December 1964

THE VARIANCE OF INTRACLASS CORRELATION INVOLVING GROUPS WITH ONE OBSERVATION

L.A. Swiger
Unviersity of Nebraska-Lincoln

W.R. Harvey
University of Nebraska-Lincoln

D.E. Everson
University of Nebraska-Lincoln

K.E. Gregory
University of Nebraska-Lincoln

THE VARIANCE OF INTRACLASS CORRELATION INVOLVING GROUPS WITH ONE OBSERVATION

L. A. SWIGER, W. R. HARVEY[1], D. O. EVERSON[2] AND K. E. GREGORY[3]

University of Nebraska and United States Department of Agriculture

SUMMARY

An approximate formula is derived for the variance of intraclass correlation when unequal numbers of observations per group occur. The effect on the variance of t of adding groups with single observations is examined using the formula and results obtained by empirically generating data on a computer. The empirical results indicate that the approximate formula is satisfactory over the range of numbers used.

Adding a group with fewer than the average number of observations per group tends to reduce Vt by increasing the degrees of freedom for groups by one, but tends to increase Vt by decreasing the average precision of estimating group means. The net effect can be either negative or positive, depending on t, s and the n_i's. Robertson [1962] pointed out that, when the ratio of the between group mean square to the within group mean square is small, exclusion of groups below half the average size will reduce the variance of the between group component. He further suggested a method for combining estimates of the between group component when n is highly variable.

Results using the formula show that the point where efficiency is lost when a group of size one is added is primarily a function of the number per group, and is affected very little by the number of groups. The value of n where groups of size one should be excluded is shown graphically for varying levels of t. Increases in Vt are demonstrated using the empirical data. The empirical results suggest that the increase in Vt may be even larger than the formula indicates, especially for large values of t. Only the addition of groups of size one is studied. Adding small groups larger than one would also tend to increase Vt when n and t are small.

INTRODUCTION

Intraclass correlation is frequently used in the field of population genetics where observations are assigned to groups on the basis of some genetic relationship. An example is estimating heritability from a paternal half-sib analysis. Often a sire is represented by only one progeny. Including such an observation in an analysis adds one degree of freedom for estimating the mean square for sires, and leaves the mean square for half-sibs unchanged. The gain or loss in precision resulting

[1]Biometrical Services, ARS, USDA, Beltsville, Md.

[2]Present address: Agricultural Experiment Station, University of Idaho, Moscow, U.S.A.

[3]Beef Cattle Research Branch, ARS, AHRD, USDA, Lincoln, Nebraska.

from adding information from groups with only one observation is examined in this paper by using an approximate formula for the variance of the intraclass correlation and by empirical results.

RESULTS FROM APPROXIMATE FORMULA
(Swiger and Gregory)

In order to evaluate the change in the variance of the intraclass correlation when a group with one observation is added, it was first necessary to develop the formula for this variance with unequal numbers of observations per group. The analysis of variance is shown in Table 1, where

N = total number of observations
s = number of groups

$$k = \frac{1}{s-1}\left(N - \frac{\sum n_i^2}{N}\right)$$

n_i = number of observations in the ith group.

TABLE 1
ANALYSIS OF VARIANCE

Source	Degrees of Freedom	Sums of Squares	Expected Sums of Squares
Groups	$s-1$	S	$(s-1)(\sigma_e^2 + k\sigma_g^2)$
Observations/Groups	$N-s$	E	$(N-s)\,\sigma_e^2$

Then the computed intraclass correlation is

$$t = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_e^2 + \hat{\sigma}_g^2} = \frac{a_1E + a_2S}{a_3E + a_4S}$$

where

$$a_1 = -\frac{1}{(N-s)k}, \quad a_2 = \frac{1}{(s-1)k}, \quad a_3 = \frac{(k-1)}{(N-s)k} \text{ and } a_4 = \frac{1}{(s-1)k}.$$

The variance of t may be estimated using the approximate formula for the variance of a ratio. Letting V stand for variance and COV for covariance.

$$V(y/x) \simeq (\mu_y/\mu_x)^2(\sigma_y^2/\mu_y^2 + \sigma_x^2/\mu_x^2 - 2\,\mathrm{COV}\ y,\ x/\mu_y\mu_x). \qquad (1)$$

Assuming that mean squares are distributed as multiples of χ^2 with d degrees of freedom

$$\begin{aligned} &V \text{ (Mean Square)} = 2\sigma^4/d \\ &V \text{ (Mean of Squares)} = 2\sigma^4 d \\ &\text{Estimated } V \text{ (Sum of Squares)} = 2 \text{ (Mean Square)}^2 d \\ &\qquad\qquad\qquad\qquad\quad = 2 \text{ (Sum of squares)}^2/d. \end{aligned} \tag{2}$$

Crump [1947] has shown that while (2) holds for the within group sum of squares with unequal numbers, the variance of the group sum of squares is a complex function depending on the variability of the n_i . Using the approximate variance of the group sum of squares computed from (2) will have little effect on the results given here, since equal numbers of observations per group are assumed prior to adding the group with a single observation.

The necessary quantities are estimated as follows:

$$\hat{\mu}_y = a_1E + a_2S,$$

$$\hat{\mu}_x = a_3E + a_4S,$$

$$\hat{\sigma}_y^2 = a_1^2 2E^2/(N - s) + a_2^2 2S^2/(s - 1) \text{ (since COV } E, S = 0),$$

$$\hat{\sigma}_x^2 = a_3^2 2E^2/(N - s) + a_4^2 2S^2/(s - 1)$$

$$\text{COV } y, x = a_1a_3 2E^2/(N - s) + a_2a_4 2S^2/(s - 1).$$

Then the approximate variance of the intraclass correlation may be obtained by substituting these quantities into (1). Hartley [1954] used this approach to estimate the variance of a heritability estimate computed from the more complex analysis involving sires, dams, and full-sibs. For the analysis in Table 1, substitution in (1) yields

$$\left[\frac{a_1E + a_2S}{a_3E + a_4S}\right]^2 \left\{ \frac{a_1^2 2E^2/(N - s) + a_2^2 2S^2/(s - 1)}{(a_1E + a_2S)^2} + \frac{a_3^2 2E^2/(N - s) + a_4^2 2S^2/(s - 1)}{(a_3E + a_4S)^2} - \frac{2[a_1a_3 2E^2/(N - s) + a_2a_4 2S^2/(s - 1)]}{(a_1E + a_2S)(a_3E + a_4S)} \right\},$$

which reduces to

$$\left[\frac{a_1E + a_2S}{a_3E + a_4S}\right]^2 \left[\frac{2(N - 1)E^2S^2(a_1a_4 - a_2a_3)^2}{(a_1E + a_2S)^2(a_3E + a_4S)^2(N - s)(s - 1)} \right]. \tag{3}$$

Since $(a_1a_4 - a_2a_3)^2$ reduces to $1/k^2(N - s)^2(s - 1)^2$, and sums of squares can be replaced by their variance-component expectations, formula (3) can be rewritten as

$$Vt \simeq \frac{2(N - 1)(1 - t)^2[1 + (k - 1)t]^2}{k^2(N - s)(s - 1)}. \tag{4}$$

It is readily observed that, by assuming equal numbers of observation per group ($k = n$, $N = ns$), the expression reduces to Fisher's [1954] formula for the variance of an intraclass correlation if a factor $(N - 1)/N$ is removed. This last term rapidly approaches unity as the total number of observations increases. If the groups are sets of paternal half-sibs and mating is at random, under certain assumptions $4t$ estimates heritability and 16 Vt estimates the variance of heritability.

The change (Δ) in the variance of t when a group with a single observation is added may be examined to determine the gain or loss of precision. Letting ($'$) refer to the estimate including the added datum,

$$\Delta Vt = Vt - Vt'$$

which equals

$$\frac{2(1 - t)^2}{k^2k'^2(N - s)s(s - 1)} \cdot \{k'^2(N - 1)s[1 + (k - 1)t]^2 - k^2(s - 1)N[1 + (k' - 1)t]^2\}.$$

The last term determines the sign. The change in the variance of t may be evaluated allowing t, s and k to vary one at a time. The point where the change in the variance changes from negative to positive governs the addition of the datum in question. The problem may be examined rather easily if equal numbers are assumed prior to the addition of the single observation, so that $\sum n_i^2$ is not an additional variable. Then

$$k' = n - n(n - 1)/(ns + 1).$$

The change in the variance of t was evaluated, letting t vary from .01 to .99, k from 2 to 100 and s from 10 to 1000. Figure 1 is a graph of ΔVt for $t = .10$. The number of groups has little effect on the value of k where the sign of ΔVt changes, especially for large values of t. When s ranges from 10 to 1000, the value of k at which ΔVt changed sign rounds to the same whole number until t gets as small as .05. The relationship between t and the value of k when ΔVt changes sign is shown in Figure 2. The values graphed are for $s = 1000$: however, as pointed out above, changes in s have little effect on these values. The number per group is 101 when $t = .01$, and falls below 2 at about $t = .35$.

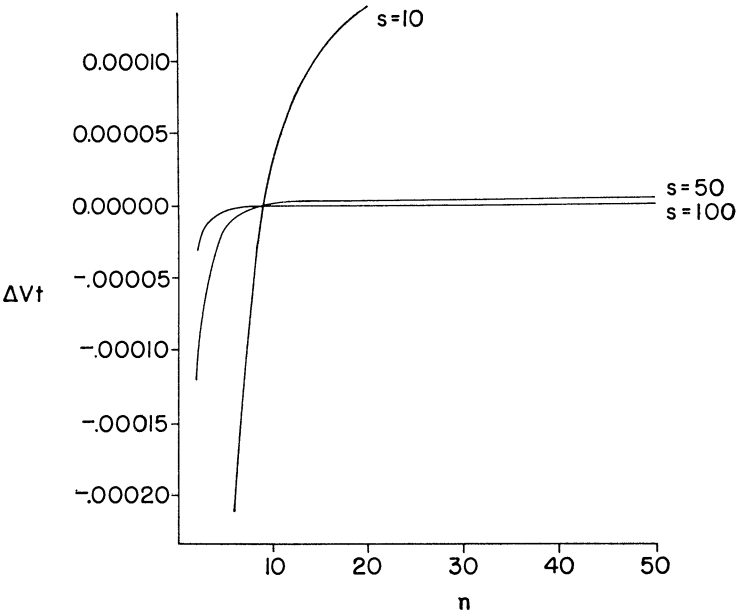

FIGURE 1
ΔVt WHEN A GROUP WITH ONE OBSERVATION IS INCLUDED, FOR VARYING s AND n, $t = .10$

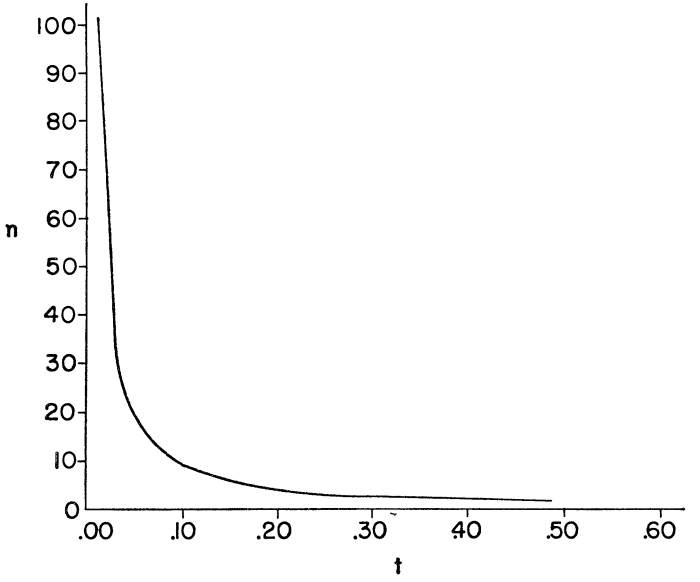

FIGURE 2
THE VALUE OF n FOR WHICH ΔVt CHANGES SIGN PLOTTED AGAINST t FOR $s = 1000$

RESULTS FROM EMPIRICAL STUDY
(Harvey and Everson)

Replicate sets of data were generated by Monte Carlo techniques on an IBM 1620 computer. Each set of data was then analyzed by conventional methods to obtain the intraclass correlation. The variance in intraclass correlations among replicate sets of data gives an empirical estimate of the variance of the intraclass correlation. A selected set of 3750 random standardized normal deviates was stored in the computer, and these were chosen at random to generate the observations. The same set of random normal deviates was used in all runs. The set used was selected from 15 such sets, which had been computed from generated random uniform numbers, as having the distribution that most nearly approximated the theoretical distribution.

The observations (y_{ij}) were generated from the random normal deviates as follows:

$$y_{ij} = g_i + e_{ij}$$

where g_i is the random normal deviate selected for the ith group, and e_{ij} is the random normal deviate selected for the jth observation in the ith group, multiplied by σ_e . Hence, the variance component for groups was always expected to be unity and the within group variance component was expected to be σ_e^2 .

The numbers of groups generated were 25, 50, 100, and 200. The numbers of observations generated per group were 2, 4, 6, 10, 15, 20 and 50. For each of these 28 combinations of numbers of groups and numbers per group, replicate sets of data were generated separately for .05 and .10 intraclass correlation, both with equal numbers and with the addition of single observation groups. The numbers of groups with single observations that were added were 6, 12, 25 and 50, for the numbers of groups with equal frequency of 25, 50, 100 and 200, respectively. (In preliminary runs, it was found that the change in the variance of the intraclass correlation due to the addition of only one group with a single observation could not be established accurately without an extremely large number of replications.)

A comparison of the standard errors obtained for the intraclass correlation by using the approximation formula (4) with those obtained empirically is given in Table 2. The number of replications on which the empirical estimates are based is given in the last column of Table 2. The number of replicates made in each case was the number required to give a standard error of approximately .003 for the mean empirically-calculated intraclass correlation when the parameter (t) was .050.

TABLE 2

COMPARISON OF THE STANDARD ERRORS OF INTRACLASS CORRELATIONS (σ_t)

No. of Groups	Obs. per Group	Standard Errors of Intraclass Correlations								No. Replicates
		Approximation Formula				Empirical Estimates				
		$t = .05$		$t = .10$		$t = .05$		$t = .10$		
		Equal	Unequal*	Equal	Unequal	Equal	Unequal	Equal	Unequal	
25	2	.202	.210	.200	.206	.195	.205	.197	.203	6500
	4	.091	.095	.097	.100	.091	.093	.097	099	1300
	6	.062	.065	.071	.072	.061	.065	.068	.071	622
	10	.042	.043	.052	.052	.041	.046	.053	.054	280
	15	.032	.033	.043	.042	.031	.031	.040	.040	165
	20	.027	.027	.039	.037	.029	.030	.035	.039	120
	50	.019	.018	.031	.029	.019	.022	.028	.030	54
50	2	.142	.148	.141	.145	.140	.148	.140	.141	3250
	4	.064	.067	.068	.070	.064	.064	.069	.073	646
	6	.044	.046	.050	.051	.042	.050	.052	.051	306
	10	.029	.030	.036	.036	.029	.030	.035	.036	137
	15	.022	.023	.030	.029	.023	.022	.029	.034	81
	20	.019	.019	.027	.026	.021	.019	.025	.027	59
	50	.013	.013	.022	.020	.012	.014	.019	.017	29
100	2	.100	.104	.099	.103	.099	.103	.102	.102	1600
	4	.045	.047	.048	.050	.048	.050	.049	.050	321
	6	.031	.032	.035	.036	.031	.033	.036	.031	152
	10	.021	.021	.026	.026	.020	.021	.025	.028	68
	15	.016	.016	.021	.021	.016	.014	.019	.026	40
	20	.014	.014	.019	.018	.014	.013	.019	.021	29
	50	.009	.009	.015	.014	.011	.009	.021	.014	14
200	2	.071	.074	.070	.073	.070	.074	.070	.073	798
	4	.032	.033	.034	.035	.032	.034	.034	.035	160
	6	.022	.023	.025	.025	.022	.022	.024	.028	76
	10	.015	.015	.018	.018	.012	.014	.022	.018	34
	15	.011	.011	.015	.015	.011	.010	.015	.020	20
	20	.010	.010	.013	.013	.011	.010	.013	.014	15
	50	.007	.006	.011	.010	.005	.008	.008	.008	7

*See text for type of unequal frequency considered.

The difference between the standard error of the intraclass correlation as computed with the approximation formula (4) and the Monte Carlo estimate is given in Table 3 for each case considered. In addition, Table 3 gives the change expected in the standard error when single observation groups are added as estimated by the two methods. These differences clearly show that, for the combinations of numbers chosen for study, the approximate formula for the standard error of the intraclass correlation is entirely satisfactory. No appreciable bias exists, for these combinations of numbers, in the standard error calculated with the approximation formula.

The empirical results in the last two columns of Table 3 support the conclusion reached from the calculations made with the approximate formula that the addition of single observation groups will actually increase the variance of the intraclass correlation in many cases. There is a suggestion from the empirical results that this increase in variance is even more important than indicated by the formula, especially with larger intraclass correlations.

TABLE 3

DIFFERENCES BETWEEN STANDARD ERRORS OBTAINED WITH FORMULA AND EMPIRICALLY AND A COMPARISON OF THE CHANGE IN THE STANDARD ERROR WHEN SINGLE OBSERVATION GROUPS ARE ADDED

No. of Groups	Obs. per Group	Formula σ_t − Empirical σ_t				Equal σ_t − Unequal σ_t			
		t = .05		t = .10		Formula		Empirical Est.	
		Equal	Unequal*	Equal	Unequal	t = .05	t = .10	t = .05	t = .10
25	2	.007	.005	.003	−.003	−.008	−.006	−.010	−.005
	4	.000	.002	.000	.001	−.004	−.003	−.002	−.002
	6	.001	.000	.003	.001	−.003	−.001	−.004	−.003
	10	.000	−.003	−.001	−.002	−.001	.000	−.005	−.001
	15	.001	.002	.003	.002	−.001	.001	.000	.000
	20	−.002	−.003	.004	−.002	.000	.002	−.001	−.004
	50	.000	−.004	.003	−.001	.001	.002	−.003	−.002
50	2	.002	.000	.001	.004	−.006	−.004	−.008	−.001
	4	.000	.003	−.001	−.003	−.003	−.002	.000	−.004
	6	.002	−.004	−.002	.000	−.002	−.001	−.008	.001
	10	.000	.000	.001	.000	−.001	.000	−.001	−.001
	15	−.001	.001	.001	−.005	−.001	.001	.001	−.005
	20	−.002	.000	.002	−.001	.000	.001	.002	−.002
	50	.001	−.001	.003	−.003	.000	.002	−.002	.002
100	2	.001	.001	−.003	.001	−.004	−.004	−.004	.000
	4	−.003	−.003	−.001	.000	−.002	−.002	−.002	−.001
	6	.000	−.001	−.001	.005	−.001	−.001	−.002	.005
	10	.001	.000	.001	−.002	.000	.000	−.001	−.003
	15	.000	.002	.002	−.005	.000	.000	.002	−.007
	20	.000	.001	.000	−.003	.000	.001	.001	−.002
	50	−.002	.000	−.006	.000	.000	.001	.002	.007
200	2	.001	.000	.000	.000	−.003	−.003	−.004	−.003
	4	.000	−.001	.000	.000	−.001	−.001	−.002	−.001
	6	.000	.001	.001	−.003	−.001	.000	.000	−.004
	10	.003	.001	−.004	.000	.000	.000	−.002	.004
	15	.000	.001	.000	−.005	.000	.000	−.001	.005
	20	−.001	.000	.000	−.003	.000	.000	.001	−.001
	50	.002	−.002	.003	.002	.001	.001	−.003	.000

*See text for type of unequal frequency considered.

REFERENCES

Crump, S. Lee [1951]. The present status of variance component analysis. *Biometrics 7*, 1–16.

Fisher, R. A. [1954]. *Statistical methods for research workers.* 12th ed. Oliver and Boyd, Edinburgh; Hafner Publishing Co., Inc., New York.

Hartley, H. O. [1954]. *Unpublished derivation:* see G. M. Farnsworth, [1955]. *Ph. D. Thesis,* Iowa State University, Ames.

Robertson, Alan. [1962]. NOTE: Weighting in the estimation of variance components in the unbalanced single classification. *Biometrics 18*, 413–17.

Published with approval of the Director as Paper No. 1313, Journal Series, Nebraska Agricultural Experiment Station, Contribution from the North Central Regional Project NC-1, Improvement of Beef Cattle Through Breeding Methods.