放送大学研究年報　第5号（1987）81-89 頁
Journal of the University of Air, No. 5 (1987) pp. 81-89

Cross-Tabulation and Statistical Inference

Chikio HAYASHI

相関分析と統計的推論

林　知己夫

要　旨

統計学の基本的方法として，相関表による分析がある．ものごとの関係を問題にするとき必ず用いられる方法である．これは，推論の基本的方法として，進んだ統計的理論によって精密化されている．

しかし，これには大きな陥穽がある．これについて述べたのが本論文である．どうしたことか，数理統計学において，この問題が看過されている．言及されたものを見たことはない――論文としてはもとより，教科書においてすら言及されていない――．ここでは，多次元的統計的データ分析の立場から，このパラドックスと見られる現象を取扱って，その解決を与える．内容は，1．個人表章と集団表章，2．多次元的相関表と推論，3．考えの筋道による分析，4．二分法論理と多次元的推論，となっている．

I. Introduction

Statisticians frequently encounter the problems I intend to discuss in this article. The ideas involved are fundamental to rudimentary statistical data analysis (before the application of sophisticated statistical tests and estimations), and they also may play an important role in the interpretation of statistical data. I don't know why these points should be disregarded in modern textbooks of statistics. It may be that they are considered too elementary and perhaps too trivial. However, many cases of " lying with statistics " are due to misuse of the most basic statistical methods.

I offer here some considerations of cross-tabulation and statistical inference. Although the population in question must always be taken into consideration, I will not focus on this aspect of data analysis in this paper.

II. Inference concerning the collective vs. the individual

We often encounter the following situation in social survey data analysis. First, we define " majority opinion " as an opinion which is supported not only by more than 2/3 of the total sample, but also by more than 2/3 of each of the breakdowns in sex,

Table 1. Percentage (%) of categories in the Total

	category A in question 1	category B in question 2	category C in question 3	category D in question 4
1973	73	81	90	79
1968	78	84	91	83

age and education. Though it is often known which opinions are supported by more than 2/3 of the total, majority opinions, as defined here, are rarely found. I refer to actual surveys in which opinions A, B, C and D for example, were found to be " majority opinions " in four questions, 1, 2, 3 and 4, respectively (see Table 1).

It can be said that such majority opinions characterize the collective. We may call these majority opinions, " typical opinions " in the collective. This concept of majority opinion applies only to the collective entity as a whole. Those individuals in the majority responding in a certain way to a particular attitudinal question will not necessarily respond with the majority on any other attitudinal question. Indeed, the percentage of those individuals who responded with the majority on all four questions shown in Table 1 amounted to only 44% in each time period. The typical individuals who respond only with opinions typical to the collective are not always typical in the collective, i.e. they do not represent the majority but some minority. In addition, the number of individuals who respond with the majority decrease as the number of " majority opinions " found in the survey increases. Therefore, in the data interpretation, great care must be taken to observe the difference between an inference based on the collective entity and that based on the individual.

III. Chain of reasoning by cross-tabulation

For the sake of simplicity, we will discuss analysis by breakdown of dichotomies. For example, categories such as " agree " and " disagree " are adopted, as in the table by sex and the table by age, shown in Table 2. Let the size of the sample be 400. In consideration of the date, we see that " agree " is a majority opinion for the " male " category and for the " young " (<35 years old). Laymen are tempted to conclude that young males " agree " with a higher percentage than either the " young " or " male " categories.

Table 2 is calculated from Table 3 and they are consistent. However, in the breakdown of young men, " disagree " is a majority opinion. (See Table 3). The percentage of agreement is 66 in the male category and 66 in the young category, while it is 41 in the young male category. Table 2 is calculated from Table 4 too and these tables are also consistent. In this table, in the breakdown of young men, " agree " is a majority opinion. The percentage is 74 and very high.

Table 2. Analysis by sex and age

category / sex	agree	disagree	Total
male	132 (66)	67 (34)	199 (100)
female	93 (46)	108 (54)	201 (100)
Total	225	175	400

category / age	agree	disagree	Total
<35 years old	131 (66)	67 (34)	198 (100)
≧35 years old	94 (47)	108 (53)	202 (100)
Total	225	175	400

(··) means percentage.

Table 3. Analysis by sex×age

category		agree	disagree	Total
sex	age			
male	<35	41 (41)	58 (59)	99 (100)
	≧35	91 (91)	9 (9)	100 (100)
female	<35	90 (91)	9 (9)	99 (100)
	≧35	3 (3)	99 (97)	102 (100)
Total		225	175	400

(··) means percentage.

Table 4. Analysis by sex×age

category		agree	disagree	Total
sex	age			
male	<35	73 (74)	26 (26)	99 (100)
	≧35	59 (59)	41 (41)	100 (100)
female	<35	58 (59)	41 (41)	99 (100)
	≧35	35 (34)	67 (66)	102 (100)
Total		225	175	400

(··) means percentage.

These seeming contradictions show that the chain of reasoning based on separate single tabulation by breakdowns is not consistent with the results of a single tabulation by combination of breakdowns. The latter is, of course, correct. The additive reasoning of careless laymen is sometimes correct and sometimes incorrect. This is

a problem of the existence or non existence of interaction between sex and age.

IV. Cross-tabulation and ways of thinking

In some comparative studies, it is not sufficient merely to compare the marginal distributions of responses, because this cannot clarify the ways of thinking which express characteristics of various groups. Differences in the ways of thinking form a barrier to mutual understanding and consequently result in the lack of communication. Frequently we meet situation in cross-societal or cross-cultural surveys.

Let us consider the following simple example as a case in which the ways of thinking are different even though the marginal distributions are the same. Suppose we have two questions. Ways of thinking may be clarified by the cross tabulation of these two questions. Let the questions be I and II, and the responses be dichotomous, (α_1, α_2) and (β_1, β_2) respectively. Let A and B be two groups, each group consisting of 1,000 peoples.

The marginal distributions are exactly the same and there is no difference between groups A and B on either question I or II (Table 5). If we take two cross tabulations, however, quite different patterns are obtained, as we see in Table 6. In group A, α_1 and β_1, and α_2 and β_2 are closely related, whereas in group B, α_1 and β_2, and α_2 and β_1 are closely related. That is to say, a strong relation exists between α_1 and β_1, and α_2 and β_2 in group A, and between α_1 and β_2, and α_2 and β_1 in group B. In a case like this, mutual understanding between groups A and B is extremely difficult, as the way of thinking, which is revealed by the cross-tabulation of two questions, is different.

It can be recognized that the way of thinking will be realized through the

Table 5. Marginal distribution in A, B group

	I		II		Total
	α_1	α_2	β_1	β_2	
A-group	600	400	600	400	1,000
B-group	600	400	600	400	1,000

Table 6. Cross-tabulation in A, B group

I \ II	β_1	β_2	Total
α_1	500	100	600
α_2	100	300	400
Total	600	400	1,000

I \ II	β_1	β_2	Total
α_1	200	400	600
α_2	400	0	400
Total	600	400	1,000

Table 7. Time series data

1960

I \ II	+	−	Total
+	1,000	0	1,000
−	0	1,000	1,000
Total	1,000	1,000	2,000

1965

I \ II	+	−	Total
+	750	250	1,000
−	250	750	1,000
Total	1,000	1,000	2,000

1970

I \ II	+	−	Total
+	500	500	1,000
−	500	500	1,000
Total	1,000	1,000	2,000

1975

I \ II	+	−	Total
+	250	750	1,000
−	750	250	1,000
Total	1,000	1,000	2,000

1980

I \ II	+	−	Total
+	0	1,000	1,000
−	1,000	0	1,000
Total	1,000	1,000	2,000

analysis of response patterns of people on many question items. We call the group structure, which is revealed by the analysis of response patterns of individuals, the way of thinking of that group. In this way, the idea of cross-tabulation analysis may be regarded as a useful method to reveal ways of thinking.

In the discussion above, a simple example is given of the difference in ways of thinking between groups A and B. In order to explore ways of thinking through the relations among many questions, a factor analytic method in qualitative or categorical data may be applied. This method is called quantification on response pattern by this author, and the analysis of correspondences by Professor J.-P. Benzécri (University VI of Paris).

Suppose that we have the time series survey data as shown in the longitudinal study in Table 7. Two dichotomous questions, I and II, are considered. Responses are either agree (+) or disagree (−).

The ways of thinking have gradually changed even though the marginal distributions of questions I and II have remained constant over the twenty years. The ways of thinking have rotated and reversed themselves. I have experienced a dramatic change such as this one in Table 7, in time series data, in a 30 year longitudinal study of Japanese national character (ad. hoc. Committee of Institute of Statistical Mathe-

matics), conducted through social surveys which were repeated every 5 years.

V. Meaning of inference by cross-tabulation

This topic refers to marginal distribution of groups, association analysis by cross-tabulations of individuals in each group and follow-up study individuals. An example of this type of analysis was found in the prospective epidemiological study of cardio vascular disease, outlined below.

For the sake of discussion, we take an artificial example. We assume two questions, each dichotomous. Let us consider two groups A and B. The categories are represented by + and −. For illustrative purposes, I will use epidemiological terms. Question I refers to in-take of salt in the diet where + means high in-take of salt in daily meals (as evidenced by a health care survey) and − means a low in-take of salt. Question II refers to diagnosis of hypertension where + means hypertension and − means no hypertension.

Table 8 shows the results of cross-tabulation of question II by question I in groups A and B. The conclusion is that there is not a correlation between I and II in either group. In neither group is the relation between the high in-take of salt and the outcome of hypertension confirmed. In other words, the hypothesis of no correlation is not rejected. The dependent relation between the two questions does not emerge by any association test.*

But the percentage of high in-take of salt in question I and the outcome of hypertension in question II are high in group A, while the percentage of high in-take of salt and the outcome of hypertension are low in group B. The comparison between the two groups by marginal distribution alone suggests a closed relationship between

Table 8. Cross-tabulation by I and II

A-group

I \ II	+	−	Total
+	400	200	600
−	200	100	300
Total	600	300	900

B-group

I \ II	+	−	Total
+	100	200	300
−	200	400	600
Total	300	600	900

* Outcome of hypertension is not only induced by high in-take of salt but by many factors and their interactions. It is well-known that high in-take of salt is only one of important factors and does not always make a cause of hypertension under the interaction with other factors. If one single factor (salt in-take) is taken and two-way analysis (outcome of hypertension × salt in-take) is used without regarding the influences of other relevant factors, we often meet this situation. This suggests the necessity of the idea of multidimensional data analysis.

Table 9. Changing pattern

in-take of salt*	hypertension**				
+	+	400	−	−	$400\times\frac{4}{5}=320$
			−	+	$400\times\frac{1}{5}=80$
−	+	200 →	−	+	200 (no change)
+	−	200	+	+	$200\times\frac{2}{5}=80$
			+	−	$200\times\frac{3}{5}=120$
−	−	100 →	−	−	100 (no change)

* + means high in-take and − means low in-take
** + means hypertension and − means normal blood pressure

high in-take of salt and the outcome of hypertension. Some careless epidemiologists may conclude that a closed relationship exists. Other careless epidemiologists, checking the results of only a single cross-tabulation in group A(B), are apt to conclude that strong association may not be observed between high in-take of salt and the outcome of hypertension. Epidemiologists in general may be perplexed as to what conclusion should be drawn when more than two groups are studied and results similer to those mentioned above are obtained.

It is well known that the blood-pressure of patients with hypertension often falls with adherence to a low daily in-take of salt. Doctors consider that low in-take of salt lowers the blood-pressure. General epidemiologists are concerned in light of the above three findings.

To investigate this problem statistically, we assume that the blood pressure of 4/5 (four fifths) of the patients with hypertension will change to normal, and the blood pressure of the remaining one fifth will remain unchanged, with a program of low in-take of salt. We further assume that the blood pressure of 2/5 (two fifths) of the population with normal blood-pressure and high in-take of salt will change to a condition of hypertension. Suppose that the patients in group A with hypertension (600 in Table 8) adhere to a program of low in-take of salt. According to the above assumption, this results in the pattern illustrated in Table 9. Table 10 is obtained by cross-section survey, after patients have received the guidance of the low salt in-take program, using data from groups A of Table 8 and from Table 9.

It must be remarked that Table 10 also shows the correlation between question I and question II. Cross-tabulations in the cross-section data for the two periods of before and after adherence to a low salt diet show no pattern of correlation between

Table 10. Cross-tabulation by I and II after guidance

I \ II	+	−	Total
+	0+80 80	120 120	200
−	200+80 280	100+320 420	700
Total	360	540	900

Table 11. Another case of cross-tabulation

A-group

I \ II	+	−	Total
+	300	100	400
−	100	400	500
Total	400	500	900

B-group

I \ II	+	−	Total
+	250	150	400
−	150	350	500
Total	400	500	900

Table 12.

A-group

I \ II	+	−	Total
+	210	340	550
−	340	10	350
Total	550	350	900

B-group

I \ II	+	−	Total
+	50	300	350
−	300	250	550
Total	350	550	900

the two questions, even though there has been a strong effect on individuals. In the comparison of marginal distributions for group A for the two periods, it is found that the frequency of the condition of hypertension decreased after following the diet. Epidemiologically speaking, this suggests the existence of serious problems in data analysis. What does analysis of the cross-tabulation of individuals mean? In this case, the marginal distributions tell the truth.

Statistically speaking, this is no problem, if we follow the above logic. We conclude that such cases do exist. The point is not to reach false conclusions by simple data analysis.

In contrast to the situation revealed in Table 8, we often find situations such as that revealed in Table 11. In this case, the data analysis by cross-tabulation tells the true story. In both group A and group B, the relationship between question I and question II is confirmed, although the marginal distributions are always the same

in the two groups, and therefore not informative.

Another crucial example is shown in Table 12, where the inference on marginal distribution and that on cross-tabulation imply an inverse conclusion.

According to marginal distributions on I and II, A-group is higher in I+ and II+, while B-group is lower in I+ and II+. The conclusion is that + in I implies + in II and − in I implies − in II.

According to cross-tabulation, the conclusion is that the relationships (+ in I and − in II) and (− in I and + in II) are dominant both in A-group and B-group, i.e. the relationships (+ −) or (− +) in I and II are confirmed.

We can find this apparently paradoxical situation when we take up only one factor and analyze by a simple cross-tabulation (as a contingency table), disregarding that the outcome (i.e. hypertension in this case) is brought out by many factors (including salt high in-take as only one factor) and their combinations.

Although I have used epidemiological terms here, similar situations may also appear in general analysis of social surveys. In these cases, question I is regarded as an independent variable, and question II as a dependent variable. The hypothesis being tested is that of question I being a cause and question II being a result.

Analysis by marginal distribution in groups, association analysis by cross-tabulation of indviduals in those groups, and follow-up studies of individuals are analyzed on different levels of statistical logic, and they are not always consistent. Investigation of the features of these three kinds of data analysis is informative. And comparison and examination of the results, obtained by the statistical logic mentioned above, with the actual data and other relevant information, can be very fruitful in medical science and social science research. This suggests the necessity of multidimensional data analysis as a tool of exploration. This leads to the development of statistical data analysis for optimal process control. The construction of well-organized data base and the development of well-designed statistical and data analytic soft ware are indispensable for this realization.

In any case, in order to explore reality through the design and analysis of data, it is necessary to consider the characteristics of the data, to apply suitable methods of data analysis, and to compare the statistical findings. Simple data analysis is often misleading.

(昭和62年9月11日受理)