Unequal Size, Two-Way Analysis of Variance for Categorical Data

Han-Yong Chung*

1. Introduction

The techniques about the analysis of variance for quantitative variables have been well-developed. But when the variable is categorical, we must switch to a completely different set of varied techniques. R.J. Light and B. H. Margolin[1] presented one kind of techniques for categorical data in their paper, where there are G unordered experimental groups and I unordered response categories.

This note is an extension of one of the technique to a two-way table, where there are I unordered response categories, J unordered experimental levels crossed by another K unordered experimental levels, with unequal size of observations in each of JK cells. For terminology and notation, we follow[1].

For n responses, each in one and only one of I possible categories, the data can be summarized with a vector Φ of category counts $\Phi=(n_1,\cdots, n_I)$, where n_i is the number of responses in the ith category, $i=1,\cdots, I$, so that $\sum_{i=1}^{I} n_i=n$. Then the variation of these responses is:

$$\frac{1}{2n}[\sum_{i\neq j} n_i n_j]=\frac{1}{2n}[n^2-\sum_{i=1}^{I} n_i^2]$$

To further motivate this definition of variation, we need the following known lemmas[1]:

Lemma 1 The variation of n categorical responses is minimized if and only if they all belong to the same category.

*Associate Professor, Department of Computer Science and Statistics, Seoul National University.

Lemma 2 The variation of n responses, where $n=IS+L$, $0\leq L<I$, is maximized for any vector Φ of category counts such that L counts equal $S+1$, and $I-L$ counts equal S.

2. The Model and Variation Components

We construct the two-way table where there are I unordered response categories, J unordered experimental levels crossed by another K unordered experimental levels with an unequal size of observations in each JK cells. Each response is in one and only one of the I categories. Denote the number of responses in category i, jth level (of the second index), kth level (of the third index) by n_{ijk}.

We assume that responses in different cells are stochastically independent, and that each cell's responses $(n_{1jk}, n_{2jk}, \cdots n_{Ijk})$ follow a multinomial law:

$$Pr(n_{1jk},\cdots,n_{Ijk})=\binom{n_{\cdot jk}}{n_{1jk},\cdots,n_{Ijk}}\prod_{i=1}^{I}(p_{ijk})^{n_{ijk}}$$

where $\sum_{i=1}^{I}p_{ijk}=1$, $p_{ijk}>0$, $i=1,\cdots,I$, $j=1,\cdots,J$, and $k=1,\cdots,K$.

If we let

$$V=(n_{111},n_{211},\cdots,n_{I11},n_{121},n_{221},\cdots,n_{I21},\cdots,n_{1J1},n_{2J1},\cdots,n_{IJ1},n_{112},n_{212},\cdots,n_{I12},\cdots,n_{1JK},n_{2JK},\cdots,n_{IJK})',$$

Then,

$$E(V)=Y=(n_{\cdot 11}p_{111},n_{\cdot 11}p_{211},\cdots,n_{\cdot 11}p_{I11},\cdots,n_{\cdot J1}p_{1J1},n_{\cdot J1}p_{2J1},\cdots,n_{\cdot J1}p_{IJ1},n_{\cdot 12}p_{112},\ n_{\cdot 12}p_{212},\cdots,n_{\cdot 12}p_{I12},\cdots,n_{\cdot JK}p_{1JK},n_{\cdot JK}p_{2JK},\cdots,n_{\cdot JK}p_{IJK})',$$

$$\mathrm{Cov}(V)=Z=Z_{11}\oplus Z_{21}\oplus\cdots\oplus Z_{J1}\oplus Z_{12}\oplus Z_{22}\oplus\cdots\oplus Z_{J2}\oplus\cdots\oplus Z_{1K}\oplus\cdots\oplus Z_{JK}$$

where

$$Z_{jk}=n_{\cdot jk}\begin{pmatrix} p_{1jk}(1-p_{1jk}) & -p_{1jk}p_{2jk}\cdots\cdots & -p_{1jk}p_{Ijk} \\ & p_{2jk}(1-p_{2jk})\cdots & -p_{2jk}p_{Ijk} \\ & \vdots & \\ & \cdots\cdots\cdots & p_{Ijk}(1-p_{Ijk}) \end{pmatrix}$$

and $\oplus$ denotes the direct sum operation (see[2]).

With the two-way table introduced as our model we define the following variations:

The total variation in the response variable (TSS) is

$$\mathrm{TSS}=n/2-\sum_{i=1}^{I}n_{i..}^{2}/2n;$$

the within-2nd index level variation (WSS_1) is

$$\mathrm{WSS}_1=\sum_{j=1}^{J}(n_{.j.}/2-\sum_{i=1}^{I}n_{ij.}^{2}/2n_{.j.});$$

the between-2nd index level variation (BSS_1) is

$$\mathrm{BSS}_1=\mathrm{TSS}-\mathrm{WSS}_1;$$

the within-3rd index level variation (WSS_2) is

$$\mathrm{WSS}_2=\sum_{k=1}^{K}(n_{..k}/2-\sum_{i=1}^{I}n_{i.k}^{2}/2n_{..k});$$

the between-3rd index level variation (BSS_2) is

$$\mathrm{BSS}_2=\mathrm{TSS}-\mathrm{WSS}_2;$$

the within-cell variation (WSS_3) is

$$\mathrm{WSS}_3=\sum_{k=1}^{K}\sum_{j=1}^{J}(n_{.jk}/2-\sum_{i=1}^{I}n_{ijk}^{2}/2n_{.jk});$$

the between-cell variation (BSS_3) is

$$\mathrm{BSS}_3=\mathrm{TSS}-\mathrm{WSS}_3;$$

where

$$n_{.jk}=\sum_{i=1}^{I}n_{ijk},\quad n_{i.k}=\sum_{j=1}^{J}n_{ijk},\quad n_{ij.}=\sum_{k=1}^{K}n_{ijk},$$

$$n_{i..}=\sum_{j=1}^{J}\sum_{k=1}^{K}n_{ijk},\quad n_{.j.}=\sum_{i=1}^{I}\sum_{k=1}^{K}n_{ijk},\quad n_{..k}=\sum_{i=1}^{I}\sum_{j=1}^{J}n_{ijk},$$

$$n=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}n_{ijk}$$

3. Definitions

Definition 1 The interaction between the 2nd index level and the 3rd index level is defined as $I=\mathrm{BSS}_3-\mathrm{BSS}_1-\mathrm{BSS}_2$.

Definition 2 Ω is the space where $I=0$.

Definition 3 p_i is the probability of an element belonging to ith category. $p_{ij\cdot}$ is the probability of an element belonging to ith category and ith level, regardless of the 3rd index level. $p_{i\cdot k}$ is the probability of an element belonging to ith category and kth level, regardless of the 2nd index level.

Definition 4 The hypothesis H_1 is $p_{ij\cdot}=p_i$ for all j. The hypothesis H_2 is $p_{i\cdot k}=p_i$ for all k. The hypothesis H_3 is $p_{ijk}=p_i$ for all j and k.

4. Testing of the Hypothesis

Theorem 4-1 (a) Under the hypothesis H_1,

$$(n-1)(I-1)\mathrm{BSS}_1/\mathrm{TSS}$$

is asymptotically approximated as $\chi^2(I-1)(J-1)$.

(b) Under the hypothesis H_2,

$$(n-1)(I-1)\mathrm{BSS}_2/\mathrm{TSS}$$

is asymptotically approximated as $\chi^2(I-1)(K-1)$.

Proof The above facts can be proved as in the case of one-way table (see [1]). To prove (a), since there are I categories and J levels the degree of freedom is $(I-1)(J-1)$. (b) can be proved in the similar way.

Theorem 4-2 With large $n_{\cdot jk}=n_{\cdot j\cdot}n_{\cdot\cdot k}/n$ for all j,k, BSS_1 and BSS_2 are asymptotically independent under the hypothesis H_3.

Proof With large $n_{\cdot jk}$, V is asymptotically multivariate normal, i.e., $V\sim\mathcal{N}(Y,Z)$. Under the hypothesis H_3, Z can be reduced as

$$Z=Z_{11}\oplus Z_{21}\oplus\cdots\oplus Z_{jk}\oplus\cdots\oplus Z_{JK},$$

where

$$Z_{jk}=n_{\cdot jk}\begin{pmatrix} p_1(1-p_1) & -p_1p_2\cdots\cdots-p_1p_I \\ & p_2(1-p_2)\cdots-p_2p_I \\ & \vdots \quad \vdots \\ & \cdots\cdots\cdots\cdots p_I(1-p_I) \end{pmatrix}$$

Let

$$T=-(U_{JK}\otimes I_I)/2n,\ A=Y_K\otimes I_{IJ},\ A'=X_K\otimes I_{IJ},$$

$$W_1=-\frac{1}{2}\left(\frac{1}{n_{\cdot1\cdot}}I_I\oplus\frac{1}{n_{\cdot2\cdot}}I_I\oplus\cdots\oplus\frac{1}{n_{\cdot j\cdot}}I_I\right)$$

$$B=I_K\otimes(Y_J\otimes I_I),\ B'=I_K\otimes(X_J\otimes I_I),$$

and

$$W_2=-\frac{1}{2}\left(\frac{1}{n_{\cdot\cdot1}}I_I\oplus\frac{1}{n_{\cdot\cdot2}}I_I\oplus\cdots\oplus\frac{1}{n_{\cdot\cdot K}}I_I\right)$$

where U_r is a $r\times r$ matrix of ones, I_r is a $r\times r$ identity matrix, X_r is a $l\times r$ matrix of ones, and Y_r is a $r\times l$ matrix of ones.

Then

$$\text{TSS}=\frac{n}{2}+V'TV,\ \text{WSS}_1=\frac{n}{2}+V'AW_1A'V,$$

$$\text{WSS}_2=\frac{n}{2}+V'BW_2B'V,\qquad BSS_1=V'(T-AW_1A')V,$$

$$BSS_2=V'(T-BW_2B')V,$$

Now to prove that BSS_1 and BSS_2 are independent, it suffices to show that

$$(T-AW_1A')Z(T-BW_2B')=0$$

(see[3]).

$$AW_1A'=-\frac{1}{2n}\left[U_K\otimes\left(\frac{n}{n_{\cdot1\cdot}}I_I\oplus\frac{n}{n_{\cdot2\cdot}}I_I\oplus\cdots\oplus\frac{n}{n_{\cdot J\cdot}}I_I\right)\right]$$

$$BW_2B'=\left(U_J\otimes\frac{1}{n_{\cdot\cdot1}}I_I\right)\oplus\left(U_J\otimes\frac{1}{n_{\cdot\cdot2}}I_I\right)\oplus\cdots\oplus\left(U_J\otimes\frac{1}{n_{\cdot\cdot K}}I_I\right)$$

$$(T-AW_1A')Z(T-BW_2B')=Y_K\otimes(e_{XY}),$$

$X=1,2,\cdots,IJ,$ and $Y=1,2,\cdots,IJK.$

Here

$$e_{XY}=\begin{cases} p_{s'}(1-p_{t'})\left(\dfrac{n^2n_{\cdot st}}{n_{\cdot s\cdot}n_{\cdot\cdot t}}-n\right) & \text{if } s'=t', \\ p_{s'}p_{t'}\left(\dfrac{n^2n_{\cdot st}}{n_{\cdot s\cdot}n_{\cdot\cdot t}}-n\right) & \text{if } s'\neq t', \end{cases}$$

where

$$s'=X-I\left[\frac{X-1}{I}\right],\ t'=Y-I\left[\frac{Y-1}{I}\right],$$

$$s=\left[\frac{X-1}{I}\right]+1-J\left[\frac{\frac{X-1}{I}}{J}\right],\ t=\left[\frac{Y-1}{IJ}\right]+1.$$

Since $n_{\cdot jk}=n_{\cdot j\cdot}n_{\cdot\cdot k}/n$ for all $j,\ k,\ \frac{n^2 n_{\cdot st}}{n_{\cdot s\cdot}n_{\cdot\cdot t}}=n$, i.e.,

$e_{XY}=0$ for all X,Y.

Therefore, BSS_1 and BSS_2 are asymptotically independent.

Theorem 4-3 With large $n_{\cdot jk}$, in the space Ω, and under the hypotheses H_1, H_2, and H_3,

$$(n-1)(I-1)BSS_3/TSS$$

is approximated as $\chi^2(I-1)(J-1+K-1)$.

Proof If $I=0$, then $BSS_3=BSS_1+BSS_2$. Hence,

$$\frac{(n-1)(I-1)BSS_3}{TSS}=\frac{(n-1)(I-1)BSS_1+(n-1)(I-1)BSS_2}{TSS}$$

With large $n_{\cdot jk}$ and under the hypotheses H_1 and H_2, the distributions of $\frac{(n-1)(I-1)BSS_1}{TSS}$ and $\frac{(n-1)(I-1)BSS_2}{TSS}$ are approximated as $\chi^2(I-1)(J-1)$ and $\chi^2(I-1)(K-1)$ respectively. With large $n_{\cdot jk}$ and under the hypothesis H_3, BSS_1 and BSS_2 are asymptotically independent. So $(n-1)(I-1)BSS_3/TSS$ is approximated as $\chi^2(I-1)(J-1+K-1)$.

REFERENCES

[1] Richard J. Light and Margolin, Barry H., "An Analysis of Variance for Categorical Data", *Journal of the American Statistical Association,* Vol. 66, No. 335, (1971), pp. 534-544.

[2] Graybill, F.A., *Introduction to Matrices with Applications in Statistics,* Wadsworth Publishing Co. Inc., 1969.

[3] Graybill, F.A., *An Introduction to Linear Statistical Models,* New York; McGraw-Hill Book Co., 1961.