Journal of Applied Statistics
Vol. 35, No. 4, April 2008, 465–474

Ranking responses in multiple-choice questions

Hsiuying Wang*

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

In many studies, the questionnaire is a common tool for surveying. There are two kinds of questions designed: single-choice questions and multiple-choice questions. For single-choice questions, the methodology for analyzing it has been provided in the literature. However, the analyses of multiple-choice questions are not established as in depth as those for single-choice questions. Recently, there has been a lot of literature published about testing the marginal independence between two questions involving at least one multiple-choice question. However, another important problem regarding this topic is to rank the responses in a multiple-choice question. The issue is whether there are significant differences in the popularity of particular responses within the same question. In this paper, methodologies for ranking responses are proposed.

Keywords: single-choice question; multiple-choice question; survey; likelihood ratio test; Wald test; ranking consistency

1. Introduction

The questionnaire method is a widely-used tool for researchers in any field to collect information. The researchers list questions that they are interested in, in a questionnaire, and analyze the survey data collected from interviewing the respondents. There are two kinds of questions: single-choice questions and multiple-choice questions. The analyses of single-choice questions have been investigated in the literature and textbooks. Approaches of analyzing multiple-choice questions have been lacking until recently. Umesh [5] first discussed the problem of analyzing multiple-choice questions. Agresti and Liu [1] discuss the modeling of multiple-choice questions. Loughin and Scherer [4], Decady and Thomas [3] and Bilder *et al.* [2] propose several methods for testing marginal independence between a single-choice question and a multiple-choice question.

The above-mentioned papers focus on the analyses of dependence between a single-choice question and a multiple-choice question. However, for most researchers, they are also interested in ranking the responses in a question according to the probabilities of responses being chosen. We have not found any literature discussing this problem. Thus, in this paper, methods of ranking

*Email: hywang@stat.sinica.edu.tw

ISSN 0266-4763 print/ISSN 1360-0532 online

DOI: 10.1080/02664760801924533
http://www.informaworld.com

responses are provided. We first focus on ranking two specific responses that we are interested in. For example, a company is designing a marketing survey to help develop an insect killer. The researchers list several factors, including high quality, price, packaging and smell that could affect the sales market. Especially, they are interested in comparing the two factors of high quality and sales price. Since a high quality product has a higher manufacturing cost, to balance the profits, the sales price needs to be raised. The possibility exists that most consumers are more concerned about the price than the quality of the product. Thus, researchers want to know which factor of the two is more important for most consumers. Suppose that a group of individuals are surveyed on purchasing an insect killer. They are asked to fill out questionnaires which list all the questions that we wish to address to each respondent. The following is a multiple-choice question in the questionnaire:

Question 1 Which reasons are important to you when considering the purchase of an indoor insect killer? (1) price (2) high quality (3) packaging (4) smell (5) others.

In this example, besides being interested in ranking the two factors of sales price and high quality, the researchers also want to investigate the importance of the other responses.

First, we consider ranking two specific responses. For the general case, assume that a multiple-choice question has k responses, $v_1, \ldots, v_k$, and we interview n respondents. Each respondent is asked to choose at least one and at most s answers for this question, where $0 < s \leq k$. If $s = 1$, it is a single-choice question. There are a total of $c = C_1^k + \cdots + C_s^k$ possible kinds of answers that respondents will choose. Let $n_{i_1,\ldots,i_k}$ denote the number of respondents selecting the responses v_h and not selecting $v_{h'}$ if $i_h = 1$ and $i_{h'} = 0$, and $p_{i_1,\ldots,i_k}$ denotes the corresponding probability. For example, when $k = 7$, $n_{0100100}$ denote the number of respondents selecting the second and the fifth responses and not selecting the other responses. Thus, the pmf function of $n_{i_1,\ldots,i_k}$ is

$$f_s(n_{i_1\cdots i_k}) = I\left(\sum_{j=1}^{k} i_j \leq s\right) \frac{n!}{\prod_{i_j=0 \text{ or } 1} n_{i_1\cdots i_k}!} \prod_{i_j=0 \text{ or } 1} p_{i_1\cdots i_k}^{n_{i_1\cdots i_k}}, \tag{1}$$

where $I(\cdot)$ denotes the indicator function. Let m_j denote the sum of the number $n_{i_1,\ldots,i_k}$ such that the jth response is selected, and π_j denote the corresponding probability, that is $m_j = \sum_{i_j=1} n_{i_1,\ldots,i_k}$ and $\pi_j = \sum_{i_j=1} p_{i_1,\ldots,i_k}$. Note π_j is called a marginal probability of response j. Also let m_{jl} denote the sum of the number $n_{i_1,\ldots,i_k}$ such that the jth and lth responses are selected, and π_{jl} denote the corresponding probability. Then $m_{jl} = \sum_{i_j=i_l=1} n_{i_1,\ldots,i_k}$ and $\pi_{jl} = \sum_{i_j=i_l=1} p_{i_1,\ldots,i_k}$.

For ranking the importance of two specified responses, say response 1 and response 2 in Question 1 from the survey data, we will consider the two-sided test:

$$H_0 : \pi_1 = \pi_2 \; vs \; H_1 : \pi_1 \neq \pi_2, \tag{2}$$

which is equivalent to

$$H_0^* : \pi_1 - \pi_{12} = \pi_2 - \pi_{12} \; vs \; H_1^* : \pi_1 - \pi_{12} \neq \pi_2 - \pi_{12}. \tag{3}$$

If (2) is rejected, then we can rank the response with larger m_j first.

In Section 2, we will propose several methods to test (2). In Section 3, simulation results of comparing the rejection rates and the powers of the methodologies are presented. Besides ranking two responses, a rule for ranking all responses is proposed in Section 4. A ranking consistency property is also introduced in Section 4.

2. Testing approach

In this section, we will propose three methods for testing (2).

2.1. *Wald test*

A Wald test is a test based on a statistic of the form

$$Z_n = \frac{W_n - (\pi_1 - \pi_2)}{S_n},$$

where W_n is an estimator of $\pi_1 - \pi_2$, and S_n is a standard error for W_n. An unbiased estimator of $p_{i_1,\ldots,i_k}$ is $n_{i_1,\ldots,i_k}/n$, which is also a maximum likelihood estimator (MLE). Let $\hat{\pi}_1 = m_1/n$, $\hat{\pi}_2 = m_2/n$ and $\hat{\pi}_{12} = m_{12}/n$. We can use $\hat{\pi}_1 = m_1/n$ and $\hat{\pi}_2 = m_2/n$ as estimators of π_1 and π_2 respectively. Note that $\hat{\pi}_1 - \hat{\pi}_2 = (\hat{\pi}_1 - \hat{\pi}_{12}) - (\hat{\pi}_2 - \hat{\pi}_{12})$. By the facts $E(m_1) = n\pi_1$, $E(m_2) = n\pi_2$ and $\mathrm{cov}(\hat{\pi}_1 - \hat{\pi}_{12}, \hat{\pi}_2 - \hat{\pi}_{12}) = -(\pi_1 - \pi_{12})(\pi_2 - \pi_{12})/n$, we have

$$E\left(\hat{\pi}_1\right) = \pi_1,\ E\left(\hat{\pi}_2\right) = \pi_2,\ E\left(\hat{\pi}_{12}\right) = \pi_{12},$$

and

$$\begin{aligned}
\mathrm{Var}\left(\hat{\pi}_1 - \hat{\pi}_2\right) &= \mathrm{Var}\left((\hat{\pi}_1 - \hat{\pi}_{12}) - (\hat{\pi}_2 - \hat{\pi}_{12})\right) \\
&= (\pi_1 - \pi_{12})(1 - \pi_1 + \pi_{12})/n + (\pi_2 - \pi_{12})(1 - \pi_2 + \pi_{12})/n \\
&\quad + 2(\pi_1 - \pi_{12})(\pi_2 - \pi_{12})/n \\
&= (\pi_1 - \pi_{12})(1 - \pi_1 + 2\pi_2 - \pi_{12})/n \\
&\quad + (\pi_2 - \pi_{12})(1 - \pi_2 + \pi_{12})/n.
\end{aligned} \tag{4}$$

When $s = 1$, π_{12} is zero, which leads to

$$\mathrm{Var}\left(\hat{\pi}_1 - \hat{\pi}_2\right) = \begin{cases} \pi_1(1 - \pi_1)/n + \pi_2(1 - \pi_2)/n + 2\pi_1\pi_2/n & \text{if } s = 1, \\ (\pi_1 - \pi_{12})(1 - \pi_1 + 2\pi_2 - \pi_{12})/n + & \\ (\pi_2 - \pi_{12})(1 - \pi_2 + \pi_{12})/n & \text{otherwise.} \end{cases} \tag{5}$$

Under the null hypothesis H_0 in (2) and based on the central limit theorem, the statistics

$$\frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\mathrm{Var}(\hat{\pi}_1 - \hat{\pi}_2)}} \tag{6}$$

converges in distribution to a standard normal random variable when n is large. Since π_1, π_2 and π_{12} are unknown, we can use $\hat{\pi}_1$, $\hat{\pi}_2$ and $\hat{\pi}_{12}$ to substitute π_1, π_2 and π_{12} in (5). Thus, for testing (2), H_0 is rejected if the absolute value of (6) is greater than $z_{\alpha/2}$, where $z_{\alpha/2}$ is the upper $\alpha/2$ cutoff point of the standard normal distribution.

2.2. *Generalized score test*

In Section 2.1, π_1, π_2 and π_{12} in $\mathrm{Var}(\hat{\pi}_1 - \hat{\pi}_2)$ are replaced by $\hat{\pi}_1$, $\hat{\pi}_2$ and $\hat{\pi}_{12}$ in the test statistic. In this section, we consider the variance under the null hypothesis in (2), that is, $\pi_1 = \pi_2$. Thus, we have

$$\mathrm{Var}_{\pi_1=\pi_2}\left(\hat{\pi}_1 - \hat{\pi}_2\right) = \begin{cases} 2\pi_1/n & \text{if } s = 1, \\ 2(\pi_1 - \pi_{12})/n & \text{otherwise.} \end{cases} \tag{7}$$

By the central limit theorem, under H_0, the statistic

$$(\hat{\pi}_1 - \hat{\pi}_2)/\sqrt{\mathrm{Var}_{\pi_1=\pi_2}\left(\hat{\pi}_1 - \hat{\pi}_2\right)}$$

coverges to a standard normal distribution when n is large. We can use $(\hat{\pi}_1 + \hat{\pi}_2)/2$ and $\hat{\pi}_{12}$ as substitutes for π_1 and π_{12} in the variance. Hence, for testing (2), when $1 < s \leq k$, the null

hypothesis is rejected if

$$\frac{\sqrt{n} * |\hat{\pi}_1 - \hat{\pi}_2|}{\sqrt{(\hat{\pi}_1 + \hat{\pi}_2 - 2\hat{\pi}_{12})}} > z_{\alpha/2}.$$

When $s = 1$, the null hypothesis is rejected if

$$\frac{\sqrt{n} * |\hat{\pi}_1 - \hat{\pi}_2|}{\sqrt{\hat{\pi}_1 + \hat{\pi}_2}} > z_{\alpha/2}.$$

This approach is similar to the score test of testing a marginal probability equal to a specified value. Hence we call this approach a generalized score test.

2.3. *Likelihood ratio test*

The third approach is the likelihood ratio test (LRT). For testing $H_0 : \pi_1 = \pi_2$, let

$$\Lambda_{12} = \frac{L(\hat{\hat{p}}_{i_1,\ldots,i_k})}{L(\hat{p}_{i_1,\ldots,i_k})}, \tag{8}$$

where L is the likelihood function, and $\hat{\hat{p}}_{i_1\cdots i_k}$ and $\hat{p}_{i_1,\ldots,i_k}$ denote the MLE of $p_{i_1,\ldots,i_k}$ under the restricted parameter space $\pi_1 = \pi_2$ and the whole parameter space, respectively. Thus, we have

$$\hat{p}_{i_1,\ldots,i_k} = n_{i_1,\ldots,i_k}/n.$$

When $s = 1$,

$$\hat{\hat{p}}_{i_1,\ldots,i_k} = \begin{cases} (n_{100\cdots0} + n_{010\cdots0})/(2n) & \text{if } i_1 = 1, \\ (n_{100\cdots0} + n_{010\cdots0})/(2n) & \text{if } i_2 = 1, \\ n_{i_1,\ldots,i_k}/n & \text{otherwise,} \end{cases} \tag{9}$$

which is easily to be interpreted because under $\pi_1 = \pi_2$, $\hat{\hat{p}}_{10\cdots0}$ and $\hat{\hat{p}}_{010\cdots0}$ should be equal to $(\hat{p}_{10\cdots0} + \hat{p}_{010\cdots0})/2$.

When $1 < s \leq k$, by solving the equations of derivatives of the likelihood ratio functions with respect to $p_{i_1\cdots i_k}$ being zero, we have

$$\hat{\hat{p}}_{i_1,\ldots,i_k} = \begin{cases} S \cdot n_{i_1,\ldots,i_k}/(2n(\sum_{i_1=1,i_2=0} n_{i_1,\ldots,i_k})) & \text{if } i_1 = 1, i_2 = 0, \\ S \cdot n_{i_1,\ldots,i_k}/(2n(\sum_{i_1=0,i_2=1} n_{i_1,\ldots,i_k})) & \text{if } i_1 = 0, i_2 = 1, \\ n_{i_1,\ldots,i_k}/n & \text{otherwise,} \end{cases} \tag{10}$$

where

$$S = \sum_{i_1=1,i_2=0} n_{i_1,\ldots,i_k} + \sum_{i_1=0,i_2=1} n_{i_1,\ldots,i_k}.$$

Equation (10) is not as easy as (9) to be interpreted. However, using the fact that the equation $\sum_{i_1=1} n_{i_1,\ldots,i_k} = \pi_1 = \pi_2 = \sum_{i_2=1} n_{i_1\cdots i_k}$ is equal to

$$\sum_{i_1=1,i_2=0} n_{i_1,\ldots,i_k} = \sum_{i_1=0,i_2=1} n_{i_1,\ldots,i_k}, \tag{11}$$

the maximum likelihood estimators of $p_{i_1,\ldots,i_k}$ which are not in the set $A = \{p_{i_1,\ldots,i_k} : i_1 = 1, i_2 = 0\}$ and $B = \{p_{i_1,\ldots,i_k} : i_1 = 0, i_2 = 1\}$ should be the same as $\hat{p}_{i_1,\ldots,i_k}$ because they are not affected by the restriction (11).

For the other $p_{i_1,\ldots,i_k}$, in the case of $i_1 = 1$ and $i_2 = 0$ or the case of $i_1 = 0$ and $i_2 = 1$, under (11), $\hat{\hat{p}}_{i_1,\ldots,i_k}$ can be interpreted as $n_{i_1,\ldots,i_k} / \sum_{i_1=1,i_2=0} n_{i_1,\ldots,i_k}$ the proportion of $S/(2n)$.

According to the asymptotic theory of the likelihood ratio test, $-2\log\Lambda_{12}$ has a limiting distribution with one degree of freedom. For testing (2), H_0 is rejected if

$$-2\log\Lambda_{12} > \chi^2_{1,\alpha},$$

where $\chi^2_{1,\alpha}$ is an upper α cutoff point of the chi-square distribution with one degree of freedom.

3. Simulation result

In this section, we will use Question 1 as an example to compare the three methods proposed in Section 2. Assume that we interview n respondents. Each respondent is allowed to choose at least one answer and at most five answers in Question 1.

Example 3.1 Assume that the true probabilities $p_{i_1,\ldots,i_k}$ as given in Table 1, which leads to $\pi_1 = \pi_2 = 0.6$ and $\pi_3 = \pi_4 = \pi_5 = 0.5143$. For testing (2), Table 2 lists the rejection rates of the three methods when the level of the tests is 0.05. Here the replication is 10,000.

Example 3.2 Assume that the true probabilities $p_{i_1,\ldots,i_k}$ are as given Table 3, which leads to $\pi_1 = \pi_2 = 0.48$, $\pi_3 = \pi_4 = \pi_5 = 0.5086$. For testing (2), Table 4 lists the rejection rates of the three methods when the level of the tests is 0.05. Here the replication is 10,000.

Example 3.3 In this example, we list two cases of $\pi_1 \neq \pi_2$ and compare the powers of the three tests when the level of the tests is 0.05. Here we only show the probability of π_i instead of $p_{i_1,\ldots,i_2}$. Table 5 lists the powers of three tests corresponding to a set of probabilities satisfying $\pi_1 = 0.658$, $\pi_2 = 0.578$, $\pi_3 = 0.481$, $\pi_4 = 0.505$, and $\pi_5 = 0.479$. Table 6 lists the powers of

Table 1. The probability of $p_{i_1 i_2 i_3 i_4 i_5}$.

p_{00000}	p_{10000}	p_{01000}	p_{00100}	p_{00010}	p_{00001}	p_{11000}	p_{10100}
0	0.025	0.025	0.0286	0.0286	0.0286	0.05	0.025
p_{10010}	p_{10001}	p_{01100}	p_{01010}	p_{01001}	p_{00110}	p_{00101}	p_{00011}
0.025	0.025	0.025	0.025	0.025	0.0286	0.0286	0.0286
p_{11100}	p_{11010}	p_{11001}	p_{10110}	p_{10101}	p_{10011}	p_{01110}	p_{01101}
0.05	0.05	0.05	0.025	0.025	0.025	0.025	0.025
p_{01011}	p_{00111}	p_{11110}	p_{11101}	p_{11011}	p_{10111}	p_{01111}	p_{11111}
0.025	0.0286	0.05	0.05	0.05	0.025	0.025	0.05

Table 2. Assume the true probabilities is Table 1 which satisfy $\pi_1 = \pi_2$. For testing $H_0 : \pi_1 = \pi_2$, the rejection rates of three methods for n respondents are listed.

	n			
	100	300	500	1000
Wald	0.0418	0.0383	0.0423	0.0379
score	0.0458	0.0457	0.0523	0.0478
LRT	0.0485	0.0471	0.0524	0.0478

Table 3. The probability of $p_{i_1 i_2 i_3 i_4 i_5}$.

p_{00000}	p_{10000}	p_{01000}	p_{00100}	p_{00010}	p_{00001}	p_{11000}	p_{10100}
0	0.05	0.05	0.01714	0.01714	0.01714	0.01	0.05
p_{10010}	p_{10001}	p_{01100}	p_{01010}	p_{01001}	p_{00110}	p_{00101}	p_{00011}
0.05	0.05	0.05	0.05	0.05	0.01714	0.01714	0.01714
p_{11100}	p_{11010}	p_{11001}	p_{10110}	p_{10101}	p_{10011}	p_{01110}	p_{01101}
0.01	0.01	0.01	0.05	0.05	0.05	0.05	0.05
p_{01011}	p_{00111}	p_{11110}	p_{11101}	p_{11011}	p_{10111}	p_{01111}	p_{11111}
0.05	0.01714	0.01	0.01	0.01	0.05	0.05	0.01

Table 4. Assume the true probabilities as given in Table 1 which satisfy $\pi_1 = \pi_2$. For testing $H_0 : \pi_1 = \pi_2$, the rejection rates of three methods for n respondents are listed.

	n			
	100	300	500	1000
Wald	0.0318	0.0327	0.032	0.0339
score	0.0465	0.0494	0.0513	0.0513
LRT	0.0465	0.0494	0.0514	0.0513

Table 5. For testing $H_0 : \pi_1 = \pi_2$, powers of the three tests corresponding to data with the true probabilities $\pi_1 = 0.658, \pi_2 = 0.578, \pi_3 = 0.481, \pi_4 = 0.505$, and $\pi_5 = 0.479$ are listed.

	n			
	100	300	500	1000
Wald	0.153	0.356	0.548	0.843
score	0.172	0.406	0.614	0.873
LRT	0.176	0.407	0.614	0.873

Table 6. For testing $H_0 : \pi_1 = \pi_2$, powers of the three tests corresponding data with the true probabilities satisfying $\pi_1 = 0.73$, $\pi_2 = 0.69$, $\pi_3 = 0.564$, and $\pi_4 = \pi_5 = 0.546$ are listed.

	n			
	100	300	500	1000
Wald	0.1769	0.4317	0.6653	0.9278
score	0.2072	0.4962	0.7184	0.9475
LRT	0.2072	0.4994	0.72	0.9475

three tests corresponding to a set of probabilities satisfying $\pi_1 = 0.73, \pi_2 = 0.69, \pi_3 = 0.564$, and $\pi_4 = \pi_5 = 0.564$.

From Examples 3.1–3.3, the performances of the score and likelihood ratio tests are more similar than the Wald test. The rejection rates of the score and likelihood ratio tests are closer to the test level 0.05. However, the rejection rate of the Wald test is less than 0.05 for all cases in

Examples 3.1 and 3.2. In Example 3.3, the powers of the score and likelihood ratio tests are higher than the Wald test. The powers are increasing in sample size. When the sample size is greater than 1000, the powers of the three tests are greater than 0.84.

In these examples, we choose the cases for π_i's such that the performances of the three tests are more distinguishable. If π_1 is chosen to be significantly greater than π_2, then the powers of the three tests would be 1. In this case, the performances of the three tests are very good, and it is hard to compare them. Thus, in Example 3.3, π_1 and π_2 are chosen to be close to each other such that we can compare the three tests. It indicates that the performances of the proposed methods in many cases can be better than those in the cases we presented here.

4. Rank

The previous sections mainly discuss ranking two responses: high quality and sales price. We are also interested in investigating which one of the five responses is the most important, and interested in ranking the influence of the five responses.

Assume that we have k responses. For seeking the most influential response, it is necessary to compute each m_j, $j = 1, \ldots, k$. Let $m_{(j)}$ be the order statistics of m_j, that is, $m_{(1)} \leq \cdots \leq m_{(k)}$. Let $v_{(j)}$ be the response corresponding to $m_{(j)}$. It is natural to rank the influence of responses in order of $m_{(j)}$. That is, the most influential response is $v_{(k)}$, and the second influential response is $v_{(k-1)}$. However, basing this only on the order of m_j is risky. The proposed tests in Section 2 can be used to rank the responses. If the hypothesis $\pi_{(k)} = \pi_{(k-1)}$ is rejected, where $\pi_{(r)}$ denotes the marginal probability corresponding to $v_{(r)}$, then we may claim that $v_{(k)}$ is the most influential response. If it is not rejected, then we compare $v_{(k)}$ with $v_{(j)}$, $j \leq k-2$ sequentially. For example, if $H_0{:}\pi_{(k)} = \pi_{(a+1)}$ is not rejected, but $H_0{:}\pi_{(k)} = \pi_{(l)}$, $l \leq a$ is rejected, then the responses, $v_{(k)}$ is ranked first, and $v_{(a)}$ is ranked second. Response $v_{(j)}$, $a+1 \leq j \leq k-1$ is also ranked first if $H_0{:}\pi_{(j)} = \pi_{(a)}$ is rejected, and is ranked between first and second if $H_0{:}\pi_{(j)} = \pi_{(a)}$ is not rejected. By a similar argument, all the responses can be ranked. According to the rule of ranking responses, a reasonable test should have the following property: if $\pi_{(j)} = \pi_{(i)}$, $i \leq j$ is rejected by test, then $\pi_{(j)} = \pi_{(g)}$, $g < i$ should also be rejected by the test with the same level because $|m_{(j)} - m_{(i)}| < |m_{(j)} - m_{(g)}|$. We call this property ranking consistency. If a test has ranking consistency property, we call it a ranking consistency test.

In this section, it will be shown that the three tests proposed in Section 2 are ranking consistent when $s = 1$. However, the ranking consistency property is not valid when $1 < s \leq k$. An example will be given to show that the tests are rank inconsistent when $s = k = 5$. It is also possible to find ranking inconsistent examples for the other cases when $s > 1$.

THEOREM 4.1 *The Wald test is ranking consistent when* $s = 1$.

Proof Let $j > i > d$. For testing

$$H_0 : \pi_{(j)} = \pi_{(i)} \quad vs \quad H_1 : \pi_{(j)} \neq \pi_{(i)}, \tag{12}$$

and

$$H_0^* : \pi_{(j)} = \pi_{(d)} \quad vs \quad H_1^* : \pi_{(j)} \neq \pi_{(d)}, \tag{13}$$

the Wald test statistics for testing H_0 and H_0^* are

$$\frac{|\hat{\pi}_{(j)} - \hat{\pi}_{(i)}|}{\sqrt{\dfrac{\hat{\pi}_{(j)}(1-\hat{\pi}_{(j)}) + \hat{\pi}_{(i)}(1-\hat{\pi}_{(i)}) + 2\hat{\pi}_{(j)}\hat{\pi}_{(i)}}{n}}} \tag{14}$$

and

$$\frac{|\hat{\pi}_{(j)} - \hat{\pi}_{(d)}|}{\sqrt{\frac{\hat{\pi}_{(j)}(1-\hat{\pi}_{(j)}) + \hat{\pi}_{(d)}(1-\hat{\pi}_{(d)}) + 2\hat{\pi}_{(j)}\hat{\pi}_{(d)}}{n}}}, \tag{15}$$

respectively, when $s = 1$. The numerator $|\hat{\pi}_{(j)} - \hat{\pi}_{(i)}|$ of (14) is smaller than $|\hat{\pi}_{(j)} - \hat{\pi}_{(d)}|$. For a fixed j, let

$$W(x) = \hat{\pi}_{(j)}(1 - \hat{\pi}_{(j)}) + x(1-x) + 2\hat{\pi}_{(j)}x. \tag{16}$$

Taking the derivative of (16) with respect to x, we have

$$\frac{\partial W(x)}{\partial x} = 1 - 2 * x + 2\hat{\pi}_{(j)}, \tag{17}$$

which is equal to zero when $x = (1 + 2\hat{\pi}_{(j)})/2$. Hence, (16) is increasing in x for all $x \leq (1 + 2\hat{\pi}_{(j)})/2$. By definition, we have $\hat{\pi}_{(i)} \leq (1 + 2\hat{\pi}_{(j)})/2$ and $\hat{\pi}_{(d)} \leq (1 + 2\hat{\pi}_{(j)})/2$. The above arguments imply $W(\hat{\pi}_{(i)}) \geq W(\hat{\pi}_{(d)})$, which lead to

$$\frac{|\hat{\pi}_{(j)} - \hat{\pi}_{(i)}|}{\sqrt{\frac{\hat{\pi}_{(j)}(1-\hat{\pi}_{(j)}) + \hat{\pi}_{(i)}(1-\hat{\pi}_{(i)}) + 2\hat{\pi}_{(j)}\hat{\pi}_{(i)}}{n}}} \leq \frac{|\hat{\pi}_{(j)} - \hat{\pi}_{(d)}|}{\sqrt{\frac{\hat{\pi}_{(j)}(1-\hat{\pi}_{(j)}) + \hat{\pi}_{(d)}(1-\hat{\pi}_{(d)}) + 2\hat{\pi}_{(j)}\hat{\pi}_{(d)}}{n}}}.$$

Thus, if H_0 is rejected, then H_0^* is also rejected, which implies the Wald test is ranking consistent when $s = 1$. ■

By a similar argument, when $s = 1$, the ranking consistency property of generalized score test is given in Theorem 4.2.

THEOREM 4.2 *The generalized score test is ranking consistent when $s = 1$.*

It is the same as the Wald test and the generalized score test that the likelihood ratio test is ranking consistent in the single-choice question case.

THEOREM 4.3 *The likelihood ratio test is ranking consistent when $s = 1$.*

Proof To prove the ranking consistency of LRT, we need to show, for a fixed i, if the null hypothesis in (12) is rejected by LRT with level α, then the null hypothesis in (13) would also be rejected by LRT, with the same level. That is, we need to show

$$-2\log\Lambda_{(d)(j)} > -2\log\Lambda_{(i)(j)}, \tag{18}$$

where

$$\Lambda_{(i)(j)} = \frac{((m_{(i)} + m_{(j)})/(2n))^{m_{(i)}+m_{(j)}}}{(m_{(i)}/n)^{m_{(i)}}(m_{(j)}/n)^{m_{(j)}}}.$$

Taking the derivative of $\Lambda_{(i)(j)}$ with respect to $m_{(i)}$, we have

$$\frac{\partial \log\Lambda_{(i)(j)}}{\partial m_{(i)}} = \log(m_{(i)} + m_{(j)}) - \log(2m_{(i)}),$$

which is greater than zero because $m_{(j)} \geq m_{(i)}$. Thus, $\Lambda_{(i)(j)}$ is an increasing function in $m_{(i)}$, which implies (18) because $m_{(i)} > m_{(d)}$. Thus, the proof is completed. ■

In the multiple-choice question case, this ranking consistency property is not valid for all data. Example 4.1 gives a set of data such that all of the three tests can be shown to be ranking inconsistent when $s = k = 5$.

Example 4.4 Assume that a multiple-choice question has five answers ($k = 5$). We have survey data: $n_{10000} = 12$, $n_{01000} = 1$, $n_{00100} = 25$, $n_{00010} = 9$, $n_{00001} = 12$, $n_{11000} = 29$, $n_{10100} = 0$, $n_{10010} = 1$, $n_{10001} = 1$, $n_{01100} = 1$, $n_{01010} = 0$, $n_{01001} = 0$, $n_{00110} = 1$, $n_{00101} = 1$, $n_{00011} = 1$, $n_{11100} = 0$, $n_{11010} = 1$, $n_{11001} = 1$, $n_{01110} = 1$, $n_{01101} = 0$, $n_{01011} = 0$, $n_{00111} = 1$, $n_{10110} = 0$, $n_{10101} = 0$, $n_{10011} = 1$, $n_{11110} = 0$, $n_{11101} = 0$, $n_{11011} = 1$, $n_{10111} = 0$, $n_{01111} = 0$ and $n_{11111} = 0$. Then $m_{(5)} = 47$, $m_{(4)} = 35$, $m_{(3)} = 30$, $m_{(2)} = 17$, $m_{(1)} = 19$. For testing

$$H_0 : \pi_{(5)} = \pi_{(4)} \text{ vs } H_1 : \pi_{(5)} \neq \pi_{(4)}, \tag{19}$$

the values of the three statistics with respect to the Wald test, generalized score test and the likelihood ratio test are 2.91, 2.83 and 8.73. The upper 0.05 cutoff point of standard normal distribution and χ^2_1 distribution are 1.96 and 3.84, respectively. Hence (19) is rejected by all the three tests with level 0.05. Then, we expect that the hypothesis

$$H_0 : \pi_{(5)} = \pi_{(3)} \text{ vs } H_1 : \pi_{(5)} \neq \pi_{(3)}, \tag{20}$$

should also be rejected by the tests due to $|\pi_{(5)} - \pi_{(3)}| > |\pi_{(5)} - \pi_{(4)}|$. However, for testing (20), the values of statistics corresponding to the Wald test, generalized score test and the likelihood ratio test are 1.81, 1.94 and 3.78, which does not lead one to reject (20) in any one of the three tests.

Although all the three tests are ranking inconsistent when s is greater than 1, the ratio of the number of the data that ranking inconsistency phenomenon occurs to the total number of data is low according to some simulation results. Therefore, in a real application, the three tests still can be utilized when the data does not lead to ranking inconsistency phenomenon.

5. Conclusion

The aim of this paper is to rank the marginal probabilities of the responses in a multiple-choice question. Three methods are proposed for solving this problem. For ranking any two specified responses, the simulation results in Section 3 show that the proposed methods can achieve good performance. The rejection rates and powers of LRT and score tests are similar and their rejection rates are closer to the level of the test than the Wald tests. In Section 4, the three tests are shown to be ranking consistent when they are utilized in ranking the responses of a single-choice question. However, they are ranking inconsistent for the multiple-choice question case. Although, these tests are ranking inconsistent in the multiple-choice question case, they still can be used to rank the responses under the circumstances when the data does not lead the ranking inconsistency phenomenon. The situation like Example 4.1 does not always happen. Thus, the three tests can still be useful tools for solving the problem. Since most researchers in designing a multiple-choice question are interested in this problem and there is seldom literature discussing it, the methodologies proposed in this paper can provide a good way for analyzing this problem.

Acknowledgements

The author would like to thank the referee for helpful comments.

References

[1] A. Agresti and I.-M. Liu, *Modeling a categorical variable allowing arbitrarily many category choices*, Biometrics 55 (1999) pp. 936–943.

[2] C.R. Bilder, T.M. Loughin, and D. Nettleton, *Multiple marginal independence testing for pick any/c variables*, Comm. Statist. Simulation Comput. 29 (2000) pp. 1285–1316.

[3] Y.J. Decady and D.H. Thomas, *A simple test of association for contingency tables with multiple column responses*, Biometrics 56 (2000) pp. 893–896.

[4] T.M. Loughin and P.N. Scherer, *Testing for association in contingency tables with multiple column responses*, Biometrics 54 (1998) pp. 630–637.

[5] U.N. Umesh, *Predicting nominal variable relationships with multiple responses*, J. Forecasting 14 (1995) pp. 585–596.