Statistical Techniques Research Group
Section of Mathematical Statistics
Department of Mathematics
Princeton University, Princeton, N.J.

Technical Report 41
March, 1961

DISCUSSION, EMPHASIZING THE CONNECTION BETWEEN ANALYSIS OF VARIANCE AND SPECTRUM ANALYSIS*

by

John W. Tukey**

** Also with Bell Telephone Laboratories, Inc.

Department of Army Project No. 5B 99-01-004
Ordnance R and D Project No. PB2-0001
OOR Contract No. 1715
Contract No. DA 36-034-ORD-2297

INTRODUCTION

The contents of this Technical Report was prepared for publication in Technometrics to accompany papers by Gwilym Jenkins (1) and Emanuel Parzen (2) which formed the core of a discussion of spectral analysis of time series for statisticians at the Stanford meeting in August 1960. The present account is based on an oral contribution to that discussion, but extends the treatment in a number of directions.

It is hoped that it will serve as a sueful introduction to statisticians wishing to acquire a better understanding of spectral analysis.

John Tukey

(1) Gwilym M. Jenkins 1961, "General Considerations in the Analysis of Spectra", to appear in Technometrics.

(2) Emanuel Parzen 1961, "Mathematical Considerations in the Estimation of Spectra", to appear in Technometrics.

This session was to be expository and to be directed to statisticians. Accordingly, the discussants have a responsibility to provide such comments as may tend to make both the two papers and the general subject more understandable to statisticians, particularly by relating spectrum analysis to statistical techniques and to fields of application more widely familiar to them. Fortunately, the connection between spectrum analysis and those aspects of the analysis of variance which emphasize variance components is extremely close.

One essential in this close connection is that, as emphasized by Jenkins, all practical time series problems can be treated as if time were discrete and the available data came at equally-spaced intervals. Since most problems can also be treated as if time were continuous, there will be little need for us to distinguish continuous time from equi-spaced discrete time. When we come to computation, time always can, and usually will, be discrete.

To make this connection evident, however, we shall have to analyze the implications and foundations of our procedures and thinking in classical analysis of variance more deeply than usual. It is fair to say that the spectrum analysis of a single time series is just a branch of variance component analysis, but only if one

describes its main difference from the classical branches as a requirement for explicit recognition of what is being done and why. In classical (i.e. single-response analysis-of-variance) variance component analysis, one can (and most of us do) analyze data quite freely and understandingly with little thought about what is being done and why it is being done. This is, perhaps unfortunately, not the case for the time series analysis branch of variance component analysis.

I

VARIANCE COMPONENTS AND SPECTRUM ANALYSIS

When variance components?

When conducting analyses of data in conventional analysis-of-variance patterns, we sometimes pay attention to individual values of main effects, interactions, and the like. At other times, we pay attention to estimates of variance components. The controlling factor in this choice is the character of the sets of data which would be considered to be other realizations of the same experiment (or of the same patterned observation). Thus, if we were comparing the times taken by the five outstanding runners of the world to run 1500 meters, another realization of the experiment would reasonably involve the same runners,

and it would be appropriate to pay attention to individual main effects. If, however, we were considering the speeds for a standard assembly operation as shown by five assemblers drawn at random from a pool of 250 assemblers in a large factory, another realization of the same experiment would almost certainly involve a different group of assemblers, since our concern would have been with assemblers as a whole, rather than with 5 particular assemblers. Consequently, in analyzing such data, we would pay attention to the estimated variance component for assemblers. (We are here concerned with the direct issue of what aspect of the classification concerned receives attention, not with the indirect, but perhaps equally important, issue of how the character of this classification affects the proper error term for other main effects -- the question sometimes discussed in terms of "fixed, mixed, or random models".) There is a clear analog to this choice in the Fourier-oriented analysis of time series.

Let us first consider the case of a function of time which is periodic with known period. If we may choose the time unit for convenience, the period may as well be 2π, and the function will then have (in practice) a Fourier series representation of the form

$$y(t) = a_0 + \Sigma_j(a_j \cos jt + b_j \sin jt)$$

Let us lay aside for the moment questions of errors of measurement, numbers of (and spacings between) times at which observations are made, and whether j has a finite or infinite range. Since we are statisticians, concerned with a statistical problem, the coefficients $a_0, a_1, b_1, a_2, b_2, \ldots$ are not to be thought of as constant, but rather as having some joint distribution. This joint distribution reflects the functions corresponding to "all the realizations" of the same experiment or observational program. At one extreme, the functions of time representing different realizations might all be very nearly the same. If this is the case, then, given a single realization, it is clearly appropriate to concentrate our attention upon the estimated values of $a_0, a_1, b_1, a_2, b_2, \ldots$. This is, of course, the situation envisaged in classical harmonic analysis. One opposite extreme, one which you may claim only a statistician would think of, occurs when there are parameters $\sigma_0^2, \sigma_1^2, \sigma_2^2, \ldots$ and the a's and b's are independent normal deviates with ave a_0 = ave a_j = ave b_j = 0, var $a_0 = \sigma_0^2$, var a_j = var $b_j = \sigma_j^2/2$. Given one realization of such an experiment, it is only reasonable to look at quadratic functions of the observations, and to regard them as telling us about $\sigma_0^2, \sigma_1^2, \sigma_2^2, \ldots$. Specifically it is appropriate to look at $a_0^2, a_1^2 + b_1^2, a_2^2 + b_2^2, \ldots$ and at certain linear combinations of these quantities. In contrast to classical harmonic analysis, this sort of periodic-time-function problem is a <u>variance component</u> problem

The model which lies behind the classical tests of significance in harmonic analysis, a line of development finally completed by Fisher [1929], is an incomplete mixture of the two we have just described, in which

$$y_{observed}(t) = y_{fixed}(t) + y_{random}(t).$$

In this decomposition the "fixed" component is usually thought of as involving only one, two, or perhaps three values of j, while, both most importantly and most dangerously, the "random" component is thought of as having $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_j^2 = \ldots = \sigma^2$.

Equality of the σ_j^2, the analog for periodic functions of being a "white noise", is exactly what would hold if the "random" component consisted only of independent (or merely uncorrelated) observational errors in observations equally spaced through $(0, 2\pi)$. It is also, unfortunately, exactly what is most unlikely to occur in practice (for reasons to be discussed in a moment). As a consequence, the practical applications of such "largest value against all the rest" tests of significance in harmonic analysis is, to say the least, extremely limited. (If only our estimates, $a_j^2 + b_j^2$, of σ_j^2 had more than <u>two</u> degrees of freedom, we could

improve the classical tests of significance by fitting some sort of reasonable dependence of σ_j^2 upon j, before proceeding to the construction of a significance test. Even with only two degrees of freedom, some such replacement may be possible.)

Thus, even in the case of periodic time functions, we have some situations which should be treated almost entirely in terms of means, others which should be treated entirely in terms of variance components, and still others where both descriptions should be used together.

The character of time

Time is connected. And functions of time reflect this fact in their structure, not only in the tendency toward continuity shown by individual time functions, but even more obviously in the associated probability structures. When a time function is wisely regarded as generated from constituents coming from different sources, as most are, the individual constituents are not likely to be "white noises." (Not even the measurement error constituent!) And, even more crucially, the processes by which these constituents are combined are not likely to treat different frequencies alike, so that even if the constituents were white noises, their resultant would not be one. Both in the periodic case and the more usual and general case of a continuous spectrum, a random time functions is rarely a "white noise".

Another characteristic of time is that it is quite frequently measured from an arbitrary origin. To be sure, if the simple periodic case has an annual period, we may place the computing origin of time where we will, but that will not make 1 January and 1 July the same. But if we are examing the harmonics of a 400-cycle electrical voltage, there is no equally necessary or special relation between local time and 400-cycle time. In a repetition of the same experiment, the generator phase at zero local time may well be equally likely to have any value between 0 and 2π. And if this is so, the situation is a stationary one. (This example may help to emphasize that stationarity is a condition "across the ensemble", a condition relating one realization to another, a condition on a whole ensemble, that it is not a condition on single realizations, and, most specifically, is not a condition of steadiness within individual realizations.)

Finally, phenomena in time are rarely periodic. (In fact, when examined under a microscope, no known phenomenon is precisely periodic.) Consequently, an effective Fourier description of real phenomena can rarely be a periodic description. We must allow all frequencies to contribute, and hence, as Jenkins has explained, must turn to a continuous spectrum.

The statistically vital contrast between situations appropriately describable by means and situations appropriately describable by variances continues here, as we should have expected. The motions of a springboard from which a diver has just leaped require all frequencies for their description. The motions following successive leaps by a single careful and precise diver will be relatively similar. They will, as a whole, probably be most appropriately described "by means", by a description of the typical time history of board motion. But if no diver is present, if the springboard is vibrating through a very small amplitude because the wind is blowing on the board and its supports, and because the ground itself is vibrating because of vehicle traffic and factory machinery, the situation is likely to be quite different. The characteristics of this "noise-like" motion of the springboard which are maintained from one realization to another are of the nature of variance components rather than means. And of course (as when a big grasshopper jumps off a small, wind-and-traffic-vibrated springboard) there are intermediate situations whose description appropriately combines both means and variance components.

Which variance components?

Discussion has proceeded, up to this point, as though the statement of a problem automatically fixed a set of variance components. When we think matters over

carefully, we find that this is far from being the case. In an abstract problem, where only the pattern of the observations and the symmetries of their distribution are specified, without any indication of their interpretation or understanding, there is no unique set of variance components. Instead there are many sets, each interconvertible by prescribable formulas into each other. Abstractly, the best we can do is to say that any set of quantities such that each of the second moments (pure and mixed) of the observations can be expressed as a linear combination of the quantities of the set (together with, say, the square of the average of some general mean) can play the formal role of a system of variance components. (If the quantities in some set do not behave like variances we might prefer to call them (together with the squared average) second-moment components rather than variance components, though we shall not be concerned with this particular precision of language here.) Still one set of variance components may be more convenient, and far more useful, than another. Why?

Replicated double classifications

If we examine one of the most classical patterns, a replicated double classification into rows and columns, we can learn why. Let us, then, consider a classical

analysis of variance, based on a pattern involving d observations in each of the r·c cells formed by crossing r rows with c columns. The analysis of variance breakdown into sums of squares, degrees of freedom, and mean squares is standard, as are the definitions of variance components. The well-known formulas for the average values of mean squares are, if all population sizes are infinite:

$$\text{ave}\{\text{MS} \mid \text{rows}\} = \sigma^2 + d\cdot\sigma^2_{RC} + dc\cdot\sigma^2_C$$

$$\text{ave}\{\text{MS} \mid \text{cols}\} = \sigma^2 + d\cdot\sigma^2_{RC} + dr\cdot\sigma^2_R$$

$$\text{ave}\{\text{MS} \mid \text{int}\} = \sigma^2 + d\cdot\sigma^2_{RC}$$

$$\text{ave}\{\text{MS} \mid \text{dup}\} = \sigma^2$$

Why did we choose σ^2, σ^2_{RC}, σ^2_C and σ^2_R as <u>the</u> variance components in terms of which we are to write out such formulas? We could for example, have used as variance components such average values of differences between differently related pairs of observations as, taking $i \neq I$, $j \neq J$, $k \neq K$:

$$\text{ave}(y_{ijk}-y_{ijK})^2$$

$$\text{ave}(y_{ijk}-y_{iJK})^2$$

$$\text{ave}(y_{ijk}-y_{IJK})^2$$

$$\text{ave}(y_{ijk}-y_{IJK})^2$$

Before trying to answer these questions we must look back at some of the implications of the way in which they were asked.

The term "variance component" can be, and is, appropriately used in two different senses. These senses differ in effect, but only when the underlying situations differ, so that no contradictions arise. When the underlying situation is such that it is appropriate to consider means in the first instance (the pigeonhole model of Cornfield and Tukey 1956 includes such extreme examples), variance components are means over more specific quadratic quantities. In particular, the within-cell or "duplication" variance component σ^2 is the average of the variances of all the cell populations. If these cell-population variances differ from cell to cell, so too do the values of

$$\text{ave}(y_{ijk}-y_{ijK})^2$$

since these averages will always be twice the variance of the population in the corresponding cell.

When the underlying situation is at the other extreme, so that only variance components should be considered, then the labels upon the rows and columns can wisely be regarded as purely arbitrary. This means that if the same "individual" were to appear as a row in each in two realizations of the same experiment, the numbers labeling the two rows would be quite unrelated. Such lack of relationship could be in the nature of the situation, or could have been enforced by our insistence on a randomization of the row numbers, separately for each realization, before the data was made available for analysis. But if the labels are arbitrary, we connot think of one cell, considered by itself, as different from another. Similarly, there will be only four kinds of pairs of cells: identical; in same column but not in same row; in same row but not in same column; in different rows and columns. And the four corresponding average square differences would have the following values:

$$\text{ave}(y_{ijk}-y_{ijK})^2 = 2\sigma^2$$

$$\text{ave}(y_{ijk}-y_{ijK})^2 = 2\sigma^2 + 2\sigma^2_{RC} + 2\sigma^2_{C}$$

$$\text{ave}(y_{ijk}-y_{IjK})^2 = 2\sigma^2 + 2\sigma^2_{RC} + 2\sigma^2_{R}$$

$$\text{ave}(y_{ijk}-y_{IJK})^2 = 2\sigma^2 + 2\sigma^2_{RC} + 2\sigma^2_{R} + 2\sigma^2_{C}$$

Knowing either set of four quantities, either the 4 average squared differences, or σ^2, σ_C^2, σ_R^2, and σ_{RC}^2, the other set is very easily calculated.

Why then do we prefer the first set, since they are arithmetically equivalent? It must be because of some matter of interpretation. And the interpretation must involve not the realizations of a single experiment but the comparison of two or more different experiments. In fact, we feel that, for example, the sort of change of circumstances which halves or doubles σ_C^2 while leaving σ^2, σ_{RC}^2, and σ_R^2 unaffected is easier to understand than the sort which changes $\text{ave}(y_{ijk}-y_{iJK})^2$ without affecting its three fellows. The prime criterion for selecting useful variance components is that we should be more easily able to understand the changes in the situation which would change some variance components while leaving others alone.

Known-period time functions

Let us now consider periodic time functions with a fixed period and a stationary joint distribution. One variance component description has already been given in terms of σ_0^2, σ_1^2, σ_2^2,... . (Normality is a matter of indifference to us in the present instance.) Another can be given in terms of Jowett's serial variance function [Jowett 1955]:

$$V_h = \frac{1}{2}\,\text{ave}(y(t+h) - y(t))^2$$

which, on account of stationarity, must be the same for all values of t. The formal relations between these two schemes is easily found to be:

$$V_h = \Sigma_j (\sin^2 \tfrac{jh}{2}) \cdot \sigma_j^2$$

The formal similarities between the two pairs of mutally related variance-component schemes, one for the replicated two-way table, and the other for stationary periodic time series, are very striking, but the actual similarities go deeper.

What are the simplest changes which we can contemplate making in a situation involving stationary periodic time functions? They are the results of such simple linear operations as the result of passing an electrical voltage through a simple circuit consisting of resistances, condensers, and inductances, or the result of passing a mechanical motion through a simple linkage of springs, masses, and dash pots. (Such processes occur, in particular, in almost every physical or chemical measuring instrument.) Any such linear process will affect the amplitude and phase of each harmonic in a characteristic way. If its effect on a pure jth harmonic would be to multiply amplitude by $|L_j|$, then the jth variance component of any stationary

ensemble of periodic time series (with period 2π) will be multiplied by $|L_j|^2 = L_j L_j^*$. There is no correspondingly simple result for the serial variance function. Consequently, the frequency-related variance components are much more useful than serial variance functions in dealing with stationary ensembles of fixed-period time functions.

(In highly mathematical language, the frequency variance components are a basis for second moments which simultaneously diagonalize the effects of all operations that are linear and time-shift variant--all black boxes on the sense of pp. xyz-uvw.)

It can be done with covariances!

The discussion just given stressed the analogy between classical analysis of variance and the analysis of stationary periodic time series by using averages of squares of differences of observations in both situations. It would have been possible to have stressed the analogy almost equally to have used covariances in both situations. In the replicated row-by-column pattern, we have, when the covariances are taken across the specification, from one realization to another, WITH AN ENTIRE NEW SAMPLE OF ROWS AND COLUMNS IN EACH REALIZATION:

$$\text{cov}\{y_{ijk}, y_{ijk}\} = \sigma^2 + \sigma^2_{RC} + \sigma^2_R + \sigma^2_C,$$

$$\text{cov}\{y_{ijk}, y_{ijK}\} = \sigma^2_{RC} + \sigma^2_R + \sigma^2_C,$$

$$\text{cov}\{y_{ijk},\ y_{iJK}\} = \sigma_R^2,$$

$$\text{cov}\{y_{ijk},\ y_{IjK}\} = \sigma_C^2,$$

These covariances across the ensemble are quite analogous to the serial covariances in the time series case, which are given by

$$R(h) = \text{cov}\{y(t),\ y(t+h)\}$$

where the covariance is again across the ensemble, from realization to realization, and whose relation to the frequency variance components is, formally,

$$R(h) = \sigma_0^2 + \Sigma_j(\cos jh)\cdot\sigma_j^2.$$

The main reason for approaching the analogy in terms of averages of squared differences is a pedagogical one. It seems to be easier to think about the averages of squared differences, when working from one realization to another. After all, as statisticians we are quite used to thinking about the average value of some quantity we have managed to measure only once. But it is a much further cry to think about a covariance of two quantities, each of which has been measured only once.

The qualitative nature of this distinction between covariances and averages squared differences is notably different for the replicated double classification and for stationary ensembles of periodic time series. This is due, in large part, to our tendency to expect the versions of classifications to have names, to try to think in terms of situations where means and main effects are more important than variance components. We feel that if, for example, i is a subscript identifying persons, that i = 3 should refer to a particular person, not to the third row of some randomly arranged data array.

Yet in a situation where a pure variance component approach is appropriate, the process of randomly rearranging the rows of the data array generates what we may think of, without doing too much violence to the situation, as a new (but clearly not independent) repetition of the experiment. If we fix our eyes on particular values of i, j, k, I, J, and K, consider all admissible rearrangements of the data array, and then average the simplest quadratic expressions, we are led to suitable symmetric functions of the original data array which are natural estimates of the covariances across the ensemble, provided the latter are given an <u>averaged</u> interpretation.

The usual practice in the spectrum analysis of a single stretch of time series is entirely analogous to such a procedure. Let us, for example, consider estimating $\mathrm{cov}(y_1, y_4)$. We have the original observations $y_1, y_2, y_3, y_4, y_5,$ --- . The results of shifting the time origin, one unit at a time, and always dropping observations at negative times, are first $y_2, y_3, y_4, y_5, y_6,$ --- , then $y_3, y_4, y_5, y_6, y_7,$ --- and so on. The pairs (y_1, y_4), (y_2, y_5), (y_3, y_6), --- (y_t, y_{t+3}) are "equivalent" (either because stationarity is assumed or because we want an averaged covariance) and we can calculate a "sample" covariance from these pairs. Such processes of imitating the sought-for covariance across the ensemble with a sample "covariance" wandering around the data pattern are inevitable when only a single realization is available, be it in an analysis-of-variance situation or a time series situation.

(In the time series situation, if and when we look more deeply into the details of the situation, we may find that the averages of squares of differences indeed, as Jowett has suggested [1955, 1957, 1958], have real advantages over covariances, insofar as problems associated with trends and very low frequencies are concerned. But this is for the future to reveal.)

Black boxes and the general case

A discussion exactly analogous to the one just given for stationary ensembles of period-2π time series can be given for the general case of a stationary ensemble of time series. We shall not attempt to give details here, trying only to hit the high points.

There are many circumstances under which it is convenient to call any procedure or process (be it computational, physical, or conceptual) which converts an input to an output a black box. In dealing with time series it is convenient to restrict the term black box to procedures or processes which satisfy two further conditions:

(1) The output corresponding to the superposition of two inputs is the superposition of the corresponding outputs.

(2) The only effect of delaying an input by a fixed time is to delay the output by the same time.

If the procedure or process departs from one or both it is conveniently called a colored box, with specific colors for specific sorts of departure.

Some examples of black boxes include:

(a) moving averages, such as

$$z_t = \frac{1}{h}\{y_{t-k+1} + y_{t-h+2} + \ldots + y_t\}$$

(b) time delays

$$z_t = y_{t-h}$$

(c) differences

$$z_t = y_t - y_{t-h}$$

(d) more general moving linear combinations

$$z_t = a_0 y_t + a_1 y_{t-1} + \cdots + a_h y_{t-h}$$

(e) linear electric networks (which may include amplifiers, transmission lines, and wave guides),

(f) linear mechanical systems,

(g) linear economic systems,

(h) differentiation with respect to time,

(i) integration with respect to time.

Clearly many of the most important computational, physical, and conceptual processes are black boxes in this sense.

It is easy to show (if we grant a small amount of continuity and a sufficient lack of dependence of present output on what happened at $t = -\infty$) that, if the input to a black box is $A \cdot \cos(\omega t+\delta)$, then the output has to take the form $G(\omega) \cdot A \cdot \cos(\omega t+\delta+\varphi(\omega))$, where the amplification $G(\omega)$ and the phase shift $\varphi(\omega)$ depend only upon ω.

This brings every black box into the framework discussed by Jenkins, so that

$$(\text{spectrum of output}) = [G(\omega)]^2 . (\text{spectrum of input}).$$

The important thing about this relation, for our present purposes, is that the variance component associated with a single frequency (or narrow band of frequencies) in the output is determined by the corresponding variance component of the input. There is no mixing up of frequency variance components. This is simultaneously true for all black boxes, and is the basic reason why the user, be he physicist, economist, or epidemiologist, almost invariably finds frequency variance components the most satisfactory choice for any time series problem which should be treated in terms of variance components.

II

OTHER ANALOGIES

I hope that Part I has made the close relationship between spectrum analysis of a single time series and variance component analysis very much clearer. There are similar analogies to other classical techniques. These are worthy of mention here, even though we cannot take the space to describe them in detail.

Even though the cross-spectrum analysis of two or more time series was not discussed in this session (in part because an understanding of the spectrum analysis of one time series is an essential prerequisite), it is important to point out that probably the most important aspects of cross-spectrum analysis are cases of (complex-valued, frequency-dependent) regression analysis in which the analog of a regression coefficient is the ratio of a (complex-valued) cross-spectrum density to a spectrum density, and is estimated by the corresponding ratio of estimates of averaged densities. (This fact will not surprise those who recall that a simple regression coefficient is estimated as the ratio of a sample covariance to a sample variance, or that a structural regression coefficient is sometimes estimated as the ratio of a sample covariance component to a sample variance component.) In studying time series, as in its more classical situations, regression analysis, whenever there is a suitable regression variable, is a more sensitive and powerful form of analysis than variance component analysis. As a consequence, one major reason for learning about spectrum analysis is as a foundation for learning about cross-spectrum analysis.

The other approaches to data associated, directly or indirectly, with the analysis of variance and the name of R. A. Fisher also have their analogs in the analysis of time series. We have already noted, for example, how classical harmonic analysis is the appropriate approach to known-period time functions when the over-all situation is such that one should look at means rather than at variances.

In dealing with the mean-like behavior of non-periodic time functions from a Fourier point of view, a natural and effective approach is furnished by <u>complex demodulation</u> in which the given stretch of data $\{X_j\}$ is first converted into two stretches of (real) values, viz.

$$\{X_j \cos \omega_o t\} \quad \text{and} \quad \{X_j \sin \omega_o t\}$$

which can usefully be regarded as the real and (+ or -) imaginary parts of one or the other of the complex stretches of data

$$\{X_j e^{i\omega_o t}\} \quad \text{or} \quad \{X_j e^{-i\omega_o t}\} .$$

The second step is to smooth the two real-valued stretches, smoothing both in the same way. The simplest smoothing process is the formation of equally-weighted "moving averages,"

but it is often desirable to use weights which taper down at each end appropriately. The final step is to display the result in various ways, including:

(1) Plotting individual stretches of smoothed values against time.

(2) Plotting corresponding smoothed values against one another, using time as a parameter.

(3) Plotting against time the phase or the magnitude of the complex number whose real and imaginary parts are the corresponding smoothed values.

The interpretation of such plots is usually guided by an understanding of what happens if a particular single frequency or band of frequencies are prominent in the original data. If the original data were simply $X_j = A \cos(\omega t+\varphi)$, then the values of the two modulation-product stretches would be

$$X_j \cos \omega_o t = \frac{1}{2} A \cos \left[(\omega-\omega_o)t + \varphi\right] + \frac{1}{2} A \cos \left[(\omega+\omega_o)t + \varphi\right]$$

$$X_j \sin \omega_o t = -\frac{1}{2} A \sin \left[(\omega-\omega_o)t + \varphi\right] + \frac{1}{2} A \sin \left[(\omega+\omega_o)t + \varphi\right]$$

and the result of smoothing these would be to nearly eliminate both terms if ω was not near ω_o, and to nearly eliminate the terms in $(\omega+\omega_o)t + \varphi$ if ω is near ω_o. The results of smoothing, then, would, if ω is near ω_o, be close to

$$\left[\tfrac{1}{2} A \cdot G(\omega-\omega_o)\right] \cos\left[(\omega-\omega_o)t + \varphi\right]$$

and

$$\left[\tfrac{1}{2} A \cdot G(\omega-\omega_o)\right] \sin\left[(\omega-\omega_o)t + \varphi\right]$$

where $G(\omega-\omega_o)$ is the magnitude of the transfer function of the smoothing process (which we have assumed to use symmetrical weights and thus not to affect phase). In this simple case, a cosinusoidal variation of angular frequency ω in the original, which may have been quite effectively concealed by larger contributions at other frequencies, has been demodulated, and appears as a cosinusoidal variation at the very much reduced angular frequency $\omega-\omega_o$, which is likely to be much more evident to the eye. (Complex demodulation, the calculation and smoothing of two stretches of modulation-products, is necessary if we are to distinguish the results of demodulating $\cos(\omega_o+\delta)t$ from the results of demodulating $\cos(\omega_o-\delta)t$.)

This technique is the natural extension to the nonperiodic case of the ideas underlying the classical Buys-Ballot table [Stumpff 1937, pp. 132ff, or Burkhardt 1904, pp. 678-679], the so-called secondary analysis, and Bartels's summation dial [Chapman and Bartels 1940, pp. 593-599 or Bartels 1935, pp. 30-31]. It has to be tried out on actual data before its incisiveness and power is adequately appreciated.

Problems involving the simultaneous behavior of more than two time series have not been worked on in a wide variety of fields of application, but enough has been done to point the way and suggest the possibilities. There will be an increasing number of instances where the corresponding nontime-series problems would be naturally approached by multiple regression. These can be effectively approached by multiple cross-spectrum and spectrum techniques which will be precise analogs of multiple regression in spirit and, if care is taken in choice, in the algebraic form of their basic equations. The differences which will arise in the development will stem from:

(1) the fact that regression goes on separately at each frequency (which produces merely an extensive parallelism of results), and

(2) the fact that regression coefficients will now take complex values rather than real values (which enable us to learn a little bit more about the underlying situation).

To my knowledge the multiple-time-series analogs of discriminant functions and canonical variates have not yet arisen in practice. But there would seem to be no difficulty in analogizing either or both.

III

PARSIMONY AND ERROR TERMS

Parsimony

It appears to be natural to try to set up statistical problems in such a way that the numerical values of only a few characteristics, each easily estimated from the observations, suffice to complete the fixing of a probability model for the situation. And it appears all too natural to feel that such presuppositions as normality or constancy of variance are important, since, if they failed to hold, the whole situation would not be completely fixed by the values of those characteristics which are easily estimated. But, for all such naturalness, the working statistician knows that it is often useful to estimate the mean of a population whose variance is unknown, and, similarly, that it is often useful to estimate

the variance of a population that is non-normal (frequently without trying to assess the nature and amount of its non-normality). For characteristics to be usefully estimated, it is not necessary that their values complete a precisely stated model.

It is frequently the case, that results about designing an experiment are only precise when the characteristics to be estimated complete a precisely stated model. Thus the famous telephone query, "I'm going to do an experiment, how many sheep should I use?" cannot be answered when all else that is known is that the experimenter wants to compare the means of two treatments to a precision of ±1.5 pounds of body weight, or that he wants to assess a simple variance of ±10% of itself. In the first of these instances, precise design would require a precise variance of observation. In the second, precise design would require precise knowledge of distributional shape. Yet experiments can be, and are, wisely, if not optimally, designed and validly analyzed in the absence of such precise information.

Insofar as normality is needed only (i) to ensure that knowledge of the spectrum would leave nothing else to learn, or (ii) to ensure that pre-experimental assessments of variability are precise, and these are the only reasons why Jenkins is concerned with normality, normality is not of great practical importance in spectrum analysis.

(It is fortunate that normality is moderately closely approximated to in certain applications, since there are further branches of time series analysis, for example those dealing with numbers of upcrosses or numbers of maxima, for which normality is of crucial importance. Sequences of zeroes and ones represent one ultimate expression of non-normality. In some instance, such sequences are usefully studied by spectrum analysis, in others they are not. The difference has to do with which aspects of their behavior is important.)

Indeed there is a very general principle of data analysis upon which all examiners of main effects (in analyses of variance) lean, whether they know it or not. This can be boldly stated as the Principle of Parsimony, viz., IT MAY PAY NOT TO TRY TO DESCRIBE IN THE ANALYSIS THE COMPLEXITIES THAT ARE REALLY PRESENT IN THE SITUATION. Every time that one pays attention to main effects alone, whether because they are so much larger than interactions, or because the interactions cannot be estimated with sufficient precision, or for almost any other reason, one is behaving in accord with this principle. Thus this principle is widely, though usually implicitly, adopted. The same principle applies to the quadratic analysis of time series, to spectrum analysis and its relatives, not just in a single way, but in some three or four separate and distinct ways:

Normality

The first application is to the need, or lack of need, for estimation to a complete specification, for either assuming normality or estimating more complex matters than the spectrum. In most practical situations this need is non-existent. Knowledge about the spectrum of a probably non-normal ensemble of time-functions can be useful, just as knowledge about the mean of a population of imprecisely known variance can be useful. (In either case, once the data has been gathered, consistency of repetition is the appropriate basis for judging the stability of the result, not assumptions about normality or known variance.)

Stationarity

The second application of the general principle is to the assumption of stationarity, the analog in time series situations to the assumption of constancy of variance in more classical situations. The assumption of stationarity is one at which the innocent boggle, sometimes even to the extent of failing to learn what the data would tell them if asked. Yet I have yet to meet anyone experienced in the analysis of time series data (Gwilym Jenkins is an outstanding example) who is over-concerned with stationarity. All of us give some thought to both possible and likely deviations from stationarity in planning how to collect or work up data, but no one of us will allow the possibility of non-stationarity to keep us from making estimates of an average spectrum, any more than working

analysis-of-variance statisticians will refrain from estimating a variance component because the variability thus assessed may well have to be an average.

The fact that the spectrum is changing with time (or elevation, or azimuth) need not make it unwise to estimate one, or several, average spectra. The detection of waves 1 millimeter high, 1 kilometer long, with a 10,000 kilometer fetch [Munk and Snodgrass 1957] was based upon estimates of spectra averaged over four-hour periods. The crucial point in identifying the length of the fetch was the rate of change of the center frequency of this distinctive, but very small peak, from one four-hour period to another. Once we admit that we are estimating an average spectrum, we have admitted that there may well be other relevant characteristics of the situation beyond the spectrum, that estimation is not completing specification. Such an admission, as this example shows, is a good thing rather than a bad one.

There seems to be extra reluctance to consider an average spectrum. It is hard to be sure of the principal reasons for this, but a well-founded desire for replication as a basis of security is likely to be one. If only one time series is available for analysis, as is far too often the case in so many economic instances, it is comforting to believe that, somehow, stationarity makes it possible to have "replication" from one time period of another. The truth is not so comforting. Stationarity is frequently absent. Even when stationarity holds, something like "replication"

can only occur within the limits of a single stretch of moderate length if the true spectrum is devoid of detailed features (is sufficiently smooth in the small). And it is surely not wise to trust in "replication" that may not be there.

Harry Press notes (private communication) that average spectra may hide an important departure from stationarity. In an entirely similar way, the use of analysis of variance on the results of an experiment comparing 12 treatments in randomized blocks may hide a substantial dependence of variability upon treatment, or a substantial dependence of treatment effect upon block. These things can, and do happen. The possibility of their occurrence must be carefully kept in mind. But this fact is not relevant to the point we have just been discussing.

Surely, if one has both adequate data and scientific or insightful ground to fear non-stationarity, it will be wise not to average spectra over too long a time. But the urge to choose the averaging time wisely is strengthened by an understanding that all data analyses estimate average spectra.

Wisely-chosen resolution

The third application of the general principle is to the question of the narrowness of the frequency ranges for which we should seek spectrum estimates. There are infinitely many frequencies. The number of separate frequencies over which we could seek estimates from a given body of data

is limited by the extent of the data, and grows without limit as longer and longer pieces of data become available. But it does not follow that we should always, or even usually, work close to this limit. The analogy with an interaction mean square in a row-by-column table is close and persuasive. There are r·c individual estimates of the interaction mean square, each based on just one of the residuals which remain after fitting rows and columns, each involving just one degree of freedom. How often does it pay us to calculate and compare all these separate estimates? Only very rarely. (It is often useful to calculate and compare a few estimates of an interaction mean square, each based on a reasonable portion of the available degrees of freedom.) The position with spectrum estimates is analogous and similar; to be effective we must estimate averages over well-selected frequency ranges. (This is in addition to the averaging over time necessitated by lack of perfect stationarity.) In both instances, interaction mean square and spectral estimate, it does not pay to try to estimate too much detail, even if the detail is really there.

Proper error terms

The question of the proper error term is a classic of the analysis of variance, often relied upon to separate the men from the boys and the pastry cooks. It is well recognized that, for example, the plot-to-plot error of an agricultural experiment is almost certain to be too small, specifically because it rules out place-to-place and

year-to-year components of variation. It is not too great a stretch to consider this question, which arises for time series in an only slightly different form, a fourth example of the general principle of parsimony. For while it will not be costly to estimate plot-to-plot variance, it is likely to be costly to trust it, to use such estimates as error estimates. Even its estimation may be costly, in the agricultural situation, if the result is to expend too much effort on choosing the optimum plot size, on doing one's best to reduce what may be a minor source of variation. As Jenkins points out at the very end of his paper, it is not uncommon for spectrum estimates based upon different experimental repetitions to differ more than might be expected from their internal behavior. (Statisticians familiar with any of a wide variety of other situations would be surprised if this were not so, if external error were not larger than internal error.) As a consequence, it is not likely to be worth while to expend too much effort in using estimates whose windows have optimum widths and optimum detailed shapes, since this may mean exerting a large effort to minimize a minor component of variability.

One way to describe matters is in terms of alternative ensembles. In each repetition of the experiment, the time series which is actually realized is drawn from

a different ensemble (from a different population each element of which is a whole time series). Such a description is entirely analogous to a description of an agricultural experiment in which each local comparison of two treatments is drawn from a population, but the populations for different "places" or "years" differ. The fact that matters may be appropriately described in such a way often affects what we wish to estimate. If an average comparison, in the agricultural situation, depends upon the "place" in a way, or for reasons, that we do not understand, we are usually driven to estimate, not average responses at individual places, but rather average responses for all places. (These are the natural "main effects".) There are situations, however, as for example when studying a cheaper substitute to see if it causes occasional deleterious effects, where we may need, because of variation from place to place, to estimate the value of the least favorable average response and, perhaps, the frequency with which similarly unfavorable situations will arise in more extended practice. The situation with time series is exactly similar.

Most of the time we shall be driven to estimation of a spectrum averaged over repetitions, where the pattern, or the causes, of the changes in spectrum from repetition

to repetition are not understood. This averaging over repetitions, forced on us by alternate ensembles, is superposed upon the averaging over time within repetition, partially forced upon us by non-stationarity, and upon the averaging over frequency bands, forced upon us by the limited extent and amount of our data. What we estimate, then, is an average of averages of averages. We have come a long way from the idea of a tight specification-estimation relationship, where everything which is not presupposed should be estimated. But it is well that we have done so. And no one who has considered carefully what is estimated by a main effect in a reasonably complex analysis of variance can maintain that so much averaging is surprising or unusual.

Just as in more conventional areas of statistical application, there are situations, the comparison of vibration intensity with structural strength being perhaps the most obvious, where we shall need to estimate not the average spectrum but some upper limit, perhaps an upper 99% limit, for the spectra in the various replications, for the spectra of the various alternative ensembles. But such instances are the exception, not the rule.

Effects upon balance between stability and resolution

In any case, the presence of true differences between repetitions, of differences between the spectra of the alternative ensembles, will surely force a readjustment of the balance between stability and resolution. The main reason for estimating average spectral densities over relatively broad frequency bands is to assure moderate stability of estimate. If variation within ensembles should be small compared to variation between ensembles, such within-ensemble stability is of little value to us. Thus we can afford, in such circumstances, to improve our frequency resolution by estimating spectral densities averaged over narrower bands. (There will still remain a natural limitation on resolution, however, associated with the limited duration of the individual ensembles.)

IV

SPECIAL PROBLEMS OF TIME SERIES

Resolution

The notion of resolution, as applied in optics and other branches of physics, is a well-recognized and useful physical concept. It does not have any single definition in numerical terms, and it is well that it does not. For the

general idea that "higher resolution" means "capable of detecting more detail" is clear, while any one way of making it quantitative would not be universally satisfactory. (If you like, "resolution" is not "unidimensional". But whether you like this fact or not, it would be unwise to make it unidimensional by a fiat of definition.) Jenkins and Parzen have introduced us to a number of definitions of bandwidth. There are, and will be, other such definitions. The value of any of them lies in what the values of the variously defined bandwidths tell us about "resolution". No one definition, nor even all the definitions so far given, can tell us all about resolution. As Goodman pointed out in his verbal discussion, such matters as "rejection slope in db/octave away from the major lobe" or "db of rejection at a particular frequency" can be important in particular circumstances. Thus numerical values of bandwidths according to any definition closely related to "resolution" can help us, but they will help us most if we regard them as telling us part, not all, of the story.

Choice of resolution

There is one matter upon which I should not like to have my views misunderstood: the desirability in exploratory work of making spectral analyses of the same data with

different resolutions (usually represented in packaged systems of calculation of spectrum analysis by the use of varying numbers of lags in the initial computing step, which is the calculation of sums of lagged products). Let me be quite clear that, in my judgment and according to my experience, it definitely is very often desirable in exploratory work, and sometimes essential, to make analyses of the same data at differing resolutions. Moreover, it may be equally important to use different window shapes and different prewhitenings.

The place where Jenkins and I differ seriously, at least verbally (and I suspect the difference is more verbal than actual) is in the utility of examining some sequence of mean lagged products as a firm basis for choosing the number of such values to be inserted in an appropriate Fourier transformer, and transformed into spectral estimates. Our difference is greater still in connection with the adequacy of the point of apparent "damping down" of these values as a basis for choosing this number. It is not that knowledge of the "damping down" lag is not useful, but rather that, at least in my view, its unthinking use may be dangerous.

On the one hand, I have known of cases where the useful estimates of power spectra came from stopping well short of the damping-down point. On the other hand, if the spectrum were to contain one very large, very broad, very

smooth peak, and a close group of small, narrow peaks, the mean lagged products would appear to damp down at a lag associated with the width of the large broad peak, so that a spectrum whose resolution was associated with this damping-down point would fail to resolve the close group of small peaks. Here, as in all sorts of data analysis, there is no substitute for careful thought combined with trial of various alternatives.

It is natural to be tempted into calculating more spectrum estimates than the number of mean lagged products used as their basis. This temptation need not be a dangerous one, once it is realized that, given the mean lagged products and the shape of the window, all the possible spectrum estimates lie on a cosine polynomial of degree equal to the number of lags used. Once the usual number of spectrum estimates have been calculated, they are enough to determine this polynomial, and the calculation of further estimates is equivalent to a process of cosine-polynomial interpolation. This does not mean that calculating more estimates is useless, or that the results of further calculation will lie close to the results of straight-line interpolation between the poins already calculated. But it does mean that the additional estimates provide no new information, only more detailed exposition of information already present. And it means that drawing smooth

freehand curves through the original spectral estimates is often much more useful than connecting them by segments of straight lines.

Blurred estimands

In discussing the general principle of parsimony we emphasized the need to estimate averages over bands of frequencies. This point is so central to spectrum analysis as to make its heuristic and intuitive understanding worth considerable effort. Let us begin with classical situations. If one has more degrees of freedom than variance components, then one can find estimates of some (and perhaps all) of these variance components whose average values do not depend upon the other variance components. But once there are more variance components than degrees of freedom, this need not be the case. Consider a two-way r-by-c array of observations in which there are r·c+2 variance components, viz. a rows variance component, a columns variance component, and one variance component for each of the r·c cells. (This is a natural model when the variance of the cell contributions varies irregularly from cell to cell.) In this situation there is no estimate of any of the r·c cell variance components whose average value is free of all the other variance components.

In the time series case there are very many more variance components than degrees of freedom. For, unless some periodicity assumption holds perfectly (and I know of not a single instance where it does), a contribution of the form

$$A \cos \omega t + B \sin \omega t$$

is permissible for any value of ω in some interval. And as statisticians know from bitter experience, at least all the things that are permissible will happen. Thus, in principle, there are infinitely many variance components, one for each possible ω. And, when the realities of band-limiting and of finite duration of data are faced, there are only a finite number of observations available, and hence only a finite number of degrees of freedom. There is no hope of estimating all variance components here, even by using impractically unstable estimates.

Bracketing undesired effects

Let us return, for the moment, to a situation with a finite number of variance components, only four of which will enter our discussion. Let us suppose that we are interested in estimating a particular one of these variance

components, σ_1^2, and that our choice has narrowed down to three quadratic functions of the observations, whose average values are

$$\text{ave}\{A\} = \sigma_1^2 + 0.04\ \sigma_2^2 - 0.02\ \sigma_3^2 + 0.01\ \sigma_4^2$$

$$\text{ave}\{B\} = \sigma_1^2 + 0.06\ \sigma_2^2 + 0.04\ \sigma_3^2 + 0.02\ \sigma_4^2$$

$$\text{ave}\{C\} = \sigma_1^2 - 0.08\ \sigma_2^2 - 0.05\ \sigma_3^2 - 0.03\ \sigma_4^2$$

So long as we insist on using only a single quadratic function of the observations, the choice of A, whose average value is least affected by σ_2^2, σ_3^2, and σ_4^2 has a real advantage. But if we were willing to look at two quadratic functions of the observations together, then B and C are a more effective choice, at least so far as average values go. For, on the average, one is raised by the other variance components, while the other is lowered. If, for example, the observations are replicated m times, so that there are m A's, m B's, and m C's, and so that, consequently,

$$\bar{B} + ts_B/\sqrt{m}$$

is an upper confidence limit for ave B, while

$$\overline{C} - ts_C/\sqrt{m}$$

is a lower confidence limit for ave C, then the interval

$$(C - ts_C/\sqrt{m},\ B + ts_B/\sqrt{m})$$

is a confidence interval for σ_1^2, without regard for the values of σ_2^2, σ_3^2, and σ_4^2. (No such confidence interval can be based upon the m values of A.) Whenever we cannot get estimates (of what we want to estimate) whose average values are wholly free of what we do not want to estimate, the use of such paired estimates, one underestimating and the other overestimating, is likely to be useful and, perhaps, even necessary.

When we make estimates of spectrum densities, the window which relates the average value of our estimate to the spectrum is (for the apparently inescapable case of equally-spaced data) inevitably a cosine polynomial (of degree no larger than the index of the longest lag used). It can vanish at only a finite number of points. Consequently its main lobe, which points out the band of frequencies over which we seek to estimate some average spectrum density, is inevitably accompanied by minor lobes which allow <u>leakage</u>

from the parts of the spectrum outside the desired band to affect the average value of our estimate, and hence to affect its individual values. Even if we are willing to accept the blurring due to averaging within the major lobe, as we must, like it or not, we are rightly reluctant to face unknown possibilities of leakage from other parts of the spectrum. The cure is the same as for the example with four variance components: use two estimates. (This time one estimate should have all minor lobes negative while the other has all minor lobes positive.) This general situation is discussed more fully elsewhere [Tukey 1961(?)], and it is to be hoped that some suitable pairs of estimates will soon be explicitly available. (For one pair see Wonnacott 1961.)

Kinds of asymptosis

The purpose of asymptotic theory in statistics is simple: to provide usable approximations before passage to the limit. Consequently asymptotic results and asymptotic problems are likely to be of limited utility when the finiteness of a sample size or of some other quantity is of overwhelming importance. (Thus, for example, the theorem that maximum likelihood estimates are asymptotically normally distributed with a certain variance-covariance matrix is

rarely of any use when there are only 1 or 2 degrees of freedom for error.) It is sometimes hard, but almost always important, to remember this fact.

Time series analysis follows its usual pattern, "like most statistical areas, only more so!", insofar as asymptosis is concerned. For there are <u>three</u> distinct ways in which time series data could tend toward a simplifying limit:

(1) The total extent of all the stretches of data available could become more nearly infinite.

(2) The extent of each individual stretch of data could become more nearly infinite.

(3) The bandwidth of the measurement could become more nearly infinite (requiring a more nearly vanishing interval between times of recording).

The consequences of these three, which are quite distinct, depend upon whether the resolution of the estimates to be made (a) remains constant, (b) increases as fast as the total extent, extent, or bandwidth of the data, or (c) behaves in an intermediate manner.

If (1) occurs without (2) or (3), the possible resolution does not increase, so that (a) is the only relevant situation. The stability of individual estimates of (averaged) spectrum density then increases essentially proportionally to the total extent of data.

If (2) proceeds, (1) must also. If (2) and (1) proceed without (3), the range of (aliassed) frequencies to be considered will not change, so that a constant number of estimates corresponds to constant resolution, and to an increase in stability essentially proportional to total extent of data. If, on the other hand, the resolution is increased proportionally to the total extent of data, the stability of individual estimates will remain constant.

If (3) proceeds without (1) or (2), we may make estimates over a wider and wider frequency range, but we cannot obtain higher and higher resolution. For constant resolution, we obtain constant stability.

In practice, where there are several repetitions, several stretches of data, it may be that we can wisely treat the <u>total</u> extent of all data stretches asymptotically (especially when the additional variability in external error should be considered), but I know of <u>no single practical instance</u> where an asymptotic treatment of <u>either</u> stretch length <u>or</u> band-limitation gives useful results.

The limitation on ultimate resolution due to limited extent of data stretches, and the limitation on frequency ranges for which estimates can be made due to band-limiting, always seem to behave like small-sample phenomena, and must be faced in detail. They do not at all behave like large-sample phenomena, where everything can be "smoothed out" and treated in a limiting, continuous way.

V

THE MORAL

To analyze time series effectively we must do the same as in any other area of statistical technique: "Fear the Lord and Shame the Devil" by admitting that:

(1) The complexity of the situation we study is greater than the complexity of that description of it offered by our estimates.

(2) Balancing of one ill against another in choosing the way data is either to be gathered or to be initially analyzed always requires knowledge of quantities which cannot be merely hypothesized, and which, in many cases, we cannot usefully

estimate from a single body of data, such as ratios of (detailed) variance components or extents of non-normality. Theoretical optimizations based upon specific values of such quantities may be useful guides, but only when the failure of past experience (and the present data) to give precise values for these quantities is recognized and allowed for.

(3) There is no substitute for some sort of repetition as a basis for assessing stability of estimates and establishing confidence limits.

(4) Asymptotic theory must be a tool, and not a master.

The only difference is that one must be far more conscious of these acceptances in time series analysis than in most other statistical areas.

In a single sentence, the moral is: ADMIT THAT COMPLEXITY ALWAYS INCREASES, FIRST FROM THE MODEL YOU FIT TO THE MODEL YOU USE TO THINK AND PLAN ABOUT THE EXPERIMENT, AND THENCE TO THE TRUE SITUATION.

VI

THREE MYSTERIES

Up to this point, we have been concerned with the fundamentals of time series analysis and with the close and cogent analogies between time series analysis and other areas of statistics. As a consequence our remarks have related most closely to the first of the two papers. It is now time to turn to the second paper, which grapples with some of the more detailed aspects of time series analysis. Here it seems best to try to shed light on a few of the aspects which are likely to seem most mysterious. Our attention will be given to the mysterious importance of dividing sums of lagged products by n rather than by n-k, to the mystery of how new window patterns are sought, and to the mysterious importance of choosing a window.

Does the divisor matter?

The major computational effort, as measured in millions of multiplications or minutes of machine time, of any conventional careful spectral analysis is expended on the calculation of the sums of lagged products

$$\Sigma_{(k)} = \sum_{i=1}^{n-k} X_i X_{i+k}$$

(If these are calculated for k=0, 1, 2, ..., m, some $(m+1)n - m(m-1)/2 \sim m \cdot n$ multiplications will be required.) The X_i in this calculation will be raw, or prewhitened, or otherwise modified observations, from which means, fitted polynomials, or other fitted trends may or may not have been subtracted. Unless unusually careful preparatory steps for the elimination of very low frequencies were already taken in the preparation of the X_i, the next step after calculating these sums of lagged products will be adjustment of these sums of lagged products for means or trends. It is vital to deal in practice with such adjusted sums of lagged products, as almost everyone who enters upon time series analysis seems to have to learn for himself. (However, it will save space and, hopefully, promote clarity if we omit the word "adjusted" during the remainder of this discussion. We shall omit it.) Having been told of <u>sums</u> of lagged products, every analyst of variance expects us to go on to <u>mean</u> lagged products. Going on is inevitable.

There is a question of the appropriate divisor. If we had not corrected for the mean (or any trend) there are cases to be made for both n and n-k. If we had corrected for, say, a general linear trend (which absorbs 2 degrees of freedom), there are cases to be made for n, for n-2, for n-k and for n-k-2. Parzen gives attention, between his (4.6) and (4.7), to some of the reasons for choosing n or n-2 rather

than n-k or n-k-2. By analogy with the analysis of variance we might feel that n-k-2 (or, when no adjustment is made, n-k) would be desirable because unbiasedness is good. The unbiasedness argument is found not to be a strong one in the time series situation.

Is this choice an important one for the analyst or investigator whose concern is with the spectrum? You should be happy to be told that the answer is "no". If one's concern is with the spectrum, then the most important thing about any quadratic function of the observations is the spectrum window which expresses the average value of the estimate in terms of the spectrum of the ensemble. (The next most important thing is, of course, the variability of the quadratic function.) This is just what we should expect for a variance-component problem, where means and other <u>linear</u> combinations of the observations are without direct interest. For if, in some very complex (probably unbalanced to begin with, and then peppered with missing plots) analysis of variance, one is given the values of certain mean squares (or other quadratic functions of the observations), the first question one concerned with variance components asks is "How are the average values of these mean squares expressible in terms of our variance components?". (The question about stability "How many degrees of freedom should be assigned to each?" is important but secondary.)

If we know the windows associated with our spectrum estimates, we need not be concerned, in the first instance, with how these estimates were obtained. And, moreover, any linear combination of the results of dividing the sums of lagged products by n is also a linear combination of the results of dividing the sums of lagged products by n-k, and vice versa.

The practicing spectrum analyst need not be concerned with division by n or n-k, so long as he doesn't misassemble formulas by combining some which are appropriate for one divisor with others appropriate for the other.

However, those interested in the theory of spectrum analysis do need to give some attention to this choice, partly because of the reasons given by Parzen, partly because this choice affects just what functions of frequency the mean lagged products are Fourier transforms of, partly for various other reasons. The man who has a practical interest in the autocovariance function, if there really be such, clearly also has to take an interest in alternative estimates.

Unlikely though it may seem at first, there is a moderately close analogy between the biased estimates supported by Parzen and biased estimates which are reasonable in classical analysis of variance. Consider data in a single classification with r observations in each class, so that the between mean square has average value $\sigma^2 + r\sigma_1^2$, where σ^2 is the error variance component, and σ_1^2 is the between variance component.

If we wish to estimate the population average corresponding to a particular classification, there is little doubt that the sample mean for that classification is the most reasonable estimate. But if we wish to depict the pattern of the population averages corresponding to all classifications, we should do something about the inflation of this pattern by error variance; we should replace the pattern of observed means by a suitably shrunken pattern. (In the simplest cases it may suffice to shrink each classification mean toward the grand mean by the factor $[r\sigma_1^2/(\sigma^2+r\sigma_1^2)]^{1/2}$. In others the method developed by Eddington for dealing with stellar statistics [Trumpler and Weaver 1953, pp 101-104] may need to be applied.) The analogy with the time series case is reasonably, in fact surprisingly, close. If we wanted to estimate just one autocovariance, we should undoubtedly use the unbiased estimate. But if we are concerned with the pattern made by the estimated values, with the nature of the autocovariance function, we may, as Parzen points out, do better to use the biased estimate.

(The extreme instance of the problem underlying this choice in the time series case arises when one 5-minute record is "cross-correlated" [really cross-covarianced] with another 5-minute stretch of the same time series, as recorded an hour, a day, or a week later. If the spectrum of the ensemble is relatively sharp, the average

value of the covariance will still tend to zero, but the average value of its square will tend, not to zero, but to a value depending upon the product of the 5-minute duration with the width of the spectral peak. Thus if one calculates autocovariances at lags from 24 hours 0 minutes to 24 hours 5 minutes one will almost certainly find an apparently systematic wavy pattern in the unbiased estimates of autocovariances or autocorrelations computed for a particular realization. It is natural to believe that this pattern is "real", although the true average values of the autocovariances are actually very, very much smaller in magnitude than the values found from a single realization. Such patterns can be so regular as to mislead investigators into an unwarranted belief that the presence of a strikingly accurate underlying clock has been demonstrated.)

How can I construct a window?

If we leave aside a few matters which really do not matter here, although some of them are very important elsewhere (such as adjustment for the mean, other devices for rejection of very low frequencies, and division by n-k not n), the function of lag by which the mean lagged products are multiplied before Fourier transformation, and the window (expressed in terms of $\omega-\omega_o$ and $\omega+\omega_o$ separately, where ω_o is the center frequency of the estimate) through which the

spectrum determines the average value of the estimate, are Fourier transforms of one another. (If you have never followed a derivation of this, just take it on faith.) Since every lag must be a multiple of the data interval, one of these functions is a finite array of spikes, spaced one data interval apart. The other function is a polynomial in $\cos(\omega-\omega_o)$ of an appropriate degree.

While the discreteness of time is generally an important aspect of the data, it is not important for our present purposes, so that we may replace the spiky lag window by a smooth function of a continuous variable without altering its Fourier transform in any way which is essential to the present discussion. (Provided that we began with, say, at least 10-20 spikes.) Since we are going to calculate mean lagged products for only a finite number of lags, this continuous lag window must vanish outside a finite interval. If it were possible, we would like to have its Fourier transform, the corresponding spectrum window, also vanish outside a finite interval, for then the average value of the corresponding spectrum estimate would only involve contributions from a restricted part of the spectrum.

It is, however, well known that a function and its Fourier transform cannot <u>both</u> vanish outside finite intervals. Indeed, they cannot <u>both</u> go to zero too rapidly as their arguments tend to infinity. The standard example of a

function which, together with its Fourier transform, goes to zero rapidly at infinity is the standard normal density function, which together with its Fourier transform, goes to zero as the negative exponential of half the square of its argument. Unfortunately, we cannot make use of the normal density as a lag window, because it does not vanish outside a finite interval.

Every statistician knows, however (or so the phrase goes), how to approximate a normal distribution by a bounded distribution. It is only necessary to consider the distribution of means of simple random samples from any bounded parent distribution. And what parent distribution could be simpler than the rectangular (uniform) distribution? If we take samples of size k, the Fourier transform of the distribution of means will be of the form $(\sin u/u)^k$, where u is a multiple of $\omega-\omega_o$, depending upon k and the number of lags used. The larger is k, the smaller are the minor lobes of this window in comparison with the main lobe, and the more lags are required to give a main lobe of prescribed narrowness. If k=1, which corresponds to a raw Fourier transform of the mean lagged products, the minor lobes adjacent to the main lobe are about 1/3 the height of the main lobe (and negative), which proves to be impractical. If k=2, which corresponds to line 1 in Parzen's Table 1, the minor lobes are at most 1/9 the height of the

main lobe, and the resulting spectral window, often called the Bartlett window, is everywhere positive. If k=4, which corresponds to line 8 in Parzen's Table 1, and to $h_3(u)$ in his Table 2, the minor lobes are at most 1/81 the height of the main lobe, and the resulting spectral window, as Parzen shows, is quite effective.

It would be perfectly possible to use k=8 or k=16 if we wished even lower minor lobes. The cost to us of doing this would be twofold. There would have to be an increase in computational effort in order to provide mean lagged products for the additional lags required to give a main lobe of comparable width. And the shapes of the main lobes would be somewhat less favorable, since the process of raising the window to a higher and higher power will make both the minor lobes and the lower portions of the main lobe still lower. As a result the main lobe will "occupy" a smaller and smaller part of the frequency band between the zeroes (of the window) which define it, and, consequently, the variability of the corresponding estimate (leakage aside) will be greater than that of an estimate with a more "blocky" spectrum window.

As is clear from Parzen's paper, these are not the only useful lag windows, the "cosine-arch" or "hamming" lag window which is proportional to "one plus cosine" being also

of practical interest. This latter window was "discovered" by empirical observation, and the best reason for considering it are the properties it is found to have.

(Two further easily understandable types of window which may sometimes prove useful may be obtained respectively, (i) by taking a truncated normal distribution as the lag window, (ii) by taking a Čebyšev polynomial for the spectral window. This last choice makes all minor lobes of equal height, and as small in comparison with the main lobe as is possible for a given number of lags. This equality of height, which makes the minor lobes adjacent to the main lobes lower than those of most other windows but makes minor lobes far away from the main lobes relatively higher than those of most other windows, seems to prove to be a disadvantage rather more often than it proves to be an advantage.)

How important is window choice?

We have discussed window carpentry briefly. Now we need to ask what does it buy us, how much better can we do with a specially constructed window than with a rather routine one. This question has opposite answers, depending on whether one relies upon his window to do everything for him, or not.

If one relies solely upon windows, faces a peaky or steeply slanting spectrum, and is concerned with the behavior of the spectrum where the density is noticeably

below its highest values, then the quality of workmanship and polish of the window used can easily be of the utmost importance. (During the early '50s I spent considerable effort on a variety of ways to improve windows. The results have never been published because it turned out, as will shortly be explained, to be easier to avoid the necessity for their use.)

If one applies his windows, actually or effectively, not necessarily to the original data but, whenever useful, to the results of simple linear modifications of the original data, chosen so as to depress peaks, to raise valleys, and, where necessary, to remove narrow peaks (which may appear to be "lines"), he will rarely, if ever, find any need for anything beyond a window of routinely good quality, such as the hamming or cosine arch window (or, if a slight increase in variance of estimate and a substantial increase in computational effort are worth bearing, the k=4 window described above). (For discussion of techniques of linear modification see Blackman and Tukey 1959, Holloway 1938, and, perhaps, the work of the Labroustes referred to by Chapman and Bartels [1940, p. 992] and Blackman and Tukey [1959, p. 180].) In my own experience this sort of approach to the problem, which corresponds [Blackman and Tukey 1959, p. 42] to using different window shapes in different frequency bands, is much easier than seeking out explicit forms for very special

windows to meet each special situation. Moreover [e.g. Blackman and Tukey 1959, pp. 62-63; Tukey 1959, pp. 315-316], consideration of this technique leads to very helpful insights into how the data is best gathered in the first place.

But each of us is entitled to do his calculations as he pleases, so long as he does adjust his techniques to provide the amounts of precision and stringency his problems require.

VII

COMPUTATIONAL CONSIDERATIONS

It is important to say something about the role of computational efficiency and computational choices as considerations in time series analysis. Computational considerations are particularly important in time series analysis, in part because of the relatively large amounts of data processed, in part because of the very many multiplications involved in obtaining sums of lagged products, and in part for more subtle reasons. And it is sometimes hard, especially for the novice, to separate computational, statistical, and aims-and-purposes considerations, one from another. Yet if they are not separated, neither sound practices nor sound advice can be understood as such, rather than being taken on faith.

Computational considerations depend very much on the equipment available. Crude spectral analysis is possible with paper and pencil [Blackman and Tukey 1959, pp. 151-169], and modestly refined computations have been done on hand calculators. The beginning of effective spectrum calculation probably involves the use of punched-card tabulators to obtain sums of lagged products (by applying progressive digiting to cards obtained by off-set reproduction [Hartley 1946] and the conduct of all further computation on hand calculators. The steps from this to fully automatized spectrum analysis on machines of the capacity and speed of an IBM 7090 or CDC 1604 are many and long. The reluctance or eagerness with which one faces another hundred thousand multiplications depends very strikingly on the equipment available.

And, consequently, so does one's attitude toward using many more lags to improve window shape or increase resolution, or toward recomputing mean lagged products whenever new spectrum estimates (estimates differing in resolution, in window shape, in prewhitening, or in rejection filtration) are to be obtained from the same data. In the economy of abundance which goes with modern electronic computers, I prefer to recompute mean lagged products when a new set of spectrum estimates are required, but others feel quite differently. Some of the reasons for this difference can be made manifest, and their mention may serve to illuminate a variety of computational issues.

To recompute or not to recompute?

First, recomputation when necessary allows the use of packaged, unified machine programs, which require only values for a few constants and the data in order to provide the desired spectrum estimates. This makes it much easier for those unsophisticated in time series analysis, whether investigators or technical aides, to process data more easily and effectively. Most data analysis is going to be done by the unsophisticated. As statisticians we have a responsibility to package as many techniques as possible for safe and effective use by those who will analyze data, and who will not understand why the choices in the package were made wisely or unwisely.

Next, and perhaps more important for the present, is the absence of adequate facilities for data analysis. There is no data-analytic language analogous to FORTRAN or ALGOL, in whose terms it is easy to describe the operations of data-analysis, and, what is far more crucial, I know of no large machine installation whose operations are adapted to the basic step-by-step character of most data analysis, in which most answers coming out of the machine will, after human consideration, return to the machine for further processing. Neither programming languages or computer center operations are adapted to stepwise operation, and all of us who use big machines for data analysis are thus forced to more unified operation than might otherwise be desirable.

Third, and this consideration is not related or restricted to big machines, stepwise computation tends to produce stepwise thinking. I believe that stepwise thinking led to the classical Schuster periodogram, and hence to decades of ineffective quiescence for frequency oriented analysis of time series. The individual steps from data through intermediate results to periodogram ordinates seemded reasonable each by itself. And while Stumpff's book recognized the nature of the corresponding spectral window before 1940 [Stumpff 1937, pp. 98-100], nothing was done to provide more useful estimates until people began to relate average values of estimates to the spectrum of the ensemble of which the data is one realization. What security we can have in frequency-oriented time-series analysis comes from over-all thinking, while many of the most threatening dangers come from step-by-step thinking. Thus we often do very much better to apply over-all processes (which have been thought through over-all, not merely stepwise) to data than to apply the individual steps separately. This view does not deny the great desirability of "try, look, and try something a little different" as the typical pattern of data analysis. It merely asks that each trial, unless it is extremely exploratory, be thought through as a unit. It does not even say that it is unwise to calculate sums of lagged products once and for all. It only calls on those

who do so to be sure that the total processes they apply to data have been thought through as wholes. It does, however, note that using preplanned packages increases the chances that such thinking will have been done.

Precision may matter

Finally, there is a question of required precision of arithmetic. Let us approach this somewhat indirectly. In friendly conversation, James Durbin recently brought firmly to my attention that there was an alternative to first prehitening the observations and then calculating sums of lagged products for these modified values, remarking that one might, instead calculate rather more sums of lagged products for the original observations, and then calculate the suitable simple linear combinations of these sums which would be identically equal to the sums of lagged products for the modified observations. This remark is surely well taken. The results are algebraically identical. And if spectrum estimates for the results of enough different prewhitenings of the same data are going to be required, then the computational path suggested by Durbin will surely have real advantages. But it behooves us equally to consider the possible disadvantages of this alternate approach. Perhaps the greatest of these is the likely requirement of greater precision of arithmetic (although it is interesting

to note that, if only one set of spectrum estimates is to be calculated, prewhitening first will even save some multiplications).

This statement about accuracy sounds a little peculiar at first to one familiar with more classical statistical computations, but when he recalls the advantages of postponing divisions in calculating sums of squares of deviations (and in more general analysis of variance computations) he becomes aware of the practical inequivalence of algebraically identical forms of computation.

An adequately prewhitened time series, at least one that is a realization from an ensemble which produces spectrum estimates which are even a quarter as variable as those provided by a Gaussian ensemble (most ensembles arising in practice will produce estimates more variable than those), requires the observations to be recorded to, at most, only the precision offered by 1.5 to 2 decimal digits [Tukey 1959b, pp. 319-320]. But one that is far from adequately prewhitened may require several decimal digits. This happens because the spread between the maximum and minimum observations is determined by the (areas of) peaks in the spectrum, while the precision necessary to avoid serious loss of information about the spectrum is determined by the depths of its valleys.

A similar difficulty can arise in so simple a situation as fitting a quadratic polynomial, though there most statisticians would see the difficulty coming and evade it. Thus if

$$y_i = 12.71 + 1{,}000{,}000\ x_i + 0.03(x_i^2 - 1/3) + \varepsilon_i$$

where x_i ranges from -1 to +1, var $\varepsilon_i = 10^{-5}$, and we seek to find the quadratic term by ordinary quadratic regression, it will not suffice to use y-values with only 7-decimal digits of precision, because rounding to units introduces deviations of up to 0.50 (which is large compared to the maximum quadratic effect of +0.02) and increases the effective error variance by a factor of more than 8000.

Similarly, in the time series case, if one is not prepared to prewhiten first, when desirable, it is necessary to make provision for moderate to high precision in input data, and correspondingly higher precision in accumulating sums of lagged products. The most likely result is a program which computes sums of lagged products in double-precision arithmetic, perhaps even floating-point double-precision arithmetic. This means extra effort at many stages of the computation.

No one of these four considerations rule out calculating sums of lagged products once and for all, but each exerts pressure. The combined effect influences me very much, but I must admit that they might not be as potent if the calculations with which I was concerned were to be made on quite other computing equipment.

VIII

OTHER INTRODUCTORY REFERENCES

Where is the statistician to seek further enlightenment about spectral analysis? It is hard to give extensive lists of highly informative sources, bu some guidance may be helpful.

One useful route for many statisticians will be to turn to instances where the technique has been applied. A list of references to recent applications can be found in either Tukey 1959a (pp. 408-411) or Tukey 1959b (pp. 327-330). These lists unfortunately omitted the 1957 Symposium at the Royal Statistical Society on the Analysis of Geophysical Time Series [Craddock 1957, Charnock 1957, Rushton and Neumann 1957, and discussion], where further references to geophysical applications can be found.

REFERENCES

JULIUS BARTELS, 1935, Random fluctuations, persistence, and quasipersistence in geophysical and cosmical periodicities, 40 Terr. Magnetism 1-60.

R. B. BLACKMAN and J. W. TUKEY, 1959, The measurement of power spectra from the point of view of communications engineering, New York, Dover, x + 190 pp. (Reprinted from 37 Bell System Technical Journal (1958) with added preface and index.)

H. BURKHARDT, 1904, Trigonometrische interpolation, IIA9a Encyklopadie der Math. Wiss. 642-693.

SYDNEY CHAPMAN and JULIUS BARTELS, 1940, Geomagnetism Oxford, Univ. Press (2 Vols). Especially chap. 16, Periodicities and harmonic analysis in geophysics, pp. 545-605 (in vol. 2). (Second Edition 1951, photographic reprint with (?) additions.)

H. CHARNOCK, 1957, Notes on the specification of atmospheric turbulence, A120 J. Roy. Statist. Soc. 398-408 (discussion 425-439).

JEROME CORNFIELD and J. W. TUKEY, 1956, Average values of mean squares in factorials, 27 Annals Math. Statist. 907-949.

J. M. CRADDOCK, 1957, An analysis of the slower temperature variations at Kew Observatory by means of mutually exclusive band pass filters, A120 J. Roy. Statist. Soc. 387-397 (discussion 425-439).

R. A. FISHER, 1929, Tests of significance in harmonic analysis, A125 Proc. Roy. Soc. London 54-59. (Reprinted as paper 16 in his Contributions to Mathematical Statistics, New York, Wiley, 1950.)

N. R. GOODMAN, 1957, On the joint estimation of the spectra, co-spectrum and quadrature spectrum of a two-dimensional stationary Gaussian process, Scientific Paper No. 10, Engineering Statistics Laboratory, New York University 1957 (also Ph. D. Thesis, Princeton University).

ULF GRENANDER and MURRAY ROSENBLATT, 1957, Statistical Analysis of Stationary Time Series, New York, Wiley; Stockholm, Almqvist + Wiksell, 300 pp.

H. O. HARTLEY, 1946, The application of some commercial calculating machines to certain statistical calculations, 8 Suppl. J. Roy. Stat. Soc. 154-183. (Especially pp. 167-168).

J. LEITH HOLLOWAY, JR., 1958, Smoothing and filtering of time series and space fields, 4 Advances in Geophysics (Ed. H. E. Landsberg) pp. 351-389, New York, Academic Press.

G. H. JOWETT, 1955, Sampling properties of local statistics in stationary stochastic series, 42 Biometrika 160-169.

G. H. JOWETT, 1957, Statistical analysis using local properties of smooth heteromorphic stochastic series, 44 Biometrika 454-463.

G. H. JOWETT and WENDY M. WRIGHT, 1958, Jump analysis, 45 Biometrika 386-399.

W. H. MUNK and F. E. SNODGRASS, 1957, Measurements of southern swell at Guadalupe Island, 4 Deep-sea Research 272-286.

H. PRESS and J. W. TUKEY, 1956, Power spectral methods of analysis and application in airplane dynamics, AGARD Flight Test Manual (Ed. E. J. Durbin) Vol. IV, Part IVC, pp. IVC:1-IVC:41, North Atlantic Treaty Organization. (Also Bell System Monograph 2606.) (2nd edition, London, Pergamon Press 1959.)

S. RUSHTON and J. NEUMANN, 1957, Some applications of time series analysis to atmospheric turbulence and oceanography, A120 J. Roy. Statist. Soc. 409-425 (discussion 425-439).

KARL STUMPFF, 1937, Grundlagen und Methoden der Periodenforschung, Berlin, Springer, vii + 332 pp. (Reprinted, Ann Arbor, Edwards, 1945.)

ROBERT J. TRUMPLER and HAROLD F. WEAVER, 1953, Statistical Astronomy, Berkeley, Univ. Press.

JOHN W. TUKEY, 1959a, The estimation of (power) spectra and related quantities, On Numerical Approximation (Ed. Radolph E. Langer), pp. 389-411, Madison, Wisc., Univ. Press.

JOHN W. TUKEY, 1959b, An introduction to the measurement of spectra, Probability and Statistics: The Harold Cramér Volume. (Ed. Ulf Grenander), pp. 300-330, Stockholm, Almqvist + Wiksell; New York, Wiley.

JOHN W. TUKEY, 1961(?), Curves as parameters and touch estimation, to appear in the Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. (Also circulated as Technical Report No. 39 of the Statistical Techniques Research Group, Princeton University.)

THOMAS WONNACOTT, 1961(?), Spectral analysis combining a Bartlett window with an associated inner window, submitted to Technometrics.