AN APPROXIMATE DISTRIBUTION OF ESTIMATES OF VARIANCE COMPONENTS

F. E. SATTERTHWAITE

General Electric Company, Ft. Wayne, Indiana

1. INTRODUCTION

In many problems, only simple mean square statistics are required to estimate whatever variances are involved. If the underlying populations are normal, these mean squares are distributed as is chi-square and may therefore be used in the standard chi-square, Student's t and Fisher's z tests. Frequently, however, the variances must be estimated by linear combinations of mean squares. Crump (1) has recently discussed a problem of this type, based on the following data:

ANALYSIS OF VARIANCE OF TOTAL EGG PRODUCTION OF 12 FEMALES *(D. melanogaster)* FROM 25 RACES IN 4 EXPERIMENTS

Source of Variation	*Degrees of Freedom*	*Mean Square*	*Average Value of the Mean Square*
Experiments	3	MS_e = 46,659	$\sigma_z^2 + 12\,\sigma_{er}^2 + 300\,\sigma_e^2$
Races	24	MS_r = 3,243	$\sigma_z^2 + 12\,\sigma_{er}^2 + 4\,\sigma_r^2$
E × R	72	MS_{er} = 459	$\sigma_z^2 + 12\,\sigma_{er}^2$
Within Subclasses	1,100	MS_z = 231	σ_z^2

The variance of the mean of the i th race is shown in his paper to be estimated by

$$(1) \qquad V_{.i.} = \frac{1}{e}\left(\hat{\sigma}_e^2 + \hat{\sigma}_{er}^2\right) + \frac{1}{en}\left(\sigma_z^2\right)$$

$$= \frac{1}{e}\left\{\frac{MS_e - MS_{er}}{300} + \frac{MS_{er} - MS_z}{12}\right\} + \frac{1}{en}\left(MS_z\right)$$

$$(2) \qquad = \frac{1}{e}\left\{\frac{MS_e}{300} + \frac{24\,MS_{er}}{300}\right\} + \left(\frac{1}{n} - \frac{1}{12}\right)MS_z\Bigg\}$$

where e is the number of experiments and n is the number of females in each experiment. Variance estimates such as (2) have been called *complex estimates* (2). Thus a complex estimate of variance is a linear function of *independent* mean squares.

It is stated in (1) that "increasing the number of females indefinitely still leaves us with

$$(3) \qquad V(\bar{x}_{.i.}) = \frac{MS_e + 24\,MS_{er} - 25\,MS_z}{300\,e} = \frac{173}{e}\,."$$

Conclusions are then reached without analysis of the sampling errors involved. Now the standard deviation of $V(\bar{x}_{.i.})$ is very large

$$(4)\qquad \{\hat{V}[\hat{V}(\bar{x}_{.i.})]\}^{\frac{1}{2}} = \frac{\sqrt{2}}{300\,e}\left[\frac{(MS_e)^2}{5} + \frac{(24\,MS_{er})^2}{74} + \frac{(25\,MS_z)^2}{1102}\right]^{\frac{1}{2}}$$
$$= 0.57\,\hat{V}(\bar{x}_{.i.})\,;$$

and further analysis leading to confidence limits for $V(\bar{x}_{.i.})$ should be helpful in choosing a course of action.

The writer has studied the distribution of complex estimates of variance in a paper (2) in *Psychometrika*. Since this paper may not be readily available to biometricians, the principal results are outlined below and a few applications are given.

2. The Distribution of Complex Estimates of Variance

The exact distribution of a complex estimate of variance is too involved for everyday use. It is therefore proposed to use, as an approximation to the exact distribution, a chi-square distribution in which the number of degrees of freedom is chosen so as to provide good agreement between the two. This is accomplished by arranging that the approximating chi-square have a variance equal to that of the exact distribution. If MS_1, MS_2, . . . are independent mean squares with r_1, r_2, . . . degrees of freedom and

$$(5)\qquad \hat{V}_s = a_1(MS_1) + a_2(MS_2) + \ldots$$

is a complex estimate of variance based on them, the number of degrees of freedom of the approximating chi-square is found to be given by

$$(6)\qquad r_s = \frac{[a_1E(MS_1) + a_2E(MS_2) + \ldots]^2}{\dfrac{[a_1E(MS_1)]^2}{r_1} + \dfrac{[a_2E(MS_2)]^2}{r_2} + \ldots}$$

where $E(\ \ )$ denotes mean or expected values.

In practice, the expected values of the independent mean squares will not be known. The observed values will usually be substituted in (6), giving, as an estimate of r_s,

$$(7)\qquad \hat{r}_s = \frac{[a_1(MS_1) + a_2(MS_2) + \ldots]^2}{\dfrac{[a_1(MS_1)]^2}{r_1} + \dfrac{[a_2(MS_2)]^2}{r_2} + \ldots}.$$

An approximation of this type for a slightly simpler problem was first suggested by H. Fairfield Smith (3). In his problem, there were only two mean squares and $a_1 = a_2 = 1$. This approximation does not support the use of $r + 2$ in place of r as a correction for bias [(1) formula 3].

The writer has checked the accuracy of the suggested approximations by calculating the exact distribution for a number of special cases. Typical results are as follows:

r_1	r_2	$\frac{E(MS_1)}{E(MS_2)}$	r_s	$\chi^2(95\%)$		$\chi^2(99.9\%)$	
				exact	approx.	exact	approx.
4	2	4	100/33	7.9	8.0	16.2	17.3
8	4	1	32/ 3	19.4	19.5	30.5	31.0
6	4	2	54/ 7	15.1	15.3	26.0	27.2
20	4	2	180/21	16.2	17.0	27.7	29.0
4	2	1	16/ 3	11.5	11.7	21.3	22.3

The above discrepancies between the exact and the approximate chi-squares, even for the extreme 99.9 percent case, are very small compared with their sampling errors. Thus it appears that the approximation may be used with confidence. Furthermore, we know from general reasoning that if r_s is large, both the approximate and the exact distributions approach the same normal distribution; if r_s is small, the sampling errors in the chi-squares are large and refinement is superfluous.

Some care must be taken in the cases where one or more of the a's in (5) are negative. If it is possible for $\hat{V}_s$ to be negative with a fairly large probability, the approximate distribution will become rather poor since it can not allow negative estimates. However, here again the sampling errors in $\hat{V}_s$ will be quite large compared with its expected value so that only the sketchiest of conclusions can be drawn in any case.

3. Further Analysis of Crump's Example

The distribution of Crump's estimate of the residual variance of the race means,

$$\hat{V}(\bar{x}_{\cdot i \cdot}) = \frac{(MS_e) + 24(MS_{er}) - 25(MS_z)}{300e} \tag{3}$$

can now be approximated. Thus

$$\hat{V}(x_{\cdot i \cdot}) = \frac{1}{e}\left[\frac{46{,}659}{300} + \frac{(24)(459)}{300} - \frac{(25)(231)}{300}\right] \tag{8}$$

$$= \frac{1}{e}[155 + 37 - 19] = \frac{173}{e}.$$

From (7) we have

$$\hat{r}_s = \frac{[155 + 37 - 19]^2}{\frac{(155)^2}{3} + \frac{(37)^2}{72} + \frac{(19)^2}{1{,}100}} \tag{9}$$

$$= \frac{29{,}929}{8{,}008 + 19 + 1} = 3.7$$

From chi-square tables interpolated for 3.7 d.f. at the 5 percent and 95 percent points we find that, with a high degree of probability,

$$0.60 < \frac{3.7\,\hat{V}}{V} < 9.0 \tag{10}$$

or

$$\frac{(3.7)(173)}{(9.0)e} < V < \frac{(3.7)(173)}{(0.60)e}, \tag{11}$$

$$\frac{71}{e} < V < \frac{1{,}067}{e},$$

Thus if it were necessary to reduce V to 9 and if time were important so that a second series of experiments could not be made, we should run

$$e = \frac{1{,}067}{9} = 119 \tag{12}$$

experiments in the first series for confidence that V would be properly reduced. On the other hand, if the experiments were expensive and time not important, we might run

$$e = \frac{(3.7)(173)}{(5.6)(9)} = 13 \tag{13}$$

experiments and then get a more accurate estimate of V to determine how many additional experiments should be run (5.6 obtained from the 20 percent point for chi-square, 3.7 d.f.).

4. Difference of Means

The usual estimate of variance used in Student t tests for the difference of two means is

$$V_t = \left[\frac{r_1(MS_1) + r_2(MS_2)}{r_1 + r_2}\right]\left[\frac{1}{r_1+1} + \frac{1}{r_2+1}\right] \tag{14}$$

with

$$r_t = r_1 + r_2 \tag{15}$$

degrees of freedom. This assumes that both populations have the same variance. Seldom do we have positive evidence that this is so and often we have evidence that the variances are different. For example, $F = (MS_1)/(MS_2)$ may be significant. Note that a non-significant F is not evidence that the variances are equal, especially if one of the MS's has a small number of degrees of freedom.

The assumption of equal variances can be avoided by use of a complex estimate of variance,

$$\hat{V}_s = \frac{MS_1}{r_1+1} + \frac{MS_2}{r_2+1} \tag{16}$$

with

(17) $$\hat{r}_s = \frac{\{[MS_1/r_1+1)] + [MS_2/(r_2+1)]\}^2}{\frac{[MS_1/(r_1+1)]^2}{r_1} + \frac{[MS_2/(r_2+1)]^2}{r_2}}$$

degrees of freedom.

For example, consider the numerical case:

$$MS_1 = 100,\ r_1 = 99,$$
$$MS_2 = 90,\ r_2 = 9.$$

By the standard analysis one would obtain

(18) $$\hat{V}_t = \left[\frac{(99)(100) + (9)(90)}{108}\right]\left[\frac{1}{100} + \frac{1}{10}\right] = 10.9$$
$$r_t = 99 + 9 = 108$$

The complex estimate gives

(19) $$\hat{r}_s = \frac{100}{100} + \frac{90}{10} = 10.0,$$
$$\hat{r}_s = \frac{(1+9)^2}{\frac{1^2}{99} + \frac{9^2}{9}} = 11.1.$$

One will sometimes reach different conclusions with 108 degrees of freedom from those he will reach with 11 degrees of freedom.

If from general reasoning or other *a priori* considerations it is believed that both MS_1 and MS_2 are independent estimates of the same variance, then the use of 108 degrees of freedom is justified. On the other hand, if the given data are the entire admissible knowledge, then the use of more than 11 degrees of freedom is not valid.

5. Conclusion

In many practical problems the most efficient estimate of variance available is a linear function of two or more independent mean-squares. Usually the exact distribution of such estimates is too complicated for practical use. A satisfactory approximation can be based on the chi-square distribution with the number of degrees of freedom determined by (7).

Many problems, such as the difference of means, can be more conservatively analyzed by use of complex estimates of variance. Assumptions regarding homogeneity of variance can then be avoided.

References

(1) Crump, S. Lee. The estimation of variance components in the analysis of variance. *Biometrics Bulletin* 2 :1 :7–11. February 1946.

(2) Satterthwaite, Franklin E. Synthesis of variance. *Psychometrika* 6 :309–316. October 1941.

(3) Smith, H. Fairfield. The problem of comparing the results of two experiments with unequal errors. *Journal of the Council of Scientific and Industrial Research* 9 :211–212. August 1936.