Bull. Korean Math. Soc.
Vol. 13, No. 1, 1976.

UNEQUAL SIZE, TWO-WAY ANALYSIS OF VARIANCE FOR CATEGORICAL DATA

BY 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 experimertal groups and I unordered response categories.

Assume that there are n respenses $X_1, \cdots, X_n$. Each X_i is the name of one of I possible categories.

Definition; The variation for categorical responses $X_1, \cdots, X_n$ is

$$\frac{1}{2n}\sum_{j=1}^{n}\sum_{i=1}^{n}d_{ij}{}^2=\frac{1}{2n}\sum_{j=1}^{n}\sum_{i=1}^{n}d_{ij}$$

where $d_{ij}=1$ if X_i and X_j name different categories.

$=0$ if X_i and X_j name the same category.

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 $\sum_{i=1}^{I} n_i=n$. The variation of these responses is:

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

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.*

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.

This note is a 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].

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 reponses 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 $(n_{1jk}, n_{2jk}, \cdots, n_{Ijk})$ obey a multinomial law:

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

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

$$\boldsymbol{V}=(n_{111}, n_{211}, \cdots, n_{I11}, n_{121}, n_{221}, \cdots, I_{21}, \cdots, n_{1J1}, n_{2J1}, \cdots, n_{IJ1}, n_{112}, n_{212}, \cdots, n_{I12}, \cdots, n_{1JK}, n_{2JK}, \cdots, n_{IJK}),$$

then

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

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

where

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

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

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

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

$$WSS_1=\sum_{j=1}^{J}(n_{.j.}/2-\sum_{i=1}^{I}n_{ij.}^{2}/2n_{.j.});$$

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

$$BSS_1=TSS-WSS_1;$$

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

$$WSS_2=\sum_{k=1}^{K}(n_{..k}/2-\sum_{i=1}^{I}n_{i.k}^{2}/2n_{..k});$$

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

$$BSS_2=TSS-WSS_2;$$

the within-cell variation (WSS_3) is

$$WSS_3=\sum_{k=1}^{K}\sum_{j=1}^{J}(n_{.jk}/2-\sum_{i=1}^{I}n_{ijk}^{2}/2n_{.jk});$$

the between-cell variation (BSS_3) is

$$BSS_3=TSS-WSS_3;$$

where

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

3. Definitions

DEFINITION 1. The interaction between the 2nd index level and the 3rd index level is defined as $I=ESS_3-ESS_1-BSS_2$.

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

DEFINITION 3. p_i is the probability of an element belonging to ith category. $p_{ij.}$ is the probability of an element belonging to ith category and jth level, regardless of the 3rd index level. $p_{i.k}$ is the probability of an element be-

longing to ith category and kth level, regardless of the 2nd index level.

DEFINITION 4. The hypothesis H_1 is $p_{ij.}=p_i$ for all j. The hypothesis H_2 is $p_{i.k}=p_i$ 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,

$$(n-1)(J-1)BSS_1/TSS$$

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

(*b*) *Under the hypothesis* H_2,

$$(n-1)(I-1)BSS_2/TSS$$

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* $n_{.jk}=n_{.j.}n_{..k}/n$ *for all* j, k, BSS_1 *and* BSS_2 *are asymptotically independent under the hypothesis* H_3.

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

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

where

$$\boldsymbol{Z}_{jk}=n_{.jk}\begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_I \\ & p_2(1-p_2) & \cdots & -p_2p_I \\ & & \vdots & \vdots \\ & & \cdots & p_I(1-p_I) \end{pmatrix}$$

Let

$$\boldsymbol{T}=-(\boldsymbol{U}_{JK}\otimes\boldsymbol{I}_I)/2n,\quad \boldsymbol{A}=\boldsymbol{Z}_K\otimes\boldsymbol{I}_{IJ},\quad \boldsymbol{A}'=\boldsymbol{X}_K\otimes\boldsymbol{I}_{IJ},$$

$$\boldsymbol{W}_1=-\frac{1}{2}\left(\frac{1}{n_{.1.}}\boldsymbol{I}_I\oplus\frac{1}{n_{.2.}}\boldsymbol{I}_I\oplus\cdots\oplus\frac{1}{n_{.j.}}\boldsymbol{I}_I\right),$$

$$\boldsymbol{B}=\boldsymbol{I}_K\otimes(\boldsymbol{Y}_J\otimes\boldsymbol{I}_I),\quad \boldsymbol{B}'=\boldsymbol{I}_K\otimes(\boldsymbol{Y}_J\otimes\boldsymbol{I}_I),$$

and

$$\boldsymbol{W}_2=-\frac{1}{2}\left(\frac{1}{n_{..1}}\boldsymbol{I}_I\oplus\frac{1}{n_{..2}}\boldsymbol{I}_I\oplus\cdots\oplus\frac{1}{n_{..K}}\boldsymbol{I}_I\right),$$

where $\boldsymbol{U}_r$ is a $r\times r$ matrix of ones, $\boldsymbol{I}_r$ is a $r\times r$ indentity matrix, $\boldsymbol{X}_r$ is a $1\times r$ matrix of ones, and $\boldsymbol{Y}_r$ is a $r\times 1$ matrix of ones.

Then

$$TSS=\frac{n}{2}+\boldsymbol{V}'\boldsymbol{T}\boldsymbol{V}, \qquad WSS_1=\frac{n}{2}+\boldsymbol{A}'\boldsymbol{V}\boldsymbol{W}_1\boldsymbol{A}'\boldsymbol{V},$$

$$WSS_2=\frac{n}{2}+\boldsymbol{V}'\boldsymbol{B}\boldsymbol{W}_2\boldsymbol{B}'\boldsymbol{V}, \quad BSS_1=\boldsymbol{V}'(\boldsymbol{T}-\boldsymbol{A}\boldsymbol{W}_1\boldsymbol{A}')\boldsymbol{V},$$

$$BSS_2=\boldsymbol{V}'(\boldsymbol{T}-\boldsymbol{B}\boldsymbol{W}_2\boldsymbol{B}')\boldsymbol{V}.$$

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

$$(\boldsymbol{T}-\boldsymbol{A}\boldsymbol{W}_1\boldsymbol{A}')\boldsymbol{Z}(\boldsymbol{T}-\boldsymbol{B}\boldsymbol{W}_2\boldsymbol{B}')=\boldsymbol{0}$$

(see[3]).

$$\boldsymbol{A}\boldsymbol{W}_1\boldsymbol{A}'=-\frac{1}{2n}\left[\boldsymbol{U}_K\otimes\left(\frac{n}{n_{\cdot 1\cdot}}\boldsymbol{I}_I\oplus\frac{n}{n_{\cdot 2\cdot}}\boldsymbol{I}_I\oplus\cdots\oplus\frac{n}{n_{\cdot J\cdot}}\boldsymbol{I}_I\right)\right],$$

$$\boldsymbol{B}\boldsymbol{W}_2\boldsymbol{B}'=\left(\boldsymbol{U}_J\otimes\frac{1}{n_{..1}}\boldsymbol{I}_I\right)\oplus\left(\boldsymbol{U}_J\otimes\frac{1}{n_{..2}}\boldsymbol{I}_I\right)\oplus\cdots\oplus\left(\boldsymbol{U}_J\otimes\frac{1}{n_{..K}}\boldsymbol{I}_I\right),$$

$$(\boldsymbol{T}-\boldsymbol{A}\boldsymbol{W}_1\boldsymbol{A}')\boldsymbol{Z}(\boldsymbol{T}-\boldsymbol{B}\boldsymbol{W}_2\boldsymbol{B}')=\boldsymbol{Y}_K\otimes(e_{XY}),$$

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

Here

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

where

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

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

Since $n_{\cdot jk}=n_{\cdot j\cdot}n_{\cdot\cdot k}/n$ for all j, j, $\dfrac{n^2 n_{\cdot st}}{n_{\cdot s\cdot}n_{\cdot\cdot t}}=n$, i.e., $e_{XY}=0$ for all X, Y.

Therefore, BSS_1 and BSS_2 are asymptotically independent.

THEOREM4-3. *With large $n_{.jk}$, 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 $n_{.jk}$ 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 $n_{.jk}$ 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 Barry H. Margolin, *An analysis of Variance for Categorical Date, Journal of the American Statistical Association*, Vol. **66**, No. 335 (1971), 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; Mcraw-Hill Book Co., 1961.

Seoul National University