COMPARING MEANS OF TWO NON-HOMOGENEOUS NORMAL POPULATIONS

Mohd. Nawi Abd – Rahman
School of Management and Accountancy, UUM

The researchers often avoid the maximum likehood (M.L) procedure that is known to have a number of desirable properties, because the solutions are complicated. However, it can be attractive to many if some efficient algorithm is available. This paper intends to give an alternative approach for testing the means of two normal populations having unequal variances but whose coefficient of variations are homogeneous. The M.L. method is appropriate. The statistic is a funtion of the three M.L. estimates. A simple algorithm which can be implemented on a microcomputer is proposed for calculating the required values. The statistic is shown to be a one-degree chi-square variable. An application on the proposed technique using an experimental data is given.

For comparing the means between two normal populations N (μ_i , σ_i^2), i = 1,2; the model under consideration must satisfy the following assumptions: homogeneity of the error variance, and the independence between the samples. Oftentimes, economics data are dependent on occasions, such as when measurements are collected over an extended span of time, or on locations, like rural and urban classifations. Under such circumstances we are usually interested for the two given conditions, not in detecting wheter the two means differ but rather by how much they differ, or even, by how many times.

Let the sample values of the associated random variables Y_1 and Y_2 be represented by y_{11} , y_{12}, ..., y_{1n_1} and y_{21} ,y_{22} ,...,y_{2n_2} respectively, where n_1 and n_2 may or may not be equal. Then we envisage that most of the values in the first sample are smaller than those in the second set and that the first sample mean will be less than the second's, i. e. , $\bar{y}_1 < \bar{y}_2$, where $\bar{y}_i = n_i^{-1} \sum_{j}^{n} y_{ij}$, i = 1,2. We shall consider only the positive samples. Generalization is achieved by a simple transformation of observations.

Some conditions may influence the values of the observations in a manner that results in more variablility in the case of the second set of values. In other words, the sample variance s_2^2 that is obtained for the second group is larger than s_1^2 , the sample variance calculated for the first group. In this situation the homegeneity of variance assumption for testing the hypothesis may no longer hold. But as we have observed the increase in the variance is indicated by the increase in the mean. This variation may follow a desirable pattern where the ratio of the standard error to the mean is constant. That is, the coefficient of variations (C.V.) remain homogeneous.

The homogeneous C.V. model is now becoming popular but at a very slow phase, perhaps, due to the complexity of its statistics, both in its expression and the associated distribution for use in the hypothesis testing. An assumption of this nature is valid and possibly is more convenient in a number of problems involving economics, psychology and agriculture variables. In one instant, for example, this can be achieved by a proper combination of conditions such as on household expenses (see Theil, 1971, p. 245).

As another example, one may be interested to study the estimated income elasticity, after considering the result displayed on page 111 of Philips (1974), between tea (0.68 ± 0.08) and coffee (1.42 ± 0.20) for the Middle Class group, where the number in each parenthesis indicate the estimate for the mean followed by its standard error. The C.V. for that of tea is approximately equal to that of coffee.

In our model we assume that the two random positive samples are distributed as N (μ_i, $c^2\mu_i{}^2$), where $c = \sigma_i / \mu_i$, $i = 1,2$; is the common C.V. for both populations. Maximum likehood (M.L.) solutions for the three parameters μ_1 , μ_2 and c in the case of equal sample sizes, $n_1 = n_2$ are given by Lohrding (1969). General solutions for the case in which $n_1 \neq n_2$ are indicated by Gerig and Sen (1980). Numerical solution algorithm for the model in its most general form, where the number of groups may exceed two and in different sample sizes, is available from the author (see also Abd Rahman and Gerig 1983). A simpler model assumes that c is known (Azen and Reed 1973).

In this paper M.L solutions are given for the case where $n_1 \neq n_2$ with indication of relative sample sizes to ensure that the M.L. solution always exist. An asymptotic chi-square test is proposed.

THE M.L. SOLUTIONS

Following the derivations indicated in Gerig and Sen (1980) or as a special case shown in Abd – Rahman and Gerig (1983), the three M.L. equations in the unknown μ_1 , μ_2 and c are given as under, dropping the "cap" signs for simplicity.

$$c^2\mu_1^2 + \mu_1\bar{y}_1 - (s_1^2 + \bar{y}_1^2) = 0 \qquad (1)$$

$$c^2\mu_2^2 + \mu_2\bar{y}_2 - (s_2^2 + y_2^2) = 0 \qquad (2)$$

$$c_2\bar{y}_2\mu_1 + n_1\bar{y}_1\mu_2 - n\mu_1\mu_2 = 0 \qquad (3)$$

By expressing c in terms of other variabless in (1) and μ_2 in terms of others in (3), equation (2) can be expressed free of c and μ_2 so that it becomes a quadratic equation of the form a $\mu_2{}^2$ + b μ_2 + d = 0, where

$$a = n(n - n_2)\,\bar{y}_2^2 + n^2 s_2^2$$

$$b = (n_2^2 + n_1 n_2 - 2nn_1)\,\bar{y}_1 \bar{y}_2{}^2 - 2nn_1 \bar{y}_1 s^2$$

$$d = n_1^2 \bar{y}_1^2 s_2^2 - n_2^2 \bar{y}_2^2 s_1^2$$

Note that $s_i{}^2 = \sum^{n_i} (y_{ij} - \bar{y}_i) / n_i$, $i = 1,2$ and $n = n_1 + n_2$.

Now, the expression for $b^2 - 4ad$ can be arranged to contain the sum of squared terms, except for the term

$$4nn_1\,(nn_1 - n_2^2)\,\bar{y}_1^2 \bar{y}_2^2 s_2^2$$

This term may contribute a negative quantity unless the sample sizes are confined to a certain range of values, namely according to the relation

$$n_1 \geq n_2^2 / n. \qquad (4)$$

It is easy to see this. Suppose that $n_1 = 2$, $n_2 = 4$ so that $n = 6$, a case violating relation (4), then for, say, $\bar{y}_1 = 5$, $\bar{y}_2 = 1$, $s_1 = s_2 = 1$, $b^2 - 4ad = -1728 < 0$.

The statistics chosen above, however, are relatively extreme under the present model assumptions. In fact, by further rearrangging the terms in $b^2 - 4ad$ above we would have an expression of the form

$$4nn_1\, n_2^2\, \bar{y}_1^2\, (s_1{}^2 / \bar{y}_1^2 - s_2^2 / \bar{y}_2^2) + \text{“squared terms”}.$$

If the samples are proper then the rations $s_1{}^2 / \bar{y}_1{}^2 \cong s_2^2, / \bar{y}_2{}^2$, and the contribution of the terms above is small and may not cause the discriminant to be negative.

For completeness we further observe that the relation

$$n_2 \geq n_1{}^2 / n \qquad (5)$$

must hold for solutions of μ_2. As a consequence of the restrictions (4) and (5) we have the following. Let $r_i = n_i / n$, = i 1,2. Then these imply $r^2{}_1 - 3r_1 + 1 \leq 0$ and $r_1 + r_2 = 1$, that is $.382 \leq r_1 \leq .618$. By symmetry, $.382 \leq r_2 \leq .618$. This means that both restrictions are satisfield if n_1 and n_2 are chosen so that $.382\,n \leq n_1 \leq .618\,n$ and $.382\,n \leq n_2 \leq .618\,n$, where $n = n_1 + n_2$.

Finanlly we come to the expressions for $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{c}$. As solutions for quadratic equations there seems to exits a pair of values for each the mean estimates. However, it can be deduced that only the solutions with the plus signs in front of the redicals are admissible (see e.g. Gerig and Sen 1980).

The solutions for $\hat{\mu}_1$ and $\hat{\mu}_2$ are

$$\hat{\mu}_1 = \bar{y}_1 \frac{(2nn_1 - n_2^2 - n_1n_2)\,\bar{y}_2^2 + 2nn_1s_2^2)}{2nn_1y_2^2 + 2n^2s_2^2}$$

$$+ \frac{\bar{y}_2\{4nn_1(nn_1^2 - n_2^2)\,\bar{y}_1^2\,s_2^2 + (n_2^2 + n_1n_2 - 2nn_1)^2\bar{y}_1^2\,\bar{y}_2^2 + 4nn_1n_2^2\bar{y}_2^2s_1^2 + 4n^2n_2^2s_1^2s_2^2\}^{1/2}}{2nn_1\bar{y}_1^2 + 2n^2\,s_2^2}$$

and, by symmetry,

$$\hat{\mu}_1 = y_2 \frac{(2nn_2 - n_1^2 - n_1n_2)\bar{y}_1^2 + 2nn_2s_1^2)}{2nn_2\bar{y}_1^2 + 2n^2s_1^2}$$

$$+ \frac{\bar{y}_1\{4nn_2(nn_2^2 - n_1^2)\bar{y}^2s_1^2 + (n_1^2 + n_1n_2 - 2nn_2)^2\bar{y}_1^2\bar{y}_2^2 + 4n_1^2n_2^2\bar{y}_1^2s_2^2 + 4n^2n_1^2s_1^2s_2^2\}^{1/2}}{2nn_2^2\bar{y}^2{}_1 + 2n^2s_1^2}$$

Also,

$$\hat{c} = [(s_1^2 + \bar{y}_1^2) - \mu_1\,\bar{y}_1]/\,\hat{\mu}_1\,,\ \text{or}$$

$$= [(s_1^2 + y_2^2) - \hat{\mu}_2\bar{y}_2]/\hat{\mu}_2 \tag{6}$$

Suppose $n_1 = n_2 = m$, then $n = 2m$, and $m \geq m^2 / 2m = m/2$, i.e., $n_1 \geq n_2{}^2 / n$ and $n_2 \geq n^2{}_1 / n$ in both cases, and thus the solutions for $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{c}$ always exist and are given by, cf: Lohrding (1969),

$$\hat{\mu}_1 = \tfrac{1}{2}\,\bar{y}_1 + \tfrac{1}{2}\,\bar{y}_2\,\{(\bar{y}_1^2 + 2\,s_1^2)/(\bar{y}_2^2 + 2\,s_2^2)\}^{1/2}$$

$$\hat{\mu} = \tfrac{1}{2}\,y_2 + \tfrac{1}{2}\,y_1\,\{(y_2^2 + 2\,s_2^2)/(y_1^2 + 2\,s_1^2)\},^{1/2}$$

and

$$\hat{c} = \frac{[(s_1^2 + \bar{y}_1^2) - [\tfrac{1}{2}\bar{y}_1 + \tfrac{1}{2}\bar{y}_2\{(\bar{y}_1 + 2s_1^2)/(\bar{y}_2^2 + 2s_2^2)\}^{1/2}]]^{1/2}}{\tfrac{1}{2}\bar{y}_1 + \tfrac{1}{2}\bar{y}_2\,\{\bar{y}_1^2 + 2s_1^2)\,/(\bar{y}_2^2 + 2s_2^2)\}^{1/2}}$$

say.

For testing the hypothesis about the means we shall need the asymptotic variance-corariance matrix for $\mu = (\mu_1,\mu_2)$. This is given in Abd-Rahman and Gerig (1983) as follows:

$$\begin{bmatrix} \dfrac{c^2\mu_1^2(1 + 2\,n_1\,c^2 / n)}{n_1\,(2c^2 + 1)} & \dfrac{n\,(2\,c^2 + 1)}{2\,\mu_1\mu_2\,c^4} \\ & \dfrac{c^2\mu^2\,(1 + 2\,n_2\,c^2 / n)}{n_2\,(2\,c^2 + 1)} \end{bmatrix} \quad (7)$$

Its estimate is obtained by substituting $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{c}$ in (7) above. Note that it is symmetric.

COMPUTING ALGORITHM

The algorithm for computing the M.L. solution in the general model, $y_{ij} \sim N(\mu_i, c^2\mu^2_i)$, $j = 1,2,.., n_i$ and $i = 1,2,..., k$; using the method of bisection (see e.g. Kelly 1967), has been developed by the author. Since this is not readily available the following algorithm may be applied for the model under consideration.

For $\hat{\mu}_1$ we calculate a, b and d of the quadratic equation a $\hat{\mu}^2_1$ + b$\hat{\mu}_1$ + = 0. Simple programs can easily be obtained or developed for an equation of this form, noting that the solution is one that has the plus sign in front of the redicals. To find $\hat{\mu}_2$, the corresponding symmetric values of a, b and d are calculated and the solution is arrived by the same routine as above.

We the use either of equations (6) to obtain $\hat{c}$. With all these values the estimate for the asytmptotic var-cov (μ'_1, μ_2) can be formed.

TESTING OF HYPOTHESIS ABOUT THE MEANS

Our problem is a special case of the following well known result. If $\Sigma\mu$ denotes the asymptotic variance-covariance matrix for $\underline{\mu} = (\hat{\mu}_1, \hat{\mu}_2)$, then for the testing a hypothesis of the form $H_0 : C'\mu = \underline{\Upsilon}$, a test statistic for H_0 can be obtained from the consideration that

$$C'\,\hat{\underline{\mu}} \sim A\,N_r\,(C'\underline{\mu}\,,\; C'\,\Sigma\mu\,C)$$

where rank (C) = r, and AN_r indicates the asymptotic normal distribution. Since $\hat{\Sigma}\mu$, the estimate for $\Sigma\mu$, is consistent, then for

$$L_r = (C'\underline{\hat{\mu}} - \underline{\gamma})'(C'\hat{\Sigma}\mu C)^{-1}(C'\underline{\mu} - \underline{\gamma}),$$

$L_r \sim A\chi^2(r)$, i.e., L_r is a value of a chi-square random variable with r.d.f. (asymptotically).

For testing a hypothesis of the form $H_0 : \mu - \alpha\mu_2 = \delta_0$, the matrix C is given by the vector $C' = (1, -\alpha)$ and γ is only the scaler δ_0. $\hat{\Sigma}\mu$ is obtained from (7). The rank (C) is 1. For example, if $H_0 : \mu_1 - \mu_2 = \delta$ then $C' = (1, -1)$, and if H_0: $\mu_1 - 2\mu_2 = \delta_1$, that is, the null hypothesis that μ_1 differs from twice that of μ_2 by δ_1 . then $C' = (1, -2)$, and so on. Then we can show that

Let $\Delta = | \hat{\mu}_1 - \alpha\hat{\mu}_2 - \delta_0 |$ and the estimates of the elements in (7) be e_{11} ,. e_{12}, e_{21} and e_{22} where $e_{12} = e_{21}$. Then we can show that

$$L_1 = \Delta^2 / (e_{11} + \alpha e_{22} - (1 + \alpha) e_{12}).$$

Since each of the quantities in the above expression is known then L_1 can be computed and compared with the $\chi^2(1)$ critical regions.

As an example of application we analyse the data on the "Absorbance values of a standard substances compared to the second level of enzyme concentration", as given in Azen and Reed (1973). Suppose it is believed that the values for the standard solution is twice as much as the latter, then we set $H_0 : \mu_1 - 2\mu_2 = \delta_0 = 0$. Here it is found that $\hat{\mu}_1 = 120.21$, $\hat{\mu}_2 = 69.73$ and $\hat{c} = 0.0351$. The asymptotic var – cov $(\hat{\mu}_1, \hat{\mu})$ is given by the elements $e_{11} = 0.9352$, $e_{22} = 0.3147$ and $e_{12} = e_{21} = 0.004$. Hence : $L_1 = 19.258 / 1.563 = 12.321$, and H_0 is rejected at less than 0.5%.

CONCLUSION

A test for comparing two means of the normal distributions having equal coefficient of variations has been developed for large samples. The test statistic is show to have asymptitically a chi-square distribution with one degree of freedom. Such a test is more powerful than other classical procedures. This is also true for small samples (see, e.g. Lohrding, 1969).

The values that appear in the statistic are obtained through the M.L. procedures. If the model departs to a certain degree from the homogeneous c.v. assumption then there is a risk that the solution does not exist, unless either of the samples sizes is within 39 to 61 percent of the total size.

The algorithm for computation is easy to implement even on a microcomputer. We first calculate the values of $\bar{y}_1$, $\bar{y}_2$, s_1^2 and s_2^2. The coefficients and the constant of the two quadratic equations are functions of these values. A simple program leads to the solution for $\hat{\mu}_1$ and then by the same routine the solution for $\hat{\mu}_2$ follows.

Finally, the solution for $\hat{c}$ is calculated. Immediately, we obtain the three elements of the symmetric asymptotic variance-convariance matrix for $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2)$, and hence the chi-square value for a given null hypothesis. The example demonstrates an application for a real data.

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 Reed, A.H. (1973): Maximum likelihood estimation of correlation between variates having equal coefficients of variation. *Technometrics,* 15(13) :457–462.

Gerig, T.M. and Sen, A.R. (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.

Kelly, L.G. (1967): *Handbook of Numerial Methods and Applications.* California: Addison-Wesley, 86–88.

Lohrding, R.K. (1969): A test of equality of two normal population means assuming homogeneous coefficient of variation. *Ann. Math. Statist.,* 40(3): 1374–1385.

Phlips, L. (1974): *Applied Consumption Analysis.* Oxford: North-Holland.

Theil, H. (1971): *Principles of Econometrics.* New York: John Wiley.