THE EFFECT OF THE STRUCTURE OF CONTINGENCY TABLE DATA ON STUDENTS' INFERENCES

Kai-Lin Yang, Ting-Ying Chu, and Xin-Yi Liu
National Taiwan Normal University, Taiwan
60540015s@ntnu.edu.tw

The aim of this research is to understand the students' spontaneous inferences on different kinds of contingency tables. The sign of the association, the type of tables and the number structure are considered to design problems. The 10th graders of this study have not received any school instruction on contingency tables prior to the test. The results show that the contingency table problems can be classified by three factors composed of the interaction between the sign of the association and the number structure. The three factors are named as (1) correlated data, (2) uncorrelated data with the similar ratios between column values, (3) uncorrelated data with the same ration.

INTRODUCTION

With the information society and the age of big data, statistics, as an important tool, is widely used in many fields. Hence, being a good statistical citizen must be able to explain, decide, judge, evaluate, and make decisions about the information (Rumsey, 2002). In the case of senior high school, they have learned not only the reasoning of distribution, but also about the inference in the two variables. The inference of the two variables is involved in the course in many countries, such as Australia, Spain, China, etc. However, the course focuses mainly on the two numerical variables, rarely on the two categorical variables. Batanero et al. (1996) have analyzed secondary school students' inference and strategies in contingency tables at the level that the topic of association is introduced in their curriculum. They made an approach to the three variables, such the type of tables, sign of association (including direct, inverse and independence) and relationship between context and prior belief, and these variables may affect the students' inferences on contingency tables. In addition, they also found that their students are lack the concept of proportional reasoning.

In view that the structure of numbers effect students' proportional reasoning (Steinthorsdottir, 2006), we further consider this variable to investigate the effect on senior high school students' inference on contingency tables. We wonder the proportional reasoning may not only benefit but also disadvantage students' inference on contingency tables in the case of independence. Hence, we developed the questionnaires involved three factors - the type of tables, the sign of the association and the number structure - to justify whether they would affect the students' inferences on contingency tables or not.

METHOD

Participants

The participants in this study were 536 tenth-graders of eight Taiwan senior high schools. None of them had received any school instruction on contingency tables prior to the test. The questionnaires were delivered as a paper-and-pencil test.

Questionnaire

In this study, we used 2×2 and 2×3 contingency tables that varied with respect to the cell frequencies. According to Obersteiner et al. (2015) study, we used the content about the two bags and two different color balls, that is, the contingency tables represented the number of the balls with two different colors (blue and red) which be randomly drawn out from one of the two different bags (A and B). However, in our study, contingency tables were presented in numbers rather than iconic representations. The problems in the questionnaire are also different from Obersteiner et al. (2015). The problems in our study were to decide whether "what bags that you draw out from" and "what color you draw out" have association or not.

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2018 ISI/IASE]

The items in the questionnaire involved three task variables, including the type of tables, the sign of the association and the number structure. The test was divided into two parts by the type of tables - 2×2 and 2×3 contingency tables - and both of them were composed of 9 items. Moreover, we designed five different types of the number structure:

(A) data with the different ratios between column values which the ratio is larger than 1
(B) data with the different ratios between column values which the ratio is smaller than 1
(C) data with the same differences of the diagonal values
(D) data with the same ratios between column values
(E) data with the same sum of the diagonal values.

The items that we designed were shown in Table 1.

Table 1. Items of the questionnaire used in this study

The type of tables	2×2		2×3	
The sign of the association	correlated data	uncorrelated data	correlated data	uncorrelated data
(A)	20 160 20 30 (Item 1)	30 120 30 135 (Item 5)	15 30 160 15 25 35 (Item 15)	30 60 120 30 65 125 (Item 11)
(B)	80 60 80 10 (Item 8)	80 40 80 45 (Item 2)	100 70 60 100 65 5 (Item 17)	120 60 30 120 65 35 (Item 10)
(C)	30 90 80 40 (Item 6)	30 30 90 90 (Item 9)	20 45 70 60 35 10 (Item 13)	30 30 30 90 90 90 (Item 18)
(D)		80 60 40 30 (Item 3)		120 60 20 60 30 10 (Item 14)
(E)	70 130 10 70 (Item 4)	60 80 40 60 (Item 7)	70 80 130 10 20 70 (Item 12)	50 60 70 30 40 50 (Item 16)

Data analysis

In the proportional reasoning, if the values of columns are proportional, they are uncorrelated actually. But we conjectured that students may make mistakes because of the number structure with the relation of proportion. Hence, we analyzed the items with correlated data and the items with uncorrelated data respectively. As for the items with correlated data, we find 8 items (item 1, 4, 6, 8, 12, 13, 15, 17), and as for the items with uncorrelated data, we find 10 items (item 2, 3, 5, 7, 9, 10, 11, 14, 16, 18). Randomly drew 200 valid samples running item analysis and exploratory factor analysis, and the other 336 samples running confirmatory factor analysis for the whole items.

RESULTS

We showed the result of item analysis and factor analysis as for the items with correlated date as well as the uncorrelated data respectively.

Correlated data

For item analysis, a T-test was performed to compare mean scores of each items. The result showed statistically significant difference in the mean scores, except the item 12, 15, 17 due to the standard deviations were 0 (Table 2). According to Hair et al. (1998), the Cronbach's alpha result of 0.60-0.70 was at the lowest limit of acceptability for the internal consistent reliability coefficient based on correlation between variables. The correlation coefficient of these items were higher than 0.3, and the Cronbach's alpha was 0.914 that means these items measured the same concepts. Thus, exploratory factor analysis (EFA) was then run with all data from these eight items.

Table 2. Items analysis of 8 items with correlated data

Item	Mean	SD	t-value	Correlation Coefficient	Factor Loading
1	0.60	0.491	62.746*	0.720	0.793
4	0.58	0.494	72.744*	0.698	0.773
6	0.64	0.481	40.978*	0.656	0.735
8	0.57	0.496	127.000*	0.759	0.825
12	0.53	0.500	-	0.719	0.790
13	0.58	0.494	127.000*	0.710	0.782
15	0.58	0.495	-	0.754	0.822
17	0.56	0.497	-	0.728	0.800

**p*< 0.005

Use SAS 9.4 to run EFA of 8 items. EFA yielded one factor with eigenvalues greater than 1.0. The middle column (Factor 1) in Table 3 presents the results of EFA.

Table 3. Exploratory factor analysis of 18 items

The sign of the association	Item	Factor		
		1	2	3
Correlated date	1	0.875		
	4	0.892		
	6	0.886		
	8	0.926		
	12	0.907		
	13	0.894		
	15	0.916		
	17	0.868		
Uncorrelated date	2		0.951	
	5		0.890	
	7		0.682	
	10		0.951	
	11		0.896	
	16		0.788	
	3			0.825
	9			0.868
	14			0.902
	18			0.746

Uncorrelated data

For item analysis, a T-test was performed to compare mean scores of each items. The result showed statistically significant difference in the mean scores (Table 4). According to Hair et al. (1998), the Cronbach's alpha result of 0.60-0.70 was at the lowest limit of acceptability for the internal consistent reliability coefficient based on correlation between variables. The correlation coefficient of these items were higher than 0.3, and the Cronbach's alpha was 0.866 that means these items measured the same concepts. Thus, EFA was then run with all data from these 10 items. Use SAS 9.4 to run EFA of 10 items. EFA yielded two factors with eigenvalues greater than 1.0. The fourth and fifth columns (Factor 2 & 3) in Table 3 presents the results of EFA.

Table 4. Items analysis of 10 items with uncorrelated data

Item	Mean	SD	t-value	Correlation Coefficient	Factor Loading
2	0.51	0.500	34.319*	0.664	0.778

3	0.63	0.482	14.603*	0.501	0.563
5	0.48	0.500	27.027*	0.623	0.745
7	0.44	0.497	25.305*	0.588	0.686
9	0.65	0.477	13.996*	0.477	0.535
10	0.51	0.500	34.021*	0.653	0.770
11	0.49	0.500	45.153*	0.686	0.791
14	0.59	0.493	17.394*	0.523	0.583
16	0.48	0.500	26.024*	0.603	0.710
18	0.64	0.482	13.859*	0.482	0.543

*$p< 0.005$

Use SAS 9.4 to run confirmatory factor analysis (CFA) of 18 items. Table 5 presents the results of CFA, confirming the factor structure found in the exploratory stage.

Table 5. Confirmatory factor analysis of 18 items

χ^2/df	GFI	AGFI	NFI	RMR	RMSEA
2.938	0.863	1.000	0.968	0.054	0.076

CONCLUSION

According to the results of exploratory factor analysis (EFA) of 18 items, we totally obtain three factors, one of them is obtained from correlated data and the others are obtained from uncorrelated data. It means that the contingency table problems can be classified by three factors composed of the interaction between the sign of the association and the number structure. Hence, we name these three factors on the basis of features of the items. The three factors are named as (1) correlated data, (2) uncorrelated data with the similar ratios between column values, (3) uncorrelated data with the same ration.

Comparing to the result with Batanero et al. (1996), we made the following discussions. First, we found that the sign of correlation is an important factor that affect students' inferences on contingency tables. Moreover, this factor from Batanero et al. (1996) was divided into direct and inverse, but our result did not. The reason may be that they didn't control the variable of the content, for example, each content dealt with only one sign of correlation, namely they didn't experiment both signs with the same content. On the other hand, because of using the same content, our study found that the sign of association was not affect students' inferences. Secondly, in the uncorrelated data, we divided it into two factors, such as with the similar ratios between column values, and uncorrelated data with the same ration, which was different from Batanero et al. (1996).

Based on our findings, further studies could investigate whether the learning of correlation affect students' inferences on contingency tables.

REFERENCES

Batanero, C., Estepa, A., Godino, J. D., & Green, D. R. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education*, 151-169.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). *Multivariate data analysis*, *5*(3), 207-219.

Obersteiner, A., B., M., & Reiss, K. (2015). Primary school children's strategies in solving contingency table problems: the role of intuition and inhibition. *ZDM*, *47*(5), 825-836.

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, *10*(3).

Steinthorsdottir, O. B. (2006). Proportional reasoning: Variable influencing the problems difficulty level and one's use of problem solving strategies. *Proceedings of the 30th conference of the international group of Psychology of Mathematics Education*, *5*, 169-176.