Heredity **73** (1994) 544–548
©The Genetical Society of Great Britain

Received 21 March 1994

The estimation of the genetic correlation: the use of the jackknife

DEREK A. ROFF* & RICHARD PREZIOSI
Department of Biology, McGill University, 1205 Dr Penfield Ave., Montréal, Québec, Canada H3A 1B1

To understand fully the process of evolution of quantitative traits it is necessary to be able to estimate the genetic correlation and its associated standard error. At present, estimation methods are available only for relatively simple designs. An alternative procedure is to use the correlation of family means as an estimate of the genetic correlation. We evaluate the utility of the family mean method and that of the more general procedure, the jackknife. The family mean method is shown to be potentially very biased unless family sizes are very large (≈ 20), and therefore its general utility is questionable. However, the jackknife method does provide valid estimates of both the correlations (phenotypic and genetic) and their standard errors.

Keywords: bias, confidence limits, genetic correlation, heritability, jackknife, phenotypic correlation.

Introduction

Because of genetic correlations among characters, selection, even if it operates directly only on a single character, also affects many other characters (Lande & Arnold, 1983; Falconer, 1989). For this reason the estimation of the genetic correlation is of considerable importance both for the application of quantitative genetic theory in artificial selection and for the understanding of evolutionary processes in natural populations. In any estimation procedure it is necessary to estimate both the value of the statistic itself and also the associated confidence limits. The estimation of confidence limits for the genetic correlation is difficult and even in the restricted cases where estimation methods have been worked out the statistical behaviour is not well understood (Robertson, 1959, 1960; Van Vleck & Henderson, 1961; Van Vleck, 1968; Hammond & Nicholas, 1972; Grossman & Norton, 1974; Becker, 1985). An alternative approach suggested by Via (1984) is to use the Pearson product-moment correlation between family means, for which the usual methods of estimating confidence intervals on correlations can be applied. It is known that such an estimate is biased, the amount of bias declining with family size. However, the magnitude of this bias has not been estimated.

Clearly, it would be useful to have a general method of estimating the genetic correlation and its associated confidence interval. A likely candidate is the jackknife method. The jackknife is one of a number of methods collectively known as robust techniques because they are relatively insensitive to the underlying distributional properties of the statistics being estimated (Potvin & Roff, 1993). Nevertheless, these methods are not panaceas in the sense that they can be applied indiscriminately, and it is necessary to demonstrate in each particular circumstance that they are valid approaches (Miller, 1974; Potvin & Roff, 1993). This can be done using simulation. The efficacy of the jackknife in estimating the heritability and its standard error has already been demonstrated (Simons & Roff, 1994), and thus *a priori* it seems likely that the method will also work for the estimation of the genetic correlation and its standard error.

In this paper we present a theoretical analysis of the family mean method, deriving the relationship between the bias and family size. Having demonstrated that this approach is, in general, unacceptable, we apply the jackknife technique to a set of simulated data, showing that it correctly estimates the genetic correlation and its standard error.

The simulation model

Because the simulation model is used to verify the theoretical analysis of the family mean method as well as to verify the utility of the jackknife method, we present its construction first. The model is based upon the procedure worked out by Ronningen (1974) and

*Correspondence.

Olausson & Ronningen (1975). The case simulated is that of full-sibs in which each family is split between two cages. The basic method of estimating the heritabilities and genetic correlation is through a series of nested ANOVAs (Becker, 1985). Because the nested design is identical in structure to that of the half-sib design without cage effects, the results should also apply to this design. The values of the two characters X and Y are given by

$$X_{i,j,k} = a_{x,i}\sqrt{\frac{1}{2}\,h_x^2} + b_{x,i,j,k}\sqrt{1-\frac{1}{2}\,h_x^2} + \sigma c_{x,i,k}$$

$$Y_{i,j,k} = r_g a_{x,i}\sqrt{\frac{1}{2}\,h_y^2} + a_{y,i}\sqrt{\frac{1}{2}(1-r_g^2)h_y^2} + r_e b_{x,i,j,k}\sqrt{1-\frac{1}{2}\,h_y^2} + b_{y,i,j,k}\sqrt{\left(1-\frac{1}{2}\,h_y^2\right)(1-r_e^2)} + \sigma c_{y,i,k},$$

where:

$X_{i,j,k}$, $Y_{i,j,k}$ are the values of traits X and Y, respectively, for individual j in family i of cage k;
$a_{x,i}$, $a_{y,i}$ are random standard normal values, N(0,1), common to the ith family;
$b_{x,i,j,k}$, $b_{y,i,j,k}$ are random standard normal values, N(0,1), of the jth individual from family i in cage k;
$c_{x,i,k}$, $c_{y,i,k}$ are random standard normal values, N(0,1), common to individuals in cage k of family i;
h_x^2, h_y^2 are the heritabilities of traits X and Y, respectively;
r_g is the genetic correlation between traits X and Y;
r_e is the environmental correlation between traits X and Y, given by

$$r_e = \frac{r_p - \frac{1}{2}\,r_g\sqrt{h_x^2 h_y^2}}{\sqrt{\left(1-\frac{1}{2}\,h_x^2\right)\left(1-\frac{1}{2}\,h_y^2\right)}}$$

where r_p is the phenotypic correlation; and
σ is the standard deviation of the cage effect.

For each variable combination, 1 000 simulations were run. In all simulations the number of families was kept constant at 50.

The method of family means

Because standard errors are not generally available for complex designs, Via (1984) suggested using as an approximation of the genetic correlation, the correlation of family means:

$$r_m = \frac{cov_{m(xy)}}{\sqrt{var_{m(x)}\ var_{m(y)}}},$$

where cov_m and var_m are calculated from family means of phenotypic values. This is an approximation because the variance and covariance terms contain a fraction of the within-family error term:

$$cov_m = cov_{\text{among}} + \frac{cov_{\text{within}}}{n},$$

where n is the family size. The potential advantage of the above method is that because they are simply standard product-moment correlations the usual significance tests can be applied and confidence intervals computed. Specifically, if the number of families is N, the confidence interval is computed by first transforming the correlation to the z scale

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right).$$

The standard error is then approximately $1/\sqrt{(N-3)}$, and confidence limits on r can be estimated by computing the confidence limits on z and back-transforming (for more details see Sokal & Rohlf (1981, pp. 583-591)). Note that the family size, n, does not appear in the formula, and thus any bias due to an insufficiently large family size will not be reflected in the confidence interval.

It can be shown by use of the equations derived for the simulation model, that the family mean correlation is equal to

$$r_m = \frac{r_g + \frac{1}{n}\left(\frac{2r_p}{h_x h_y} - r_g\right)}{\sqrt{\left(1+\frac{1}{n}\left[\frac{2}{h_x^2}-1\right]\right)\left(1+\frac{1}{n}\left[\frac{2}{h_y^2}-1\right]\right)}},$$

where r_p is the phenotypic correlation. Note that, unlike the estimated standard error which contains only N, the family mean correlation contains only n, the family size. To examine the potential effect of changing family size on this approximation for the genetic correlation we assumed $h_x^2 = h_y^2 = r_p = 0.5$, $r_g = 0$, and no cage effect ($\sigma = 0$). Under these conditions $r_m = 1/(n+3)$. The predicted and observed family mean correlations agree very closely (Table 1). It is evident that for family sizes less than 20 there is a considerable bias in the family mean correlation as an approximation to the genetic correlation. Further, the estimated 95 per cent confidence intervals are far from correct when family sizes are small (Table 1). The bias and error in the confidence interval will decline as the genetic correlation approaches the phenotypic correlation. However, given that these are the parameters we are attempting to estimate it would certainly be unwise to make any *a priori* assumptions. If the genetic correlation and phenotypic correlation were even more different than assumed above then even a family size of 20 might not be sufficient. The formula given above can be used as a guide for the bias resulting from a particular family size. Nevertheless it would be certainly preferable to have a method of estimation that is not potentially confounded as this method can be.

The jackknife

The jackknife method works as follows: first, the genetic and phenotypic correlations are estimated using the usual ANOVA approach (see Becker (1985) for the appropriate formulae). A sequence of N pseudovalues is computed by dropping in turn each of the families, estimating the resulting correlations and using the formula

$$S_{N,i} = Nr_N - (N-1)r_{N-1,i},$$

where: $S_{N,i}$ is the ith pseudovalue, r_N is the correlation estimated using all N families, and $r_{N-1,i}$ is the correlation obtained by dropping the ith family alone.

Table 1 The predicted and observed family mean correlation for different family sizes, given $h_x^2 = h_y^2 = r_p = 0.5$, $r_g = \sigma = 0$

	Family mean correlation		
Number per family	Predicted	Observed	P
3	0.333	0.334	0.34
5	0.250	0.256	0.57
10	0.154	0.149	0.83
15	0.111	0.112	0.88
20	0.087	0.082	0.92

Results from 1 000 simulations using 50 families per run.
P is the proportion of times the estimated 95 per cent confidence interval actually enclosed the true genetic correlation.

The jackknife estimate of the correlation, r_j, is simply the mean of the pseudovalues

$$r_j = \frac{\sum_{i=1}^{i=N} S_{N,i}}{N}.$$

An estimate of the standard error, SE, is given by

$$\mathrm{SE} = \frac{\sum_{i=1}^{i=N} (S_{N,i} - r_j)^2}{N(N-1)}.$$

For the suite of variable values used previously the confidence limits estimated using the jackknife for both the phenotypic and genetic correlations are not significantly different from the required 95 per cent (Table 2). With the possible exception of the genetic correlation for $N = 3$, the mean estimates show no evidence of significant bias. Though there is a possible bias for $N = 3$ the confidence limits are still acceptable ($P = 0.93$) and the mean standard error is so large (SE = 0.457) that the estimate is essentially meaningless in any case. To test for a potential bias given such a small family size we ran the model changing the genetic

Table 2 The genetic and phenotypic correlations and their associated standard errors predicted using the jackknife method for different family sizes, given $h_x^2 = h_y^2 = r_p = 0.5$, $r_g = \sigma = 0$ (one set with $r_g = -0.50$)

	Genetic correlation means			Phenotypic correlation means		
N	r_g	SE	P	r_p	SE	P
3	0.151	0.457	0.93	0.500	0.070	0.93
5	0.017	0.277	0.95	0.501	0.060	0.94
10	0.005	0.198	0.95	0.502	0.050	0.94
15	−0.001	0.177	0.94	0.500	0.047	0.93
20	0.005	0.169	0.94	0.500	0.045	0.94
5 ($r_g = -0.5$)	−0.470	0.318	0.97	0.501	0.069	0.94

Results from 1000 simulations using 50 families per run.
P is the proportion of times the estimated 95 per cent confidence interval actually enclosed the true correlation.

correlation to -0.5. In this case there is no evidence of a marked bias (Table 2).

Falconer (1989) suggested the approximate formula derived by Reeve (1955) and Robertson (1959)

$$\mathrm{EstSE}(r_g)=\frac{1-r_g^2}{\sqrt{2}}\sqrt{\frac{\mathrm{SE}_{h_x^2}\mathrm{SE}_{h_y^2}}{h_x^2h_y^2}},$$

where $\mathrm{SE}_{h_x^2}$, $\mathrm{SE}_{h_y^2}$ are the estimated standard errors of the heritabilities, h_x^2, h_y^2, respectively. However, this approximation produces confidence intervals that are smaller than the required 95 per cent (for $N=3, 5, 10, 15, 20$, $P\%=79, 85, 87, 85, 84$, respectively).

The above variable set is ideal in the sense that there are no cage effects and family size is constant. More generally there will be cage effects (or the design will be half-sib) and family size will vary. To examine this situation we examined a range of values, with family size for each cage being a uniform random variable between 3 and 7, and two cages per family, giving a mean family size of 10 (Table 3). The cage effect on trait X may be in some degree correlated with the cage effect on trait Y. We examined the two extremes, namely, zero correlation of the cage effects on X and Y and a correlation of one ($c_{x,i,k}=c_{y,i,k}$).

As in the other simulations there is no evidence of bias in the estimate, and the estimated confidence limits do enclose the actual value in the required 95 per cent of cases. Thus we conclude that the estimation of the genetic correlation and its associated standard error by the method of the jackknife is valid.

Discussion

The analysis of the method of family means shows it to be inappropriate unless the family sizes are very large (≈ 20) and/or the genetic and phenotypic correlations are very similar. Unless these conditions are satisfied the estimate can be strongly biased and the confidence intervals much smaller than the supposed 95 per cent (Table 1). The approximate formula for the computation of the standard error of the genetic correlation given by Falconer (1989, p. 317) also substantially underestimates the standard error. However, it can still be used as a guide to the magnitude of the standard error expected given rough estimates of the relevant parameters. Klein *et al.* (1973) provide tables based on the aforementioned approximation for several types of breeding design.

The jackknife method provides both statistically accurate estimates of the phenotypic and genetic correlations and their standard errors (Tables 2, 3). The general design used in this analysis is that of nested ANOVA, and thus although the particular model is specifically that of full-sibs with cage effects it should also be applicable to the half-sib design. The statistical model simulating the pattern of inheritance is quite complex and it may not always be possible to write a model suitable for more complex designs. However, because the jackknife cannot be guaranteed to work in other novel situations it is essential that its performance be evaluated before it is used. An alternative method to that of constructing a simulation model as done in the present paper is given by Mueller (1979).

Table 3 Jackknife estimates of the genetic and phenotypic correlations for different heritabilities and correlations

			Genetic correlation means			Phenotypic correlation means		
h_x^2, h_y^2	r_p, r_g	σ	r_g	SE	P	r_p	SE	P
No correlation between cage effects								
0.50, 0.50	0.5, 0.0	0.30	0.006	0.247	0.95	0.498	0.061	0.94
0.25, 0.25	0.5, 0.0	0.15	0.073	0.373	0.95	0.499	0.045	0.95
0.50, 0.25	0.5, 0.0	0.15	0.018	0.280	0.94	0.499	0.049	0.95
0.50, 0.50	0.5, 0.5	0.30	0.493	0.194	0.94	0.498	0.052	0.94
0.50, 0.50	0.5, −0.5	0.30	−0.485	0.268	0.96	0.499	0.073	0.93
Correlation between cage effects = 1								
0.50, 0.50	0.5, 0.0	0.30	0.019	0.269	0.94	0.500	0.056	0.93
0.25, 0.25	0.5, 0.0	0.15	0.094	0.401	0.94	0.500	0.049	0.94
0.50, 0.25	0.5, 0.0	0.15	0.029	0.290	0.94	0.499	0.048	0.94
0.50, 0.50	0.5, 0.5	0.30	0.507	0.184	0.92	0.500	0.049	0.94
0.50, 0.50	0.5, −0.5	0.30	−0.467	0.314	0.96	0.500	0.067	0.93

In all cases the number per cage is a uniform random variable between 3 and 7, and there are two cages per family.
All estimates based on 1000 simulations per combination.
P is the proportion of times the estimated 95 per cent confidence interval actually enclosed the true correlation.

In this approach a particular data set is used as the 'real' population. From this population samples are drawn with replacement, thereby generating simulated data sets. The statistical methods of interest are applied to these data sets in the same manner as done here. The actual values of the statistics are those obtained from the original data set and thus the ability of the statistical methods to estimate the relevant parameters can be evaluated. The disadvantage of this method is that it is restricted to the available empirical data sets. However, it is certainly better to test the method on a restricted set than simply to assume that it is correct.

Acknowledgement

This work was supported by a grant from the Natural Sciences and Engineering Council of Canada.

References

BECKER, W. A. 1985. *Manual of Quantitative Genetics.* McNaughton and Gunn, Ann Arbor, MI.

FALCONER, D. S. 1989. *Introduction to Quantitative Genetics,* 3rd edn. Longman, New York.

GROSSMAN, M. AND NORTON, H. W. 1974. Simplification of the sampling variance of the correlation coefficients. *Theor. Appl. Genet.,* **44**, 332.

HAMMOND, K. AND NICHOLAS, F. W. 1972. The sampling variance of the correlation coefficients estimated from two-fold nested and offspring-parent regression analysis. *Theor. Appl. Genet.,* **42**, 97–100.

KLEIN, T. W., DEFRIES, J. C. AND FINKBEINER, C. T. 1973. Heritability and genetic correlation: standard error of estimates and sample size. *Behav. Genet.,* **3**, 355–364.

LANDE, R. AND ARNOLD, S. J. 1983. The measurement of selection on correlated characters. *Evolution,* **37**, 1210–1226.

MILLER, R. G. 1974. The jackknife — a review. *Biometrika,* **61**, 1–15.

MUELLER, L. D. 1979. A comparison of two methods for making statistical inferences on Nei's measure of genetic distance. *Biometrics,* **35**, 757–763.

OLAUSSON, A. AND RONNINGEN, K. 1975. Estimation of genetic parameters for threshold characters. *Acta Agr. Scand.,* **25**, 201–208.

POTVIN, C. AND ROFF, D. A. 1993. Distribution-free and robust statistical methods: viable alternatives to parametric statistics? *Ecology,* **74**, 1617–1628.

REEVE, E. C. R. 1955. The variance of the genetic correlation coefficient. *Biometrics,* **11**, 357–374.

ROBERTSON, A. 1959. The sampling variance of the genetic correlation coefficient. *Biometrics,* **15**, 469–485.

ROBERTSON, A. 1960. Experimental design on the measurement of heritabilities and genetic correlations. In: Kempthorne, O. (ed.) *Biometrical Genetics,* pp. 101–106. Pergamon Press, Oxford.

RONNINGEN, K. 1974. Monte Carlo simulation of statistical-biological models which are of interest in animal breeding. *Acta Agr. Scand.,* **24**, 135–142.

SIMONS, A. M. AND ROFF, D. A. 1994. The effect of environmental variability on the heritabilities of traits of a field cricket. *Evolution* (in press).

SOKAL, R. R. AND ROHLF, F. J. 1981. *Biometry.* W. H. Freeman, San Francisco.

VAN VLECK, L. D. 1968. Selection bias in estimation of the genetic correlation. *Biometrics,* **24**, 951–962.

VAN VLECK, L. D. AND HENDERSON, C. R. 1961. Empirical sampling estimates of genetic correlations. *Biometrics,* **17**, 359–371.

VIA, S. 1984. The quantitative genetics of polyphagy in an insect herbivore. II. Genetic correlations in larval performance within and among host plants. *Evolution,* **38**, 896–905.