NEW LIGHT ON TRANSFORMATIONS INVOLVING CORRELATIONS*

BY

M. SANKARAN

The Institute of Mathematics and Precision Technology, Dhinoor, Bangalore-560032

(Received : September, 1981)

SUMMARY

The paper throws light and added insight on the transforms involving correlations of the author and others based on the theoretical results of Borges [1]. The familiar Z transformation of the multiple correlation coefficient results in definite improvement in its accuracy. The arc-sine transformation V of the author comes out very favourably, as it should be, for the multiple correlation coefficient as well like the Z transform.

INTRODUCTION

Works of Sankaran [7] [8] [9] on transformations involving correlation coefficient are extended and placed in proper perspective based on the theoretical investigation of Rudolf Borges [1]. It provides statistical algorithms throwing fresh light and additional insight on the transforms enabling one to see the one connected whole. New applications include those of the multiple correlation coefficient and serial correlation coefficient. Fisher's Z transform of the correlation coefficient is applied to the multiple correlation coefficient resulting in definite improvement in its accuracy. Arc-sine transformation of the correlation coefficient when applied to the multiple correlation coefficient compares very favourably with Z.

2. ALGORITHMS ON BORGES'S RESULT

Following Borges, let $f_n(\hat{\theta}; \theta)$ be the *p.d.f.* of the statistic $\hat{\theta}$ involving the distribution parameter θ (whose estimate is $\hat{\theta}$) and

*In the present paper, correlations refer to the product moment correlation, multiple correlation and serial correlation coefficients.

another parameter n usually the sample size. Let, further, logarithm of $f_n(\hat{\theta};\theta)$ have the representation (U, V and W are, generally, functions of $\hat{\theta}$ and θ)

$$\log f_n\hat{\theta};\theta)=-nU+V+n^{-1}W+\tfrac{1}{2}\log[nU''(\theta)/2\pi] \qquad ...(1)$$

in some non-degenerate interval $|\hat{\theta}-\theta|\leqslant C_1$, $c_1>0\forall\hat{\theta}$ and n, θ being fixed, $n\in N$ with inf $N>0$. As $n\to\infty$, it is well known that $\hat{\theta}$ is asymptotically normal with mean θ and variance $\sigma^2(\theta)/n$ (say).

Further assume, as required in the following sequal, that U is differentiable continuously thrice, V once differentiable continuously. All primes denote differentiation *w.r.t.* $\hat{\theta}$. Now state the main result on refinement of normalising transforms with an error of order $\frac{1}{n}$ in the form of *statistical algorithms* which proceed sequentially and which process terminates the moment the desired objective is realised with either variance stabilised or with improved order of normalisation.

Algorithm I: Write $\log f_n(\hat{\theta};\theta)$ in the form (I) with proper choice of U, V and W, U, V and W being certainly not arbitrary. In the sequal usually W is not required explicitly.

Algorithm II: Compute U', U'', U''' and V'. Compute also $n\mho''(\hat{\theta})]^{-1/2}$. At this stage, the familiar variance stabilising transformations are obtained. Borges does not mention this expicitly. With the computation of $\sigma_{\hat{\theta}}=V'(\theta)[n\mho''(\theta)]^{-1/2}$ the required constanc y of σ_Ψ independent of θ, leads to the variance stabilising transform $\Psi(\hat{\theta})$. It so, terminate the process, otherwise proceed to.

Algorithm III: Compute $\dfrac{\Psi''(\hat{\theta})}{\Psi'(\hat{\theta})}=\tfrac{1}{3}\dfrac{U'''(\hat{\theta})}{U''(\theta)}$. Hence comput $\Psi(\hat{\theta})$. Terminate the process here. Now either at the stage of Algorithm *II* or at this stage, proceed to

Algorithm IV: Compute $k(\hat{\theta})=\dfrac{\Psi''(\hat{\theta})-\Psi'(\hat{\theta})V'(\hat{\theta})}{U''(\hat{\theta})}$

and lastly

Algorithm V: Compute the transformations

$$y_1=\frac{\Psi(\hat{\theta})+n^{-1}k(\hat{\theta})-\Psi(\theta)}{\sigma_\Psi} \qquad ...(2)$$

and $$y_2=\frac{\Psi(\hat{\theta})-\Psi(\theta)+n^{-1}k(\theta)}{\sigma_\Psi} \qquad \ldots(3)$$

Notice y_1 and y_2 differ only in the term $k(\theta)$. σ_Ψ is, as stated in

Algorithm III' Do not consider here the transformation y_3 of Borges.

According to Borges, the relative error is of order $\frac{1}{n}$ for the transformations y_1 and y_2 of (2) and (3). It is worth recalling now that $U' \sim \frac{dlogf}{d\hat{\theta}}$ and $U'' \sim \frac{d_2 logf}{d\hat{\theta}^2}$ to order n. From general result on the bound for the variance of any statistic (Sankaran [6]), $E(U')=0$, Var $U=E\left(\frac{d\,log\,f}{d\hat{\theta}}\right)^2=E(-U'')=H(\theta)$. Such linking of scattered results is illuminating.

Consider the applications of the algorithms to the theoretical (sampling) distributions of the multiple correlation and serial correlation coefficients.

3. Multiple correlation coefficient

The distribution of the multiple correlation coefficient was first derived by Fisher [2] in the form

$$\left.\begin{aligned} f_n(R^2;\,\rho)=C(1-\rho^2)^{\frac{n}{2}}\;(1-R^2)^{\frac{1}{2}(n-p-2)}(R^2)^{(\frac{1}{2}p-2)} \\ \int_0^{\pi}\int_{-\infty}^{\infty}\frac{sin^{p-2}\phi\,dz\,d\phi}{(coshz-\rho R cos\phi)^n} \end{aligned}\right\} \qquad \ldots(4)$$

with $$C=\frac{n-1}{2}\frac{1}{B\left(\frac{n-1}{2},\,\frac{p-1}{2}\right)}$$

The double integral can be alternately expressed in terms of the $I_n(x)$ function involving the hypergeometric series. For these algebraic and other details refer to the paper of Yoong-Sin-Lee [4]. Fisher intuitively suggested his Z transform for the multiple correlation coefficient as well. Lee, while working on approximations to the

distribution of the multiple correlation coefficient, noted only the inadequacy of the Z transform for ρ large.

Now, from Algorithm *II* (to order $n-p$ rather than n)

$$\left.\begin{aligned} U&=\log(1-\rho R)-\tfrac{1}{2}\log(1-R^2)-\tfrac{1}{2}\log(1-\rho^2)\\ V&=(-p-\tfrac{1}{2})\log(1-\rho R)-\log(1-R^2)\\ &\quad+\left(\frac{p}{2}-1\right)\log(1-\rho^2). \end{aligned}\right\} \quad ...(5)$$

The rest of $\log f_n(R^2;\rho)$, being absorbed in the function W.

$$\text{Hence } U'=\frac{\rho}{1-\rho R}+\frac{R}{1-R^2}\sim\frac{R-\rho}{1-\rho R},$$

$$\left.\begin{aligned} U''&=-\frac{\rho^2}{(1-\rho R)^2}+\frac{1}{1-R^2}+\frac{2R^2}{(1-R^2)^2},\\ U'''&=-\frac{2\rho^3}{(1-\rho R)^3}+\frac{6R}{(1-R^2)^2}+\frac{8R^3}{(1-R^2)^3}, \end{aligned}\right\} \quad ...(6)$$

$$\text{and } V'=(p-\tfrac{1}{2})\left\{\frac{\rho}{1-\rho R}+\frac{2R}{(1-R_2)}\right.=\tfrac{1}{2}\frac{R-\rho}{1-\rho R}+(p+\tfrac{3}{2})\frac{R}{1-R^2}$$

$$\text{so that } U'(R)=0,\ U''(R)=\frac{1}{(1-R^2)^2},\ U'''(R)=\frac{6R}{(1-R^2)^3},\ \Big\} \quad ...(7)$$

$$\text{and } V'(R)=(p+\tfrac{3}{2})\frac{R}{1-R^2}.$$

Notice here U' is essentially Nair's transform $U=\dfrac{R-\rho}{1-\rho R}$ considered by the author in his series of investigations on the correlation coefficient (Sankaran [5] [7] [9] V' also involves this U to order n. Also $[(n-p)U''(R)]^{-1/2}\simeq(1-R^2)/\sqrt{n-p}$.

Now here itself from Algorithm *II* we have $\Psi'(R)=\dfrac{1}{1-R^2}$ and we have the Fisher's Z transform $\Psi(R)=\frac{1}{2}\log\dfrac{1+R}{1-R}$ for the multiple correlation coefficient as well with $\sigma_\Psi=\dfrac{1}{\sqrt{n-p}}$.

$$\text{Also with } U'\sim\frac{R-\rho}{1-\rho R}=U$$

$$\text{We have } E(U')=E\left(\frac{d\log f}{dR}\right)=0,\ \text{Var } U=E\left(\frac{d\log f}{dR}\right)^2$$

$$=E\left(-\frac{d^2\log f}{dR^2}\right)=\frac{1}{n-p}.$$

Further, if we consider the are-sine transform

$$V=\sin^{-1}\left(\frac{R-\rho}{1-\rho R}\right)$$

it is easily verified that Var $V=\frac{1}{n-p}$.

So we have the interesting fact that the transformation

$$V=\sin^{-1}\left(\frac{R-\rho}{1-\rho R}\right)$$

is a contender for the multiple correlation coefficient as well as in the case of the correlation coefficient. V is not to be confused with the function V of the Algorithm. It is easily verified that for both the transformations V and Z,

$$k(R)=(-p+\tfrac{1}{2})R \text{ and } \sigma_\Psi=\frac{1}{\sqrt{n-p}} \qquad \ldots(8)$$

Substitution of these in the y_1 and y_2 transformations and after simplification leads to

$$\left.\begin{aligned} z_1&=\sqrt{n-p}\left[\tfrac{1}{2}\log\frac{1+R}{1-R}+\frac{(-p+\frac{1}{2})}{n}\rho-\tfrac{1}{2}\log\frac{1+\rho}{1-\rho}\right],\\ \text{and}\quad z_2&=\sqrt{n-p}\left[\tfrac{1}{2}\log\frac{1+R}{1-R}+\frac{(-p+\frac{1}{2})}{n}R-\tfrac{1}{2}\log\frac{1+\rho}{1-\rho}\right].\end{aligned}\right\} \qquad \ldots(9)$$

Similar refinements to the transform $V=\sin^{-1}\left(\frac{R-\rho}{1-\rho R}\right)$ lead to

$$\left.\begin{aligned} v_1&=\sqrt{n-p}\left[V+\frac{(-p+\frac{1}{2})}{n}\rho\right],\\ \text{and}\quad v_2&=\sqrt{n-p}\left[V+\frac{(-p+\frac{1}{2})}{n}R\right].\end{aligned}\right\} \qquad \ldots(10)$$

In actual numerical calculations as shown below V only compares favourably with Z and $U=\frac{R-\rho}{1-\rho R}$ *fails* here also as noted by the author in the case of the bivariate normal distribution.

Lee assumes that B^2 is distributed as non-central $\chi'^2(p, B^2)$ where $B=\sqrt{n-p}\tanh^{-1}R$ and $\beta=\sqrt{n-p}\tanh^{-1}\rho$.

Lee reports that the Z transform based on this approximations fails in accuracy for large ρ. Similar first order approximations based on V assumes the A^2 is distributed as a χ^2 with $\mid$ *d.f.*, where

$A=\sqrt{n-p}\ V$. These two transforms A and B are comparable for ρ small or n large. Our main interest at present is ρ large and n moderately sized. Other variants of the transforms of (9) and (10)

$$\left.\begin{aligned} z_3&=\sqrt{n-p}\left[Z+\frac{(-p+\frac{3}{2})}{n-p}\rho-\zeta\right],\\ z_4&=\sqrt{n-p}\left[Z+\frac{(-p+\frac{3}{2})}{n-p}R-\zeta\right],\end{aligned}\right\}^{*} \quad \ldots(11)$$

and

$$\left.\begin{aligned} v_3&=\sqrt{n-p}\left[V+\frac{(-p+\frac{3}{2})}{n-p}\rho\right],\\ v_4&=\sqrt{n-p}\left[V+\frac{(-p+\frac{3}{2})}{n-p}R\right].\end{aligned}\right\} \quad \ldots(12)$$

In actual numerical calculations, all these four pairs of transforms are equally good and in the Tables below only transformations v_1, v_3 and z_1, z_3, whose middle term involves the known population multiple correlation coefficient ρ are considered.

TABLE I

Accuracy of approximations probability integral of R for $n=24$, $p=5$, $\rho=0.8$. 10^4 x | approximate-true | values are recorded in cols 3-7. Columns 2 and 7 are taken from Lee.

x	$P(R\leqslant x)$	v_1	v_3	z_1	z_3	*Actual Z*
0.5	0.0006	11	11	6	16	8
0.6	0.0054	34	37	22	26	54
0.7	0.0419	83	86	57	68	272
0.8	0.2609	41	5	41	5	761
0.9	0.8586	322	300	236	188	388

* $Z=\frac{1}{2}\log\frac{1+R}{1-R}$; $\zeta=\frac{1}{2}\log\frac{1+\rho}{1-\rho}$.

TABLE II

Accuracy of approximations to probability integral of R for $m=14$, $p=3$, $\rho=0.9$. 10^4x | approximate-true | values are recorded in cols. 3-7. Columns 2 and 7 are taken from Lee

x	$P(R\leqslant x)$	v_1	v_3	z_1	z_3	*Actual Z*
0.55	0.0004	7	9	0	2	4
0.60	0.0010	9	10	4	5	9
0.65	0.0023	14	31	1	10	24
0.70	0.0057	20	32	2	22	57
0.75	0.0146	23	82	18	51	135
0.80	0.0397	11	133	13	105	300
0.85	0.1131	61	189	73	178	605
0.90	0.3235	263	186	263	186	947
0.95	0.7808	437	44	357	28	625

An examination of the Tables I and II reveals that approximations are definitely better than before and the refinements based on V and Z are comparable. And *V has the theoretical advantage that its distribution is independent of the parental stewness in non-normal populations* as the investigations on V and U clearly proved. Fisher's Z does not have this property (Sankaran [7], [9]).

4. Serial Correlation Coefficient

The distribution of the serial correlation coefficient r in the circular case when the underlying process is the Markov process

$x_{n+1}=\rho x_n+\epsilon_n$, is given by

$$f_n(r;\rho)=\frac{1}{B\left(\frac{1}{2},\frac{n+1}{2}\right)}\frac{(1-r^2)^{\frac{n-1}{2}}}{(1-2\rho r+\rho^2)^{\frac{n}{2}}}$$

Here $U=\frac{1}{2}\log(1-2\rho r+\ ^2)-\frac{1}{2}\log(1-r^2)$

and $V=-\frac{1}{2}\log(1-r^2)$ so that $V^1=\frac{r}{1-r^2}=V^1(r)$.

Also $U''(r)=\frac{1}{1-r^2}$, $U'''(r)=\frac{6r}{(1-r^2)^3}$, .. (14)

and $[n\ U''(r)]^{-1/2}=\frac{\sqrt{1-r}}{n}$.

From Algorithm II, $\psi(r)=\sin^{-1} r$, $\sigma_\psi=\frac{1}{\sqrt{n}}$ ...(15)

If we stop here which is valid, we have the angular transformation $\sin^{-1} r$ of Jenkins [3]. The interesting fact to note here is that

$$k(r)=\left[\frac{r}{(1-r^2)^{3/2}}-\frac{r}{(1-r^2)^{3/2}}\right](1-r^2)=0; \quad \ldots(16)$$

no further improvement is possible to the angular transformation of Jenkins. This result along with its derivation is new.

REFERENCES

[1] Borges, Rudolf (1971) : Derivation of Normalising Transformations with an error of order $\frac{1}{n}$. *Sankhyā*, **33**, 441-60.

[2] Fisher, R.A. (1928) : The general sampling distribution of the multiple correlation coefficient. *Proc. Roy. Socy* A, **121**, 654-73.

[3] Jenkins, G.M. (1954) : An Angular transform of the serial correlation coefficient *Biometrika*, **41**, 261-5.

[4] Lee, Yoong-Sin (1971) : Some Results on the sampling Distribution of the Multiple Correlation Coefficient *J.R. Statist. Socy.* B, **33**, 117-29.

[5] Sankaran, Munuswamy (1958) : On Nair's Transform of the correlation Coefficient *Biometrika* **5**, 567-71.

[6] Sankaran, Munuswamy (1964) : On an Analogue of Bhattacharya Bound, Biometrika **51** 268-70.

[7] Sankaran, Munuswamy (1973) : On Nair's Transform of the correlation coefficient *Sankhyā* B, **35**, 317-24.

[8] Sankaran M. (1958) : The probability integral of the correlation coefficient, 45th sessions of the Indian Science Congress Association : Proceedings Pt IV late abstract 2, 4-5.

[9] Sankaran M. (1969) : On Nair's transform of the correlation coefficient. 56th session of the ISCA Proc. Pt. III abstract 5, 30.