Biometrics 30, 507–512
September 1974

A CLUSTER ANALYSIS METHOD FOR GROUPING MEANS IN THE ANALYSIS OF VARIANCE

A. J. Scott and M. Knott

University of Auckland, Auckland, New Zealand and The London School of Economics and Political Science

SUMMARY

It is sometimes useful in an analysis of variance to split the treatments into reasonably homogeneous groups. Multiple comparison procedures are often used for this purpose, but a more direct method is to use the techniques of cluster analysis. This approach is illustrated for several sets of data, and a likelihood ratio test is developed for judging the significance of differences among the resulting groups.

1. INTRODUCTION

When the result of an F-test in an analysis of variance shows that treatment means differ, it is often important to obtain some idea of where these differences are. There are many papers dealing with the large number of multiple comparison procedures that have been proposed to help the search for differences among means. For some purposes it is enough to split the means into approximately homogeneous groups. In discussing a hypothetical example, Tukey [1949] said "At a low and practical level, what do we want to do? We wish to separate the varieties into distinguishable groups as often as we can without too frequently separating varieties which should stay together". Tukey proposed a sequence of multiple comparison procedures to accomplish this grouping, each based on an intuitive criterion. Another method of grouping the treatment means is cluster analysis. The possibility of using cluster analysis in place of a multiple comparison technique was suggested by Plackett in his discussion of the review paper by O'Neill and Wetherill [1971].

In this paper we study the consequences of using a well-known method of cluster analysis to partition the sample treatment means in a balanced design and show how a corresponding likelihood ratio test gives a method of judging the significance of the differences among groups obtained.

2. A LIKELIHOOD RATIO TEST FOR TWO GROUPS

Suppose we have a set of independent sample treatment means $y_1, y_2, \cdots, y_k$ with $y_i \sim N(\mu_i, \sigma^2)$, and an independent estimate, s^2, of the common variance where $(\nu s^2)/\sigma^2 \sim \chi_\nu^2$. We could check the homogeneity of the means with an F-test in the usual way, but if we suspect that the means fall into two distinct, internally homogeneous groups then it is natural to consider the likelihood ratio test for this specific alternative. Thus, let B_0 be the maximum value, taken over all possible partitions of the k treatments into two groups, of the between groups sum of squares, and let $\hat{\sigma}_0^2$ be the maximum likelihood (ML) estimate of σ^2 when all μ_i's are assumed equal, i.e.

$$\hat{\sigma}_0^2 = \left[\sum_1^k (y_i - \bar{y})^2 + \nu s^2 \right] \Big/ (k + \nu).$$

Then it is easy to verify that the likelihood ratio test for the null hypothesis $H_0 : \mu_i = \mu$ $(i = 1, \cdots, k)$ against the alternative that μ_i is equal either to m_1 or m_2 (with at least one mean in each group), where m_1 and m_2 represent the unknown means of the two groups, is equivalent to a test that rejects H_0 if $B_0/\hat{\sigma}_0^2$ is too large. This requires the null distribution of $B_0/\hat{\sigma}_0^2$. Using the methods of Hartigan ([1972] section 4), it can be shown that, as $k \rightarrow \infty$, $k^{1/2}\left(\frac{B_0}{k\sigma^2} - \frac{2}{\pi}\right)$ converges in distribution to a normal random variable with mean zero and variance $\frac{8(\pi - 2)}{\pi^2}$ when H_0 is true. Hence, take for a modified test statistic $\lambda = \frac{\pi}{2(\pi - 2)} B_0/\hat{\sigma}_0^2$. Then it follows that $k^{1/2}\left(\frac{\lambda}{k} - \frac{1}{\pi - 2}\right)$ is asymptotically equivalent to $k^{1/2}\left(\frac{Z}{k} - \frac{1}{\pi - 2}\right)$ where Z is a χ^2 random variable with $\nu_0 = \frac{k}{\pi - 2}$ degrees of freedom (D.F.). This suggests approximating the percentage points of the null distribution of λ by those of a χ^2 distribution with ν_0 D.F. An extensive simulation was carried out to determine exact percentage points of the null distribution of λ, and the results are reported in section 5. It turns out that the χ^2 approximation is very good indeed even for k as small as two (see Table 2) and should be adequate for most practical situations.

There is a one to one relationship between this likelihood ratio test and a standard method of cluster analysis. Calculation of λ involves finding that partition of the treatments for which the between groups sum of squares is a maximum (or equivalently, the within groups sum of squares is a minimum). This partition is the one given by the method of cluster analysis suggested by Edwards and Cavalli-Sforza [1965], when applied to the univariate means $y_1, \cdots, y_k$. Whether one considers a likelihood ratio test or uses the intuitive justifications of cluster analysis, it is clear that the ML partition under H_1 gives an estimate of which means are in each of the groups postulated by H_1. We need to be able to find this ML partition easily to calculate λ. There are $2^{k-1} - 1$ possible partitions of the k means into two nonempty groups but Fisher [1958] has shown that it is enough to look at the $(k - 1)$ partitions formed by ordering the means and dividing between two successive ones. This makes the calculation feasible even by hand if k is no more than 11 or 12. In particular when $k = 3$, it is only necessary to split the ordered means at the largest gap.

TABLE 1

SIMULATION RESULTS FOR TWO GROUPS, EACH CONTAINING 5 MEANS, WITH 40 D. F.

Number of groups obtained	δ 1	 2	 3	 4
1	4407	2677	562	20
2	589	2270	4276	4660
More than 2	4	53	162	320
Percent error	88.2	54.6	14.5	6.8

TABLE 2

95% POINTS FOR THE DISTRIBUTION OF

$$\lambda = \frac{\pi}{2(\pi - 2)} \frac{B_0}{\hat{\sigma}_0^2}$$

$\frac{\nu}{k}$	k			
	2	5	10	20
0	2.75	6.60	12.11	21.74
1	4.97	9.31	15.13	26.14
2	5.44	9.77	15.92	27.14
3	5.50	10.30	16.22	27.77
4	5.50	10.44	16.61	27.93
5	5.47	10.47	16.73	28.01
∞	5.28	10.89	17.22	29.26
From $\chi^2_{\nu_0}$	5.46	10.09	16.58	28.25

3. THREE OR MORE GROUPS

In practice there are likely to be several groups of treatments so that it is not always enough to split the means into just two groups. We adopt the hierarchical splitting method suggested by Edwards and Cavalli-Sforza [1965] in their work on cluster analysis. This starts with the best split into two groups, based on the between groups sum of squares, and then applies the same procedure separately to each subgroup in turn. The subdivision process is continued until the resulting groups are judged to be homogeneous by application of the λ test of section 2. This method is simple to apply, and it is often easier to interpret the results in an unambiguous way with a hierarchical method in which the groups at any stage are related to those of the previous stage.

Choosing an appropriate value for α is difficult. If α is too small, the splitting process will terminate too soon, while if α is too large, the process will go too far and split homogeneous sets of means. It would be particularly desirable to know the probability that the method will split the means into more than p groups when in fact the treatments fall into exactly p groups. Suppose first that there are really two groups of treatments. If they are so widely separated that the true groups will almost always be recovered, then the probability that we stop at two groups is $(1 - \alpha)^2$ where α is the significance level for λ used at the second stage. If the separation is less extreme, the estimated groups will appear more homogeneous than the true groups, and the probability of stopping at two groups is greater than $(1 - \alpha)^2$. Thus the probability of splitting the treatments into more than two groups is bounded above by $\alpha^* = 1 - (1 - \alpha)^2$. More generally, if we attempt to split j homogeneous groups at some stage, the probability of getting at least one significant split is no more than $\alpha^* = 1 - (1 - \alpha)^j$. We have done some simulation with $k = 10$ means, 5 in each group, and $\nu = 40$ D.F. Five thousand samples were generated for $\delta = 1, 2, 3, 4$ where $\delta = (m_1 - m_2)/\sigma$ is the normalized distance between the group means m_1 and m_2. The χ^2 approximation was used to judge the significance of splits with $\alpha = 0.05$ so that the upper bound is $\alpha^* = 0.0975$, and the results are set out in Table 2. Effective complete separation starts at about $\delta = 5$, and the upper bound is a good approximation for values of

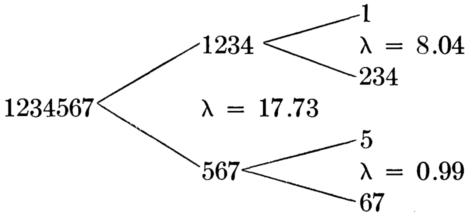
δ larger than this. If $\delta \leq 3$, most of the errors result from failure to recognize genuine splits, and the overall error rate would be reduced by choosing a larger value of α. For values of $\delta \geq 4$, on the other hand, there are too many improper splits, and the overall error rate would be reduced with a smaller value of α.

4. EXAMPLES

We illustrate our method with three examples that have been used for other multiple comparison procedures. In each case the χ^2 approximation is used to judge significance at a nominal level of 5%.

Example 1

Shulkcum (see Duncan [1955]) conducted a randomized block experiment involving six blocks of seven varieties of barley. The variety means were 49.6, 58.1, 61.0, 61.5, 67.6, 71.2, 71.3. The analysis of variance gave a numerical F value of 4.61 on 30 D.F. which suggests strongly that there are differences among the means. When the ordered means are numbered 1 to 7, our modified likelihood analysis gives the breakdown below.

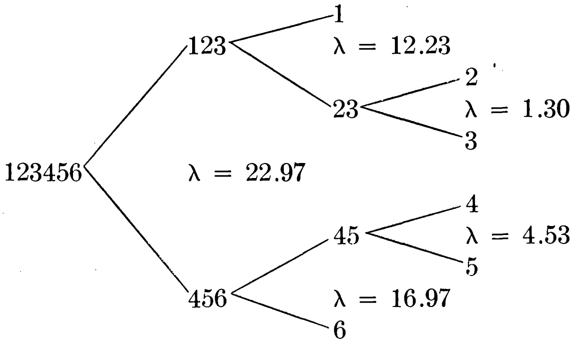


This suggests that the groups should be 1234,567. The split between 1 and 234 is on the borderline of significance, and we may want to split the groups further into 1,234,567 especially in the light of the simulation results reported in section 3. This is the result suggested by Plackett in the discussion of O'Neill and Wetherill [1971] on the basis of a normal probability plot.

Example 2

Tukey [1949] gave examples of his methods in use on the results of a 6 × 6 Latin square experiment on potatoes. The experiment was not a simple Latin square design but had a factorial structure which makes multiple comparison procedures inappropriate. This will be ignored here, as it was in Tukey's analysis. The six means were 345.0, 405.2, 426.4, 477.8, 502.2, 601.8, and the independent estimate of their variance was 254.4 with 20 D.F.

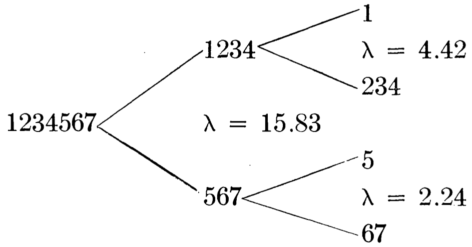
The breakdown of the means (numbered 1 to 6) obtained by our analysis is given below.



The estimated grouping is 1,23,45,6. Tukey obtained the same partition using his approach.

Example 3

Tukey [1949] had a second example from Snedecor ([1946] Example 11.28) which concerned a 7 × 7 Latin square experiment on potatoes. The means were 341.9, 360.6, 363.1, 379.9, 386.3, 387.1, and the estimate of the variance of the sample means was 90.63 with 30 D.F. The breakdown of the means (numbered 1 to 7) obtained by our analysis is given below.



Using the χ^2 approximation, we obtain the estimated grouping 1234,567. Tukey did not obtain the split between 4 and 5 using his battery of tests, but separated the first mean from the others. He remarked that if one uses an ordinary t-test for the particular grouping obtained here (1234,567) the result is significant.

5. THE DISTRIBUTION OF λ

Some information about the distribution of λ is already available. When $k = 2$, λ reduces to a multiple of $(\nu + t_\nu^2)^{-1}$, where the random variable t_ν has a Student t distribution with ν D.F., and percentage points can be obtained directly from those of the t distribution. When $\nu = 0$, there is no direct estimate of σ^2, and λ can be expressed as a monotone function of $C = B_0/W_0$ where W_0 is the minimum within groups sum of squares. Tables of the percentage points of C have been given by Engelman and Hartigan [1969]. As indicated in section 2, the approximate distribution of λ as $k \to \infty$ is $\chi_{\nu_0}^2$ with $\nu_0 = k/(\pi - 2)$.

Information about the distribution of λ for other values of ν and k was obtained by simulation using 5000 independently sampled values of λ in each case. Normal random variables were generated by the Marsaglia-Bray [1964] transform of pseudo-random uniform variates $\{\eta_i\}$ where η_i is the fractional part of $2^{-36}[\eta_{i-1}(2^6 + 1) + 54197344997]$. The histograms of the simulated values of λ were checked to make sure there were no obvious peculiarities due to regularities in the pseudo-random number generator and none was found. From the simulated values of λ the first 100 Fourier sine coefficients of $F_n(\lambda) - \lambda/a$ were calculated, where $F_n(\lambda)$ is the empirical distribution function from the $n = 5000$ simulated values of λ and $a = \frac{\pi(k+\nu)}{2(\pi-2)}$ is an upper bound for λ. A smoothed version, $F_n^*(\lambda)$, of $F_n(\lambda)$ was obtained by using the first m coefficients $\beta_1, \beta_2, \cdots, \beta_m$ in the formula

$$F_n^*(\lambda) = \sum_{j=1}^{m} \beta_j \sin \frac{\pi j \lambda}{a}.$$

Plots of $F_n^*(\lambda)$ for various values of m suggested that $m = 70$ would be the best choice because small and irregular oscillations began appearing in the plots for values of m larger than 70.

Since the distribution of λ as $k \to \infty$ is approximately χ^2 we tried a χ^2 approximation in other cases. A simple way to investigate this is through a probability plot. Wilson and Hilferty [1931] showed that if $Y \sim \chi_\omega^2$ then the distribution of $Y^{1/3}$ is closely approximated by a normal distribution. We plotted $\lambda^{1/3}$ against $\phi^{-1}(F_n^*(\lambda))$, where ϕ is the standard normal distribution function, for a grid of λ values equally spaced in $\lambda^{1/3}$. The plots were almost exact straight lines, particularly if attention was limited to values of λ with $F_n^*(\lambda)$ between 0.50 and 0.99. A final smoothing of the simulated distributions was carried out by fitting straight lines by least squares to the plots between the 50% and 99% values. The correlation coefficients for these straight line fits were always around 0.999. Table 2 shows 5% level significance points for λ estimated from the straight lines fitted to the probability plots. However, since the limiting χ^2 approximation with ν_0 D.F. worked so well for moderate values of ν/k and gave a conservative approximation for small values of ν, it appears that more precise percentage points will only be needed for a large value of ν/k in practice.

ACKNOWLEDGMENT

We appreciate the helpful comments made by an Associate Editor and referees on an earlier draft particularly the reference to the asymptotic results of Hartigan [1972].

UNE MÉTHODE D'ANALYSE EN GROUPES SUR DES MOYENNES EN ANALYSE DE VARIANCE

RESUME

Il est quelquefois utile dans une analyse de variance de répartir les traitements en des groupes raisonnablement homogènes. Des procédures de comparaisons multiples sont souvent utilisées dans ce but, mais une méthode plus directe consiste à employer les techniques de l'analyse en groupes. Ce point de vue est illustré par plusieurs ensembles de données, et un test du maximum de vraisemblance est appliqué pour déterminer si les différences entre les groupes sont significatives.

REFERENCES

Duncan, D. B. [1955]. Multiple range and multiple F tests. *Biometrics 11*, 1–42.

Edwards, A. W. F. and Cavalli-Sforza, L. L. [1965]. A method for cluster analysis. *Biometrics 21*, 362–75.

Engleman, L. and Hartigan, J. A. [1969]. Percentage points of a test for clusters. *J. Amer. Statist. Ass. 64*, 1647–8.

Fisher, W. K. [1958]. On grouping for maximum homogeneity. *J. Amer. Statist. Ass. 53*, 789–98.

Hartigan, J. A. [1972]. Direct clustering of a data matrix. *J. Amer. Statist. Ass.* 67, 123–9.

Marsaglia, G. and Bray, T. A. [1964]. A convenient method for generating normal variables. *SIAM Rev. 6*, 260–4.

O'Neill, R. and Wetherill, G. B. [1971]. The present state of multiple comparison methods. *J. R. Statist. Soc. 33*, 218–50.

Snedecor, G. W. [1946]. *Statistical Methods*. Collegiate Press, Ames, Iowa.

Tukey, J. W. [1949]. Comparing individual means in the analysis of variance. *Biometrics 5*, 99–114.

Wilson, E. G. and Hilferty, M. M. [1931]. The distribution of Chi-square. *Proc Nat. Acad. Sci. U.S.A. 17*, 684–8.

Received July 1972, *Revised November* 1973

Key Words: Multiple comparisons; Cluster analysis; Grouping means.