MISCELLANEOUS NOTES

TEST OF SIGNIFICANCE FOR INTRA-CLASS CORRELATION WHEN FAMILY SIZES ARE NOT EQUAL.

By R. P. BHARGAVA

Statistical Laboratory, Calcutta.

Introduction.

The test of significance for the intra-class correlation when family sizes are equal was given by Fisher in his book "Statistical Methods for Research Workers".

In practice, we often meet with cases where family sizes are unequal. Thus, if we consider a certain character among brothers in a family, it often happens that the number of brothers in the different families is not the same. In such cases, some of the problems that arise are,

(1) test of significance, that is, whether population correlation coefficient is equal to any assigned value ρ

(2) to find the fiducial limits for ρ

(3) to determine the best estimate of population correlation ρ.

(4) to get a suitable estimate of population mean m.

(5) to estimate intra-class correlation from family means of two unequal family sizes, the individual observations of the family being not known.

2. Tests of significance.

Let us suppose there are n families and let the k_i individuals of the ith family give the observations

$$x_{i1}, x_{i2}, \ldots x_{ik_i} \quad (i = 1, 2, \ldots n)$$

We consider $x_{i1}, x_{i2}, \ldots, x_{ik_i}$ to be a random observation from a k_i variate normal population in which each of the variates has the same mean m, and the same standard deviation σ, the correlation between any two variates being equal to ρ. It may be noted that as m, σ and ρ are not functions of i, each family has the same mean m, the same standard deviation σ and the same correlation coefficient ρ.

Let $\bar{x}_i$ be the sample mean of the ith family. The probability of occurrence of the observation in the i-th family is given by

$$\text{Const.} \cdot e^{-\frac{1}{2\sigma^2(1-\rho)}\left[\sum_{j=1}^{k_i}(x_{ij}-m)^2-\frac{\rho}{1+k_i-1\,\rho}\sum_{j,j'=1}^{k_i}(x_{ij}-m)(x_{ij'}-m)\right]}\prod_{j=1}^{k_i}dx_{ij},$$

where $\text{Const.} = (\sqrt{2\pi}\sigma)^{-k_i}(1-\rho)^{-(k_i-1)/2}(1+k_i-1\,\rho)^{-1/2}$

This can be written as

$$\text{Const.} \cdot e^{-\frac{1}{2\sigma^2(1-\rho)}\left[\sum_{j=1}^{k_i}(x_{ij}-\bar{x}_i)^2+\frac{(1-\rho)}{1+k_i-1\,\rho}k_i(\bar{x}_i-m)^2\right]}\prod_{j=1}^{k_i}dx_{ij}; \qquad (2\cdot1)$$

Hence the joint distribution of the sets of observations of our sample can be written as

$$\text{Const.} \cdot e^{-\frac{1}{2\sigma^2(1-\rho)}\left[\sum_{i,j}(x_{ij}-\bar{x}_i)^2+(1-\rho)\sum\frac{k_i}{1+k_i-1\,\rho}(\bar{x}_i-m)^2\right]}\Pi dx_{ij} \qquad (2\cdot2)$$

Let
$$\sigma^2{}_i=\frac{k_i(1-\rho)}{1+k_i-1\,\rho}, \quad \Sigma\sigma^2{}_i\bar{w}_i=\Sigma\sigma^2{}_i\bar{x} \qquad (2\cdot3)$$

Then (2.2) can be written as

$$\text{Const.}\ e^{-\frac{1}{2\sigma^2(1-\rho)}\left[\Sigma(x_{ij}-\bar{x}_i)^2+\Sigma c_i^2(\bar{x}_i-\bar{x})^2+\Sigma c_i^2(\bar{x}-m)^2\right]}\prod dx_{ij} \tag{2·4}$$

From this it is easily seen that

$$F=\frac{\Sigma c_i^2(\bar{x}_i-\bar{x})^2}{(n-1)}\cdot\frac{\Sigma(k_i-1)}{\Sigma(x_{ij}-\bar{x}_i)^2} \tag{2·5}$$

is distributed in the usual F distribution with $n_1=n-1$ and $n_2=(\Sigma k_i-1)$ degrees of freedom.

This F statistic can be used for testing whether our families (sizes being unequal) came from multivariate normal populations having population correlation ρ. Thus as a particular case if we desire to know, whether our sample of n families with unequal sizes came from multivariate normal populations having $\rho=\cdot5$, all that we have to do is, to calculate c_i^2 and $\bar{x}$ by putting $\rho=\cdot5$, and then calculate the value of F by using (2·5) and consult the F tables.

Thus we can with the help of the F distribution answer any of the following questions (*i*) whether population correlation is significantly different from ·5 (say), (*ii*) whether the population correlation coefficient is significantly greater than ·5 and (*iii*) whether the population correlation coefficient is significantly less than ·5.

These questions are the same as those which arise in testing the significance of intra-class correlation for families of equal size. But it is useful to note that for answering question (*i*) we have to consider both the tail-ends of the F distribution, while for question (*ii*) we have only to consider the tail-end on the right and for (*iii*) the tail-end on the left.

It is of interest to note that (*i*) Fisher's Z-test for testing the significance of intra-class correlation for equal family size and (*ii*) Fisher's technique of analysis of variance for testing the significance of intra-class correlation when family sizes may not be equal follow also from this test.

(*i*) Thus if in (2·5), we put $k_1=k_2=\ldots=k_n=k$ say then F becomes

$$F=\frac{1-\rho}{1+k-1\,\rho}\cdot\frac{\Sigma k(\bar{x}_i-\bar{x})^2}{n-1}\cdot\frac{n(k-1)}{\Sigma(x_{ij}-\bar{x}_i)^2}$$

$$=\frac{1-\rho}{1+k-1\,\rho}\cdot\frac{1+\overline{k-1}\,r}{1-r}\cdot\frac{n}{n-1} \tag{2.7}$$

where $\bar{x}$ is the grand mean and r is the sample intra-class correlation. And it is easily seen that Fisher's Z is the same as $\frac{1}{2}\log_e F$.

(*ii*) Again putting $\rho=0$, F is equal to

$$F=\frac{\Sigma k_i(\bar{x}_i-\bar{x})^2}{n-1}\cdot\frac{\Sigma(k_i-1)}{\Sigma(x_{ij}-\bar{x}_i)^2}$$

which is the same as that used for testing the significance of between and within variation in analysis of variance.

3. FIDUCIAL LIMITS OF ρ

Next we consider the problem of ascertaining fiducial limits of population ρ for the case of unequal family sizes. The usual procedure for finding the lower fiducial limit is to first find (say) the 5 % point of F corresponding to $n_1=n-1$, $n_2=\Sigma(k_i-1)$. Let us call this point as F_s then from the equality

$$\frac{\Sigma c_i^2(\bar{x}_i-\bar{x})^2}{n-1}\cdot\frac{\Sigma(k_i-1)}{\Sigma(x_{ij}-\bar{x}_i)^2}=F_s$$

we solve for ρ. From this we get that $\rho>\rho_s$ in 95 % cases. This provides us with the lower limit of ρ. But in our case, the solution does not come out very easily, hence we shall follow the method of trial and error and graduation. Let us calculate values of F of (2.5) for various values of ρ say $\rho=$ .1, .2, .3, .4, .5, .6, .7, .8, .9.

This can be presented in a tabular form as follows :—

ρ	$F(\rho)$	$F_5\ (n_1, n_2)$
·1	$1 < F(\rho=\cdot1)$	$> \quad F_5\ (n_1=n-1,\ n_2=\Sigma(k_i-1)$
·2	$1 < F(\rho=\cdot2)$	$> \quad F_5\ (n_1=n-1,\ n_2=\Sigma(k_i-1)$
·3	$1 < F(\rho=\cdot3)$	$< \quad F_5\ (n_1=n-1,\ n_2=\Sigma(k_i-1)$
·4	$1 = F(\rho=\cdot4)$	
·5	$1 > F(\rho=\cdot5)$	$> \quad \frac{1}{F_5}\ (n_1=\Sigma k_i-1,\ n_2=n-1)$
·6	$1 > F(\rho=\cdot6)$	$> \quad \frac{1}{F_5}\ (n_1=\Sigma k_i-1,\ n_2=n-1)$
·7	$1 > F(\rho=\cdot7)$	$< \quad \frac{1}{F_5}\ (n_1=\Sigma k_i-1,\ n_2=n-1)$
·8	$1 > F(\rho=\cdot8)$	$< \quad \frac{1}{F_5}\ (n_1=\Sigma k_i-1,\ n_2=n-1)$
·9	$1 > F(\rho=\cdot9)$	$< \quad \frac{1}{F_5}\ (n_1=\Sigma k_i-1,\ n_2=n-1)$

In the above table $F_5 = (n_1, n_2)$ in column (3) indicate the 5 % point of F table corresponding to n_1 and n_2 degrees of freedom. Column (2) gives the value of F corresponding to particular values of ρ.

In the above hypothetical table it is clear that the fiducial limits of ρ are given by

$$\cdot2 < \rho < \cdot7$$

This gives a rough idea, and some more numerical vlues of ρ between .2 and .3, and .7 and .6 will (with the help of graduation or otherwise) give closer approximation for the fiducial limits.

4. Best estimate of ρ

Next comes the problem of obtaining the best estimate of ρ. The value of ρ in the above table which makes F equal to unity is the best estimate of ρ, in the sense of maximising the likelihood of $z = \frac{1}{2}\log_e F$.

Thus from the above table we see that $\rho = .4$ gives $F = 1$. Thus our point estimate of ρ from the given sample comes as .4. This may be called unbiassed sample intra-class correlation r.

It is clear from the way we have proceeded that our method of determining fiducial limits and point estimate of ρ ; coincide with the usual methods when family sizes are equal.

5. Estimate of m

From (4) it is clear that the distribution of $\bar{x}$ is

$$\text{Const.}\ e^{-\frac{1}{2\sigma^2}\Sigma\frac{k_i}{1+k_i-1\rho}(\bar{x}-m)^2}\,d\bar{x} \qquad (5\cdot1)$$

where

$$\bar{\bar{x}} = \frac{\Sigma\frac{k_i\bar{x}_i}{1+k_i-1\rho}}{\Sigma\frac{k_i}{1+k_i-1\rho}}$$

Thus $\bar{\bar{x}}$ is a sufficient statistic to estimate m. Thus if population ρ is known, we can calculate the best estimate of m. But, as is usual, if ρ is not known, we can estimate m, substituting for ρ, the sample intra-class correlation r, and this gives us a consistent estimate for m. Thus

$$\bar{\bar{x}} = \frac{\Sigma\frac{k_i\bar{x}_i}{1+k_i-1\,r}}{\Sigma\frac{k_i}{1+k_i-1\,r}}$$

is a consistent sample estimate of m. The use of this very often arises in crop surveys, budget enquiries etc

8. ESTIMATION OF INTRA-CLASS CORRELATIONS FROM TWO SETS OF FAMILIES

Let us suppose that there are n families of size k and n' families of size k', $(k'<k)$ coming respectively from k and k' variate normal populations having the same mean m, the same standard deviation σ, and the same correlation ρ. The individual readings in a family are not known but the family means (or totals) are known. The problem is to get an estimate of ρ.

Let $\bar{x}_i$ be the mean of the ith family of size k ; and let $\bar{x}'_{i'}$ be the mean of the i'th family of size k'. Then

$$n\bar{x}=\sum_{i=1}^{n}\bar{x}_i \text{ and } n'\bar{x}'=\sum_{i'=1}^{n'}\bar{x}'_{i'} \tag{6·1}$$

also $$(n-1)s_1^2=\Sigma k(\bar{x}_i-\bar{x})^2, \text{ and } (n'-1)s_2^2=\Sigma k'(\bar{x}'_{i'}-\bar{x}')^2 \tag{6·2}$$

Then from (2.4), it is easily seen that

$$F=\frac{1+(k'-1)\rho}{1+(k-1)\rho}\cdot\frac{s_1^2}{s_2^2} \tag{6·3}$$

is distributed in the usual F distribution with $n_1=n-1$ and $n_2=n'-1$ degrees of freedom.

From this, very often we can get an estimate of ρ. The numerical value of ρ, lying between $\frac{-1}{k-1}<\rho<1$ which makes F of (6.3) equal to unity, is the best estimate of ρ from family means.

It is further seen that whenever, the ratio of the two mean squares s_1^2/s_2^2 is $>k/k'$ no such numerical value of ρ lying in $\frac{-1}{k-1}<\rho<1$ exists which will make F of (6.3) equal to unity ; but on the other hand, as is usually the case, if the ratio s_1^2/s_2^2 is $<k/k'$ such a numerical value of ρ always exists. And in such cases this estimate is given by

$$r=-1/(k-1)+\frac{s_1^2}{s_1^2-s_2^2}(k-k')$$

SUMMARY.

(1) We have given a test of significance (unknown) for testing whether our sample of n families, family size being unequal, has come from multivariate normal populations having population correlation coefficient equal to ρ.

(2) We have given a method of deriving the fiducial limits of ρ.

(3) We have given a new definition for intra-class correlation when family size is unequal. This definition reduces to the usual definition for the equal family size and is the best estimate of ρ.

(4) We have derived a consistent estimate for the population mean m.

(5) We have discussed a connected problem of estimation of intra-class correlaton from family means of two unequal family sizes when individual readings of the family are not known but only family means or totals are known.

REFERENCES.

(1) FISHER, R. A. (1944) Statistical Methods for Research workers, 9th Edition.

(2) ———— (1938) Statistical Theory of Estimation, Calcutta University Readership Lectures.

(3) KENDALL, M. G. (1943) The Advanced Theory of Statistics, Vol. I.

Paper received : 4 April 1946.