This occurs when the state of maximum probability has little chance to change; it is a so-called *stationary state* or state of *statistical equilibrium*. It would mean a great deal if we could be able to say through how many states the statistical phenomena must pass before attaining its equilibrium, or in other words, whether the ergodic hypothesis of the kinetic theory of gas applies to certain social or economic phenomena. We will not go further into this now; the results obtained here must be considered as an initial exploratory step, which does permit us, however, to end with the following conclusive statement:

> If N elements, provided N is large enough, are distributed at random in k class "intervals of energy", it is highly probable that they will approach a configuration of statistical equilibrium, a distribution of maximum probability. Pareto's and Pearson's curves represent special configurations of statistical equilibrium in a stochastic system.

REFERENCES

[1] F. P. CANTELLI, "Sulle deduzioni delle leggi di frequenza da considerazioni di probabilità," *Metron*, Vol. 1 (1921), N. 3.
[2] G. CASTELNUOVO, *Calcolo della Probabilità*, Roma, 1919.
[3] R. B. LINDSAY, *Introduction to Physical Statistics*, New York, 1941.
[4] A. L. BOWLEY, *Elements of Statistics*, London, 1926.
[5] J. L. COOLIDGE, *An Introduction to Mathematical Probability*, Oxford, London, 1925.

ON THE COMPLETELY UNBIASSED CHARACTER OF TESTS OF INDEPENDENCE IN MULTIVARIATE NORMAL SYSTEMS

BY R. D. NARAIN

Indian Council of Agricultural Research

1. Introductory. To prove the unbiassed character of likelihood ratio tests like the test of significance of the multiple correlation coefficient or Hotelling's T^2 test, Daly [1] used the non-null frequency distributions of these test criteria. This leads to obvious difficulties when tackling the general regression problem and the test of independence of several sets of variates, and Daly [1] has shown only their locally unbiassed character.

This paper demonstrates an approach which does not require an explicit knowledge of the frequency distribution of the test criteria and it has been possible to prove that the likelihood ratio test for the general regression problem and the Wilks' criterion for independence of sets of variates are completely unbiassed. The argument proceeds in a chain, the unbiassedness of the Wilks' criterion following ultimately from the unbiassedness of the t-test. The link up has been achieved by working with a chain of conditional distribution densities, a principle employed earlier by the author [3], [4] in presenting a unified distribution theory of the common statistical coefficients relevant to normal theory.

2. The t-test. As the simplest demonstration of the procedure which is applicable generally, consider the t-test for the significance of the mean of a normal population. Let the frequency function of a sample of size n be

$$(2\pi V)^{-(n/2)} \exp\left[-\frac{1}{2V}\sum_{i=1}^{n}(x_i - m)^2\right]. \tag{1}$$

The region $W - w$ complementary to the critical region w for testing the hypothesis

$$m = 0$$

is given by

$$\bar{x}^2 \leq k^2\chi^2,$$

where k is a positive constant depending on the size of w and

$$n\bar{x} = \sum_{i=1}^{n} x_i,$$

$$\chi^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

We write

$$I(m) \equiv \int_0^\infty \left[\int_{-k\chi}^{k\chi} e^{-((n/2V)(\bar{x}-m))^2}\, d\bar{x}\right] f(\chi^2)\, d(\chi^2), \tag{2}$$

where

$$f(\chi^2)\, d(\chi^2)$$

is the frequency function of χ^2 which is distributed independently of $\bar{x}$. To show that the test is completely unbiassed is equivalent to showing that

$$I(m) \leq I(0) \text{ for all values of } m.$$

We have

$$\frac{\partial I}{\partial m} = \int_0^\infty \{e^{-(n/2V)(K\chi+m)^2} - e^{-(n/2V)(K\chi-m)^2}\}\, f(\chi^2)\, d(\chi^2)$$

which is positive or negative according as m is negative or positive. Therefore

$$I(m) \leq I(0).$$

3. The E^2 and R^2 tests. Let the frequency function of n observations of a random variate x_p be

$$(2\pi V)^{-(n/2)} \exp\left[-\frac{1}{2V}\sum_{i=1}^{n}\left(x_{ip} - \sum_{r=1}^{p-1}\beta_r x_{ir}\right)^2\right]\prod_i dx_{ip}. \tag{3}$$

With the usual notation for partial variates in regression analysis, the critical region w based on the likelihood ratio test for the hypothesis

$$0 = \beta_m = \beta_{m+1} = \cdots = \beta_{p-1}, \qquad m \leq p - 1,$$

is given by

$$1 - E^2 \equiv \frac{\sum_i x^2_{ip\cdot(12\cdots p-1)}}{\sum_i x^2_{ip\cdot(12\cdots m-1)}} \leq \text{a positive constant.}$$

It can be shown [2], [3] that this ratio can be expressed in the form

$$1 - E^2 = \frac{\chi^2}{\chi^2 + \sum_{r=m}^{p-1} z_r^2},$$

where the frequency function of χ^2 and the z_r is

$$(4) \qquad (2\pi V)^{-(n/2)} (\pi)^{-(p-1)/2} \frac{1}{\Gamma\left(\frac{n-p+1}{2}\right)} \cdot \exp\left[-\frac{1}{2V}\left\{\chi^2 + \sum_{r=1}^{p-1} (z_r - \eta_r)^2\right\}\right] (\chi^2)^{((n-p-1)/2)} d(\chi^2) \prod_r dz_r.$$

The hypothesis to be tested then becomes

$$0 = \eta_m = \eta_{m+1} = \cdots = \eta_{p-1}.$$

The region $W - w$ complementary to w is given

$$\sum_{r=m}^{p-1} z_r^2 \leq k\chi^2,$$

where k is a positive constant determined by the size of w. Denote by $I(\eta_{p-1}, \eta_{p-2}, \cdots \eta_m)$ the integral of (4) over the region $W - w$. Differentiating I with respect to η_{p-1}, performing the integration with respect to z_{p-1} and arguing exactly as in section 2 above we obtain

$$I(\eta_{p-1}, \eta_{p-2}, \cdots, \eta_m) \leq (0, \eta_{p-2}, \eta_{p-3} \cdots \eta_m).$$

Note that z_{p-2} is distributed independently of z_{p-1}. Therefore starting with $\eta_{p-1} = 0$ in (4) and considering the integration with respect to z_{p-2} first, we obtain as before $I(0, \eta_{p-2} \cdots \eta_m) \leq I(0, 0, \eta_{p-3} \cdots \eta_m)$ and thus finally $I(\eta_{p-1}, \eta_{p-2}, \cdots \eta_m) \leq I(0, 0, \cdots 0)$, which proves the completely unbiassed character of the E^2-test. The test of significance of the multiple correlation coefficient with any number of the predicting variates being fixed or random may be considered as a corollary to the above. We have only to multiply the frequency function (3) by a factor dF representing the frequency function of the random predicting variates (which need not be necessarily normal). This does not affect either the test criterion or the arguments showing its unbiassedness. The test of significance of the multiple correlation coefficient is thus completely unbiassed.

4. The general regression problem. Given the distribution,

$$(2\pi)^{-\frac{1}{2}pn} \,|\, \alpha^{rs} \,|^{n/2} \exp\Big[-\tfrac{1}{2} \sum_{r,s} \alpha^{rs} \Big\{\sum_i \Big(x_{ir} - \sum_h \beta_{rh} x_{ih}\Big) \cdot \Big(x_{is} - \sum_h \beta_{sh} x_{ih}\Big)\Big\}\Big] \times \prod_{i,r} dx_{ir},$$

$$\begin{aligned} i &= 1, 2, \cdots n, \\ h &= 1, 2, \cdots l, l+1, l+2, \cdots m, \\ r, s &= m+1, m+2, \cdots p, \\ n &\geq p > m \geq l, \end{aligned} \tag{5}$$

where the matrix $\|\, x_{ih} \,\|$ is of rank m. The hypothesis H to be tested is

$$\beta_{rv} = 0, \qquad \begin{aligned} r &= m+1, m+2, \cdots p, \\ v &= l+1, l+2, \cdots m. \end{aligned}$$

The likelihood ratio test gives the critical region defined by

$$\lambda \equiv \frac{|\, a_{rs} \,|}{|\, a'_{rs} \,|} \leq \text{ a positive constant,}$$

where, with the usual regression notation for partial variates,

$$\begin{aligned} a_{rs} &= \sum_{i=1}^{n} x_{ir\cdot(12\cdots m)}\, x_{(is\cdot(12\cdots m)}, \\ a'_{rs} &= \sum_{i=1}^{n} x_{ir\cdot(12\cdots l)}\, x_{is\cdot(12\cdots l)}. \end{aligned} \qquad r, s = m+1, m+2, \cdots p,$$

Now we note that

$$\lambda = \prod_{r=m+1}^{p} (1 - E_r^2) = (1 - E_p^2) \prod_{r=m+1}^{p-1} (1 - E_r^2), \tag{6}$$

where

$$1 - E_r^2 = \frac{\sum_{i=1}^{n} x^2_{ir\cdot(12\cdots l, l+1, l+2, \cdots m, m+1, \cdots r-1)}}{\sum_{i=1}^{n} x^2_{ir\cdot(12\cdots l, m+1, m+2, \cdots r-1)}}$$

Since the statistic λ is invariant to linear transformations of the random variates $x_{m+1}, x_{m+2}, \cdots, x_p$ the distribution (5) may be simplified to

$$\prod_{r=m+1}^{p} \Big[(2\pi V_r)^{-(n/2)} \exp\Big[-\frac{1}{2V_r} \sum_i \Big(x_{ir} - \sum_h \beta_{rh} x_{ih}\Big)^2\Big] \prod_i dx_{ir}\Big],$$

$$\begin{aligned} i &= 1, 2, \cdots n, \\ h &= 1, 2, \cdots m. \end{aligned} \tag{7}$$

Denote by $I(\beta_{pv}, \beta_{p-1,v}, \cdots \beta_{m+1,v})$ the integral of (7) over the region $W - w$

complementary to the critical region w, where β_{rv} in I stands for the entire set of parameters $\beta_{r,l+1}, \beta_{r,l+2}, \cdots \beta_{rm}$. We may first integrate over a subregion of $W - w$ over which $\prod_{r=m+1}^{p-1} (1 - E_r^2)$ has a given value. Using identity (6) and the result of section 3 it follows immediately that

$$I(\beta_{pv}, \beta_{p-1,v} \cdots \beta_{m+1,v}) \leq I(0, \beta_{p-1,v}, \beta_{p-2,v}, \cdots \beta_{m+1,v}).$$

If $\beta_{pv} = 0$, the distribution of E_p^2 is independent of that of E_{p-1}^2. Hence, startng with $\beta_{pv} = 0$ in (7) and considering the integration for E_{p-1}^2 first, we obtain

$$I(0, \beta_{p-1,v}, \beta_{p-2,v}, \cdots \beta_{m+1,v}) \leq I(0, 0, \beta_{p-2,v}, \cdots \beta_{m+1,v}).$$

Thus finally

$$I(\beta_{pv}, \beta_{p-1,v}, \cdots \beta_{m+1,v}) \leq I(0, 0, \cdots 0),$$

which proves the completely unbiassed character of the test.

5. Test of independence of sets of variates. Consider n observations of q sets of random variates distributed in the multivariate normal form

$$\text{Const} \times \exp\left[-\tfrac{1}{2} \sum \alpha^{rs} \{\sum_i (x_{ir} - m_r)(x_{is} - m_s)\}\right] \prod_{i,r} dx_{ir},$$

$$(8) \qquad \begin{aligned} &i = 1, 2, \cdots n, \\ &r = 1, 2, \cdots l_1, l_1 + 1, l_1 + 2, \cdots l_2, l_2 + 1, \cdots l_3, \cdots l_q, \\ &n > l_q. \end{aligned}$$

Denote by D_j the determinant of the sample dispersion matrix of the j^{th} set of variates and by $D(j)$ the determinant of the dispersion matrix of the first j sets taken together. The Wilks' statistic used for testing the independence of the q sets is given by

$$(9) \qquad \Lambda = \frac{D(q)}{\prod_{j=1}^{q} D_j} = \lambda_q \prod_{j=2}^{q-1} \lambda_j,$$

where

$$\lambda_j = \frac{D(j)}{D_j D(j-1)}, \qquad j = 2, 3, \cdots q.$$

The region $W - w$ complementary to the critical region w is defined by

$$\Lambda \geq \text{a positive constant}.$$

The statistic W is invariant to linear transformations within each set of variates. The distribution (8) may therefore without loss of generality be written in the form

$$(10) \quad \prod_{j=1}^{q} \left[\prod_{r=l_{j-1}+1}^{l_j} \left\{ (2\pi V_r^2)^{-(n/2)} \exp\left(-\frac{1}{2V_r^2} \sum_{i=1}^{n} (x_{ir} - \sum_{u=0}^{l_j-1} \beta_{ru} x_{iu})^2 \right) \prod_{i,r} dx_{ir} \right\} \right].$$

Let B_j ($j = 2, 3, \cdots q$) stand for the set of constants

$$\beta_{ru}, \qquad \begin{array}{l} r = l_{j-1} + 1, l_{j-1} + 2, \cdots l_j, \\ u = 1, 2, \cdots l_{j-1}, \end{array}$$

and let

$$B_j = 0 \tag{11}$$

imply the vanishing of all the constants of the set B_j. The q sets of variates will be independent if (11) holds for all values of j from 2 to q. Denote by $I(B_q, B_{q-1} \cdots, \cdots, B_2)$ the integral of (10) over the region $W - w$. Integrating first over the sub-region of $W - w$ for which

$$\prod_{j=2}^{q-1} \lambda_j$$

has a given value and using the result of section 4, it follows that

$$I(B_q, B_{q-1}, \cdots B_2) \leq I(0, B_{q-1}, \cdots B_2).$$

Also if $B_q = 0$, λ_q is distributed independently of λ_{q-1}. Hence starting with $B_q = 0$ in (10) and integrating for λ_{q-1} first, we obtain

$$I(0, B_{q-1}, B_{q-2}, \cdots B_2) \leq I(0, 0, B_{q-2}, \cdots B_2).$$

Thus finally, $I(B_q, B_{q-1}, \cdots B_2) \leq I(0, 0, \cdots 0)$, which proves the completely unbiassed character of the Wilks criterion.

REFERENCES

[1] J. E. Daly, "On the unbiased character of likelihood ratio tests for independence in normal system," *Annals of Math. Stat.*, Vol. 11 (1940), p. 1.

[2] P. C. Tang, "The power function of analysis of variance tests with tables and illustrations of their use," *Stat. Res. Mem.*, Vol. 2 (1938), p. 126.

[3] R. D. Narain, "A new approach to sampling distribution of the multivariate normal theory. Part I," *Jour. Ind. Soc. Agr. Stat.*, Vol. 1 (1948), p. 59.

[4] R. D. Narain, "A new approach to sampling distributions of the multivariate normal theory. Part II," *Jour. Ind. Soc. Agr. Stat.*, Vol. 1 (1948), p. 137.

ON THE DISTRIBUTION OF THE TWO CLOSEST AMONG A SET OF THREE OBSERVATIONS[1]

By G. R. Seth

Iowa State College

1. Introduction. In this note we obtain the joint distribution of the two closest observation x', x'' ($x' < x''$) of the set x_1, x_2, x_3 ($x_1 \leq x_2 \leq x_3$) when the distribution of x_1, x_2, x_3 is given or can be obtained.[2] We will assume that in general the density function is given by $f(x_1, x_2, x_3)$ and that it is continuous in the

[1] The results in this paper were presented at a meeting of the Institute of Mathematical Statistics in Madison, Wisconsin, September 9, 1948.

[2] The author's attention was drawn to this problem while visiting the National Bureau of Standards in the Spring of 1948, by Mr. Julius Lieblein of the Statistical Engineering