# Testing for marginal independence between two categorical variables with multiple responses

*Tung Nguyen, Fangyuan Zhang*

*3/17/2020*

## Abstract

Mind and body practices may be used to improve health and well-being or to help manage symptoms of health problems. The 2012 National Health Interview Survey collected information about Americans' top 3 commonly used modalities and whether using each modality is because of the recommendation from any doctor, family member, or friend. There are statistical methods to study the association between two multiple response categorical variables, i.e. modality selection and recommendation in the survey. However, due to the way the questions were designed in the survey, no information was collected about failure recommendation. No existing efficient statistical methods can be directly applied to the scenario. In this study, we proposed a modified Pearson chi-square statistic to analyze the special data structure in the survey. Simulations were conducted to evaluate the proposed method.

## Introduction

Often, in surveys, there are questions that ask respondents to "choose all that apply" from a set of items. Multiple categorical random variables are categorical variables that summarize these types of questions in a survey. Below is an example from (Bilder-and-Loughin-2007)

**Table 1**

*Marginal table for Kansas farmer data. There are a total of* 279 *farmers who participated in the survey. The percentage of farmers picking a source of veterinary information and waste storage method pair are given next to the counts. The shaded cell corresponds to the example given in Section* 1.

| | Sources of veterinary information | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Professional consultant | | Veterinarian | | State or local extension service | | Magazines | | Feed companies & representatives | |
| Waste storage methods | | | | | | | | | | |
| Lagoon | 34 | 12.19% | 54 | 19.35% | 50 | 17.92% | 63 | 22.58% | 41 | 14.70% |
| Pit | 17 | 6.09% | 33 | 11.83% | 34 | 12.19% | 43 | 15.41% | 37 | 13.26% |
| Natural drainage | 6 | 2.15% | 23 | 8.24% | 30 | 10.75% | 49 | 17.56% | 34 | 12.19% |
| Holding tank | 1 | 0.36% | 4 | 1.43% | 4 | 1.43% | 6 | 2.15% | 2 | 0.72% |

The survey was conducted by the department of Animal Sciences at Kansas State University. There were two questions in the survey about the sources of veterinary information and swine waste storage methods. Kansas farmers were allowed to select as many responses as there were from the list of items. For instance, 34 farmers selected professional consultant as a source of veterinary information and lagoon as a waste storage method. A question of interest is whether the sources of veterinary information are independent of waste storage methods. Traditional Pearson chi-square test is not appropriate here because the multiple responses are likely dependent. Hence, An alternative test should be performed.

A simultaneous pairwise marginal independence (SPMI) test is performed to determine independence of each source of veterinary information and each waste storage methods. (Bilder, C., & Loughin, T. (2004)). Let W and Y denote the multiple-response categorical variables for an r x c table. Correspondingly, W and Y refer to

## References

Bilder and T. Loughin. Modeling association between two or more categorical variables that allow for multiple category choices. Communications in Statistics–Theory and Methods, 36(2):433–451, 2007. [p144, 146, 149]

Bilder, C., & Loughin, T. (2004). Testing for Marginal Independence between Two Categorical Variables with Multiple Responses. Biometrics, 60(1), 241-248. Retrieved March 17, 2020, from www.jstor.org/stable/3695573