Journal of Modern Applied Statistical Methods

Volume 8 | Issue 2 Article 16

11-1-2009

Confidence Interval Estimation for Intraclass Correlation Coefficient Under Unequal Family Sizes

Madhusudan Bhandary
Columbus State University, bhandary_madhusudan@columbusstate.edu

Koji Fujiwara
North Dakota State University, koji.fujiwara@ndsu.edu

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 520-525

1538 – 9472/09/$95.00

Confidence Interval Estimation for Intraclass Correlation Coefficient Under Unequal Family Sizes

Madhusudan Bhandary
Columbus State University

Koji Fujiwara
North Dakota State University

Confidence intervals (based on the χ^2-distribution and (Z) standard normal distribution) for the intraclass correlation coefficient under unequal family sizes based on a single multinormal sample have been proposed. It has been found that the confidence interval based on the χ^2-distribution consistently and reliably produces better results in terms of shorter average interval length than the confidence interval based on the standard normal distribution: especially for larger sample sizes for various intraclass correlation coefficient values. The coverage probability of the interval based on the χ^2-distribution is competitive with the coverage probability of the interval based on the standard normal distribution. An example with real data is presented.

Key words: Z-distribution, χ^2-distribution, intraclass correlation coefficient, confidence interval.

Introduction

Suppose, it is required to estimate the correlation coefficient between blood pressures of children on the basis of measurements taken on p children in each of n families. The p measurements on a family provide p(p − 1) pairs of observations (x, y), x being the blood pressure of one child and y that of another. From the n families we generate a total of np(p − 1) pairs from which a typical correlation coefficient is computed. The correlation coefficient thus computed is called an intraclass correlation coefficient. Statistical inference concerning intraclass correlations is important because it provides information regarding blood pressure, cholesterol, etc. in a family within a particular race.

The intraclass correlation coefficient ρ as a wide variety of uses for measuring the degree of intrafamily resemblance with respect to characteristics such as blood pressure, cholesterol, weight, height, stature, lung capacity, etc. Several authors have studied statistical inference concerning ρ based on single multinormal samples (Scheffe, 1959; Rao, 1973; Rosner, et al., 1977, 1979; Donner & Bull, 1983; Srivastava, 1984; Konishi, 1985; Gokhale & SenGupta, 1986; SenGupta, 1988; Velu & Rao, 1990).

Donner and Bull (1983) discussed the likelihood ratio test for testing the equality of two intraclass correlation coefficients based on two independent multinormal samples under equal family sizes. Konishi and Gupta (1987) proposed a modified likelihood ratio test and derived its asymptotic null distribution. They also discussed another test procedure based on a modification of Fisher's Z-transformation following Konishi (1985). Huang and Sinha (1993) considered an optimum invariant test for the equality of intraclass correlation coefficients under equal family sizes for more than two intraclass correlation coefficients based on independent samples from several multinormal distributions.

For unequal family sizes, Young and Bhandary (1998) proposed Likelihood ratio test, large sample Z-test and large sample Z^*-test for

Madhusudan Bhandary is an Associate Professor in the Department of Mathematics. Email: bhandary_madhusudan@colstate.edu. Koji Fujiwara is a graduate student in the Department of Statistics. Email: koji.fujiwara@ndsu.edu.

the equality of two intraclass correlation coefficients based on two independent multinormal samples. For several populations and unequal family sizes, Bhandary and Alam (2000) proposed the Likelihood ratio and large sample ANOVA tests for the equality of several intraclass correlation coefficients based on several independent multinormal samples. Donner and Zou (2002) proposed asymptotic test for the equality of dependent intraclass correlation coefficients under unequal family sizes.

None of the above authors, however, derived any confidence interval estimator for intraclass correlation coefficients under unequal family sizes. In this article, confidence interval estimators for intraclass correlation coefficients are considered based on a single multinormal sample under unequal family sizes, and conditional analyses - assuming family sizes are fixed - though unequal.

It could be of interest to estimate the blood pressure or cholesterol or lung capacity for families in American races. Therefore, an interval estimator for the intraclass correlation coefficient under unequal family sizes must be developed. To address this need, this paper proposes two confidence interval estimators for the intraclass correlation coefficient under unequal family sizes, and these interval estimators are compared using simulation techniques.

Methodology

Proposed Confidence Intervals: Interval Based on the Standard Normal Distribution

Consider a random sample of k families. Let

$$\underset{\sim}{X}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ . \\ . \\ . \\ x_{ip_i} \end{pmatrix}$$

be a $p_i x1$ vector of observations from i^{th} family; $i = 1,2,...,k$. The structure of the mean vector and the covariance matrix for the familial data is given by the following (Rao, 1973):

$$\underset{\sim}{\mu}_i = \mu \underset{\sim}{1}_i$$

and

$$\underset{p_i x p_i}{\Sigma_i} = \sigma^2 \begin{pmatrix} 1 & \rho & ... & \rho \\ \rho & 1 & ... & \rho \\ ... & ... & ... & ... \\ \rho & \rho & ... & 1 \end{pmatrix}, \qquad (2.1)$$

where $\underset{\sim}{1}_i$ is a $p_i x1$ vector of 1's, $\mu(-\infty < \mu < \infty)$ is the common mean and $\sigma^2(\sigma^2 > 0)$ is the common variance of members of the family and ρ, which is called the intraclass correlation coefficient, is the coefficient of correlation among the members of the family and $\max_{1 \le i \le k}\left(-\frac{1}{p_i - 1}\right) \le \rho \le 1.$

It is assumed that $\underset{\sim}{x}_i \sim N_{p_i}(\underset{\sim}{\mu}_i, \Sigma_i); i = 1,...,k$, where N_{p_i} represents a p_i-variate normal distribution and $\underset{\sim}{\mu}_i, \Sigma_i's$ are defined in (2.1). Let

$$\text{Let } \underset{\sim}{u}_i = \begin{pmatrix} u_{i1} \\ u_{i2} \\ . \\ . \\ . \\ u_{ip_i} \end{pmatrix} = Q \underset{\sim}{x}_i \qquad (2.2)$$

where Q is an orthogonal matrix. Under the orthogonal transformation (2.2), it can be shown that $\underset{\sim}{u}_i \sim N_{p_i}(\underset{\sim}{\mu}_i^*, \Sigma_i^*); i = 1,...,k$, where

$$\underset{\sim}{\mu_i^*}_{p_i x 1} = \begin{pmatrix} \mu \\ 0 \\ . \\ . \\ . \\ 0 \end{pmatrix}$$

and

$$\Sigma_i^* = \sigma^2 \begin{pmatrix} \eta_i & 0 & ... & 0 \\ 0 & 1-\rho & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1-\rho \end{pmatrix}$$

and $\eta_i = p_i^{-1}\{1+(p_i-1)\rho\}$. The transformation used on the data from $\underset{\sim}{x}$ to $\underset{\sim}{u}$ above is independent of ρ. Helmert's orthogonal transformation can also be used.

Srivastava (1984) gave an estimator of ρ and σ^2 under unequal family sizes which are good substitute for the maximum likelihood estimator and are given by the following:

$$\hat{\rho} = 1 - \frac{\hat{\gamma}^2}{\hat{\sigma}^2},$$

where

$$\hat{\sigma}^2 = (k-1)^{-1}\sum_{i=1}^{k}(u_{i1}-\hat{\mu})^2 + k^{-1}\hat{\gamma}^2\left(\sum_{i=1}^{k}a_i\right)$$

$$\hat{\gamma}^2 = \frac{\sum_{i=1}^{k}\sum_{r=2}^{p_i}u_{ir}^2}{\sum_{i=1}^{k}(p_i-1)}$$

$$\hat{\mu} = k^{-1}\sum_{i=1}^{k}u_{i1} \qquad (2.3)$$

and $a_i = 1 - p_i^{-1}$.

Srivastava and Katapa (1986) derived the asymptotic distribution of $\hat{\rho}$; they showed: that $\hat{\rho} \sim \mathrm{N}(\rho, Var/k)$ asymptotically, where

$$Var = 2(1-\rho)^2 \left\{(\bar{p}-1)^{-1} + c^2 - 2(1-\rho)(\bar{p}-1)^{-1}k^{-1}\sum_{i=1}^{k}a_i\right\} \qquad (2.4)$$

k = number of families in the sample

$$\bar{p} = k^{-1}\sum_{i=1}^{k}p_i$$

$$c^2 = 1 - 2(1-\rho)^2 k^{-1}\sum_{i=1}^{k}a_i + (1-\rho)^2\left[k^{-1}\sum_{i=1}^{k}a_i + (\bar{p}-1)^{-1}(k^{-1}\sum_{i=1}^{k}a_i)^2\right]$$

and $a_i = 1 - p_i^{-1}$.

Under the above setup, it is observed (using Srivastava & Katapa, 1986) that:

$$Z = \frac{\hat{\rho}-\rho}{\sqrt{\frac{Var}{k}}} \sim N(0,1), \qquad (2.5)$$

asymptotically, where, Var is to be determined from (2.4) and $\hat{\rho}$ is obtained from (2.3).

Using the expression (2.5), it is found that the $100(1-\alpha)\%$ confidence interval for ρ is

$$\hat{\rho} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{Var}{k}} \qquad (2.6)$$

Interval Based on the χ^2 Distribution

It can be shown, by using the distribution of $\underset{\sim}{u_i}$ given by (2.2):

$$\frac{\sum_{i=1}^{k}\sum_{r=2}^{p_i}u_{ir}^2}{\sigma^2(1-\rho)} \sim \chi^2_{\sum_{i=1}^{k}(p_i-1)}, \qquad (2.7)$$

where χ_n^2 denotes the Chi-square distribution with n degrees of freedom. Using (2.7), a $100(1-\alpha)\%$ confidence interval for ρ can be found as follows:

$$1 - \frac{\sum_{i=1}^{k}\sum_{r=2}^{p_i}u_{ir}^2}{\chi^2_{1-\frac{\alpha}{2};n}.\hat{\sigma}^2} < \rho < 1 - \frac{\sum_{i=1}^{k}\sum_{r=2}^{p_i}u_{ir}^2}{\chi^2_{\frac{\alpha}{2};n}.\hat{\sigma}^2} \qquad (2.8)$$

where, $n = \sum_{i=1}^{k} (p_i - 1)$ and $\chi^2_{\alpha;n}$ denotes the upper 100α % point of a Chi-square distribution with n degrees of freedom and $\hat{\sigma}^2$ can be obtained from (2.3).

Data Simulation

Multivariate normal random vectors were generated using an R program in order to evaluate the average lengths and coverage probability of the intervals given by (2.6) and (2.8). Fifteen and 30 vectors of family data were created for the population. The family size distribution was truncated to maintain the family size at a minimum of two siblings and a maximum of 15 siblings. The previous research in simulating family sizes (Rosner, et al., 1977; and Srivastava & Keen, 1988) determined the parameter setting for FORTRAN IMSL negative binomial subroutine with a mean = 2.86 and a success probability = 0.483.

Given this, the mean was set to equal 2.86 and theta was set to equal 41.2552. All parameters were set the same except for the value of ρ which took values from 0.1 to 0.9 at increments of 0.1.The R program produced 3,000 estimates of ρ along with the coverage probability and the confidence intervals given by the formulae (2.6) and (2.8) for each particular value of the population parameter ρ. The average length and coverage probability of each interval at α=0.01, 0.05 and 0.10 were noted. Results are shown in Table1.

Table 1: Coverage Probability and Length for the Confidence Interval

			Coverage Probability		Length	
rho	k	alpha	Z	Chi-square	Z	Chi-square
0.1	15	0.01	1.00000	0.99933	1.04368	1.06377
0.2	15	0.01	0.98533	0.99733	1.12548	1.05273
0.3	15	0.01	0.99233	0.99300	1.08944	0.90430
0.4	15	0.01	0.98400	0.98167	1.09402	1.01612
0.5	15	0.01	0.98333	0.95133	0.95383	0.75022
0.1	15	0.05	0.92500	0.98800	0.94959	1.16835
0.2	15	0.05	0.96433	0.97367	0.87928	0.81043
0.3	15	0.05	0.97033	0.94933	0.83507	0.67550
0.4	15	0.05	0.95800	0.92067	0.83382	0.73789
0.2	15	0.10	0.95233	0.92067	0.71398	0.57282
0.3	15	0.10	0.95433	0.91067	0.69647	0.55067
0.4	15	0.10	0.95200	0.83500	0.65522	0.46074
0.1	30	0.01	1.00000	0.99967	0.79989	0.73312
0.2	30	0.01	0.99767	0.99667	0.82135	0.68646
0.3	30	0.01	0.99533	0.98833	0.80516	0.63780
0.4	30	0.01	0.99433	0.98167	0.76184	0.59005
0.5	30	0.01	0.99400	0.96867	0.67756	0.49657
0.6	30	0.01	0.99167	0.94500	0.57519	0.40045
0.7	30	0.01	0.98967	0.91200	0.44465	0.27996
0.1	30	0.05	0.96900	0.98567	0.64870	0.63591
0.2	30	0.05	0.97867	0.97333	0.66055	0.59177
0.3	30	0.05	0.98000	0.94533	0.61955	0.48249
0.4	30	0.05	0.97600	0.91800	0.57706	0.43160
0.1	30	0.10	0.96267	0.97633	0.53021	0.51577
0.2	30	0.10	0.96100	0.93867	0.54511	0.46834
0.3	30	0.10	0.96133	0.87933	0.51242	0.38011
0.4	30	0.10	0.94400	0.86567	0.49921	0.39224

The interval based on the χ^2 distribution given by (2.8) showed consistently better results in terms of shorter average interval length compared to the interval based on the standard normal distribution given by (2.6), especially for larger sample sizes for various intraclass correlation coefficient values. The average lengths and coverage probability of both intervals are presented in Table 1. The interval based on the χ^2 distribution is recommended on the basis of shorter average interval length. The coverage probability of the interval based on the χ^2 distribution is competitive with the coverage probability of the interval based on the standard normal distribution.

Real Data Example

Two intervals using real life data collected from Srivastava and Katapa (1986) were compared. The real life data presented in Srivastava and Katapa (1986) is shown in Table 2.

Table 2: Values of Pattern Intensity on Soles of Feet in Fourteen Families

Sample	Family #	Mother	Father	Siblings
A	12	2	4	2, 4
A	10	5	4	4, 5, 4
A	9	5	5	5, 6
A	1	2	3	2, 2
A	4	2	4	2, 2, 2, 2, 2
A	5	6	7	6, 6
A	8	3	7	2, 4, 7, 4, 4, 7, 8
A	3	2	3	2, 2, 2
A	6	4	3	4, 3, 3
A	14	2	3	2, 2, 2
A	7	4	3	2, 2, 3, 6, 3, 5, 4
A	2	2	3	2, 3
A	11	5	6	5, 3, 4, 4
A	13	6	3	4, 3, 3, 3

The data is first transformed by multiplying each observation vector by Helmert's orthogonal matrix Q, where

$$\underset{p_i \times p_i}{Q} = \begin{bmatrix} \frac{1}{\sqrt{p_i}} & \frac{1}{\sqrt{p_i}} & \frac{1}{\sqrt{p_i}} & \cdots & \frac{1}{\sqrt{p_i}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{\sqrt{p_i(p_i-1)}} & \frac{1}{\sqrt{p_i(p_i-1)}} & \frac{1}{\sqrt{p_i(p_i-1)}} & \cdots & -\frac{(p_i-1)}{\sqrt{p_i(p_i-1)}} \end{bmatrix},$$

which results in transformed vectors $\underset{\sim}{u}_i$ for $i = 1,2,...,k$, here, k = 14. Srivastava's formula given by (2.3) is used to compute the intraclass correlation coefficient and variance. The computed values of the intraclass correlation coefficient and variance are $\hat{\rho} = 0.8118$ and $\hat{\sigma}^2 = 8.8578$. Using formulae (2.6) and (2.8) to obtain the lengths of the 95%, 99% and 90% confidence intervals for the intraclass correlation coefficient results in the following:

- Length of 90% confidence interval based on Z-distribution = 0.26519
- Length of 90% confidence interval based on χ^2 - distribution = 0.16096
- Length of 95% confidence interval based on Z-distribution = 0.31599
- Length of 95% confidence interval based on χ^2 - distribution = 0.19644
- Length of 99% confidence interval based on Z-distribution = 0.41528
- Length of 99% confidence interval based on χ^2 - distribution = 0.27388

It is observed that the length of the 95%, 99% and 90% confidence intervals based on the χ^2 distribution(using formula 2.8) is shorter than the length of the 95%, 99% and 90% confidence intervals respectively based on standard normal distribution (using formula 2.6).

References

Bhandary, M., & Alam, M. K. (2000). Test for the equality of intraclass correlation coefficients under unequal family sizes for several populations. *Communications in Statistics-Theory and Methods, 29*(*4*), 755-768.

Donner, A., & Bull, S. (1983). Inferences concerning a common intraclass correlation coefficient. *Biometrics, 39*, 771-775.

Donner, A., & Zou, G. (2002). Testing the equality of dependent intraclass correlation coefficients. *The Statistician*, *51*(*3*), 367-379.

Gokhale, D. V., & SenGupta, A. (1986). Optimal tests for the correlation coefficient in a symmetric multivariate normal population. *Journal of Statistical Planning Inference*, *14*, 263-268.

Huang, W., & Sinha, B. K. (1993). On optimum invariant tests of equality of intraclass correlation coefficients. *Annals of the Institute of Statistical Mathematics*, *45*(*3*), 579-597.

Konishi, S. (1985). Normalizing and variance stabilizing transformations for intraclass correlations. *Annals of the Institute of Statistical Mathematics*, *37*, 87-94.

Konishi, S., & Gupta, A. K. (1989). Testing the equality of several intraclass correlation coefficients. *Journal of Statistical Planning Inference*, *21*, 93-105.

Rao, C. R. (1973). *Linear statistical inference and its applications*. NY: Wiley.

Rosner, B., Donner, A., & Hennekens, C. H. (1977). Estimation of intraclass correlation from familial data. *Applied Statistics*, *26*, 179-187.

Rosner, B., Donner, A., & Hennekens, C. H.(1979). Significance testing of interclass correlations from familial data. *Biometrics*, *35*, 461-471.

SenGupta, A. (1988). On loss of power under additional information – an example. *Scandinavian Journal of Statistics*, *15*, 25-31.

Scheffe, H. (1959). *The analysis of variance*. NY: Wiley.

Srivastava, M.S. (1984). Estimation of interclass correlations in familial data. *Biometrika*, *71*, 177-185.

Srivastava, M. S., & Katapa, R. S. (1986). Comparison of estimators of interclass and intraclass correlations from familial data. *Canadian Journal of Statistics*, *14*, 29-42.

Srivastava, M. S., & Keen, K. J. (1988). Estimation of the interclass correlation coefficient. *Biometrika*, *75*, 731-739.

Velu, R., & Rao, M. B. (1990). Estimation of parent-offspring correlation. *Biometrika*, *77*(*3*), 557-562.

Young, D., & Bhandary, M. (1998). Test for the equality of intraclass correlation coefficients under unequal family sizes. *Biometrics*, *54*(*4*), 1363-1373.