Journal of Data Science **1**(2003), 167-183

Assessing the Effect of an Open-ended Category on the Trend in $2 \times K$ Ordered Tables

Shiva Gautam[1] and Takamaru Ashikaga[2]
[1] *Vanderbilt University School of Medicine and*
[2] *University of Vermont*

Abstract: Trend in proportion in $2 \times K$ ordered tables is evaluated by assigning scores to ordered categories. Investigators often encounter $2 \times K$ordered tables with an open-ended category. An open-ended category arises when category scores for the first $K-1$ categories are known or given a priori but the score for the last category is unknown. In such situations, an arbitrary score is often assigned to the open-ended (or the last) category before evaluating the trend. Thus two investigator analyzing the same data set may assign different scores and may arrive at different conclusions. In the spirit of preliminary data analysis it is shown through examples that there are situations where the conclusion is not affected by the choice of scores assigned to the open-ended category. The paper also explores situations where the conclusion may depend on the choice of a score for the open-ended category. In the former case, the usual trend analysis may be performed after assigning a score to the open-ended category. In the latter case, the trend may be evaluated after adjusting for the open-ended category as demonstrated in this paper. Alternately, the trend may be evaluated by Gautam's method which does not depend on a particular choice of a score.

Key words: Ordered categories, score, trend.

1. Introduction

1.1 The trend in proportion

The Cochran-Armitage-Mantel (Armitage, 1955; Cochran, 1954; Mantel, 1963) trend test is widely used for analyzing data in $2 \times K$ (or $K \times 2$) ordered tables. In a typical trend analysis or testing for trend in proportion in a $2 \times K$ table, a test is carried out to find out if the proportion in the first row (or the second row) are increasing or decreasing. For example, consider the data presented in Table 1 by Hiller et al. (1995). In this 5×2 table, the five row categories are classified according increasing level of serum zinc, and the two column categories are classified by 'Desirable' and 'Undesirable' triglyceride level. The last column expresses the percentage (proportion) of patients with undesirable level of triglyceride in each category of serum zinc level. In the context of CAM trend analysis, an investigator is generally interested in finding out if these proportions (percentages) increase (or decrease) with serum zinc level.

Table 1: Level of Triglyceride by Quintile Group of Serum Zinc

Quintile Group of Serum Zinc	Level of Triglyceride			
	Undesirable	Desirable	Total	% Undesirable
≤ 76	7	143	150	4.7
77-82	5	141	146	3.4
83-88	8	167	175	4.6
89-94	5	141	146	3.4
≥ 95	17	144	161	10.6
Total	42	736	778	5.4

Test for trend $p = 0.032$

The ordered categories are sometimes expressed in qualitative form also. For example, the five zinc level categories in Table 1 can be expressed as 'Very Low', 'Low', 'Medium', 'High and 'Very High' which are qualitative but ordered categories. Thus in the context of trend in proportion analysis in $2 \times K$ (or $K \times 2$) ordered tables, the K categories are either defined

numerically or qualitatively. In any event, K numerical scores are assigned to these ordered categories before performing CAM trend analysis.

Considered Table 2 that classifies 66 mothers who suffered the death of a newborn baby by the support level they received and the level of their grief from the loss. Although the original data set credited to Tudehope *et al.* (1986) and presented by Armitage (1994) has three support categories, 'Good', 'Adequate', and 'Poor', we combined the last two categories for demonstration purpose.

Table 2: Level of Triglyceride by Quintile Group of Serum Zinc

(a)	(b)	(c)	(d)	(e)	(f)
I	1	17	17	34	0.50
II	2	6	6	12	0.50
III	3	3	9	12	0.33
IV	4	1	7	8	0.14
Total	–	27	39	66	–

(a)= Grief state, (b)= Row score x_i, (c)= Support good, (d)= Supposrt adquate or poor, (e)=(c)+(d), (f)=Proportion of good support.

The CAM trend analysis of the data in Table 2 shows a significant ($p = 0.03$) negative relationship between grief and level of support. In other words, those with good support generally experienced less grief. This is also reflected in raw data as a larger proportion of women with good support are found in lower grief state categories.

One way to carry out the computations needed for CAM trend analysis is to regress the variable Y on X, where $Y = 1$ if an observation is from 'Good support' and $Y = 0$ if an observation is from 'Adequate or Poor support' category, and $X = j, (j = 1, 2, 3, 4)$ if an observation is from the i-th 'Grief state' category. In this formulation scores 1, 2, 3, and 4 are assigned to the 'Grief state' categories I, II, III, and IV, respectively. This regres-

sion analysis yields a negative slope indicating negative trend in proportion. Thus the proportion of people receiving good support gradually decreases with increasing grief level. An alternative analysis that compares the mean grief levels between two groups of women achieves the same statistical significance. The mean grief level in the 'Good support' group is 1.556 while in the 'Adequate or poor Support group it is 2.154. The Student's t-test shows that the means grief level is significantly lower in the group who receive good support, and the test achieves the same p-value ($p = 0.03$) as in the case of regression analysis. Furthermore, the correlation coefficient between X and Y (or Grief and Support) will also yield the same p-value if the null hypothesis of zero correlation is tested. Thus, no matter how the $2 \times K$ table is obtained (e.g., two multinomials, K binomials, one bivariate), the regression analysis as described above can be used for computational purpose to obtain the p-value. If mid-rank scores are assigned to the ordered categories in stead of $1, 2, \cdots, K$, then the p-value from the regression analysis (or trend in proportion) will be equivalent to the p-value obtained from the Mann-Whitney (or Wilcoxon-Rank Sum) test. Graubard and Korn (1987) have outlined relationships among different statistical tests in the context of a $2 \times K$ ordered table.

As mentioned above, order preserving scores are assigned to the K categories of a $2 \times K$ ordered table before evaluating the CAM trend. The scores assigned to the qualitatively classified ordered categories are often chosen arbitrarily. The scores do not necessarily reflect the relative distance between the ordered categories. Gautam (1991), and Kimeldorf, Sampson and Whitaker (1992) suggested to maximize and minimize the test statistics (e.g the t-statistic) involved over all possible order-preserving scores to deal with the issue of arbitrariness of score. But if the categories are defined numerically or quantitatively then it may be prudent to utilize those numbers as scores. For example, consider data in Table 1. The ordered categories are defined by numerical intervals. In such cases, mid-values of the intervals may be assigned as category scores for analyzing the data, and such scores may have some advantages over equally spaced scores $1, 2, \cdots, K$ or Mann-Whitney's non-parametric test (Graubard and Korn, 1987).

There are several methods available for analyzing data in $2 \times K$ ordered tables. These methods can be broadly classified into two groups (a) those

which do not utilize order-preserving scores, and (b) those which utilize order-preserving scores. The general belief that method that do not utilize scores are superior could be misleading (Graubard and Korn, 1987). The second group can further be divided into three subgroups (i) those with given or known scores (ii) those with unknown scores, and (iii) those with known and unknown scores . The CAM trend analysis belongs to the second group. This article focus on a special case of the third subgroup in the context of CAM trend analysis.

1.2 Open-ended category and trend in proportion

It is mentioned in the previous section that the CAM procedure requires numerical scores to be assigned to the columns (or the ordered categories). Investigators often encounter $2 \times K$ ordered tables where scores for all but the last (or the K-th) category is known or given a priori. Such a category for which the score is not known a priori is referred to as an open-ended category in Gautam (1997). Open-ended categories usually take the form of "greater than" or "less than." For example, suppose that the number of cigarettes smoked per day is categorized as 0-4, 5-9, 10-14, 15-19, and 20. Then midpoints 2, 7, 12, and 17 (or equivalently, scores 1, 2, 3, and 4) of these intervals can be assigned to the first, second, third, and fourth category, respectively, to evaluate the CAM trend. However, only the lower limit of the fifth category is known. Therefore, an arbitrarily chosen score of 20 or more is assigned to the fifth category. Consequently, two investigators may draw different statistical conclusions while analyzing the same data set by assigning different but plausible scores to the open-ended category. Gautam (1997) derived a test statistic to evaluate the trend in the presence of an open-ended category. The test addresses the issue of the arbitrariness by maximizing the CAM trend statistic over all possible scores for the open-ended category. But in many data sets that we have encountered, the CAM trend statistic did not change noticeably over possible scores for the open-ended category. In such situations, the open-ended category may not have a significant effect on the trend analysis. After all, such an open-ended category may have been created to lump a few extreme observations.

In this paper, we examine situations where an arbitrary assignment of a

score to the open-ended category may affect the trend analysis. Investigators generally assign a score to the open-ended category which is consistent with the scores for the rest of the categories. For the above example of number of cigarettes smoked per day (e.g., 0-4, 5-9, 10-14, 15-19, and $\geq$ 20), investigators often assign 2, 7, 12, 17, and 22 (or equivalently scores 1, 2, 3, 4 and 5) to the first, second, third, fourth, and fifth category, respectively. This paper will also help investigators determine if they would have reached different conclusions had they assigned different scores to the open-ended category.

A motivational example

Consider again data in Table 1 obtained from Table 4 of Hiller *et al.* (1995). Table 1 has two open-ended categories, but we are ignoring the open-endedness of the first category because it is bounded by zero at the lower limit. Also, this example is being used only to demonstrate some implications of the open-ended category.

Hiller and colleagues (1995) have shown that there is a significant CAM trend ($p = 0.03$). A visual inspection of Table 1 suggests that the percentage of undesired level of triglyceride is constant for the first four ordered categories. The percentage increases only for the open-ended category. If scores 73.5, 79.5, 85.5, 91.5, and 97.5 (or equivalently 1, 2, 3, 4, 5) is used then the p-value = 0.03, which is obtained by Hiller *et al.* (1995). However, if the score 95.5 instead of 97.5 is assigned to the open-ended category, then $p = 0.05$. Similarly, a score of 99.5 yields a p-value equal to 0.02. Therefore, the significant trend found by Hiller and colleagues (1995) may be due to the influence of the open-ended category alone. This is further supported by the fact that no significant ($p = 0.73$) trend is observed if the test is carried out without the open-ended category.

2. Formulation of the problem

Consider the data set displayed in Table 3 (a) consisting of 'cases' and 'controls' classified into K ordered categories. Also assume that scores for the first $K - 1$ categories are known, but the score for the K-th category

which is also an open-ended category is not known. Define a binary variable U representing the row variable of the table. Suppose that $U = 1$ if an observation is from 'cases', and $U = 0$ otherwise. Denote the known scores to be assigned to the first $K-1$ categories by $x_1, x_2, \cdots$ and x_{K-1}, respectively. Without loss of generality assume that $x_1 \leq x_{\leq} \cdots \leq X_K$. Assume that a score $\alpha \geq x_K$ is to be assigned to the open-ended categories, where α is not known but its lower limit x_K is known. Let V denote the column variable such that $V = x_j$, if an observation is from the j-th category, $j = 1, 2, \cdots, K$ and $V = \alpha$ if an observation is from the K-th (or the open-ended) category. Let $V_1 = x_j$, if an observation is from the j-th category, $j = 1, 2, \cdots, K-1$, and $V_1 = x_K$ if an observation is from the open-ended category. Let V_2 denote a binary variable such that $V_2 = 1$ if an observation is from the open-ended category, and $V_2 = 0$ otherwise. For convenience, these variables are also displayed in Table 3(a).

Table 3: An Outlay of a Given Set of Data

Category	Cases $U=1$	Controls $U=0$	Category Total	V (Category Score)	V_1	V_2
1	n_{11}	n_{12}	n_{1+}	x_1	x_1	0
2	n_{21}	n_{22}	n_{2+}	x_2	x_2	0
...	...	...	...	...	...	...
$K-1$	$n_{K-1,1}$	$n_{K-1,2}$	$n_{K-1,+}$	x_{K-1}	x_{K-1}	0
K	n_{K1}	n_{K2}	n_{K+}	$\geq x_K$	w	1
Total	n_{+1}	n_{+2}	n_{++}			

The last category K is open ended.

The CAM trend statistic is given by $n_{++}r^2$, where r is the correlation coefficient between U and V for a given value $\alpha \geq x_K$. Since any $\alpha \geq x_K$ would be legitimate candidate for a score for the open-ended category, there are more than one possible scores for the open-ended category. Note that only x_K the minimum possible score for the open-ended category is known. If the value of the CAM trend statistic yields a significant result for all

scores greater than x_K, then the statistical conclusion does not depend on the choice of a score $\alpha \geq x_K$. Below we propose to assign a score to the open ended category that maximize the trend statistic. The rationale for this is based on the suggestions made in the literature on the choice of scores in the context of ordered categories. Agresti (1990, p. 294) suggests to conduct a sensitivity analysis by assigning different possible scores to find out if the result depend on a choice of various scores. In this article we are proposing to assign all possible scores and maximizing the value of the test statistic. If this maximal value does not produce a significant result, then result will not depend on the choice of scores. When the scores are not known for any of the categories in a $2 \times K$ ordered table, then Gautam (1991), and Kimeldorf, Sampson and Whitaker (1992) also suggest this approach. It is also worth noting that Pearson's chi-square statistic obtained from a $2 \times K$ nominal table can be obtained by maximizing correlation between row and column variables over all possible column scores (Haberman, 1981; Gautam and Kimeldorf, 1999). The following theorem shows that the CAM trend is maximized by regressing U on V_1 and V_2 defined above. Note that for a given set of scores for the orderd ctegories of a $2 \times K$ table, CAM trend statistics can be expressed in several forms and the following is one of them based on the entries of Table 3(a).

$$\text{CAM Trend statistic} = \frac{n_{++}(n_{++}\sum n_{i1}x_i - n_{+1}\sum n_{i+}x_i)^2}{n_{+1}(n_{++}-n_{+1})[n_{++}\sum n_{i1}x_i^2 - (\sum n_{i+}x_i)^2]}$$

Theorem 1: Consider Table 3(a) and 3(b). The maximum CAM trend statistic is obtained by regressing U on V_1 and V_2 and is given by $n_{++}R^2$, where R is multiple correlation coefficient.

Proof: Consider Table 3(b), let α be a candidate score for the open-ended category such that $\alpha = x_K + \beta$ ($\beta \geq 0$). Let V denote the variable V when the score $x_K + \beta$ is assigned to the open-ended category. Then for a given score $x_K + \beta$, $V_\beta = V1 + \beta V_2$. Note that the CAM trend statistic is given by $n_{++}r^2$ where r is the correlation between U and V_β. Therefore, maximum value of the CAM trend statistic is obtained by

$$\max_{\beta} \text{corr}(U, V_\beta) = \max_{\beta} \text{corr}(U, V_1 + \beta V_2) = \max_{\alpha_1, \alpha_2} \text{corr}(U, \alpha_1 V_1 + \alpha_2 V_2) \quad (2.1)$$

where $\beta = \alpha_2/\alpha_1$ (say).

Next, consider the regression equation $U = \beta_0 + \beta_1 V_1 + \beta_2 V_2 + \epsilon$. If $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ are the sample regression coefficient then

$$R = \max_{\alpha_1, \alpha_2} \text{corr}(\alpha_1 V_1 + \alpha_2 V_2) = \text{corr}(\hat{\beta}_1 V_1 + \hat{\beta}_2 V_2) \tag{2.2}$$

The theorem follows from (2.1) and (2.2) and by noting that correlation is invariant of origin and scale. The theorem above shows that R^2 and therefore, the CAM trend statistic is maximized when a score of $x_K + \hat{\beta}_2/\hat{\beta}_1$ is assigned to the open-ended category. If $\hat{\beta}_2 = 0$, then the maximum correlation is given simply by the correlation between U and V_1, and this is equivalent to computing the correlation by assigning scores $x_1, x_2, \cdots, x_{K-1}$, and x_K. On the other hand, if $\hat{\beta}_2 \neq 0$, then the correlation is maximized by using the category scores $x_1, x_2, \cdots, x_{K-1}$ and $x_K + \hat{\beta}_2/\hat{\beta}_1$. Therefore, a $\hat{\beta}_2$ close to zero indicates that a score greater than x_K will not produce a significantly different trend than that obtained by assigning x_K, the minimum possible given score for the open-ended category. Hence, the problem reduces to testing the following null hypothesis in the context of the regression model given by equation (2.2).

$$H_0 : \quad \beta_2 = 0 \quad \text{versus} \quad H_1 : \quad \beta_2 \neq 0$$

If the null hypothesis is rejected then one may conclude that there is a significant effect of the open-ended category, and the choice of a score for it may affect the statistical conclusion. In such a situation, the method suggested by Gautam (1997) may be applied to evaluate the trend. If the null hypothesis is not rejected, then one may conclude that the inclusion of the variable V_2 in the model does not increase the value of R^2 significantly. This further implies that increasing the score for the open-ended category does not change the value of the trend statistic. Therefore, in this case one may assign an arbitrary score $\alpha \geq x_K$ to the open-ended category, and use the CAM trend statistic rather than Gautam's (1997) maximized trend statistic. The later is less powerful than the CAM as it's distribution is a mixture of chi-square variates with one and two degrees of freedom. Thus we could have four possibilities

(i) both β_1 and β_2 are significant
(ii) both β_1 and β_2 are non-significant
(iii) β_1 is significant and β_2 is non-significant, and
(iv) β_1 is non-significant and β_2 is significant

The open-ended category shows a significant effect on the trend in both (i) and (iv) cases, but in (iv) it is the only significant effect. Therefore, in this case a significant trend is less meaningful (see section 4 below). In the remaining two cases, the open-ended category does not have any significant effect on the result. But in case (iii) the trend itself is significant, while it is non-significant in (ii). Therefore in (iii) one may simply assign the minimum possible score (which is known) to the open-ended category and then proceed with the analysis.

3. Examples

Two examples, one with and one without a significant effect of the open-ended category are presented below. No significant effect of the open-ended is observed in the first example, while a significant effect is found in the second example. In both examples, the minimum possible score for the open-ended category is used as the value for α, the arbitrarily chosen score. The test can be performed with any value for α.

3.1 Example 1

Data in Table 4 is taken from Breslow and Day (1980). The cases and controls are classified by tobacco consumption per day. The authors used scores 0, 1, 2, and 3 which is equivalent to using the scores 4.5, 14.5, 24.5, and 34.5. Therefore, the authors assumed that the interval width of the last category is the same as other categories. However, only the lower limit of this interval is provided in the data set.

Use notations of the previous section to obtain mid-values of the intervals as $x_1 = 4.5, x_2 = 14.5, x_3 = 24.5$.] Let $w = 30$, the minimum possible score for the open-ended category. These numbers are also the values of V_1. The corresponding values of V_2 are 0, 0, 0, and 1. A regression analysis based on

Table 4: Tobacco Consumption Per Day (g/day)

	0-9	10-19	20-29	30+	Total
Cases	78	58	33	31	200
Controls	447	178	99	51	775
Total	525	236	132	82	975

the equation $U = \beta_0 + \beta_1 V_1 + \beta_2 V_2$ leads to a conclusion that the hypothesis $H_0 : \beta_2 = 0$ should not be rejected (p-value = 0.24). Therefore, there is no effect of the open-ended category. This indicates that the statistical conclusion regarding the trend will be the same for all possible scores for the open-ended category.

Since $\beta_2 = 0$, the regression equation now reduces to $U = \beta_0 + \beta_1 V_1$. In this case, testing for the trend is equivalent to testing the null hypothesis $H_0 : \beta_1 = 0$ versus the alternate hypothesis $H_1 : \beta_1 \neq 0$. The estimate of β_1 (slope) from the data is found to be positive, and the trend statistic $n_{++}r^2 = 26.03$ (p-value < 0.0001). Therefore, there is a significant positive trend which implies that proportion of cases increases with categories.

3.2 Example 2

Consider the data in Table 5 where the cases and control are classified according to the amount of alcohol consumption. This data set is also taken from Breslow and Day (1980).

In this example, if midpoints are used as given scores, then $x_1 = 19.5, x_2 = 59.5, x_3 = 99.5$, and $w = 120$, the minimum possible score for the open-ended category. Results from the regression analysis discussed in this paper show that the null hypothesis $H_0 : \beta_2 = 0$ is rejected as p-value < 0.0001. Therefore, the conclusion may be affected by the choice of a score for the open-ended category. In this case, the method suggested by Gautam (1997) is used to evaluate the trend.

Table 5: Alcohol Consumption Per Day (g/day)

	0-39	40-79	80-119	120+	Total
Cases	29	75	51	45	200
Controls	386	280	87	22	775
Total	415	355	118	67	975

Since a positive trend is expected, consider the following hypotheses:

$$H_0: \quad \beta_1 = \beta_2 = 0, \quad \text{and}$$
$$H_0: \quad \beta_1 \geq 0, \beta_2 \geq 0 \quad \text{with at least one strict inequality}$$

Since $\hat{\beta}_1 = 0.003697$ and $\hat{\beta}_2 = 0.23214$, the maximum correlation between the row and the column variable is given by the coefficient of determination from the model $U = \beta_0 + \beta_1 V_1 + \beta_2 V_2$. Equivalently, the correlation or the trend statistic is maximized when a score of $w + \hat{\beta}_2/\hat{\beta}_1 = 120 + \hat{\beta}_2/\hat{\beta}_1 = 182.79$ is assigned to the open-ended category. Since the coefficient of determination $R^2 = 0.1629$, the value of the test statistic is $n_{++}R^2 = 975(0.1629) = 158.83$. Note that, this value is also obtained by assigning scores 0, 1, 2, and 4.08 to ordered categories. Using the result from Gautam (1997), it is found that $\rho_{12} = -\text{corr}(V_1, V_2) = \text{corr}(\hat{\beta}_1, \hat{\beta}_2) = -0.5528$, and $p = Pr[\beta\beta_1 > 0, \hat{\beta}_2 > 0] = 0.15691$, and the null hypothesis of no trend is rejected in favor of the positive trend (p-value < 0.0001). Since the data set has large number of observation, one can instead of using Gautam's (1997) method may simply look at the value of $n_{++}R^2$ from the regression output and compare it the chi-square variate with 2 degrees of freedom to obtain the p-value.

4. An epidemiological interpretation

Methods discussed in this paper can be used to assess the effect of an open-ended category and to evaluate the trend after adjusting its effect. For

example, in the regression equation $U = \beta_0 + \beta_1 V_1 + \beta_2 V_2 + \epsilon$, the parameter β_1 represents the slope after adjusting for the open-ended category when the score $\alpha \geq x_K$ is assigned to it. In Example 1 above, the slope (or the trend was significant after adjusting for the open-ended category when a score of 30 (minimum possible score) was assigned to it. The effect of the open-ended category was not significant. Therefore, the CAM trend analysis does not depend on the choice of a score for the open-ended category. In Example 2, the effect of the open-ended category was significant, and the trend remained significant even after adjusting for the open-ended category after a score of 120 (the minimum possible score) was assigned to it. Choice of the number 120 as the score for the open-ended category may still be considered somewhat arbitrary. Recall that the trend statistic that included the effect of the open-ended category was evaluated by Gautam's (1997) method. This method maximizes CAM trend statistic over all possible scores for the open-ended category, and therefore, does not depend on a particular choice of score for the open-ended category.

If the score that maximizes CAM trend statistic is very large, then this maximum CAM trend statistic can also be obtained by assigning scores $0, 0, \cdots, 1$. This in turn is equivalent to collapsing first $K-1$ categories into one category and then comparing it with the open-ended category. This is equivalent to the situation of $\beta_2 = 0$ in the regression context described in the previous section.Even if there is a statistical significance, this may not be a desirable interpretation of significant trend as the observed significance may be wholly due to the open-ended category. We demonstrate this using data in Table 1 which was presented at the beginning. Since the data in Table 1 suggest that the percentage of undesired level of triglyceride is constant for the first four groups and increases for the open-ended category, it seems that the difference, if there is any, comes from comparing the open-ended category with remaining categories collapsed together. We have shown earlier that in this example various possible scores for the open-ended category yield different p-values. Note that, all these results can be obtained by using the regression equation $U = \beta_0 + \beta_1 V_1 + \epsilon$ where V_1 is the variable that takes column scores as its values. For example, considered the following hypotheses after assigning a score of $\alpha e = 97.5$ to the open-ended category,

and using the regression model $U = \beta_0 + \beta_1 V + 1 + \epsilon$:

$$H_0 : \quad \beta_1 = 0 \quad \text{vs.} \quad H_1 : \quad \beta_1 \neq 0$$

The test of the null hypothesis yields a significant positive trend ($p = 0.03$), indicating that the proportion of people with undesired level of triglyceride increases with serum zinc level. This is essentially the same result obtained by Hiller and colleagues (1995). But, when the model $U = \beta_0 + \beta_1 V_1 + \beta_2 V_2 + \epsilon$ is used V_2 becomes significant ($p = 0.013$) and V_1 becomes non-significant ($p = 0.76$). In other words, the effect of the open-ended category is significant, and when this effect is adjusted the CAM trend becomes non-significant when a score of 97.5 is assigned to it. Therefore, the significant trend found by Hiller and colleagues (1995) is apparently due to the influence of the open-ended category.

Since the open-ended category has a significant effect on the trend, one may use Gautam's (1997) method to evaluate the trend. If one uses Gautam's (1997) method in this example, then it sheds some light as it yields 0, 0, 0, 0, and 1 as the maximal scores. This method deals with the issue of arbitrariness of the score but in this example it treats the first four categories as one category and then compares it with the open-ended category to find a significant trend. Since the investigators generally want to see a trend among the non open-ended category, the significance obtained may not have meaningful interpretation as a increasing or decreasing trend.

5. Discussion

The linear trend in $2 \times K$ ordered tables is often evaluated by the CAM trend statistic. In the case of an open-ended category, investigators often assign an arbitrary score to the open-ended category. If all possible scores for the open-ended category yield the same conclusion, then the open-ended category will not have any effect on the trend analysis. Any score can be assigned to the open-ended category, but we suggest to use the minimum possible score as demonstrated in this paper. In this paper, linear regression was used to assess the effect of the open-ended category. It was shown through examples the open-ended category may or may not affect the usual CAM trend analysis. Since the open-ended category is not well defined and

is only one of the categories, perhaps it is not intended that this category influence the overall outcome of data analysis. From this view point it becomes important to assess its effect.

If there is no significant effect of the open-ended category, then one may calculate the CAM trend after assigning the minimum possible score to the open-ended category as all possible scores for the open-ended category will lead to the same conclusion as in the case of data in Table 4. If a significant effect is observed, then some or all possible scores may yield significant results. Note that, the minimum possible score assigned to the open-ended category is still somewhat arbitrary as there are other possible scores. In this case, we suggest reporting the result after adjusting for the open ended category by assigning the minimum possible score for the open-ended category.

Acknowledgements

Dr. Gautam's research was supported in part by National Cancer Institute grant R03CA68527, and Dr. Ashikaga's work was supported in part by National Cancer Institute grant P30CA22435 awarded to Vermont Comprehensive Cancer Center. The authors thank Dr. Lowell Gerson for his helpful comments on earlier drafts of the manuscript.

References

Agresti, A. (1990). *Categorical Data Analysis*. Wiley.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, **11**, 375-386.

Armitage P, Berry G. (1994). *Statistical Methods in Medical Research*. Blackwell Scientific Publications.

Breslow, N. Day, N. E. (1980). *Statistical Methods in Cancer Research*, vol. I: *The Analysis of Case-Control Studies*. Lyon: IARC.

Cochran, W. G. (1954). Some methods of strengthening the common χ^2 tests. *Biometrics*, **10**, 417-451.

Gautam, S. (1991). *Application of* t, T^2 *and* F*-test to Ordered Categorical Data*. Ph.D dissertation, Program in Mathematical Sciences, University of Texas at Dallas.

Gautam, S (1997). Test for linear trend in $2 \times K$ ordered tables with open-ended categories. *Biometrics*. **53**, 1163-1169.

Gautam, S and Kimeldorf, G. (1999). Some results on the maximal correlation in $2 \times K$ contingency table. *American Statistician*, **53**, 336-341.

Graubard, B. I. and Korn, E. L. (1987). Choice of column scores for testing independence in ordered $2 \times K$ contingency tables. *Biometrics*, **43**, 471-476.

Haberman, S. J. (1981). Tests for independence in two-way contingency tables based on canonical correlation and on linear by linear interaction. *Annals of Statistics*, **9**, 1178-1186.

Kimeldorf, G, Sampson, A. R. and Whitaker, L. R. (1992). Min and max scoring for two-sampled ordinal data. *Journal of the American Statistical Association*, **87**, 241-147.(***** page number incorrect)

Hiller, R., Seigel, D., Sperduto, R. D., Blair, N., Burton, T. C., Farber MD, (****et al. ? Did you list them all?***) (1995). Serum Zinc and Serum Lipid Profiles in 778 Adults. *Annals of Epidemiology*, **5**(6), 490-496.

Mantel, N. (1963). Chi-square test with one degree of freedom; extension of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, **58**, 690-700.

Tudehope, D. I., Iredell, J., Rodgers, D, amd Gunn, A. (1986). Neonatal death: grieving families. *Medical Journal of Australia* **144**, 290-292.

Received October 15, 2001; accepted June 30, 2002

Shiva Gautam, Ph.D
Division of Biostatistics
Department of Preventive Medicine, and
Vanderbilt-Ingram Cancer Center
Vanderbilt University School of Medicine

Nashville, TN 37232-6848, USA

Takamaru Ashikaga
Department of Medical Biostatistics
Hills Building
University of Vermont
Burlington, VT 05405, USA