PART 3: THE DISCRIMINANT FUNCTION APPROACH IN THE CLASSIFICATION OF TIME SERIES

(PART III OF STATISTICAL INFERENCE APPLIED TO CLASSIFICATORY PROBLEMS)

By C. RADHAKRISHNA RAO
Statistical Laboratory, Calcutta

1. Introduction

In a recent paper on a large sample test for moving average Wold (1949) observed that, 'From one point of view the above analysis is rather disappointing, for although the numerical difference between the parameters of the two schemes (c) and (E_3) is by no means negligible, it seems as if our test is not able to discriminate between them on the basis of a time series with as many as $n = 100$ items. On the other hand, however, such a conclusion would be quite in line with the general experience in time series analysis; *i.e.*, if quantitative results are aimed at we must have very long series'. This experience is not confined to time series studies alone. The statistician faces similar problems even in controlled experimental investigations where the residual error cannot be reduced below a certain magnitude with the result that considerable replication is needed to detect small differences. Replication implies that a fresh set of figures obtained are subject to the same probablity laws as the old set so that the information supplied by the new set can be added on to that of the first set. This, unfortunately, is not the case with 'long time series'. No single model can be expected to be applicable to the whole series. Even a series of size 100 is too much to ask for in the case of economic data. Fitting of specified theoretical models must necessarily be confined to short series and this raises a number of complicated problems some of which are considered in Parts 1 and 2 of this paper. This section is intended to show that although no statement of confidence can be made in all cases that one scheme is better than the other the problem admits a provisional decision and by this method it may be possible to specify the nature of the model applicable to a group of similar time series. This is done by comparing the likelihoods as in the case of discriminant function theory.

In this connection a closer examination of the large sample tests of Quenouille (1947) and Wold (1949) has been made and a few alterations which are necessary for the application of discriminant function theory have been suggested.

The author also derived a few results in truncated sequential procedures which have general application in classificatory problems and which were intended to be used in the experimental investigation of the earlier parts of this paper. Since this method involved heavy computations the project had to be abandoned and only the theoretical developments are considered towards the end of this paper.

2. TEST FOR AN APRIORI MOVING AVERAGE MODEL

Consider a moving average scheme of order one with an assigned value of ρ_1, the other lag correlations being zero. If the observed series conforms by hypothesis to a moving average scheme of order one then a test for the significance of the computed correlations r_1, r_2,... would indicate whether the assumption about the order is correct or not. In such a test it makes little difference as to what the true value of ρ_1 considered is. For instance, Wold (1949) found no difference in using $\rho_1=.50$ or $.489$ in the test for the lag correlations r_1, r_2,... arising out of Beveridge's wheat price index (1770–1869); in fact the smaller value of ρ_1 gave higher probablities although on the basis of the observed r_1 the value $\rho_1=.50$ appeared better than .489. A proper test for a moving average scheme with an assigned ρ_1 should then take into account all the computed correlations. Let us examine such a test by considering the first five correlations of wheat price index series. The dispersion matrix of the observed correlations in terms of ρ_1 is

$$\begin{array}{ccccc}
\frac{1-3\rho_1^2+4\rho_1^4}{n-1} & \frac{2\rho_1-2\rho_1^3}{\sqrt{(n-1)(n-2)}} & \frac{\rho_1^2}{\sqrt{(n-1)(n-3)}} & 0 & 0 \\
 & \frac{1+2\rho_1^2}{n-2} & \frac{2\rho_1}{\sqrt{(n-2)(n-3)}} & \frac{\rho_1^2}{\sqrt{(n-2)(n-4)}} & 0 \\
 & & \frac{1+2\rho_1^2}{n-3} & \frac{2\rho_1}{\sqrt{(n-3)(n-4)}} & \frac{\rho_1^2}{\sqrt{(n-3)(n-5)}} \\
 & & & \cdots & \cdots
\end{array}$$

where all the terms below the diagonal are omitted because of symmetry. To this matrix is appended the column $r_1-\rho_1$, r_2, r_3,... and the columns are swept out as indicated by Rao (1949, 1950 b) and as followed in Parts 1 and 2 of this paper. Using only the first six observed correlations of the wheat prices the following χ^2's have been obtained for $\rho_1=.500$ and $.489$

	$\rho_1=.500$		$\rho_1=.489$	
r_s	χ^2	$\Sigma\chi^2$	χ^2	$\Sigma\chi^2$
1	2.57	2.57	3.02	3.02
2	1.75	4.32	2.14	5.16
3	1.70	6.02	1.18	6.34
4	1.71	7.73	.73	7.07
5	2.20	9.93	.64	7.71
6	3.00	12.93	.64	8.35

Judged by the χ^2 values alone both values of ρ_1 appear admissible although the smaller value of ρ_1 gives a better fit to the correlogram as a whole. To examine this point further we may calculate the likelihood of ρ_1 for both the values. Considering only first six values of r, the likelihood of ρ_1 is

$$L(\rho_1|r)=\frac{1}{\sqrt{|\Lambda|}}\cdot e^{-\frac{1}{2}\Sigma\chi^2}$$

where $\Sigma\chi^2$ is same as that obtained above and $|\Lambda|$, the determinant of the dispersion matrix, is the product of the six pivotal elements obtained in sweeping out Λ

to arrive at χ^2 values. Instead of likelihood the values of log likelihood are calculated,

$$\log L(.500|r) = -\tfrac{1}{2}\log(.0^{11}2009) - \tfrac{1}{2}.(12.93).(\log_{10}e)$$
$$= 7.833994 - 2.807685 = 4.526309$$

and
$$\log L(.489|r) = -\tfrac{1}{2}\log(.0^{11}3821) - \tfrac{1}{2}.(8.85)(\log_{10}e)$$
$$= 7.208911 - 1.813161 = 5.395750$$

The value of $\rho_1 = .489$ has higher likelihood so that out of the two alternatives a provisional decision can be made in favour of .489. To determine whether one likelihood is significantly higher or lower than the other a few more complicated computations are necessary; the appropriate formulae are given in an appendix at the end of the article.

By this method two values R_1 and R_2 are calculated and when the likelihood ratio exceeds R_1 the decision is in favour of the hypothesis corresponding to the numerator in the ratio and when it falls below R_2 the decision is in favour of the other, while no decisive answer can be given when the likelihood ratio lies between R_1 and R_2. A provisional answer is to choose the hypothesis with a higher likelihood. The position has been fully explained by Rao (1950a). No apriori probabilities of the hypotheses are assumed in such an analysis.

The method described above is applicable in testing goodness of fit of any apriorily assigned moving average model. The dispersion matrix of the observed correlations is obtained first by using Bartlett's formula and the deviations $r_1 - \rho_1$, $r_2 - \rho_2$, ... where r's are the observed and ρ's the assigned lag correlations, are appended to the dispersion matrix which is swept out to yield successive χ^2 values.

In deciding between two alternatives a provisional decision can be made in favour of the alternative having a higher likelihood which is obtained as a by-product in the above computations. In the case of time series analysis one is tempted to use a sequential decision function in deciding alternatives using an extra lag correlation each time till a decision is reached. This needs a successive evaluation of log likelihood differences which can be easily obtained in the above case. The successive products of pivotal elements give the successive determinants applicable to the cumulated χ^2.

	log $L(\rho_i\|r)$		difference
r	$\rho_1 = .500$	$\rho_1 = .489$	
1	3.090265	2.987672	0.102593
2	3.418854	3.219403	0.199251
3	3.781748	3.678027	0.103719
4	4.177638	4.238066	−0.060428
5	4.427857	4.818096	−0.390439
6	4.526309	5.395760	−0.869441

Using Wald's limits the successive evaluations have to be carried out till the difference exceeds $\log(1-\beta)/\alpha$ in which case the first alternative is accepted or falls below $\log \beta/(1-\alpha)$ in which case the second alternative is accepted. If no decision is reached up to some stage the process is truncated and the alternative with the higher likelihood is accepted. The values of α and β correspond to the levels of significance for the two hypotheses and can be chosen to be 5% each.

In the above example

$$\log (1-\beta)/\alpha = 1.278754 \quad \text{and} \quad \log \beta/(1-\alpha) = \bar{3}.721233$$

and all the differences lie between these two limits so that no decision is possible with only six correlations. If we choose to truncate the test at this stage the chances are in favour of the second alternative with $\rho_1 = .489$

Two objections can be raised against this test. Firstly the correlations are not all independent so that a proof is needed to show that the sequential procedure terminates. Secondly the sequential procedure cannot be continued indefinitely because there is an upper limit to the number of serial correlations. The second objection needs some attention and a successful solution to this will answer the first objection as well. This is attempted in section 6.

8. TEST FOR A FITTED MOVING AVERAGE MODEL

One method used by Wold in fitting a moving average model is to estimate $\rho_1, \rho_2 \ldots, \rho_k$ from $r_1, r_2 \ldots, r_k$ and test for the significance of the remaining lag correlations $r_{k+1}, r_{k+2}, \ldots$. From the point of view of estimation this is not, however, the best method. An illustration is given below by choosing the example of the previous section, the method being generally applicable to moving average models of any order. It is necessary first to guess the value of ρ_1, (and more values if the order is greater than one) and set up the dispersion matrix with the appended column $r_1 - \rho_1, r_2, r_3, \ldots$. The matrix is swept out using the leading element as the pivot (the italicized elements) each time.

dispersion matrix elements multiplied by 10 | | | | | | r | ρ_1

.050505	.076143	.025511	0	0	0	.614	1
	.153062	.102570	.026774	0	0	.099	0
		.154039	.103629	.026043	0	− .156	0
			.156250	.104713	.026317	− .115	0
				.157894	.105820	− .006	0
					.159574	.003	0

We obtain:

1	1.507630	.505129				12.1572	19.800
	.038266	.064108	.026774			− .83569	− 1.50763
		.141781	.103629	.026043		− .40614	− .505118
			.156250	.104713	.026317	− .115	0
				.157895	.105820	− .006	0
					.159575	.003	0
	1	1.675320	.673640			−21.8389	−39.3986
		.034349	.060450	.026043		.93391	2.02065
			.138890	.104713	.026317	.44789	1.01546
				.157895	.105820	− .006	0
					.159575	.003	0
		1	1.759870	.758190		27.1889	58.8270
			.032506	.058885	.026317	− 1.10569	− 2.54064
				.138149	.105820	− .71498	− 1.63203
					.160575	.003	0
			1	1.811360	.809600	−36.7827	−76.1391
				.031457	.058147	1.45193	3.07037
					.138269	.07104	2.03691
				1	1.846090	46.1120	97.8123
					.030339	− 1.71023	− 3.61314
					1	−85.3969	−116.972

Each pivotal row and its reduced form obtained by dividing all its elements by the pivotal element supplies three quantities

A product of elements in the column for r,

B product of elements in the column for ρ,

C sum of cross products in these two columns, which in the above case are

r	A	B	C
1	7.4615	19.8000	24.3144
2	18.2508	59.3985	65.8497
3	25.3917	118.8058	109.8704
4	43.9323	198.5741	186.9075
5	66.0500	299.3958	283.1604
6	94.6885	422.6302	400.0017
Total	250.7278	1118.6765	1070.2001

The best estimate of ρ_1 is

$$\hat{\rho}_1 = \frac{\Sigma C}{2\Sigma B} = \frac{1070.2001}{2227.3530} = .4783$$

which is closer to the value .489 considered in the previous section. The accuracy of this estimate

$$\frac{0.1}{\Sigma B} = .0^4893011$$

the factor 0.1 being the correction for starting with 10 times the dispersion matrix. The accuracy of this is about 50 times more than the estimate based on r_1 the first observed correlation only. This result is a bit unexpected. It may be that high correlations between the serial correlations are responsible for such a high efficiency. This shows that much information is gained by using all the serial correlations in estimating the first correlation. For this fitted value of ρ the χ^2 goodness of fit is

$$10(\Sigma A - \rho\Sigma C + \rho^2\Sigma B) = 8.012$$

and this has only 5 degrees of freedom one degree being lost in the estimation of ρ. Actually for the evaluation of χ^2 a new dispersion matrix with the estimated value has to be used. But this does not make much difference if the estimate is close to the guessed value.

In the above computational scheme we can, at any stage, find an estimate of ρ by using the formula $\Sigma C/2\Sigma B$ with its associated variance formula $1/\Sigma C$. We can stop as soon as the accuracy as determined by $1/\Sigma C$ reaches a specified value smaller than the maximum accuracy attainable by using all the available correlations.

The above method is applicable to the general case where the first serial correlations have to be estimated. We start with provisional estimates and set up the dispersion matrix for $r_1 r_2, \ldots, r_h, r_{h+1}, \ldots$ and append the matrix with the columns

(r)	(ρ_1)	(ρ_2)	...	(ρ_h)
r_1	1	0	...	0
r_2	0	1	...	0
r_3	0	0	...	0
⋮			...	
r_h	0	0	...	1
r_{h+1}	0	0	...	0
...	...	...	...	...

and sweep out the dispersion matrix. Consider the pivotal row and its reduced form at any stage. Let them have the elements

a_0	a_1	a_2	...	a_h
b_0	b_1	b_2	...	b_h

in the above columns. From these we compute the quantities

$$s_{ii} = a_i b_i$$
$$s_{ij} = (a_i b_j + a_j b_i)/2$$

and take the summation over all the quantities available at each stage of reduction. Representing the sums by capital letters the estimating equations can be written

$$\rho_1 S_{11} + \rho_2 S_{12} + \ldots \rho_h S_{1h} = S_{10}$$
$$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$$
$$\rho_1 S_{h1} + \rho_2 S_{h2} + \ldots + \rho_h S_{hh} = S_{h0}$$

If (c_{ij}) is reciprocal to (S_{ij}) the estimate of ρ_h is

$$\hat{\rho}_h = c_{h1} S_{10} + c_{h2} S_{20} + \ldots + c_{hh} S_{h0}$$

with

$$V(\hat{\rho}_h) = c_{hh}$$

In case the estimates differ very much from the initially guessed values the above process may be repeated evaluating a fresh dispersion matrix. In practice it is convenient first to take only a few extra correlations to correct the initial values. Starting with these estimates and including some more lag correlations better estimates can be obtained by repeating the above process. The estimates thus obtained have a high precision. The formula for χ^2 goodness of fit with $t-h$ degrees of freedom where t is the number of correlations used and h the number estimated is

$$S_{00} - \hat{\rho}_1 S_{10} - \hat{\rho}_2 S_{20} - \ldots - \hat{\rho}_h S_{h0}$$

4. TEST FOR AN APRIORI AUTOREGRESSIVE MODEL

Let us consider an autogressive scheme

$$\xi_{t+2} + \alpha \xi_{t+1} + \beta \xi_t = \epsilon_{t+2}$$

with assigned values α_0 and β_0 of α and β. To test whether the data are in agree-

ment with the assigned values of α and β, the following computations are needed. First a and b are obtained from the equations

$$r_1 + a + br_1 = 0$$
$$r_2 + ar_1 + b = 0$$

where r_1 and r_2 are the first two computed serial correlations. The dispersion matrix of the estimates is approximately equal to

$$\frac{c}{(n-3)} \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix}^{-1}$$

where

$$\rho_1 = -\alpha/(1+\beta)$$
$$c = 1 + a\rho_1 + \beta\rho_2$$

and n is the size of the sample.

The χ^2 on two degrees of freedom for testing the agreement of the computed a and b with the assigned is (c.f. Wald, 1943)

$$\chi^2 = \frac{n-3}{c}\{(a-\alpha_0)^2 + 2\rho_1(a-\alpha_0)(b-\beta_0) + (b-\beta_0)^2\}$$

In scheme II considered in Part 2 of this paper $\alpha_0 = -.7$ and $\beta_0 = .6125$. The computed values of a, b and the χ^2 for some samples are given below.

sample no.	computed a	computed b	χ^2	d.f.
3	−.7888	.4447	3.0907	2
8	−.8349	.5067	2.6515	2
13	−.6700	.8172	3.5485	2
18	−.2515	.2840	11.4311	2
23	−.6317	.6053	.2707	2
total			20.8925	10

The total χ^2 of 20.8925 is significant on 10 degrees of freedom but the whole significance is due to one bad sample (no. 18) chosen above for illustration.

To test goodness of fit of the assigned model to the observed data it is necessary to supplement the above test with Quenouille's test for r_3, r_4, etc. In each of the above samples there are 12 computed correlations with the distribution of χ^2 for r_1, r_2 (or a, b) and $r_3, r_4, \ldots, r_{12}$ as follows.

sample no.	for 2 d.f.	for the rest 10 d.f.	total 12 d.f.
3	3.0907	7.7637	10.8544
8	2.6515	13.3546	15.9061
13	3.5485	2.7714	6.3199
18	11.4311	20.7657	39.1968
23	.2707	10.5853	10.8560

The total χ^2 on 12 degrees of freedom is significant only in one case, sample no. 18, which appears to be a particularly bad case.

For discriminatory analysis it is necessary to know the likelihood of the assigned parameters for observed values of the serial correlations. We first note that the asymptotic distribution of a and b is

$$\frac{1}{2\pi\sqrt{|\Lambda|}}\, e^{-\frac{n-3}{2c}\{(a-\alpha)^2+2\rho_1(a-\alpha)(b-\beta)+(b-\beta)^2\}} da\,db$$

where

$$|\Lambda| = \frac{(1+\alpha\rho_1+\beta\rho_2)(1-\rho_1^2)}{n-3}$$

From this the asymptotic distribution of r_1 and r_2 is derived by making use of the relations

$$r_1+a+br_1 = 0$$
$$r_2+ar_1+b = 0$$

The probability density is same as that for a and b multiplied by

$$\partial(a, b)/\partial(r_1, r_2) = (1-r_2)/(1-r_1^2)^2$$

Quenouille (1947) has shown that the forms

$$R_s = r_{s-2}+2ar_{s-1}+(a^2+2b)r_s+2abr_{s+1}+b^2r_{s+2}$$

are normally and independently distributed in large samples with variance

$$v_s = \frac{1}{n-s}\left[\frac{(1-b)\{(1+b)^2-a^2\}}{1+b}\right]^2.$$

Hence, omitting the factor involving π, the log likelihood of α and β for known values of r_1, r_2 (or a, b), r_3, r_4, ... is,

$$-\frac{1}{2}\left\{\log|\Lambda|+\log v_1+\log v_2+\ldots\right\}$$

$$-\frac{1}{2}(\log_{10}e)\left\{\frac{n-3}{2c}[(a-\alpha)^2+2\rho_1(a-\alpha)(b-\beta)+(b-\beta)^2]+R_1^2+R_2^2+\ldots\right\}$$

$$+\log(1-r_2)-2\log(1-r_1^2)$$

Using likelihood we can discriminate which of two given autoregressive models is most appropriate to the data. Since the likelihood for any assigned moving average model can be computed as shown in section 2 the possibility of discriminating between two types of time series is also open to us. It is necessary for such a discrimination to compute the likelihoods of assigned models and choose that model which has a higher likelihood at the observed values of all the serial correlation coefficients.

5. Test for a fitted autoregressive model

Some difficulty arises when we want to compare the goodness of fit of unspecified autoregressive and moving average models. Let us consider a case where the first two correlations are used to derive the constants for (i) a moving average model of order 2 and (ii) an autoregressive model of order 2. The goodness of fit χ^2 for both these reconstructed models may not be significant in which case no decision is possible as to which is the better of the two. Even for a provisional decision we need to compare the likelihoods. Since the first two correlations are used in estimation we consider only the distribution of the correlations $r_3, r_4, \ldots$ assuming that r_1, r_2 are true values of the first two serial correlations. In the case of a moving average scheme the distribution of $r_3, r_4, \ldots$ is completely specified by their dispersion matrix only, the mean values being zero. The dispersion matrix can be easily calculated since Bartlett's formula assumes a simple form in this case. The probability density of $r_3, r_4, \ldots$ supplies the likelihood of the moving average model. In the autoregressive case the determination of the probability density is a bit involved. But it does not make much difference if the probability density of $R_1, R_2, \ldots$ as defined by Quenouille is used.

6. The probability ratio test with an upper limit to the number of observations

In deriving the limits for the probability ratio test suitable for any general probability law, Wald assumes that the sequential procedure will be terminated only when a decision is reached in favour of one or the other alternative hypothesis. But occasionally it is desirable and sometimes necessary to set an upper limit, say N, to the number of samples to be drawn. Wald himself considered this problem and suggested the following method of truncating the sequential procedure. If P_{1N}/P_{0N} is the probability ratio at the Nth observation and no decision is reached before N by the use of the limits A and B (suitable for the general sequential procedure) then the rule is to accept H_1 if $P_{1N}/P_{0N} \geqslant 1$ *and accept H_0 if $P_{1N}/P_{0N} < 1$. Wald also evaluates the upper limit to the error committed by such truncation.

A question may be asked as to whether the limits A and B could be replaced by narrower limits when it is known that not more than N observations will be taken. The appropriate limits in such a case may be denoted by $A(N)$ and $B(N)$. Wald's limits A and B will then correspond to $A(\infty)$ and $B(\infty)$. The advantage of getting closer limits is that the frequency of cases, where a *decisive answer* (i.e., rejecting the null hypothesis with some confidence) is obtained before the Nth observation, is increased to some extent while keeping the frequency of wrong decisions at the same level.

*It is not necessary to include the value 1 for the ratio in one case and exclude it for the other. The case when $P_{1N} = P_{0N}$ can be decided by considering the ratio $P_{1,N-1}/P_{0,N-1}$ at the previous stage.

6a. *Notations* :

We shall first consider a sequential procedure terminating at the Nth observation and leading to three results, favouring H_0 or H_1 or neither. Using Wald's notation let

(a) α be the chance of accepting H_1 when H_0 is correct

(b) β be the chance of accepting H_0 when H_1 is correct.

These quantities α and β can be chosen by the experimenter depending on the risk he is prepared to take in accepting one when the other hypothesis is correct. Since the sequential process may end without the acceptance of either H_0 or H_1 we need introduce two more quantities

(c) δ_0, the chance of the sequential process terminating without favouring any hypothesis when H_0 is true, and

(d) δ_1, the corresponding chance when H_1 is true,

These quantities are not pre-assigned but their values associated with any test procedure can be calculated.

The probability densities of the first n observations corresponding to the hypotheses H_0 and H_1 will be represented by P_{0n} and P_{1n} respectively.

6b. *Definition of the truncated sequential test* :

With the new limits $A(N)$ and $B(N)$ the probability ratio test is defined as follows :—

(i) Continue sampling so long as

$$B(N) < \frac{P_{1n}}{P_{0n}} < A(N) \qquad \text{for} \quad n < N$$

(ii) Accept H_1 if

$$\frac{P_{1n}}{P_{0n}} \geq A(N)$$

at any stage ($n \leq N$) and discontinue sampling,

(iii) Accept H_0 if

$$\frac{P_{1n}}{P_{0n}} \leq B(N)$$

at any stage ($n \leq N$) and discontinue sampling, and

(iv) declare that the hypotheses are indistinguishable when the probability ratio lies between $B(N)$ and $A(N)$ even up to the Nth stage. At this stage we observe that provisional decisions could be made instead of (iv) and this might be useful in planning future experiments. This leads to the following classification in the case of (iv) :

(iva) accept H_1 provisionally if

$$1 < \frac{P_{1N}}{P_{0N}} < A(N)$$

(ivb) accept H_0 provisionally if

$$B(N) < \frac{P_{1N}}{P_{0N}} < 1$$

and if $P_{1N}=P_{0N}$, use the ratio of $P_{1\,N-1}$ and $P_{0\,N-1}$.

6c. *Inequalities satisfied by $A(N)$ and $B(N)$:*

Let the sets of points $(x_1,\ldots,x_N)$ leading to decisions (ii), (iii) and (iv) above be denoted by w_1, w_2 and w_3 respectively. The following relationships are true.

$$\int_{w_1} P_{1N}\,dv \geqslant A(N) \int_{w_1} P_{0N}\,dv$$

or

$$1-\beta-\delta_1 \geqslant \alpha A(N)$$

Similarly

$$\int_{w_2} P_{1N}\,dv \leqslant B(N) \int_{w_2} P_{0N}\,dv$$

or

$$\beta \leqslant (1-\alpha-\delta_0)B(N)$$

Hence

$$A(N) \leqslant \frac{1-\beta-\delta_1}{\alpha} \quad \text{and} \quad B(N) \geqslant \frac{\beta}{1-\alpha-\delta_0}$$

Since

$$\frac{1-\beta}{\alpha} > \frac{1-\beta-\delta_1}{\alpha} \quad \text{and} \quad \frac{\beta}{1-\alpha} < \frac{\beta}{1-\alpha-\delta_0}$$

the limits obtained by setting

$$A(N) = \frac{1-\beta-\delta_1}{\alpha} \quad \text{and} \quad B(N) = \frac{\beta}{1-\alpha-\delta_0}$$

are narrower than the general limits

$$A(\infty) = \frac{1-\beta}{\alpha} \quad \text{and} \quad B(\infty) = \frac{\beta}{1-\alpha}$$

It is, however, difficult to determine the exact values of δ_0 and δ_1 but lower limits d_0 and d_1 to δ_0 and δ_1 may be obtained, approximately, in almost all cases when N is large. Using these values we have the modified limits

$$A(N) = \frac{1-\beta-d_1}{\alpha} \geqslant \frac{1-\beta-\delta_1}{\alpha} \qquad \ldots \quad (6.1)$$

$$B(N) = \frac{\beta}{1-\alpha-d_0} \leqslant \frac{\beta}{1-\alpha-\delta_0} \qquad \ldots \quad (6.2)$$

6d. *Determination of d_1 and d_0* :

Suppose the hypothesis H_0 has to be tested against an alternative H_1 by the current procedure by using a sample of size N. The best critical region w of size α for this is determined from the conditions

$$\int_w P_{0N}\, dv = \alpha \qquad \ldots \quad (6.3)$$

and

$$\int_w P_{1N}\, dv \quad \text{is a maximum.}$$

Such a critical region is defined by $P_{1N} > \lambda P_{0N}$ where λ is chosen to satisfy the condition (6.3). The maximum chance of detecting H_1 is

$$\int_w P_{1N}\, dv = \gamma_1 \cdot \text{ (say)}$$

According to the sequential procedure the chance of detecting H_1 is $1-\beta-\delta_1$ and this cannot exceed γ_1 the chance associated with the best critical region. Therefore

$$1-\beta-\delta_1 \leqslant \gamma_1$$
$$\delta_1 \geqslant 1-\beta-\gamma_1$$

We may define

$$d_1 = 1-\beta-\gamma_1 \quad \text{if} \quad 1-\beta-\gamma_1 \text{ is positive}$$
$$= 0 \quad \text{if} \quad 1-\beta-\gamma_1 \text{ is negative.}$$

Similarly if γ_0 denotes the maximum value of $\int_w P_{0N}\, dv$ subject to the condition $\int_w P_{1N}\, dv = \beta$ then $d_0 = 1-\alpha-\gamma_0$ if $1-\alpha-\gamma_0$ is positive, and $d_0 = 0$ otherwise.

The solution depends on γ_1 and γ_0 whose evaluation may be difficult in the general case. If N is large and the observations are independent then $\log (P_{1N}/P_{0N})$ can be taken to be a normal variate in which case the computations become very simple.

Example 6d.1:

Consider a normal population with unit standard deviation. Let the mean value be zero under the hypothesis H_0 and 0.2 according to the alternative hypothe-

sis H_1. If the maximum number of observations is 100 what is the best sequential test?

The best test for H_0 based on 100 observations is easily seen to be $\bar{x} \geqslant 1.645/\sqrt{100}$ where $\bar{x}$ is the average of 100 observations and the level of significance is 5%. The probability of detecting the alternative hypothesis is the area of the normal curve (with zero mean and unit standard deviation) beyond the ordinate at

$$1.645 - 2.00 = -.355$$

which is about .64. Therefore by using the relation (1.1) we may set

$$A(100) = \frac{\gamma_1}{\alpha} = \frac{.64}{.05} = 12.8$$

Similarly if $\beta = .05$, then

$$B(100) = \frac{\beta}{\gamma_0} = \frac{.05}{.64} = .078$$

These limits are narrower than the general limits,

$$A(\infty) = 19 \text{ and } B(\infty) = .053$$

It is important to know that error is committed in making provisional decisions when the probability ratio lies between $B(N)$ and $A(N)$ up to the Nth stage. Let $\rho_0(N)$ be the probability according to H_0 that the sequential process leads to no decisive answer and that the ratio P_{1N}/P_{0N} lies in the interval 1 and $A(N)$. The total error committed in accepting H_1 (including the provisional decision) when H_0 is true is $\alpha + \rho_0(N) = \alpha'$ (say).

Let $\overline{P}_0(N)$ be the probability that

$$1 < P_{1N}/P_{0N} < A(N)$$

when the hypothesis H_0 is true. Since this includes cases where the probability ratio P_{1n}/P_{0n} might not lie in the interval $B(N)$, $A(N)$ for all $n < N$ it follows that $\overline{P}_0(N) > \rho_0(N)$. Hence an upper limit to the above error is $\alpha + \overline{P}_0(N)$.

In the example discussed above

$$\overline{P}_{0N} = P\{0 < 0.2 \sum_1^{100} x_i - 2 < \log A(N)\}$$

so that α' is less than 0.153. The approximation used above is very crude and the actual value of α' is perhaps, much less than the upper limit found above.

In the general case the evaluation of $\overline{P}_{0N}$ is not easy but if N is large normal approximation can be used to determine.

$$\overline{P}_{0N} = P\{0 < \log P_{1N} - \log P_{0N} < \log A(N)\}$$

as shown by Wald. In his case $A(\infty)$ is used instead of $A(N)$ so that $\overline{P}_{0N}$ is over estimated. Similarly, the error committed in accepting H_0 when H_1 is true can be evaluated.

7. SOME GENERAL COMMENTS ON DISCRIMINATORY ANALYSIS OF TIME SERIES DATA

(1) In case the original observations in time series are not known to be distributed in an apriori form then for purposes of tests of goodness of fit and discriminatory analysis it is better to consider the serial correlations which have an asymptotic normal distribution as basic data.

(2) Quenouille and Wold have not discussed the problem of determining the number of correlations to be used in their tests. In a great many problems this difficulty is not serious, for it is rather exceptional that we do not know the safe limit beyond which the serial correlations are negligible. The inclusion of such correlations which have comparatively high variances will dilute the goodness of fit test (see for example Rao, 1949) and hence it is better to stop at some stage. Before analysing the data one may decide on the number of lag correlations to be computed and then construct a χ^2 test of goodness of fit based on all the computed correlations. For instance in a long series it may often be unnecessary to go beyond 20 to 30 correlations because one cannot expect dependence of two observations which are separated by 20 to 30 years in many practical problems. On the other hand in a small series containing 15 items it is unnecessary to go beyond 7 or 8 because even large differences cannot be detected in the higher order correlation because of high variances.

(3) The difficulty can be overcome to some extent by following a sequential procedure of the null hypothesis as advocated by the author (Rao, 1950c) elsewhere with an upper limit equal to the number of computed serial correlations. This number can be equal to $n-2$ corresponding to $n-2$ possible lag correlations that can be computed from the observations but in practice a smaller number n_0 can be chosen. This number can be fixed keeping in view the order of correlations to which the test is required to detect the departure from the expected values. A high value of n_0 may be chosen to be on the safe side. In this case a bad fit will be indicated if at any stage p (in Quenouille's and Wold's tests)

$$\sum_{1}^{p} \chi_i^2 \geqslant pc(n_0)$$

where $c(n_0)$ is a suitably determined quantity depending on the level of significance chosen. The author has suggested a crude approximation to $c(n_0)$ (Rao, 1950c) pending further investigation. Some results obtained in this connexion will be published elsewhere.

(4) On the other hand, for discriminatory analysis the sequential procedure is available as indicated in section 6. The upper limit N considered in that section is equal to the number of lag correlations which one decides to compute. As before a safe number may be chosen and in fact slight variations in the choice of this number will not affect the sequential test. In a long series N may be taken as ∞.

(5) In all these situations there is a possibility that no decisive answer can be given and this is inevitable whatever may be the rule of procedure adopted. But the test always shows which hypothesis is more favourable and this is indicated by the higher value of the likelihood.

APPENDIX

Consider two p-variate normal populations defined by the mean vectors $\underline{\mu}$ and $\underline{\lambda}$ and dispersion matrices $M=(m_{ij})$ and $L=(l_{ij})$. Variables $\underline{x}=(x_1, \ldots, x_p)$ have been observed and it is desired to know as to which of the above two populations the observed vector belongs. The problem in the classification of time series is similar with $\underline{x}$ corresponding to the vector of serial correlations which are asymptotically normally distributed with mean values and dispersion matrices determined by the models from which the time series is supposed to have arisen. As suggested elsewhere a provisional decision can be made by comparing the likelihoods

$$\frac{1}{\sqrt{|M|}} \cdot e^{-\frac{1}{2}(\underline{x}-\underline{\mu})M^{-1}(\underline{x}-\underline{\mu})'}$$

and

$$\frac{1}{\sqrt{|L|}} \cdot e^{-\frac{1}{2}(\underline{x}-\underline{\lambda})L^{-1}(\underline{x}-\underline{\lambda})'}$$

The difference in the log likelihoods is equal to

$$d = \tfrac{1}{2}\{\log_e|L| - \log_e|M| + (\underline{x}-\underline{\lambda})L^{-1}(\underline{x}-\underline{\lambda})' - (\underline{x}-\underline{\mu})M^{-1}(\underline{x}-\underline{\mu})'\}$$

According to the first hypothesis

$$\delta_1 = E(d) = \tfrac{1}{2}\{\log_e|L| - \log_e|M| + (\underline{\mu}-\underline{\lambda})L^{-1}(\underline{\mu}-\underline{\lambda})' + \Sigma\Sigma l^{ij}m_{ij}\}$$

$$v_1 = V(d) \sim (\underline{\mu}-\underline{\lambda})L^{-1}ML^{-1}(\underline{\mu}-\underline{\lambda})'$$

Similarly for the second hypothesis

$$\delta_2 = E(d) = \tfrac{1}{2}\{\log_e|L| - \log_e|M| + (\underline{\lambda}-\underline{\mu})M^{-1}(\underline{\lambda}-\underline{\mu})' + \Sigma\Sigma m^{ij}l_{ij}\}$$

$$v_2 = V(d) \sim (\underline{\lambda}-\underline{\mu})M^{-1}LM^{-1}(\underline{\lambda}-\underline{\mu})'$$

The rule of procedure is to accept the first hypothesis with some confidence if

$$d - \delta_2 \geqslant \lambda\sqrt{v_2}$$

and the second if

$$d - \delta_1 \leqslant -\lambda\sqrt{v_1}$$

where $\lambda = 1.6$ (for 95% confidence) and remain in doubt or make only a provisional decision if

$$\delta_2 + \lambda\sqrt{v_2} > d > \delta_1 - \lambda\sqrt{v_1}$$

In case the inequality

$$\delta_2+\lambda\sqrt{v_2} > \delta_1-\lambda\sqrt{v_1}$$

is not satisfied for a chosen λ, it is possible to increase the value of λ till the equality

$$\delta_2+\lambda\sqrt{v_2} = \delta_1-\lambda\sqrt{v_1}$$

is achieved. With this value of λ the confidence coefficient is increased and also the doubtful region is closed.

REFERENCES

MANN, H. B. AND WALD, A. (1943): On the statistical treatment of linear stochastic difference equations. *Econometrica*, 11, 173.

QUENOUILLE, M. H. (1947): A large sample test for the goodness of fit of autoregressive schemes. *J. Roy. Stat. Soc*, 110, 123.

RAO, C. R. (1949): Some problems arising out of discrimination with multiple characters. *Sankhyā*, 9, 343.

——— (1950a): Statistical inference applied to classificatory problems. *Sankhyā*, 10, 229.

——— (1950b): A note on the distribution of $D^2_{p+q}-D^2_p$ and some computational aspects of D^2 statistic and discriminant function. *Sankhyā* 10, 257.

——— (1950c): Sequential tests of null hypotheses. *Sankhyā*, 10, 361.

WALD, A. (1947): *Sequential Analysis*. John Wiley and Sons. New York.

WOLD, H. (1949): A large sample test for moving averages *J. Roy. Stat. Soc.* 11, 297.