A MULTIPLE DECISION PROCEDURE FOR CERTAIN PROBLEMS IN THE ANALYSIS OF VARIANCE

By Edward Paulson

University of Washington

1. Introduction. In this paper we will discuss a certain type of problem which arises in many applications of the analysis of variance. We suppose that we are given K varieties, and are required to investigate the differences among them on the basis of the observed yields from a given experimental design, such as a set of randomized blocks or a latin square. The classical procedure [1] for dealing with this problem has been to test the null hypothesis that the K varieties are all equal by computing the ratio of the mean sum of squares between varieties to the residual mean sum of squares, and rejecting the null hypothesis whenever this ratio exceeded the critical value corresponding to the level of significance used. However, the standard discussions of this procedure seem to be quite vague on the question of what action should be taken after the null hypothesis has been rejected.

In a number of problems, the practical situation seems to be such that instead of testing the null hypothesis that the varieties do not differ, what is really required is a statistical rule or "decision function" which on the basis of the observed yields will classify the K varieties into a "superior" group and an "inferior" group. If the superior group consists of more than one variety, the next appropriate action will of course depend on the particular problem at hand. In some situations the varieties in the superior group might then be subject to further selection on the basis of some secondary characteristic, or additional observations might be taken to discriminate between the members of the superior group, after discarding the varieties in the inferior group. However, if all varieties happen to be classified in one group, the group will be labelled "neutral" and this result is to be interpreted as implying that the varieties are homogeneous.

In this formulation, the problem is now of a multiple decision type; it is necessary to decide on the basis of a sample which one out of the $2^K - 1$ possible decisions (or classifications) to select. We will suggest a solution which seems quite reasonable on an intuitive basis, but it is still an open question whether this solution is an optimum one.

2. A special case. In this section we will discuss the problem under the assumption that the variance σ^2 of a single observation is known *a priori*. This is a rather restrictive assumption, but it can be considered as approximately satisfied when the number of degrees of freedom available for estimating the variance is large, which will often be the case. The minor modifications necessary to secure exact results for the small sample case when σ is unknown are

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

discussed in section 3. We also assume that the experimental design has been so selected that there will be the same number (r) of observations on each of the K varieties.

Now let $x_{i\alpha}$ = the αth observation on the ith variety ($i = 1, 2, \cdots, K$; $\alpha = 1, 2, \cdots, r$), let $\bar{x}_i = \Sigma_{\alpha=1}^{r} x_{i\alpha}/r$, put $m_i = E(\bar{x}_i)$ where E stands for expected value, and take λ to be a given positive constant. The conventional assumption is made that all the observations are normally and independently distributed with the same variance σ^2. Denote by $\bar{x}_M$ the maximum of the K mean values $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_K$. The rule for dividing the varieties into superior and inferior groups is the following: the superior group is to consist of all varieties whose corresponding mean values fall in the interval $[\bar{x}_M - \lambda\sigma/\sqrt{r}, \bar{x}_M]$ and the remaining varieties constitute the inferior group. (As mentioned earlier, if all the varieties fall into one group, this group is labelled 'neutral' and the varieties are considered homogeneous.)

This rule completely determines the classification as soon as λ is determined. For a given sample size, we might select λ by considering the relative importance of different types of incorrect classifications. If H denotes the error of misclassifying the varieties when in fact they are all equal, and G denotes the error of misclassifying the varieties when they actually are unequal, then it is obvious that the greater the value of λ, the smaller the probability of an error of type H, but the greater the probability of an error of type G. Therefore for a given value of r it is necessary to adopt some sort of compromise in selecting λ.

For a given value of λ we will now derive explicit formulas for $P(H)$, the probability of not classifying all the varieties in one group when $m_1 = m_2 = \cdots = m_K$, and for $P(G_1)$ the probability that as a result of the experiment there will not be a superior group consisting only of the Kth variety when $m_1 = m_2 = \cdots = m_{K-1} = m$ and $m_K = m + \Delta (\Delta > 0)$. G_1 was selected because it appeared to be the particular kind of type G error most likely to be useful in applications. Also $P(G_1)$ may be regarded as the least upper bound of the probability of misclassifying the varieties when one variety is superior to any of the others by an amount at least equal to Δ. Now if we denote by $W = (\bar{x}_M - \bar{x}_{\min})$ the difference between the maximum and minimum values of the set $\{\bar{x}_i\}$ ($i = 1, 2, \cdots, K$), then it is obvious that

$$1 - P(H) = P\left\{W < \frac{\lambda\sigma}{\sqrt{r}}\right\}. \tag{2.1}$$

The right hand side of (2.1) is equivalent to the probability that the range of a sample of K independent observations from a normal distribution with unit variance be less than λ; this probability has already been tabulated by Pearson and Hartley [2]. From these tables it is a routine matter to find $P(H)$ corresponding to a given value of λ, and conversely. To evaluate $P(G_1)$, we have

$$1 - P(G_1) = P\left\{\bar{x}_i < \bar{x}_K - \frac{\lambda\sigma}{\sqrt{r}} \quad \text{for each } i \quad (i = 1, 2, \cdots, K-1)\right\}.$$

By evaluating the probability of this event for a fixed value of $\bar{x}_K$ and then integrating out with respect to $\bar{x}_K$, it is a simple matter to verify that

$$P(G_1) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(y^2/2)} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y+(\Delta/\sigma)\sqrt{r}-\lambda} e^{-(t^2/2)}\, dt \right]^{K-1} dy. \tag{2.2}$$

In some applications, it may be desirable to have an explicit expression for the probability that the superior group will consist of the Kth variety and not more than s inferior varieties when $m_1 = m_2 = \cdots = m_{K-1} = m$ and $m_K = m + \Delta$. If we denote this probability by $1 - P_s^*$ it is not difficult to show that

$$1 - P_s^* = \sum_{\alpha=0}^{s} \binom{K-1}{\alpha} [T_{1\alpha} + \alpha T_{2\alpha}], \quad \text{where}$$

$$T_{1\alpha} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(y^2/2)} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y+(\Delta/\sigma)\sqrt{r}-\lambda} e^{-(t^2/2)}\, dt \right]^{K-\alpha-1}$$

$$\cdot \left[\frac{1}{\sqrt{2\pi}} \int_{y+(\Delta/\sigma)\sqrt{r}-\lambda}^{y+(\Delta/\sigma)\sqrt{r}} e^{-(t^2/2)}\, dt \right]^{\alpha} dy, \quad \text{and} \tag{2.3}$$

$$T_{2\alpha} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(y^2/2)} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y-\lambda} e^{-(t^2/2)}\, dt \right]^{K-\alpha-1}$$

$$\cdot \left[\frac{1}{\sqrt{2\pi}} \int_{y-\lambda}^{y} e^{-(t^2/2)}\, dt \right]^{\alpha-1} \left[\frac{1}{\sqrt{2\pi}} \int_{y-(\Delta/\sigma)\sqrt{r}-\lambda}^{y-(\Delta/\sigma)\sqrt{r}} e^{-(t^2/2)}\, dt \right] dy.$$

3. General case. We now briefly discuss the exact treatment of the problem when σ is unknown. The notation of section 2 will be used, but in addition denote by s^2 an estimate of σ^2 resulting from the given experimental design which is based on the residual sum of squares with n degrees of freedom. It is well known that s^2 is independent of the set $\{\bar{x}_i\}$ $(i = 1, 2, \cdots K)$. Now the rule to be used in classifying the varieties into two groups is as follows: the superior group is to consist of all those varieties whose mean values fall in the interval $[\bar{x}_M - \lambda s/\sqrt{r}, \bar{x}_M]$, and the inferior group consists of the remaining varieties.

We now find that:

$$1 - P(H) = P\{W \leq \lambda s/\sqrt{r}\}. \tag{3.1}$$

The right hand side of (3.1) depends only on the distribution of the 'studentized' range and has also been tabulated by Pearson and Hartley [3] although the tabulation is considerably less complete than that of the range in [2]. It is also easy to verify that the expression for $P(G_1)$ now becomes

$$P(G_1) = 1 - \frac{n^{n/2}}{\sqrt{2\pi}\, 2^{(n-2)/2} \Gamma\left(\frac{n}{2}\right)} \int_0^{\infty} \int_{-\infty}^{\infty} w^{n-1} e^{-(nw^2+y^2)/2}$$

$$\left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y+(\Delta/\sigma)\sqrt{r}-\lambda w} e^{-(t^2/2)}\, dt \right]^{K-1} dy\, dw \tag{3.2}$$

with a similar modification for P_s^*.

4. Remarks. Any application of the ideas suggested here would be greatly facilitated if tables of $P(G_1)$ were made available. If this were done, it would be possible to decide in advance of an experiment how large r should be in order to have a fixed control over both types H and G_1 errors. It is obvious that further research both along theoretical and applied lines is needed. In conclusion, the writer would like to thank Professor Albert Bowker for several helpful suggestions.

REFERENCES

[1] R. A. FISHER, *Statistical Methods for Research Workers*, Chapters 7, 8.

[2] E. S. PEARSON AND H. O. HARTLEY, "Tables of the probability integral of the range in samples from a normal population," *Biometrika*, Vol. 32 (1941–42), pp. 301–310.

[3] E. S. PEARSON AND H. O. HARTLEY, "Tables of the probability integral of the studentized range," *Biometrika*, Vol. 33 (1943), pp. 89–99.