International Journal of Mathematics Trends and Technology
ISSN: 2231-5373/ https://doi.org/10.14445/22315373/IJMTT-V68I9P503

Volume 68 Issue 9, 11-15, September 2022

Original Article

Analysis of Variance of Categorical Data

Khaled Fattah Sheet[1], Khawla Musta Sadiq[2]

[1,2]*Department of Mathmatics, College of Education for Pure Sciences University of Mosul.*

Received: 21 July 2022 Revised: 22 August 2022 Accepted: 02 September 2022 Published: 14 September 2022

Abstract - *The present study focuses on analysis of variance of categorical data. So, it was noted that contingency tables are often subjected to the analysis of variance (abbreviated hereafter as ANOVA), using the so-called log linear model. Some references were also provided. The current study applied dual scaling to ANOVA of categorical date. One natural question is whether or is not dual scaling is any better than the log linear approach. Therefore, there does not seem to be any comparative study available, and hence our discussion must remind speculative.*

Keywords - *Analysis, mathematics, Variance, Dual scaling.*

1. Introduction

The difference between Dual scaling and Leo linear approach:

(Gabriel, 1971), the log linear approach postulates an experimental design model with respect to logarithmic transforms of cell frequencies of contingency table and its analysis is guarded with all the assumptions entertained in the so-called normal theory. The model is then amenable to, for example maximum likelihood estimation or weighted least-squares estimation of parameters, and sampling distributions of the estimates can in general be specified. One can option so-called minimum-variance estimates, which implies that the analysis is optimum in the least-squares sense.

(Anderberg, 1973), the dual scaling approach doesn't employ any assumption about the distribution of response, which may appear to present some advantage of dual scaling over the log linear approach. However, the same point makes it difficult to make any probabilistic statement about weights or, more specifically, the sampling stability of the optimal weights. In this approach, the unit of analysis is not necessarily of cell frequency, but can be a single response from a single subject. In other words, it can scale response-pattern matrix of I, s and O, s so as to maximize the contributions of the ANOVA parameters. Individual differences are effectively used in scaling the data unlike the dual scaling approach, the log linear approach assumes that subject is randomly chosen, and that individual differences are nothing but random fluctuations. Consequently, the logline approach cannot derive weight for non-numerical responses.

(Benzecri, 1992), the next difference is in the transformation of data. The logarithmic transformation is at variance-stabilizing and in some sense has the effect of reducing the contribution of higher order moment or interactions to the total variance. The ANOVA decomposition however, is unique Dual scaling in contrast, can choose any effect or combination of effects to optimization, note that the individual sums of squares of effects naturally change, depending on the choice of the optimization criterion. This choice, or scaling itself, introduces an additional step in its procedure, namely, adjustment of degrees of freedom for the error term. Although some degrees of freedom will be lost in the dual scaling approach to ANOVA as just mentioned above, dual scaling is still blessed with a much large number of degrees of freedom to manipulate, for it can deal with the response-pattern representation of data. Rather than just with the contingency table. This aspect of dual scaling suggests the possibility that this approach might almost always provide a more clear-out picture of the effects of the parameters than the log linear approach. This, however, must remain a conjecture.

The comparison of the two approach deserves scrutiny, and extensive work may be necessary. Comparisons based on a few numerical examples may be misleading and hence will not be made. This is an important problem from both theoretical and practical points of view, however, and detailed analysis is much desired.

2. The Eoritical Approach: (Fisher, 1948)

This study presented dual scaling of categorical data in the ANOVA context. Fisher (1948) did pioneer work about this approach which was generalized to multi-way designs by Nishisato (1971 a, 1972 a, b, 1972 a)

The squared correlation ratio wa derived in the context of the one-way ANOVA. To see the relation between dual scaling and ANOVA, we can present a summary table of one-way ANOVA (TABLE1). As we recall, the squared correlation ratio, ɲ2 , is defined as SSb/SSt , statistic F, which is used in the one-way AOVA, is defined by

F = MSb / MSw

Both ɲ2 and F are associated with the magnitude of group differences, and are related simply as follows :

$$F = \frac{MSb}{MSw} = \frac{\frac{SSb}{dfb}}{\frac{SSw}{dfw}} = \frac{\frac{SSb}{dfb}}{\frac{(SSt-SSb)}{dfw}} = \frac{\frac{ɲ^2}{dfb}}{\frac{(1-ɲ^2)}{dfw}} \quad \ldots(2)$$

(Aitkin, 1982), in the context of dual scaling, therefore, it follows that maximization of ɲ2 is equivalent to maximization of)) F)). it is assumed that ɲ2 $\neq$ 1.

To illustrate further the relation between the two analyses, consider the two-way ANOVA with the crossed design Table 2 is the summary table, which corresponds to the table ((1)).

The sums of squares and degrees of freedom are related as follows:

$$SSt = SSb + SSw = (SSA + SSB + SSAB) + SSw \quad \ldots(3)$$

$$dft = dfb + dfw = (dfA + dfB + dfAB) + dfw \quad \ldots(4)$$

Table 1. One-way Analysis of Variance

Source of variation	df	Sum of squares	Mean squro	F-ratio
Between groups with in group	dfb dfw	SSb SSw	MSSb= SSb /df MSSw=SSw/dfw	MSSb/ MSSw
Total	dft	SSt		

(KIERS-H.A. L, 1989), note that the between-groups sum of squares has three (disjointed) components, SSA SSB and SSAB. I dual scaling is carried out so as to maximize SSb / SSt, it has the effect of maximizing the overall F ratios. It is important to recall that maximization of SSb / SSt means minimization of SSw / SSt, and these two together lead to maximization of the F ratios. Thus, if dual scaling is carried out to maximize the relative contribution of only a subset of SSb, say SSA and SSB, it does not generally maximize the corresponding F ratios. It simply maximizes (SSA + SSB) / SSt, and not (SSA + SSB) / SSw. In other words, SSw is not necessarily minimized by the maximization of (SSA + SSB) / SSt. It may be SSAB, rather than SSw, wich is substantially reduced.

The idea of quantifying categorical data so as to maximize the effects of particular treatments is probably very appealing to researchers in applied areas. One may be interested in the effects of only a few treatents our those of particular interactions.one may not wish to include high-order interaction in the aximization simply because they are too difficult to Interpret No matter what one's interests are, the dual scaling approach offers an interesting way of performing data analysis.

When ANOVA is applied to categorical data, one may be concerned about violations of its underlying assumption (e.g., normality, homoscedasticity). However, the present approach deals with a linear combination of weighted responses as a unite of analysis, and the central limit theorem would at least ensure its asymptotic normality the ANOVA procedures are also known to be quite robust with respect to violations of the assumptions. To feel truly comfortable, However, it seems as though we need further work by statisticians.

let us look at two of the possible procedures.

Table 2. A summary table tow-way analysis of variance (crossed design)

Source of variation	df	Sum of squares	Mean sqares	
Treatment A	dfA	SSA	MSA= SSA/dfA	MSA / MSw
Treatment B	dfB	SSB	MSA= SSB/dfB	MSB / MSw
Treatment A X B	dfAB	SSAB	MSAB= SSAB/dfAB	MSAB / MSw

Between groups	dfb	SSb	MSb= SSb/dfb	
withingroups	dfw	SSw	MSw= SSw/dfw	
Total	dft	SSt		

3. Applications

It could be obtained optimal score vector yr depending on the case. yr then is subjected to the ordinary ANOVA procedures. Let us consider a pxq crossed design with the same number of subjects in each of the pq cells. suppose that N subjects answered n multiple- choice questions which have in the total m response optional. Then the ANOVA results. It can be summarized as in table 3. recall that yr is already optimally scaled, and that the projection operators are used simply to indicate orthogonal deco potions. unlike the ordinary ANOVA of y of n scores, we have Nm responses, with some constraints. The study estimates m weights with the condition with the sum of the weighted responses is zero. As mentioned in other studies this condition implies that the sume of the weighted responses of each item is zero. ln consequence, that total degrees of freedom are the number of independent responses N (n-m), minus the number independent estimates m-n, that is, dft = (N-1) (n-m). this change in dft, however effects only dfw as in table (3)

Table 3. Two – way ANOVA (method 1)

Source of variation	df	Sum of squares	Mean squares &F*
Treatment A	P-1	$Y_r^{`}\ P_\alpha Y_r$	MSA= SSA/dfA
Treatment B	q-1	$Y_r^{`}\ P_B Y_r$	MSA= SSB/dfB
Interaction A X B	(P-1) (q-1)	$Y_r^{`}\ P_{\alpha B} Y_r$	MSAB= SSAB/dfAB
Between groups	(Pq-1)	$Y_r^{`}\ (P_\alpha + P_B + P_{\alpha B})\ Y_r$	MSb= SSb/dfb
within groups	(N-1) (m-n)- Pq +1	$Y_r^{`}$ (1- $P_{\alpha B}$) Y_r	MSw= SSw/dfw
Total	(N-1) (m-n)	$Y_r^{`}\ Y_1$	

See table (2) for appropriate entries.

4. Method II

Method I is probably satisfactory and also easy enough to but to routine use, but method II presents another interesting approach to the same problem in 1971 NISHISATO considered a simple generalization of the one-way ANOVA to a multi-way ANOVA through dual scaling.

To illustrate his approach, let us consider a multiple-choice date. Obtained from N subject who was sampled according to a2x2 crossed design. Define

F: the N*m response- pattern matrix.

t: the number of groups of subjects.

generated by the two-way classification, that is t = 4

nij = the number of subjects in the group (i,j), i= 1,2 ؛ j = 1,2

D* = the t*N matrix such that, in such example

$$D^* = \begin{bmatrix} 1\ 1 \dots 1 & 0\ 0 \dots 0 & 0\ 0 \dots 0 & 0\ 0 \dots 0\ 0\ 0 \dots 0 & 1\ 1 \dots 1 & 0\ 0 \dots 0 & 0\ 0 \\ \dots 0\ 0\ 0 \dots 0 & 0\ 0 \dots 0 & 1\ 1 \dots 1 & 0\ 0 \dots 0\ 0\ 0 \dots 0 & 0\ 0 \dots 0 & 0\ 0 \dots 0 & 1\ 1 \dots 1 \end{bmatrix}$$

n11 n12 n21 n22

DN = the N*N diagonal matrix of the row totals of F;

D = the m * m diagonal matrix of the column totals of F;

F. = D*F = the t * N matrix with elements being sums of column entries within groups;

Dt = the t * t diagonal matrix of the row totals of F;

X = the m * 1 vector of weights for the m options;

ε = the h * 1 vector of ANOVA parmeters

Then the vector of optimal scores of N subjects and the vector of the cell means of optimal scores are, respectively,

y = $\frac{D_N^{-1}Fx}{\eta}$

y. = $\frac{D_N^{-1}F.x}{\eta}$ = $\frac{D_t^{-1}D^{*}Fx}{\eta}$

y can be expressed in terms of the ANOVA parameters as

y. = A ε + e = K L ε + e. = Kθ + e.

Where A is the design matrix, e. Is the vector residual, y. - A ε , called errors, θ is the vector of q linearly estimable functions of ε, that is θ = L ε and K is abas is matrix.

Finally – Assuming that e. ~ N (o $\sigma^2 D_t^{-1}$) , the unbiased minimum variance estimate θ is given by

$\hat{\theta}$ = ($\acute{K}D_tK$)-1 $\acute{K}D_t$y

= ($\acute{K}D_tK$)-1 $\frac{\acute{K}F.X}{\eta}$

ANOVA information's mention soon in another research in including the application.

5. Conclusion

Dual scaling can be ablied in many statistical approaches with special analysis of variance ablieal on each one to find the optimal solution for every approach.

References

[1] Aitkin, M, "Analysis of Categorical Date Dual Scaling and Its Applications," *Royal Statistical Society Journal Series A,* vol.145, pp.513–516, 1982.

[2] Anderberg, M, "Cluster Analysis for Applications," 1973.

[3] Benzecri, J. P, "Correspondence Analysis Handbook," *Macel Dekker.*

[4] Benali, H, "Stability of Factorial Analysis of Multiple Correspondences in the Event of Missing Data and Modalities With Small Numbers," *Doctoral Dissertation, INRIA),*1986.

[5] Benzécri, J. P, "Validity of Assessment Scales In Psychology and Psychiatry and Psychosocial Correlations," *Cahiers De L'analyse Des Données,* vol.1, no.17, pp.55–86, 1992.

[6] Blasius, J., & Greenacre, M. (Eds.), *Visualization and Verbalization of Data. CRC Press,* 2014.

[7] Bock, H. H, "Probabilistic Models In Cluster Analysis," *Computational Statistics & Data Analysis,* vol.1, no.3, pp.5–28, 1996.

[8] Carman, J. M, "Consumer Perceptions of Service Quality," *An Assessment of T. Journal of Retailing,* vol.1, no.66, pp.33, 1990.

[9] Chazal, F., & Michel, B, "An Introduction to topological Data Analysis," *Fundamental and Practical Aspects for Data Scientists,* 2017.

[10] Cooksey, R. W, "The Methodology of Social Judgement Theory," *Thinking & Reasoning, vol.3*, no.21, pp.141–174, 1996.

[11] Fisher, R. A, "The Precision of Discriminant Functions," *Annals of Eugenics,* vol.1, no.10, pp.422–429, 1940.

[12] Greenacre, M, " Correspondence Analysis In Practice," In *Hapman and Hall/Crc,* 2017.

[13] Guttman, L, "An Approach for Quantifying Paired Comparisons and Rank Order," *The Annals of Mathematical Statistics,* vol.17, no.2, pp.144–163, 1946.

[14] Li, V. O., & Qiu, X, "Personal Communication Systems (PCS," *Proceedings of the IEEE,* vol.9, no.83, pp.1210–1243, 1995.

[15] Meulman, J. J, " Optimal Scaling Methods for Multivariate Categorical Data Analysis," *SPSS White Paper . : Chicago,* 1988.

[16] Mucha, H. J, " Improvement of Stability in Cluster Analysis and Principal Components Analysis By Special Weighting the Variables," *In Operations Research,* vol.91, pp.351-354, 1992. Physica-Verlag HD.

[17] Mullet, G. M, " Correspondence Analysis: A New tool for Image Studies," *Journal of Professional Services Marketing* , vol.3, no.2, pp.44 –14, 1987.
[18] Nishisato, S., & Ahn, H, " When not to Analyze Data: Decision Making on Missing Responses In Dual Scaling ," *Annals of Operations Research*, vol.2, no.55, pp.361–378, 1995.
[19] Nishisato, S., & Clavel, J. G, " A note on Between-Set Distances In Dual Scaling and Correspondence Analysis," *Behaviormetrika*, vol.1, no.30, pp.87–98, 2003.
[20] Nishisato, S., & Gaul, W, " Marketing Data Analysis By Dual Scaling," *International Journal of Research In Marketing*, vol.3, no.5, pp.151–170, 1988.
[21] Nishisato, S., & Sheu, W. J, "A note on Dual Scaling of Successive Categories Data," *Psychometrika*, vol.4, no.49, pp.493–500, 1984.
[22] Nishisato, S, "Graphical Representation of Quantified Categorical Data: Its Inherent Problems,"J*ournal of Statistical Planning and Inference*, vol.1–2, no.43, pp.121–132, 1995a.
[23] Nishisato, S, " Ptimization and Data Structure: Seven Faces of Dual Scaling," *Annals of Operations Research,* vol.2, no.55, pp.345–359, 1995b.
[24] Nishisato, S, " Gleaning In the Field of Dual Scaling," *Psychometrika*, vol.4, no.61, pp.559–599, 1996.
[25] Ogasawara, H, " Oblique Factors and Components With Independent Clusters," *Psychometrika*, vol.2, no.68, pp.299–32, 2003.
[26] Padberg, M. W., Van Roy, T. J., & Wolsey, L. A, "Valid Linear Inequalities for Fixed Charge Problems," *Operations Research*, vol.4, no.33, pp.842–861, 1985.
[27] Sharma, R., Joshi, S., Singh, K. D., & Kumar, A, "Visual Evoked Potentials: normative Values and Gender Differences," *Journal of Clinical and Diagnostic Research: JCDR*, vol.7, no.12, pp.CC12, 2015.