1	1	1	111	7			321	7	5.4
9	Cuba	11	14	1.27		1	1	1	111	8			322	8	6.2
10	Czechoslovakia	11	13	1.18		1	1	1	111	9			331	2	1.6
11	Denmark	8	10	1.25		1	1	1	111	10			332	18	14.0
12	Finland	6	7	1.17		1	1	1	111	11			342	1	0.8
13	France	8	9	1.13		1	1	1	111	12			432	5	3.9
14	Germany, D.R.	8	9	1.13		1	1	1	111	13			442	27	20.9
15	Germany, F.R.	8	10	1.25		1	1	1	111	14			Tot	129	100.0
	*****	-	-	-		-	-	-	-	-					
	*****	-	-	-		-	-	-	-	-					
38	Yugoslavia	24	27	1.13		1	1	1	111	37	37				
39	Japan	4	6	1.50		1	1	2	112	1					
40	Singapore	8	12	1.50		1	1	2	112	2	2				
41	Korea, Rep.	24	31	1.29		1	2	1	121	1					
42	Panama	23	31	1.35		1	2	1	121	2	2				
43	Albania	25	30	1.20		2	1	1	211	1	1				
	*****	-	-	-		-	-	-	-	-					
	*****	-	-	-		-	-	-	-	-					
128	Sudan	105	175	1.67		4	4	2	442	25					
129	Tanzania,U.R.	103	173	1.68		4	4	2	442	26					
130	Yemen	116	192	1.66		4	4	2	442	27	27				

Codes: No = number, Cp.Rt. = component ratio

On the other hand, 442 and 332 patterns, i.e. high or very high levels of IMR and U5MR, together with average or above-average levels for the " 5 index, account for 20.9% and 14.0% respectively. These patterns belong to developing countries.

Among these patterns, we came across an abnormal pattern 112, which indicates low levels of IMR and U5MR but no correspondingly low level of " 5 index. Generally speaking,

children 1 - 4 years of age face less risk of death than infants (those under 1 year of age). Therefore, the " 5 index, which indicates the ratio of under fives′ mortality to infant mortality, should be close to 1 in conjunction with the development of better living conditions and the advance in medical technology etc.. Hence, pattern 111 is the normal type among developed countries as a result of mortality decline for children 1 - 4 years of age. Nevertheless, Japan and Singapore belonged to the 112 pattern, and only Japan has shown this pattern intermittently since 1985 (in 1986, 1987, 1988, 1989 and 1992). In Japan, although the IMR and U5MR were at the lowest level in the world, the " 5 index turned out to be 1.50, which was much higher than for Western and Northern European countries. It meant U5MR should decrease more there in proportion to the decline of IMR. When comparing the causes of death for children under 5 years of age in both Japan and Sweden, we found a relatively high percentage of "accidents and adverse effects" in Japan. Incidentally, the newspapers frequently report tragic cases of death among children after drowning, falling accidents, parental carelessness, or even maltreatment/killing by the mother as a result of lack of social protection or the mother′s accumulated stress and fatigue, etc.. [10] The result of this pattern analysis have added an new aspect to our approach to case studies of infant and child mortality.

4 Conclusions

The basic implication of and concrete procedure for the Statistical Pattern Analysis (SPA) have been presented in this paper. The two types of SPA have been described using a PC spread-sheet with the following examples: firstly, an integrated observation method for mixed data in the form of a cross-table; and secondly, an analysis method for large scale data with fields and records in the form of data matrices.

Although the SPA is perhaps one of the simplest and most useful methods for an integrated observation of mixed or categorical data, only the basic and general idea has been indicated here. Therefore, the SPA may well have some other applications[11], but at the same time also have some limitations not mentioned in this paper.

When using the SPA, the following three points should be given careful consideration: firstly, the criteria/threshold values[11] and number of categories for classification; secondly, the number and kinds of factors for combining these categories into a set of patterns; thirdly, the order in which the categorized indices are combined.

From the view point of safeguarding privacy, the SPA is relatively more suitable for processing individual case records than traditional methods. Nevertheless, it should be noted that a risk of privacy-invasion still remains even with using SPA, and that multi-cross data analyses which are able to specify the individual cases have now become technically possible using computer systems. Therefore, when using SPA, a deliberate procedure for safeguarding privacy is indispensable for analysts.

References

(1) Hiroshi Maruyama, *Research in Socio-Medical Science I: Infant Mortality*, (in Japanese), Tokyo, Iryotosho Publisher, 1976, pp.390-415.
H. Maruyama, *The Collected Papers of Hiroshi Maruyama 1*, (in Japanese), Tokyo, Nosan-Gyoson Bunka Kyokai, 1989, p.46.

(2) Mitsuo Fujioka, "Age and Cohort Analysis of Population by Classes", (in Japanese), *The Annual Report of the Regional Research Institute*, Asahikawa University, Asahikawa, Japan, 1988, p.80.
M. Fujioka, "A Contemporary Implication of "-index on the Study of Infant Mortality", (in Japanese), *The Journal of Asahikawa University*, No. 27, Asahikawa, Japan, 1988.
Hiroshi Maruyama, "My Fifty Years Research on Population Statistics", (in Japanese), Tokyo, *Statistics*, No.58, The Society of Economic Statistics, 1993, pp.11-12.

(3) M. Fujioka, Hiroshi Iwai, "The Methods of Pattern Research as an Analysing Tool", (in Japanese), *Statistical Study on the Employment Structure and the Stratum Structure of Labour Force in Modern Japan, Economic & Political Study Series*, No.84, Osaka, The Institute of Economic and Political Studies in Kansai University, 1993.

(4) H. Iwai, "The Changes of Employment Structure of Labour Force -Analysis on Cross Tables of the Employed by Industry and Occupation in 2 Sectors", (in Japanese), *Statistical Study on the Employment Structure and the Stratum Structure of Labour Force in Modern Japan*, (op.cit.).

(5) M. Fujioka, "Statistical Indication on the Change of Local Stratum by Age and Sex", (in Japanese), H. Iwai, *Labour Force - Class Composition and Employment Structure, Chousa to Siryo*, No.66, Osaka, The Institute of Economic and Political Studies in Kansai University, 1988, pp.14-33.
M. Fujioka, "The Composition, Change and Movement of Workers by Sex, Age, Industry and Occupation", pp.156-222, "The Migration of Stratum by Industry and Occupation in Osaka Metropolitan Area -Comparative Research by Prefecture-", pp.254-297, "The Migration of Stratum Population in Rural Area - Shimane Prefecture, the most Depopulated and Aging Region (in Japan) -", p.349, (in Japanese), *Statistical Study of the Employment Structure and Stratum Structure of the Labour Force in Modern Japan*, (op.cit.).

(6) The author has not yet considered the differences in detail between SPA and these statistical methods. Please make reference to the following books for example:
Richard O. Duda, Peter E. Hart, *Pattern Classification and Scene Analysis*, New York, John Wiley & Sons, 1973,
Alan Agresti, *Categorical Data Analysis*, New York, John Wiley & Sons, Inc., 1990.
Erling B. Andersen, *The Statistical Analysis of Categorical Data*, Berlin - Heidelberg, 1990.
Keinosuke Fukunaga, *Introduction to Statistical Pattern Recognition*, San Diego, Academic Press, Inc., 1990,
Robert J. Schalkoff, *Pattern Recognition: Statistical, Structural and Neural Approaches*, New York, John Wiley & Sons, Inc., 1992,

(7) Hiroshi Maruyama, "My Fifty Years' Research on Population Statistics", (op.cit.),.

(8) Michael R. Anderberg, *Cluster Analysis for Applications*, New York, Academic Press, 1973, pp.26-28.

(9) ibid. p.27

(10) For example, *Asahi Shinbun* (Newspaper), Japan
deaths relating from maltreatment or killing by mothers: 1995/10/16, 1995/5/28, 1995/2/18, 1995/2/15, 1993/9/21, 1993/4/8, 1992/9/23, 1991/4/9, 1989/6/29, 1989/6/1, 1987/5/22
deaths from drowning: 1995/8/9, 1994/12/18, 1994/6/5, 1993/8/13, 1993/5/16, 1992/6/6, 1991/11/18, 1991/9/2, 1991/1/5, 1990/7/18, 1989/7/24, 1989/6/5
deaths caused by falling down or being crushed: 1995/12/21,1992/8/20, 1990/11/10
deaths caused by being left alone and other reasons: 1994/10/21, 1994/9/10, 1994/2/23, 1993/1/13, 1991/12/28, 1991/8/5

(11) cf. M. Fujioka, "Statistical Pattern Analysis for Application", (in English), *Journal of Economics*, Vol.1, No.3-4, Shizuoka University, 1997.

Acknowledgment

The original idea for this method is based upon the research and private lectures of the late Dr. Hiroshi Maruyama. Furthermore, this paper is the result of several years' collaboration between Iwai and Fujioka, as well as Fujioka's research during his time as a visiting scholar with the Bureau of Statistics at the International Labor Office, Geneva (October 1993 to October 1994). Therefore, although each concrete idea for data processing of this method was dependent upon Fujioka's research and this paper was written by Fujioka himself, because Iwai and Fujioka had held consultative discussion prior to the writing of this paper, it is appropriate that the whole project should be published as a joint venture between Iwai and Fujioka.

The authors are grateful to Mr. Farhad Mehran, Director, ILO Bureau of Statistics for his help and advice on the research. They are also thankful to Mr. Chanyong Park (formerly with the ILO, Bureau of Statistics) for his cooperation in the creation of this paper, as well as to Mr. David Appleyard for his help with the language correction of this paper. Nevertheless, the authors are themselves solely responsible for the contents.