ON THE DISTRIBUTION OF THE RATIO OF VARIANCES OF TWO SAMPLES DRAWN FROM A GIVEN NORMAL BIVARIATE CORRELATED POPULATION.

By SUBHENDUSEKHAR BOSE, M.Sc.

Statistical Laboratory, Calcutta.

Introduction.

An experimental study of sampling distributions for statistics relating to a bi-variate normal correlated population is in progress in the Statistical Laboratory, Calcutta. During the course of this work the ratio of variances of the two correlated variates were formed and compared against the theoretical distribution of the ratio given by R. A. Fisher. The agreement was not satisfactory, and it was realized that, while Fisher's distribution was obtained on the assumption that the two variates were independent, in the present case they were actually correlated. It was therefore considered desirable to examine the question in greater detail.

1. *The distribution function of the ratio of variances obtained from two independent samples.*

If two samples drawn at random from a given normal bivariate population have variances s_1^2 and s_2^2, the distribution of

$$z = \frac{1}{2}\log_e \frac{s_1^2}{s_2^2}$$

has been given by Fisher (1924):

$$df(z) = \frac{2n_1^{\frac{n_1}{2}}\, n_2^{\frac{n_2}{2}}}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)} \cdot \frac{e^{n_1 z}}{(n_1 e^{2z}+n_2)^{\frac{n_1+n_2}{2}}}\, dz \qquad \dots \qquad \dots \qquad \dots \quad (1.1)$$

$$\text{where} \quad B\left(\frac{n_1}{2},\frac{n_2}{2}\right) = \frac{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}{\Gamma\left(\frac{n_1+n_2}{2}\right)}$$

and n_1 and n_2 are the degrees of freedom on which the estimates of s_1^2 and s_2^2 are based.

If $\omega = \frac{s_1}{s_2} = e^z$, the distribution function of ω is obtained in the form

$$df(\omega) = \frac{2n_1^{\frac{n_1}{2}}\, n_2^{\frac{n_2}{2}}}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)} \cdot \frac{\omega^{n_1-1}}{(n_1\omega^2 + n_2)^{\frac{n_1+n_2}{2}}}\, d\omega \qquad \dots \qquad \dots \quad (1.2)$$

Again, if $x = \dfrac{n_1 s_1^2}{n_1 s_1^2 + n_2 s_2^2} = \dfrac{n_1 \omega^2}{n_1 \omega^2 + n_2}$

the distribution of x is given by a very simple Eulerian form:

$$df(x) = \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot x^{\frac{n_1-2}{2}} (1-x)^{\frac{n_2-2}{2}} dx \qquad \ldots \qquad \ldots \qquad (1.3)$$

If s_1^2 and s_2^2 are two observed variances in two independent samples based on n_1 and n_2 degrees of freedom respectively, we can easily test whether these two samples belong to one normal population by calculating any one of the three statistics:

$$(1) \qquad z = \tfrac{1}{2} \log_e \frac{s_1^2}{s_2^2}$$

$$(2) \qquad \omega = \frac{s_1}{s_2}$$

$$(3) \qquad x = \frac{n_1 s_1^2}{n_1 s_1^2 + n_2 s_2^2}$$

and obtaining with the help of the distribution functions (1.1), (1.2) or (1.3) the probabilities of occurrence of a value of z (or ω or x as the case maybe) equal to or exceeding the observed value. If this probability is small, say less than ·01, we consider it reasonable for all practical purposes to assume that the variances s_1^2 and s_2^2 are not random estimates of some definite population variance.

Fisher has given tables of 5 p.c. and 1 p.c. values of z so that, if any observed z exceeds the 1 p.c. value, we take the two variances to be significantly different from each other. Corresponding tables of ω have been prepared by Mahalanobis (1932) for facility of calculation in using the above test of significance.

If two samples are drawn at random from two populations having variances σ_1^2 and σ_2^2 and if the variances as estimated from the samples are s_1^2 and s_2^2, it is easy to show that the distribution of

$$z' = \log_e \frac{s_1}{s_2} - \log_e \frac{\sigma_1}{\sigma_2}$$

is similar to that of z in (1·1). The distribution ot

$$\omega' = \frac{\omega}{\sigma_1 / \sigma_2}$$

is similar to that of ω in (1.2) and of

$$x' = \frac{\dfrac{n_1 s_1^2}{\sigma_1^2}}{\dfrac{n_1 s_1^2}{\sigma_1^2} + \dfrac{n_2 s_2^2}{\sigma_2^2}}$$

is similar to that of x in (1.3).

2. *Distribution of the Ratio of Variances from a Correlated Population.*

All the above distributions were obtained on the hypothesis that the two variates for which s_1 and s_2 are calculated from the two observed samples are independent. It is of some interest to investigate the form of the distribution of z, ω and x when the two samples are drawn at random from a given normal bivariate correlated population.

The joint distribution of the estimated values of the coefficient of correlation (r) and the two standard deviations (s_1 and s_2) was given by Fisher (1915) as follows :

$$df = C_0 \,.\, e^{-\frac{1}{2}\frac{s_1^2}{g_1^2}} \,.\, e^{-\frac{1}{2}\frac{s_2^2}{g_2^2}} \,.\, \left(\frac{s_1}{g_1} \,.\, \frac{s_2}{g_2}\right)^{n-2} \,.\, e^{hr} \,.\, (1-r^2)^{\frac{n-4}{2}} \,.\, ds_1 \,.\, ds_2 \,.\, dr \qquad \ldots \quad (2{\cdot}1)$$

where

$$g_1^2 \;=\; \sigma_1^2(1-\rho^2)/n', \qquad g_2^2 \;=\; \sigma_2^2(1-\rho^2)/n', \qquad h \;=\; \rho s_1 s_2/g_1 g_2$$

ρ, σ_1, σ_2 are the population values of the coefficient of correlation and standard deviations, n' is the size of sample and C_0 is a constant.

Integrating over the entire range of r, we obtain the frequency surface of s_1 and s_2

$$df = \quad y_0 \,.\, e^{-\frac{1}{2}\frac{s_1^2}{g_1^2}} \,.\, e^{-\frac{1}{2}\frac{s_2^2}{g_2^2}} \,.\, \left(\frac{s_1}{g_1} \,.\, \frac{s_2}{g_2}\right)^{n-2} \times \quad .$$

$$\left[1 + \frac{1}{(2n'-2)} \,.\, \frac{\rho^2}{1\,!} \,.\, \frac{s_1^2\, s_2^2}{g_1^2 g_2^2} + \frac{1}{(2n'-2)(2n'+2)} \,.\, \frac{\rho^4}{2\,!} \,.\, \frac{s_1^4\, s_2^4}{g_1^4 g_2^4} + \ldots\ldots\ldots\right.$$

$$\left. + \frac{1}{(2n'-2)(2n'+2)\ldots(2n'+4p-6)} \,.\, \frac{\rho^{2p}}{p\,!} \,.\, \frac{s_1^{2p}\, s_2^{2p}}{g_1^{2p} g_2^{2p}} + \ldots\ldots\ldots \right] ds_1 \,.\, ds_2 \quad \ldots \quad (2.2)$$

If $g_1 = g_2 = g$ *i.e.* the population variances of the two variates are identical

$$df = y_0 . e^{-\frac{1}{2}.\frac{s_1^2+s_2^2}{g^2}} \,.\, \left(\frac{s_1 s_2}{g^2}\right)^{n-2} . \times$$

$$\left[1 + \frac{1}{2n'-2} \,.\, \frac{\rho^2}{1\,!} \,.\, \frac{s_1^2\, s_2^2}{g^4} + \frac{1}{(2n'-2)(2n'+2)} \,.\, \frac{\rho^4}{2\,!} \,.\, \frac{s_1^4\, s_2^4}{g^8} + \ldots\ldots\ldots\right.$$

$$\left. + \frac{1}{(2n'-2)(2n'+2)\ldots..\;(2n'+4p-6)} \,.\, \frac{\rho^{2p}}{p\,!} \,.\, \frac{s_1^{2p} s_2^{2p}}{g^{4p}} \quad + \ldots\ldots \quad \right] \quad ds_1 \; ds_2$$

Let $\omega \;=\; \dfrac{s_1}{s_2}$, then for a given value of s_2, $ds_1 \;=\; s_2 d\omega$. Hence by substitution,

$$df = y_0 .\ e^{-\frac{1}{2}\frac{s_2^2(1+\omega^2)}{g^2}} \left(\frac{s_2^2}{g^2}\right)^{n'-2} .\ \omega^{n'-2} .$$

$$\left[1 + \frac{1}{(2n'-2)} \cdot \frac{\rho^2}{1\,!} \cdot \frac{s_2^4\,\omega^2}{g^4} + \frac{1}{(2n'-2)(2n'+2)} \cdot \frac{\rho^4}{2\,!} \cdot \frac{s_2^8\,\omega^4}{g^8} + \ldots\ldots \right.$$

$$\left. + \frac{1}{(2n'-2)\,(2n'+2)\,\ldots..(2n'+4p-6)} \cdot \frac{\rho^{2p}}{p\,!} \cdot \frac{s_2^{4p}\omega^{2p}}{g^{4p}} + \ldots\ \ldots\ \ldots \right] d\omega\ .\ s_2\, ds_2$$

$$= \quad y_0 .\ e^{-\frac{1}{2}\frac{s_2^2(1+\omega^2)}{g^2}}$$

$$\left[\left(\frac{s_2^2}{g^2}\right)^{n'-2}\left\{1 + \frac{1}{2n'-2} \cdot \frac{\rho^2}{1\,!} \cdot \frac{s_2^4\omega^2}{g^4} + \frac{1}{(2n'-2)(2n'+2)} \cdot \frac{\rho^4}{2\,!} \cdot \frac{s_2^8\omega^4}{g^8} + \ \ldots.. \right.\right.$$

$$\left.\left. + \frac{1}{(2n'-2)(2n'+2)\ldots\ldots(2n'+4p-6)} \cdot \frac{\rho^{2p}}{p\,!} \cdot \frac{s_2^{4p}\,\omega^{2p}}{g^{4p}} + \ \ldots\ldots\right\} s_2\, ds_2 \right] \omega^{n'-2} d\omega$$

On integration for s_2 from $s_2=0$ to $s_2=\infty$ the distribution of ω is obtained readily

$$df(\omega) = y_0 .\ 2^{n'-3} .\ g^2 . \left[\Gamma(n-1).\ \frac{\omega^{n'-2}}{(1+\omega^2)^{n'-1}} + \frac{2^2}{1\,!} \cdot \frac{\rho^2\Gamma(n'+1)}{(2n'-2)} \cdot \frac{\omega^{n'}}{(1+\omega^2)^{n'+1}} + \right.$$

$$+ \frac{2^4}{2\,!} \cdot \frac{\rho^4\Gamma(n'+3)}{(2n'-2)(2n'+2)} \cdot \frac{\omega^{n'+2}}{(1+\omega^2)^{n'+3}} + \ldots\ldots\ldots$$

$$\left. + \frac{2^{2p}}{p\,!} \cdot \frac{\rho^{2p}\Gamma(n'+2p-1)}{(2n'-2)(2n'+2)\ldots(2n'+4p-6)} \cdot \frac{\omega^{n'+2p-2}}{(1+\omega^2)^{n'+2p-1}} + \ldots\ldots \right] d\omega$$

The value of y_0 has been given by Karl Pearson (1925).

$$y_0 = \frac{(1-\rho^2)^{\frac{n'-1}{2}}}{2^{n'-3} .\ g^2 . \left\{\Gamma\left(\frac{n'-1}{2}\right)\right\}^2}$$

Substituting the value of y_0 we have

$$df(\omega) = \frac{2\Gamma(n'-1)}{\left\{\Gamma\left(\frac{n'-1}{2}\right)\right\}^2} .(1-\rho^2)^{\frac{n'-1}{2}} \cdot \frac{\omega^{n'-2}}{(1+\omega^2)^{n'-1}}$$

$$\left[1 + \frac{2n'}{1\,!} \cdot \frac{\rho^2\omega^2}{(1+\omega^2)^2} + \frac{4n'(n'+2)}{2\,!} \cdot \frac{\rho^4\omega^4}{(1+\omega^2)^4} \right.$$

$$+ \frac{8n'(n'+2)(n'+4)}{3\,!} \cdot \frac{\rho^6\omega^6}{(1+\omega^2)^6} + \ \ldots\ldots\ldots$$

$$\left. + \frac{2^p.n'(n'+2)(n'+4)\ldots(n'+2p-2)}{p\,!} \cdot \frac{\rho^{2p}\omega^{2p}}{(1+\omega^2)^{2p}} + \ldots\ldots \right] d\omega$$

$$= \frac{2\Gamma(n'-1)}{\left\{\Gamma\left(\frac{n'-1}{2}\right)\right\}^2} \cdot (1-\rho^2)^{\frac{n'-1}{2}} \cdot \frac{\omega^{n'-2}}{(1+\omega^2)^{n'-1}} \cdot \times$$

$$\left[1 + \frac{\frac{n'}{2}}{1\,!} \cdot \frac{4\rho^2\omega^2}{(1+\omega^2)^2} + \frac{\frac{n'}{2}\left(\frac{n'}{2}+1\right)}{2\,!} \cdot \frac{16\rho^4\omega^4}{(1+\omega^2)^4}\right.$$

$$+ \frac{\frac{n'}{2}\left(\frac{n'}{2}+1\right)\left(\frac{n'}{2}+2\right)}{3\,!} \cdot \frac{64\rho^6\omega^6}{(1+\omega^2)^6} + \ldots\ldots\ldots$$

$$\left. + \frac{\frac{n'}{2}\left(\frac{n'}{2}+1\right)\left(\frac{n'}{2}+2\right)\ldots\ \left(\frac{n'}{2}+p-1\right)}{p\,!} \cdot \frac{4^p \cdot \rho^{2p} \cdot \omega^{2p}}{(1+\omega^2)^{2p}} + \ldots\ldots\right] d\omega$$

$$= \frac{2\Gamma(n'-1)}{\left\{\Gamma\left(\frac{n'-1}{2}\right)\right\}^2} (1-\rho^2)^{\frac{n'-1}{2}} \cdot \frac{\omega^{n'-2}}{(1+\omega^2)^{n'-1}} \left[1 - \frac{4\rho^2\omega^2}{(1+\omega^2)^2}\right]^{-\frac{n'}{2}} d\omega \quad \ldots \quad (3{\cdot}0)$$

This is the distribution of ω $(= s_1/s_2)$ when the two samples of size n' are drawn from a bivariate normal population with coefficient of correlation ρ and standard deviation σ for both variates.

If $\rho = 0$, the distribution equation reduces to

$$df(\omega) = \frac{2\Gamma(n'-1)}{\left\{\left(\Gamma\,\frac{n'-1}{2}\right)\right\}^2} \cdot \frac{\omega^{n'-2}}{(1+\omega^2)^{n'-1}}\ d\omega$$

$$= \frac{2}{B\left(\frac{n'-1}{2}, \frac{n'-1}{2}\right)} \cdot \frac{\omega^{n'-2}}{(1+\omega^2)^{n'-1}}\ d\omega \qquad \ldots \qquad \ldots \quad (3{\cdot}01)$$

n' in equation (2·1) as well as in subsequent analysis represents the actual size of the samples and is one more than the corresponding degrees of freedom. Hence if n = number of degrees of freedom, $n = n' - 1$ and equation (3·01) may be written in the form

$$df\,(\omega) = \frac{2}{B\left(\frac{n}{2}, \frac{n}{2}\right)} \cdot \frac{\omega^{n-1}}{(1+\omega^2)^n} \cdot d\omega \quad \ldots \qquad \ldots \qquad \ldots \quad (3{\cdot}02)$$

It will be seen that putting $n_1=n_2=n$ in equation (1·2), we obtain this identical equation (3·02).

From equation (3·0), it is easy to write down the corresponding distribution equations of z and x for correlated samples.

Thus putting $x=\dfrac{\omega^2}{1+\omega^2}$, we have

$$df(x)=\frac{\Gamma(n'-1)}{\left\{\Gamma\left(\frac{n'-1}{2}\right)\right\}^2}\cdot(1-\rho^2)^{\frac{n'-1}{2}}\,x^{\frac{n'-3}{2}}\cdot(1-x)^{\frac{n'-3}{2}}\cdot\left[1-4\rho^2x(1-x)\right]^{-\frac{n'}{2}}dx \quad \dots \quad (3{\cdot}1)$$

Similarly, when $z=\log_e\omega$ the distribution of (z):

$$df(z)=\frac{2\Gamma(n'-1)}{\left\{\Gamma\left(\frac{n'-1}{2}\right)\right\}^2}\cdot(1-\rho^2)^{\frac{n'-1}{2}}\cdot\frac{e^{(n'-1)z}}{\left(1+e^{2z}\right)^{n'-1}}\cdot\left[1-\frac{4\rho^2e^{2z}}{(1+e^{2z})^2}\right]^{-\frac{n'}{2}}dz \quad \dots \quad (3{\cdot}2)$$

Expressing our equations in terms of degrees of freedom (n), we have distributions of z, ω and x corresponding to (1·1), (1·2) and (1·3) as follows:

$$df(z)=\frac{2(1-\rho^2)^{\frac{n}{2}}}{B\left(\frac{n}{2},\frac{n}{2}\right)}\cdot\frac{e^{nz}}{\left(1+e^{2z}\right)^n}\cdot\left[1-\frac{4\rho^2e^{2z}}{(1+e^{2z})^2}\right]^{-\frac{n+1}{2}}dz \quad \dots \quad (4{\cdot}1)$$

$$df(\omega)=\frac{2(1-\rho^2)^{\frac{n}{2}}}{B\left(\frac{n}{2},\frac{n}{2}\right)}\cdot\frac{\omega^{n-2}}{\left(1+\omega^2\right)^{n-1}}\cdot\left[1-\frac{4\rho^2\omega^2}{(1+\omega^2)^2}\right]^{-\frac{n+1}{2}}d\omega \quad \dots \quad (4{\cdot}2)$$

$$df(x)=\frac{(1-\rho^2)^{\frac{n}{2}}}{B\left(\frac{n}{2},\frac{n}{2}\right)}\,x^{\frac{n-2}{2}}\cdot(1-x)^{\frac{n-2}{2}}\left[1-4\rho^2x(1-x)\right]^{-\frac{n+1}{2}}dx \quad \dots \quad (4{\cdot}3)$$

If we put $\rho=0$, all these three equations correspond to the forms (1·1—1·3) obtained earlier by an altogether different procedure.

If the two population variances of the bivariate correlated populations are not identical and are say σ_1^2 and σ_2^2 $(\sigma_1\neq\sigma_2)$, we put

$$\omega'^2=\frac{s_1^2}{\sigma_1^2}\Big/\frac{s_2^2}{\sigma_2^2}=\frac{s_1^2}{s_2^2}\Big/\frac{\sigma_1^2}{\sigma_2^2}=\omega^2/\text{Const.} \quad \dots \quad \dots \quad (5{\cdot}1)$$

and

$$z'=\frac{1}{2}\log_e\frac{s_1^2}{s_2^2}-\frac{1}{2}\log_e\frac{\sigma_2^2}{\sigma_1^2}$$

$$=z-\text{Const.} \quad \dots \quad \dots \quad \dots \quad \dots \quad (5{\cdot}2)$$

The distribution of z^2, w^2 and x^2 can be easily shown to be identical. with (4·1·–4·3) except for a slight modification due to constant terms.

3. *Results of a Sampling Experiment.*

As an experimental verification of the distribution functions shown in (4·1) to (4·3), a sampling experiment was conducted from a normal bivariate population in x and y with the following population parameters, available in the Statistical Laboratory :

$$\text{Mean of } x = a_x = 0$$

$$\text{Mean of } y = a_y = 0$$

$$\text{Standard Deviation of } x = \sigma_x = 1$$

$$\text{Standard Deviation of } y = \sigma_y = 1$$

$$\text{Coefficient of correlation between } x \text{ and } y = \rho = 0{\cdot}6$$

The distribution of x being simplest in form, was chosen for the test. Samples of 5 were drawn at random and the variances s_1^2 and s_2^2 were calculated for each sample. Values of $x = s_1^2/(s_1^2+s_2^2)$ were obtained for each case.

The actual observed frequencies for 240 such samples were compared with expected values calculated from equation (4·3). With 9 cells the observed value of χ^2 was 4.24 with P=0.83, giving an excellent fit. Using R. A. Fisher's original distribution for uncorrelated samples (equation (1·3)) χ^2 was 15.12 with P=0.06. The details of the sampling experiments will be published in a subsequent issue.

4. *Discussion.*

The difference between the two distributions will be more and more marked as ρ increases Some caution is therefore necessary in testing the significance of variances of two samples particularly where an association is suspected. The 5 p.c. and 1 p.c. points of z or w or x when ρ is not zero is different from the corresponding points given by Fisher for uncorrelated samples. Thus, in the present example for $n_1 = n_2 = 4$, we have the following fiducial values of w^2

	$\rho=0$	$\rho=0{\cdot}6$
5 p.c.	6·388	4·095
1 p.c.	15·978	11·531

Taking the correlation into consideration, an observed ratio of variances of 11·531 reaches the 1 p. c. level of significance. If we neglect the correlation, an observed value of 15·978 is required for this level of significance. Thus if we neglect the existence of a correlation, a much higher value of the observed ratio of variances will be required to reach the same degree of certainty. In other words, Fisher's *z*-test is too stringent for this case of correlated samples and this stringency increases with the magnitude of the correlation.

In conclusion, I acknowledge my indebtedness to Prof. P. C. Mahalanobis for his general guidance and valuable criticism in the preparation and presentation of this paper.

References.

(1) R. A. Fisher (1915): Frequency Distribution of the values of the Correlation Coefficient in Samples from an Indefinitely Large Population *Biometrika* Vol. 10, 507—521.

(2) R. A. Fisher (1924): On a Distribution Yielding the Error Function of several well-known Statistics. *Proceedings of the International Mathematical Congress*, Toronto, 805—813.

(3) P. C. Mahalanobis (1932): Auxiliary Tables for Fisher's Z-Test in Analysis of Variance. *Indian Jour. Ag. Sc. Vol.* II, Part VI, 679-693.

(4) K. Pearson (1925): Further Contributions to the Theory of Small Samples. *Biometrika*, Vol. 17. 176—199.

(Paper Received January, 1935.)

AN EDITORIAL CORRECTION.

In an Editorial Note to S. S. Bose's "Tables for Testing the Significance of Linear Regression in the case of Time series and other Single-valued Samples" on p. 284 of *Sankhyā: The Indian Journal of Statistics*, Vol. 1, Parts 2 & 3. 1934, the need for caution in using results based on small samples was emphasized. The argument was not however clearly stated, and the Editor is indebted to Dr. E. S. Pearson for pointing this out in a letter dated 30th October 1934 :—

"If assumptions regarding normality and randomness are justified, surely the sampling distribution of the ratio in equation (4), p. 278 is exact, quite apart from lack of knowledge of σ. It is the same as in Student's Test where, if we accept "t" as the appropriate criteria, its sampling distribution is known without any approximation, although σ is not known. Trouble may arise because with so few observations we cannot be sure that our assumptions are justified, but I should have thought not for any reason of our using only an estimate of a standard error. The test allows for this automatically, with the result of course that its power of discrimination is less than if we know σ, but it does not tell us anything false if the initial assumptions are justified."

Dr. Pearson has explained the position very clearly. Only one remark may be added. Student's z (or R. A Fisher's t) and R. A. Fisher's z are both ratios, and the distributions are completely independent of the population variance. There is no difficulty so long as the significance of these ratios is being tested. But the difficulty appears when one of the items in the ratio (the sample mean itself or the sample variance) is being investigated. In the present case also the Tables are based on the distribution of a ratio, and a similar difficulty will arise if the significance of the regression coefficient itself is being tested. This difficulty increases rapidly as the size of the sample is decreased.

P. C. Mahalanobis.