STATISTICAL RESEARCH REPORT No 5
Institute of Mathematics 1971
University of Oslo

JOINT CONFIDENCE INTERVALS FOR ALL LINEAR FUNCTIONS OF THE MEANS IN THE ONE-WAY LAYOUT WITH UNKNOWN GROUP VARIANCES

by

Emil Spjøtvoll

ABSTRACT

The one-way layout in the analysis of variance with unknown group variances is considered. A family of joint confidence intervals for all linear functions in the means with the property that the probability is $1 - \alpha$ that all confidence intervals covers the true values of the linear functions is found. Each confidence interval is natural in the sense that for a given linear function it is equal to an estimate of this function plus and minus a constant times an estimate of the variance of the estimate. Hence the results are analogous to Scheffé's S-method of multiple comparison.

1. STATEMENT OF THE PROBLEM AND THE METHOD

Consider the one-way layout with unequal group variances in the analysis of variance. Let the random variables y_{ij} , $j = 1,\ldots,n_i$, $i = 1,\ldots,k$ be independent with

$$E\, y_{ij} = \mu_i \ , \ \mathrm{Var}\, y_{ij} = \sigma_i^2 \ .$$

The means and the variances are all unknown. The problem is that of finding joint confidence intervals for all linear functions of the μ_i,

$$\psi = \sum_{i=1}^{k} c_i \mu_i \ ,$$

where the c_i are known constants. A solution to this problem in the case when all σ_i are equal was given by Scheffé (1953). For other solutions see, e.g., Miller (1966) and Scheffé (1959). We shall now derive a solution which takes care of the possibility that the σ_i may be unequal.

A natural estimator of ψ is

$$\hat{\psi} = \sum_{i=1}^{k} c_i y_{i\cdot} \ ,$$

where $y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}/n_i$. The variance of $\hat{\psi}$ is

$$\sigma_{\hat{\psi}}^2 = \sum_{i=1}^{k} \frac{c_i^2 \sigma_i^2}{n_i} \ .$$

An estimate of this variance is

$$\hat{\sigma}_{\hat{\psi}}^2 = \sum_{i=1}^{k} \frac{c_i^2 s_i^2}{n_i} \ ,$$

where $s_i^2 = \sum_{j=1}^{n_i} (y_{ij} - y_{i\cdot})^2/(n_i - 1)$.

We shall prove that there exists a constant A such that the probability is $1-\alpha$ that the values ψ of <u>all</u> the linear functions satisfy

$$\hat{\psi} - A\hat{\sigma}_{\hat{\psi}} \leq \psi \leq \hat{\psi} + A\hat{\sigma}_{\hat{\psi}} \tag{1}$$

The constant A will depend upon α and the n_i, but not upon unknown parameters. It is determined by the following. Let $z(n_1-1),\ldots,z(n_k-1)$ denote k independent F-distributed random variables all with one degree of freedom in the numerator and $n_1-1,\ldots,n_q-1$ degrees of freedom in the denominator, respectively. Then A is determined by

$$P\left[\sum_{i=1}^{k} z(n_i-1) \leq A^2\right] = 1-\alpha .$$

Since the exact distribution of $\sum_{i=1}^{k} z(n_i-1)$ is difficult to calculate, a simple approximation is proposed in Section 3. The approximate value of A is given by

$$A^2 \approx a\, F_{\alpha}(k,b) \tag{2}$$

where

$$b = \frac{(k-2)\left(\sum_{i=1}^{k} \frac{n_i-1}{n_i-3}\right)^2 + 4k \sum_{i=1}^{k} \frac{(n_i-1)^2(n_i-2)}{(n_i-3)^2(n_i-5)}}{k \sum_{i=1}^{k} \frac{(n_i-1)^2(n_i-2)}{(n_i-3)^2(n_i-5)} - \left(\sum_{i=1}^{k} \frac{n_i-1}{n_i-3}\right)^2} \tag{3}$$

and

$$a = \left(1 - \frac{2}{b}\right) \sum_{i=1}^{k} \frac{n_i-1}{n_i-3}, \tag{4}$$

and $F_{\alpha}(k,b)$ is the upper α-point of the F-distribution with k and b degrees of freedom.

2. PROOF OF THE METHOD

To proof the main result of the previous section we need the following lemma.

LEMMA. Let $d_1,\dots,d_k$, $z_1,\dots,z_k$ and c (>0) be given real numbers. Then

$$\sum_{i=1}^{k} d_i z_i^2 \leq c^2 \tag{5}$$

if and only if

$$\left|\sum_{i=1}^{k} c_i z_i\right| \leq c \left(\sum_{i=1}^{k} \frac{c_i^2}{d_i}\right)^{1/2} \tag{6}$$

for all real numbers $c_1,\dots,c_k$.

Proof. If (6) holds, it follows by using Schwarz's inequality that

$$\left(\sum_{i=1}^{k} c_i z_i\right)^2 = \left(\sum_{i=1}^{k} \frac{c_i}{d_i^{1/2}} d_i^{1/2} z_i\right)^2 \leq \left(\sum_{i=1}^{k} \frac{c_i^2}{d_i}\right)\left(\sum_{i=1}^{k} d_i z_i^2\right)$$

$$\leq c^2\left(\sum_{i=1}^{k} \frac{c_i^2}{d_i}\right) ,$$

from which we obtain (6). Conversely, if (6) holds for all c_i , it holds in particular for $c_i = d_i z_i$, from which we get (5).

Using the Lemma with $d_i = n_i/s_i^2$, $z_i = y_{i\cdot} - \mu_i$ and $c = A$ we obtain the following theorem.

THEOREM. $P\left[\sum_{i=1}^{k} \frac{n_i(y_{i\cdot}-\mu_i)^2}{s_i^2} \leq A^2\right] =$

$$P\Bigg[\sum_{i=1}^{k} c_i y_{i\cdot} - A\left(\sum_{i=1}^{k} \frac{c_i^2 s_i^2}{n_i}\right)^{1/2} \leq \sum_{i=1}^{k} c_i \mu_i \leq \sum_{i=1}^{k} c_i y_{i\cdot}$$

$$+ A\left(\sum_{i=1}^{k} \frac{c_i^2 s_i^2}{n_i}\right)^{1/2} \quad \text{for all } c_1,\dots,c_k\Bigg] .$$

Since the distribution of $\sum_{i=1}^{k} \frac{n_i(y_i-u_i)^2}{s_i^2}$ is the same as the distribution of $\sum_{i=1}^{k} z(n_i-1)$, the statement above (1) in Section 1 is true.

3. THE APPROXIMATION

We will approximate the distribution of the random variable

$$v = \sum_{i=1}^{k} z(n_i-1)$$

by the distribution of $aF(k,b)$, where $F(k,b)$ is an F-distributed random variable with k and b degrees of freedom. The constants a and b are determined so that the first two cumulants of v and $aF(k,b)$ are equal. The approximation gives the exact distribution when all n_i tend to infinity. Furthermore, Morrison (1971) has compared the exact and approximate distribution in the case $k = 2$, $n_1 = n_2$, and shown that the approximation is excellent.

The cumulants of v are

$$\kappa_1 = \sum_{i=1}^{k} \frac{n_i-1}{n_i-3}$$

$$\kappa_2 = \sum_{i=1}^{k} \frac{2(n_i-1)^2(n_i-2)}{(n_i-3)^2(n_i-5)}$$

while those of $aF(k,b)$ are

$$\kappa_1^* = \frac{ab}{b-2}$$

$$\kappa_2^* = \frac{2a^2(k+b-2)b^2}{k(b-2)^2(b-4)} .$$

Solving a and b from the equations $\kappa_i = \kappa_i^*$, $i = 1,2$, we get the solutions (3) and (4).

4. AN EXAMPLE

We shall use the data in Pearson and Hartley (1958, p. 27). Here $k = 2$, $n_1 = 10$, $n_2 = 15$, $y_{1.} = 73.4$, $y_{2.} = 47.1$, $s_1^2 = 51$, $s_2^2 = 141$. We shall suppose that we want to find confidence intervals for the difference $\mu_1 - \mu_2$, as well as for μ_1 and μ_2 separately. We find $a = 2.06$, $b = 12.51$ and $A = 2.31$. Note that the values obtained for a and b seem very reasonable. Using $\alpha = .10$ we find that the confidence interval for $\mu_1 - \mu_2$ is $[17.5 , 35.1]$, for μ_1 it is $[68.2 , 78.6]$ and for μ_2 it is $[40.0 , 54.2]$.

The 90 percent confidence interval for $\mu_1 - \mu_2$ obtained by Pearson and Hartley using a method due to Welch (1947) is $[19.8 , 32.8]$. As should be expected, since this method is aimed only at that difference $\mu_1 - \mu_2$, this interval is smaller than the one obtained by the method of Section 1.

If we actually wanted 3 confidence intervals and did not use the simultaneous method of Section 1, we could still do this and still have 90 percent probability that all three intervals were correct by increasing the confidence coefficient of each interval to 96.67 . Doing this we find that the confidence intervals for μ_1 and μ_2 , using ordinary t-intervals, are $[67.8 , 79.0]$ and $[39.7 , 54.5]$, respectively. The interval for $\alpha_1 - \alpha_2$ becomes $[17.7 , 34.9]$ (to find this I had to interpolate in the available tables). It is seen that two intervals are slightly wider while one is slightly narrower than the intervals obtained by the method of Section 1.

REFERENCES

MILLER, R.G. Jr. (1966). Simultaneous Statistical Inference. New York : McGraw-Hill.

MORRISON, D.F. (1971). On the distribution of linear functions of independent F variates. J. Amer. Statist. Ass. 66, 383-85.

PEARSON, E.S. & HARTLEY, H.O. (1958). Biometrika Tables for Statisticians, 1. Cambridge University Press.

SCHEFFÉ, H. (1953). A method for judging all contrasts in the analysis of variance. Biometrika, 40, 87-104.

SCHEFFÉ, H. (1959). The Analysis of Variance. New York : John Wiley and Sons.

WELCH, B.L. (1947). The generalization of 'Students' problem when several different population variances are involved. Biometrika, 34, 28-35.