

## How to Test the Means among Heteroscedastic Normal Populations

MOHD. NAWI ABD. RAHMAN  
 School of Management and Accounting  
 Universiti Utara Malaysia  
 Bandar Darulaman  
 06000 Jitra, Kedah, Malaysia.

**Key Words:** Coefficient of variation (c.v.); maximum likelihood; chi-square test.

### ABSTRAK

*Suatu ujian alternatif telah dicadangkan untuk beberapa populasi normal tak homogen dengan pekali ubahan malar. Untuk memudahkan penggunaan kaedah ini satu algoritma penganggaran yang cekap telah ditunjukkan bagi min-min, dan bagi matriks varians-kovarians sepadannya. Usaha selanjutnya hanya dalam membentuk matriks ujian. Kemudian ujian khi kuasa dua dilakukan dengan nilai yang diperoleh hasil daripada olahan matriks yang mudah.*

### ABSTRACT

*An alternative test is proposed for some heteroscedastic normal populations with constant coefficient of variations. To facilitate the application of the procedure, an efficient estimation algorithm is shown for its means and the corresponding variance-covariance matrix. The next task is only to form the test matrix. The chi-square test is then applied using the value obtained through simple matrix manipulations.*

### INTRODUCTION

Suppose that an experiment has been carried out concerning  $k$  normal populations. It is expected that, due to the conditions of the experiment, the means differ appreciably between one another, and that there is no reason to test whether there exist differences among them, but rather by how much they would be different, or possibly by what multiples. The experimenter may also realise from the samples that the usual homogeneous variance assumption between the populations may be violated. However, it may be observed that the ratios between the variances and the respective means for the samples are nearly constant throughout the  $k$  sets of observations.

The problem above may fit the model known as the constant coefficient of variations model. For the model of this form, suppose that  $y_{ij}$  represents a random  $j$ -th observation for the  $i$ -th

population, where  $j = 1, 2, \dots, n_i$  and  $i = 1, 2, \dots, k$ . Here it is assumed that  $y_{ij} \sim N(\mu_i, \sigma_i^2)$  such that  $c = \sigma_i/\mu_i$  is constant for all populations. Under this set up, the test using the analysis of variance approach (p. 128 Steel and Torrie, 1960) is not proper. We propose now a possible alternative test for the means.

For simplicity we consider only the positive random variables. The extension to the negative variables is by a simple addition of a suitable constant to each observation.

### THE ESTIMATION

The joint density function is given by

$$f(\underline{y}) = \prod_{i=1}^k \prod_{j=1}^{n_i} (2\pi c^2 \mu_i^2)^{-1/2} \exp \left[ -\frac{1}{2} (y_{ij} - \mu_i)^2 / c^2 \mu_i^2 \right]$$

A reasonable method of estimation of the parameters  $\underline{\theta} = (\mu_1, \mu_2, \dots, \mu_k, c)$  is by the maximum likelihood (m.l.) procedure.

There are  $k + 1$  m.l. equations to be solved simultaneously. It is impossible to obtain any closed expressions for the estimates, and even numerical solutions are tedious. A numerical solution technique has been proposed in Abd. Rahman and Gerig (1983). It is very accurate and is shown to be reliable according to the Monte Carlo evaluations.

A somewhat simpler version of computing algorithm is achieved as follows.

The equations to be solved are

$$(c^2 + 1)\mu_i^2 - \mu_i\bar{y}_i - \psi_i^2 = 0,$$

$i = 1, 2, \dots, k$ ; and

$$c^2 = n^{-1} \sum_i n_i \psi_i^2 / \mu_i^2,$$

where  $\bar{y}_i = n_i^{-1} \sum y_{ij}$ ,  $\psi_i^2 = n_i^{-1} \sum_j (y_{ij} - \mu_i)^2$  and  $n = \sum_i n_i$ . The 'cap' signs are omitted to ease writing at this stage. Let  $s_i^2 = n_i^{-1} \sum_j (y_{ij} - y_i)^2$

and  $t_i^2 = s_i^2 y_i^{-2}$ ,  $i = 1, 2, \dots, k$ . By eliminating suitable variables, we can form an equation containing only the unknown  $c$ , namely,

$$(n/2) / \sum_i n_i / [-1 + \{1 + 4c^2(t_i^2 + 1)\}^{1/2}] = c^2. \quad (1)$$

Here only the positive root of  $\{1 + 4c^2(t_i^2 + 1)\}^{1/2}$  is considered. Its admissibility is discussed in Gerig and Sen (1980) for the two-population case and we assume the same result applies in the present situation.

Equation (1) is of the form  $F(c^2) = c^2$ , and if we let  $f(x) = F(x) - x$ , with  $x = c^2$ , then we can show that  $f'(x) < 0$  about its zero. We can also prove that, if  $t_{(1)} = \min \{s_i/\bar{y}_i, i = 1, 2, \dots, k\}$  and  $t_{(k)} = \max \{s_i/y_i, i = 1, 2, \dots, k\}$ , then  $t_{(1)} < c < t_{(k)}$ . Hence  $t_{(1)}$  and  $t_{(k)}$  can be taken as a lower and an upper bound for  $c$ , respectively. We are now in a position to apply the method of bisection to solve equation (1).

The algorithm for computing the zero of  $f(x)$  is given below. It is also true that  $f(x)$  is negative to the left of its zero and positive on the other side.

(a) Let  $x_l$  and  $x_u$  be the respective bounds for  $c^2$ , and put  $x_m = (x_l + x_u)/2$ .

(b) Check for  $f(x_m) = 0$ , by  $F(x_m)/x_m = 1.0 \pm 10^{-16}$

If this is satisfied then we say that the solution has been achieved.

(c) Otherwise determine whether  $f(x_m)$  is negative or positive. Set  $x_u = x_m$  if  $f(x_m) < 0$ , and set  $x_l = x_m$  if  $f(x_m) > 0$ .

(d) Now put the new  $x_m = (x_l + x_u)/2$ , and repeat (b) and then (c) in a loop until the criterion of convergence above is found.

The number of iterations usually does not exceed 10 and it may be as low as 1. The latest  $x_m$  is taken to be the solution for the square of of c.v., i.e.  $\widehat{c^2} = x_m$ . Using this value we then calculate  $\mu_i$  from

$$\widehat{\mu}_i = \bar{y}_i \left\{ -1 + [1 + 4\widehat{c^2}(t_i^2 + 1)]^{1/2} \right\} / 2\widehat{c^2}, \quad (2)$$

$i = 1, 2, \dots, k$ .

To test the hypothesis about the means, the estimate of the variance-covariance matrix between them is necessary. In the present situation only the asymptotic form is available. This is obtained by inverting the information matrix, whose  $(r, s)$ -th element is  $E_{\theta} (-\delta^2 \ell / \delta \theta_r \delta \theta_s)$ , where  $\ell$  is the log-likelihood of  $\theta$ , given the random samples (Huzurbazar, 1949). The elements of the submatrix,  $V(k \times k)$  say, corresponding to the  $k$  means, are given by

$$v_{rs}(\theta) = \begin{cases} c^2 \mu_r^2 (1 + 2n_r c^2 / n) n_r (2c^2 + 1) & \text{if } r = s, \\ 2\mu_r \mu_s c^4 / n (2c^2 + 1) & \text{if } r \neq s; r, s = 1, 2, \dots, k. \end{cases}$$

At the end of the iterations this submatrix can be estimated easily, that is, by substituting the estimates  $\widehat{\mu}_1, \widehat{\mu}_2, \dots, \widehat{\mu}_k$ , and  $c$  into the elements of the above matrix.

### THE TEST

The hypothesis is of the form  $H_0: C'\underline{\mu} = \gamma$ , where  $C$  is called the test matrix. Since it can be shown that  $C'\widehat{\underline{\mu}} \sim AN(C'\underline{\mu}, C'VC)$ , and that  $V$  is consistent, then by a well known result, it is true under  $H_0$  that

$$L = (C'\hat{\underline{\mu}} - \underline{\gamma})' (C'VC)^{-1} (C'\hat{\underline{\mu}} - \underline{\gamma}) \quad (3)$$

is asymptotically chi-square with  $r$  d.f., where  $r$  is the rank of  $C$ .

Most mathematical or statistical packages have subroutines to help solve the matrix algebra that would lead to the value of  $L$  above, including the rank of  $C$ . This value may be compared with the tabulated chi-square ( $r$  d.f.) and hence all appropriate decisions on  $H_0$  could be made.

The application of the proposed test is demonstrated using the data given by Azen and Reed (1973), concerning the absorbance values of an enzyme, leucine amino peptidase, where three concentration levels are compared. Consider the following hypothesis:

$$\begin{aligned} \mu_1 - 2\mu_2 &= -10 \\ \mu_2 - \mu_3 &= 0. \end{aligned}$$

Then the form for  $H_0: C'\underline{\mu} = \underline{\gamma}$  is given by

$$\begin{bmatrix} 1 & -2 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} -10 \\ 0 \end{bmatrix},$$

and the rank of the test matrix is 2.

In the above problem,  $n_1 = n_2 = n_3 = 19$ . Using the algorithm indicated earlier we find that

$$\hat{\underline{\mu}}_1 = 120.21, \hat{\underline{\mu}}_2 = 70.42, \hat{\underline{\mu}}_3 = 69.73 \text{ and } \hat{c} = 0.035.$$

The estimate of the variance-covariance matrix corresponding to the means is

$$\begin{bmatrix} 0.9352 & & \\ 0.0004 & 0.3210 & \\ 0.0004 & 0.0003 & 0.3147 \end{bmatrix}$$

By (3) it is found that  $L = 19.257$  with 2 d.f. From the chi-square table we conclude that  $H_0$  is rejected at  $\alpha < 0.005$ .

### CONCLUSION

If the model fits the constant c.v. assumption, i.e. the ratios between the variances and the means among the various groups of observations seem to remain homogeneous, then the proposed procedure is a very logical alternative. Although popular statistical packages do not provide computing faci-

lities for this kind of model, the outlined computing algorithm is simple enough to be implemented using a number of routines available in those packages.

The method of bisection used in the estimation is a very efficient procedure. The result is very accurate and is achieved in only a few iterations. The initial statistics involve only the sample means and the sample variances. When the function of the form  $F(x) = x$  has been fixed then the computation is merely a simple routine.

On convergence, the estimate for c.v. is just the square root of the zero of  $f(x) = F(x) - x$ . The estimates for the means follow. The variance-covariance matrix is formed using the formulae provided.

From the null hypothesis the test matrix can be constructed. Finally, the value of the test statistic is calculated and compared with the values of the chi-square from within the package itself, or from the table.

### ACKNOWLEDGEMENT

The author wishes to thank the referees for their useful suggestions and comments.

### REFERENCES

- ABD-RAHMAN, M.N. and GERIG, T.M. (1983): An efficient maximum likelihood solution in normal model having constant but unknown coefficient of variation. *Pertanika*, 6(1):57-62.
- AZEN, S.P. and A.H. REED (1973): Maximum likelihood estimation of correlation between variates having equal coefficients of variation. *Technometrics*, 15(13):457-462.
- GERIG, T.M. and A.R. SEN (1980): MLE in two normal samples with equal but unknown population coefficient of variation. *J. Am. Statist. Assoc.*, 75:704-708.
- GRAYBILL, F.A. (1961): *An Introduction to Linear Statistical Models*, New York: McGraw-Hill, 82-92.
- HUZURBAZAR, V.S. (1949): On a property of distributions admitting sufficient statistics. *Biometrika*, 36:71-74.
- KELLY, L.G. (1967): *Handbook of Numerical Methods and Applications*. California: Addison-Wesley, 86-88.
- STEEL, R.G.D. and J.H. TORRIE (1960): *Principles and Procedures of Statistics*. New York: McGraw-Hill, 99-131.

(Received 28 January, 1987)