Significance of statistical relations derived from geophysical data

By JACK NORDØ, *Massachusetts Institute of Technology*[1,2]

(Manuscript received June 1, revised version September 13, 1965)

ABSTRACT

Most geophysical phenomena have typical interrelations in time and space. But these restrictions are frequently forgotten by investigators processing data in order to verify, or to detect, the laws of nature. For example, the claimed relations between geophysical events and cosmical data are numerous, but rather few such relationships survive the next decade. One main reason for this is certainly the use of random sampling techniques when investigating geophysical data, in spite of the fact that statistics of related terms were studied by Markov half a century ago.

If long series of records are available, the sampling complications due to serial correlation can be removed by selecting dates separated by a proper interval of time. But in most cases we have data only for relatively short intervals, and we are therefore forced to use all data to get optimum determination of the statistical parameters. The significance of these parameters then depend very much on the serial correlations involved.

The purpose of this paper is to derive tests of significance, which may be applicable to a variety of such investigations when serial correlations are present in the data.

List of symbols

a_i	regression coefficient in the autoregression equation of $\varepsilon(t)$, see eq. (26)
b_i	sample regression coefficient in eq. (1)
C_m	constant defined by eq. (29)
c_i	sample regression coefficient defined by eq. (9a)
c_{ir}	transformation constant defined by eq. (6)
c'_{ir}	transformation constant defined by eq. (8)
c_{p1}	regression coefficient of $x_p(t) = \sin(\alpha_p t + \beta_p)$
c_{p2}	regression coefficient of $x_p(t) = \cos(\alpha_p t + \beta_p)$
$e(t)$	sample residual at time t
f_0	a constant, usually equal to one
$f_i(t)$	independent variable number i
$G(p)$	spectral amplitude of $\sin(\alpha_p t + \beta_p)$, defined by eq. (48)
i	subscript
j	subscript
k	subscript
k_p^2	amplitude of the α_p-component in the autoregression correlation of a spectral band, see eq. (49)
m	subscript
n	number of sets of data per sample
p	subscript
q	subscript
R	coefficient of multiple correlation referring to eq. (1)
R_ε	coefficient of multiple correlation referring to eq. (26)
$R_{B,\tau}$	autocorrelation of a spectral band at τ lags, see eq. (49)
$R_{i,u-v}$	correlation coefficient, defined below eq. (22)
$R'_{i,q}$	sample "correlation" coefficient defined by eq. (35)
R_m	root of algebraic equation, see eq. (29)
$\mathcal{R}_{i,u-v}$	expectation value defined by eq. (22)
r_k	correlation of $\varepsilon(t)$ and $\varepsilon(t+k)$
r_{u-v}	correlation of $\varepsilon(u)$ and $\varepsilon(v)$
r'_k	correlation of $e(t)$ and $e(t+k)$
s	number of independent variables in the regression equation
t	time variable, except in relation (52)
u	time variable
v	time variable
$x_i(t)$	independent variable, first introduced by eq. (6)
$x_B(t)$	independent variable, defined as a finite sum of harmonics by eq. (48)

[1] On leave from the Norwegian Meteorological Institute.

[2] This research was supported by the Air Force Cambridge Research Laboratories, under contract No. AF 19 (628)-2409.

$y(t)$	dependent variable
z	variable
α	frequency, see eq. (44)
α_p	frequency first used in eq. (43a)
β_p	phase angle first used in eq. (43a)
β_i	universe regression coefficient in eq. (5)
γ_i	universe regression coefficient in eq. (10)
δ_t	residual at time t, see eq. (26)
δ_{ij}	discrete constant defined in connection with eq. (7)
∂	partial differentiation symbol
$\varepsilon(t)$	residual of the universe regression equation
$\varphi(z_*)$	function of a_i and z, see eq. (37)
χ_I^2	a variance quantity referring to residual variance, see eq. (39a)
χ_{II}^2	a variance quantity referring to all regression coefficients, see eq. (39b)
χ_i^2	a variance quantity referring to the i'th regression coefficient, see eq. (39b)
χ_p^2	as above, but referring specifically to the p'th regression coefficient
χ_I^{*2}	value of χ_I^2 if random sampling
χ_{II}^{*2}	value of χ_{II}^2 if random sampling
χ_i^{*2}	value of χ_i^2 if random sampling
ν	degrees of freedom, subscripts as above for χ^2
——	(a bar) denotes sample mean
·····	(dotted rule) denotes universe mean
····n	(dotted rule plus n) denotes universe mean for all samples of size n

1. Introduction

Although the limitations of random sampling are well described in almost any textbook on statistics, the restrictions are frequently neglected when statistical methods are applied to geophysical data, which normally possess pronounced interrelations in time and space. If we derive statistical estimates from such data, a result which would appear highly significant if the data were random, may very well be insignificant in view of the interrelations. The best way of avoiding such misuse of statistics is probably to postpone the random sampling theory until the laws of interrelated data are more thoroughly understood by the investigator. The random sampling theory is only a limiting case at best.

Some of the most significant contributions to the theory of related terms are given by MARKOV (see DYNKIN, 1961). Markov's theories are well recognized and have been extended and applied to several fields of science. G. WALKER (1931) has discussed the relative significance of periodicities in series of related terms. Our contributions to the field are given by the following references, NORDØ (1953, 1959, 1960). In this report we shall present a tentative approach to the problem of testing significance of statistical estimates based on non-random data samples.

Before we proceed to a more involved analysis, we may consider the following relation:

$$y(t) = \sum_{i=0}^{s} b_i f_i(t) + e(t). \tag{1}$$

Our dependent variable $y(t)$ and our independent variables $f_i(t)$, $i = 1, ..., s$, may be either prespecified analytic functions of time or stochastic variables, perhaps describing some physical process; for convenience, however, $f_0(t)$ will denote a prespecified constant. Let us suppose that the constants b_i have been derived by the least squares procedure, and that $e(t)$ is the residual, or that part of $y(t)$ which cannot be explained by the functions f_0 through $f_s(t)$. As a special application we may e.g. consider the case when $y(t)$ is the precipitation at a given station, and $f_i(t)$ the surface pressure at gridpoint i.

Denoting a sample average by a bar, the equations determining b_i may be written as follows:

$$\overline{f_i(t)\, e(t)} = 0, \quad i = 0, 1, 2, ..., s. \tag{2}$$

As f_0 is a constant, we notice that $\overline{e(t)} = 0$. Consequently the sample residual is not correlated with the independent variables, but this does not imply that $e(t)$ is independent of $f_i(t)$. A certain nonlinear combination of the $f_i(t)$ functions, e.g. a term occurring in the nonlinear dynamic equations, might very well be a quantity to which $y(t)$ is related.

The "goodness" of relation (1) is generally measured by the fraction of var y which is explained by $\sum_{i=0}^{s} b_i f_i(t)$. This fraction is by definition equal to the square of the coefficient of multiple correlation, R. Using this notation, the residual variance becomes

$$\operatorname{var} e = (1 - R^2) \operatorname{var} y. \tag{3}$$

In samples of moderate size it is quite possible to obtain high values of R^2 just by chance. In most cases we will be disappointed if we

apply the derived relations to another set of data. To prevent scientists from spending their time explaining such sporadic relationships, various theories of sampling evaluations have been developed. The random sample procedures are explained in most textbooks, see e.g. KENDALL (1946). But the random sample theories are based on the assumption that the residuals $e(t)$, $t = 1, 2, \ldots, n$ are uncorrelated and mutually *independent* quantities. These restrictions are rarely fulfilled when physical data are used. We shall later on see that the mere presence of a moderate yearly trend, or a diurnal variation, may cause considerable deviations from the random sample theories. It is therefore highly desirable to make an attempt to analyze the general case when all variables, including $e(t)$, have various scales in time, as e.g. diurnal and yearly variations.

2. Some theorems concerning expectation of sample residual variances, and reduction of degrees of freedom due to serial correlations

Let us return to relation (1), and assume that it is established from a particular sample consisting of n consecutive observations of each variable. Then $\overline{f_i(t)}$ denotes the sample mean of $f_i(t)$. Let $\dot{f}_i(t)$ be defined as the universe mean of $f_i(t)$. Such a mean may be assumed to exist, even though the available data do not determine it when $f_i(t)$ is stochastic. The constant f_0 is conveniently put equal to one. The sample constants b_i are derived from the equations minimizing the squares of the residuals,

$$\frac{\partial}{\partial b_i} \sum_{t=1}^{n} \{e(t)\}^2 = 0, \quad i = 0, 1, 2, \ldots, s, \tag{4}$$

which are equivalent to the relations (2).

For the universe we may correspondingly derive a similar regression equation,

$$y(t) = \sum_{i=0}^{s} \beta_i f_i(t) + \varepsilon(t), \tag{5}$$

the constants β_i being defined by $\dot{f}_i(t)\,\varepsilon(t) = 0$, whence $\dot{\varepsilon}(t) = 0$. So far we have put no restrictions on the functions $f_1(t), f_2(t), \ldots, f_s(t)$. But the procedure is much simplified if we transform the functions $f_i(t)$ into a set of orthogonal functions $x_i(t)$, where $(i \geqslant 1)$

$$x_i(t) = \sum_{r=1}^{i} c_{ir} \{f_r(t) - \dot{f}_r(t)\}. \tag{6}$$

We shall especially choose $x_0 = f_0 = 1$. Then we notice that the $\frac{1}{2}s(s+1)$ constants $c_{ir} (i \geqslant 1)$ are uniquely determined by the following sets of equations:

$$\dot{x}_i(t)\, x_j(t) = \delta_{ij}, \tag{7}$$

where $\delta_{ii} = 1$, and $\delta_{ij} = 0$ when $i \neq j$. Repeated use of eq. (6) will relate all lag correlations of $x_i(t)$ to similar lag and cross-lag correlations of the functions $f_r(t)$. Inverting eq. (6) we may express $f_i(t) - \dot{f}_i(t)$ as a function of the orthogonal functions $x_i(t)$,

$$f_i(t) - \dot{f}_i(t) = \sum_{r=1}^{i} c'_{ir}\, x_r(t), \quad i \geqslant 1. \tag{8}$$

Introducing relation (8) for the independent variables of the sample regression eq. (1), we derive that

$$y(t) = \sum_{i=0}^{s} c_i x_i(t) + e(t), \tag{9a}$$

where

$$c_i = \sum_{r=1}^{s} b_r c'_{ri}, \quad i \geqslant 1, \tag{9b}$$

and

$$c_0 = \overline{y(t)} - \sum_{i=1}^{s} b_i \{\overline{f_i(t)} - \dot{f}_i(t)\}.$$

For the universe regression equation we derive correspondingly that

$$y(t) = \sum_{i=0}^{s} \gamma_i x_i(t) + \varepsilon(t), \tag{10}$$

where

$$\gamma_i = \sum_{r=1}^{s} \beta_r c'_{ri}, \quad i \geqslant 1, \tag{11}$$

and

$$\gamma_0 = \beta_0.$$

From eqs. (9a) and (10) we may deduce the following relation:

$$\varepsilon(t) = e(t) + \sum_{i=0}^{s} (c_i - \gamma_i)\, x_i(t), \tag{12}$$

or

$$e(t) = \varepsilon(t) - \sum_{i=0}^{s} (c_i - \gamma_i)\, x_i(t). \tag{13}$$

As the normal equations, determining b_i and β_i, may be written as $\overline{f_i(t)\, e(t)} = 0$ and $\overline{f_i(t)\, \varepsilon(t)} = 0$, we deduce from eq. (6) that

$$\overline{x_i(t)\, e(t)} = 0, \tag{14}$$

and

$$\overline{x_i(t)\, \varepsilon(t)} = 0. \tag{15}$$

Squaring relations (12) and (15), then averaging and using eqs. (7), (14) and (15), we derive the following relations when neglecting the term $\sum_{i \neq j} (c_i - \gamma_i)\,(c_j - \gamma_j)\, \overline{x_i(t)\, x_j(t)}$:

$$\overline{\{\varepsilon(t)\}^2} \approx \overline{\{e(t)\}^2} + \sum_{i=0}^{s} (c_i - \gamma_i)^2\, \overline{\{x_i(t)\}^2}, \tag{16}$$

and

$$\overline{\{e(t)\}^2} = \overline{\{\varepsilon(t)\}^2} + \sum_{i=0}^{s} (c_i - \gamma_i)^2. \tag{17}$$

Multiplying (12) by $x_i(t)$ and averaging, we notice that with a similar approximation

$$c_i - \gamma_i \approx \overline{x_i(t)\, \varepsilon(t)} \left[\overline{\{x_i(t)\}^2}\right]^{-1}. \tag{18}$$

This relation permits us to eliminate the empirical constants from eqs. (16) and (17), i.e.

$$\overline{\{\varepsilon(t)\}^2} \approx \overline{\{e(t)\}^2} + \sum_{i=0}^{s} \overline{\{x_i(t)\, \varepsilon(t)\}^2} \left[\overline{\{x_i(t)\}^2}\right]^{-1} \tag{19}$$

and

$$\overline{\{e(t)\}^2} \approx \overline{\{\varepsilon(t)\}^2} + \sum_{i=0}^{s} \overline{\{x_i(t)\, \varepsilon(t)\}^2} \left[\overline{\{x_i(t)\}^2}\right]^{-2}. \tag{20}$$

Let us expand the second term on the right hand side of eq. (19),

$$\sum_{i=0}^{s} \frac{\overline{\{x_i(t)\, \varepsilon(t)\}^2}}{\overline{\{x_i(t)\}^2}} = n^{-2} \sum_{i=0}^{s} \sum_{u,v=1}^{n} \frac{\overline{x_i(u)\, x_i(v)}}{\overline{\{x_i(t)\}^2}}\, \varepsilon(u)\, \varepsilon(v). \tag{21}$$

The magnitude of this multiple sum depends on the product of the quantities

$$\frac{\overline{x_i(u)\, x_i(v)}}{\overline{\{x_i(t)\}^2}}$$

and $\varepsilon(u)\, \varepsilon(v)$. We shall replace the product by its expectation value, i.e. by

$$\frac{\overline{x_i(u)\, x_i(v)}}{\overline{\{x_i(t)\}^2}}\, \varepsilon(u)\, \varepsilon(v) = \left(\frac{\overline{x_i(u)\, x_i(v)}}{\overline{\{x_i(t)\}^2}}\right) \overline{\varepsilon(u)\, \varepsilon(v)} + \mathcal{R}_{i,\,u-v}, \tag{22}$$

where

$$\mathcal{R}_{i,\,u-v} \equiv \frac{\overline{x_i(u)\, x_i(v)}}{\overline{\{x_i(t)\}^2}}\, \varepsilon(u)\, \varepsilon(v) - \left(\frac{\overline{x_i(u)\, x_i(v)}}{\overline{\{x_i(t)\}^2}}\right) \overline{\varepsilon(u)\, \varepsilon(v)}.$$

We notice that $\mathcal{R}_{i,\,u-v} = 0$ if $x_i(u)\, x_i(v) / \overline{\{x_i(t)\}^2}$ and $\varepsilon(u)\, \varepsilon(v)$ are independent quantities. We may proceed by replacing

$$\left(\frac{\overline{x_i(u)\, x_i(v)}}{\overline{\{x_i(t)\}^2}}\right) \text{ by } \frac{\overline{x_i(u)\, x_i(v)}}{\left[\overline{\{x_i(u)\}^2}\, \overline{\{x_i(v)\}^2}\right]^{\frac{1}{2}}} = R_{i,\,u-v},$$

where $R_{i,\,u-v}$ *is the universe correlation coefficient of values of* $x_i(t)$ *separated by* $(u-v)$ *units of time.* In a similar way we may replace $\overline{\varepsilon(u)\, \varepsilon(v)}$ by $r_{u-v}\, \overline{\{\varepsilon(t)\}^2}$, *where* r_{u-v} *is the universe correlation coefficient of residuals at* $(u-v)$ *lags.*

Introducing these definitions in (21), we derive

$$\sum_{i=0}^{s} \overline{\{x_i(t)\, \varepsilon(t)\}^2} \left[\overline{\{x_i(t)\}^2}\right]^{-1} = \frac{\overline{\{\varepsilon(t)\}^2}}{n^2} \sum_{i=0}^{s} \sum_{u,v=1}^{n} \left[r_{u-v}\, R_{i,\,u-v} + \mathcal{R}_{i,\,u-v} \overline{\{(\varepsilon(t))^2\}^{-1}}\right]. \tag{23}$$

In the general case there is no reason *a priori* to neglect the second term within the square brackets. But in most applications it seems fair to assume that there should not be much preference of $\varepsilon(u)\, \varepsilon(v)$ for a given value $x_i(u) x_i(v)$. We shall therefore proceed with our analysis, assuming that the integrated contributions of $\mathcal{R}_{i,\,u-v}$ may be neglected in comparison to the remaining terms. If we in eq. (20) replace $\left[\overline{\{x_i(t)\}^2}\right]^{-2}$ by $\left[\overline{\{x_i(t)\}^2}\right]^{-1}$ and perform the same procedure as above, we derive the following approximate relations:

$$\text{var } \varepsilon = \overline{\{\varepsilon(t)\}^2} \approx \overline{\{e(t)\}^2}\, \frac{n}{n - n^{-1} \sum_{i=0}^{s} \sum_{u,v=1}^{n} r_{u-v}\, R_{i,\,u-v}} \tag{24}$$

and

$$\overline{\{e(t)\}^2}^n \approx \overline{\{e(t)\}^2}^n \frac{n + n^{-1}\sum_{i=0}^{s}\sum_{u,v=1}^{n} r_{u-v} R_{i,u-v}}{n - n^{-1}\sum_{i=0}^{s}\sum_{u,v=1}^{n} r_{u-v} R_{i,u-v}}. \tag{25a}$$

Eq. (24) gives a relation between the expected sample variance $\overline{\{e(t)\}^2}^n$ and the variance of the universe.

Eq. (25a) defines a relation between the expected sample variance and the variance to be expected, if we apply the constants of eq. (9) (or eq. (1)) when deriving $y(t)$ from an independent sample of data.

The formula (24) indicates that the estimated degrees of freedom of the residual variance is approximately equal to

$$\nu_{\mathrm{I}} = n - n^{-1}\sum_{i=0}^{s}\sum_{u,v=1}^{n} r_{u-v} R_{i,u-v} \equiv n - \nu_{\mathrm{II}}.$$

We shall later on apply this estimate in connection with the evaluation of the sampling distributions defined by the relations (40) through (42).

If $s = 0$, the residual is just the deviation of $y(t)$ from its mean value:

$$e(t) = y(t) - \overline{y(t)} = y(t) - \overline{\overline{y(t)}} - \{\overline{y(t)} - \overline{\overline{y(t)}}\}$$
$$= \varepsilon(t) - \{\overline{y(t)} - \overline{\overline{y(t)}}\},$$

or
$$\operatorname{var}\varepsilon \approx \overline{\{e(t)\}^2}^n \frac{n}{n - n^{-1}\sum_{u,v=1}^{n} r_{u-v}}.$$

This relation corresponds to (24) if we put $R_{0,u-v}$ equal to one.

We shall illustrate the theorems (24) and (25a) by some simple cases. First of all, let us consider the random sampling case, i.e. $r_{u-v} = 0$ when $u \neq v$. Then $\nu_{\mathrm{I}} = n - s - 1$, and relation (24) reduces to Gauss' theorem on residuals (see KENDALL, 1946). The following random sampling versions of relations (24) and (25a) are also derived by LORENZ (1956),

$$\operatorname{var}\varepsilon \approx \overline{\{e(t)\}^2}^n \frac{n}{n - s - 1}$$

and
$$\overline{\{e(t)\}^2}^n \approx \overline{\{e(t)\}^2}^n \frac{n + s + 1}{n - s - 1}.$$

When $s = 0$, $\nu_{\mathrm{I}} = n - 1$, which is the degrees of freedom usually given to the sample variance.

Next we will consider the simple Markov case when $R_{i,u-v} = R_i^{u-v}$ and $r_{u-v} = r^{u-v}$. Introducing these assumptions, we observe that the summations in eqs. (24) and (25a) become truncated geometric series, so that these equations then reduce to

$$\operatorname{var}\varepsilon \sim \overline{\{e(t)\}^2}^n \times \frac{n}{n - \sum_{i=0}^{s}\frac{1+R_i r}{1-R_i r} + \frac{2}{n}\sum_{i=0}^{s} R_i r \frac{1-R_i^n r^n}{(1-R_i r)^2}} \tag{25b}$$

and

$$\overline{\{e(t)\}^2}^n \sim \overline{\{e(t)\}^2}^n \times \frac{n + \sum_{i=0}^{s}\frac{1+R_i r}{1-R_i r} - \frac{2}{n}\sum_{i=0}^{s} R_i r \frac{1-R_i^n r^n}{(1-R_i r)^2}}{n - \sum_{i=0}^{s}\frac{1+R_i r}{1-R_i r} + \frac{2}{n}\sum_{i=0}^{s} R_i r \frac{1-R_i^n r^n}{(1-R_i r)^2}}. \tag{25c}$$

If $R_i r$ is not close to one, and n is moderately large, we may neglect the final summations in the denominators and numerator, i.e., we may treat the geometric series as infinite in length. If we put $r = 0$, we derive the random sample relations discussed above.

The first order Markov process is a fairly good approximation to the statistical behavior of many atmospheric phenomena, especially when we have properly eliminated very short (e.g. diurnal) periodic time scales, and also very long periods, as e.g. the annual variation. But if we intend to detect a very weak, say external, influence in our data, the higher order Markov schemes must be taken into account, see e.g. BAUR (1944) and NORDØ (1953). The serial correlation of surface pressure has been studied by STUMPFF (1936) and HISDAL *et al.* (1956). Serial correlation of temperature involves certainly higher order Markov schemes, cf. NORDØ (1959) and GODSKE (1962). Other elements like precipitation, cloudiness and wind do all obey various kinds of Markov schemes. Let us therefore conclude this part of our study by looking briefly into the case when

$$R_{i,u-v} = \sum_{p=1}^{J} C_{ip} R_{ip}^{|u-v|}$$

and
$$r_{u-v} = \sum_{q=1}^{j} C_q R_q^{|u-v|}.$$

We may then verify that in this general case we have

$$v_{II} = n^{-1} \sum_{i=0}^{s} \sum_{p=1}^{J} C_{ip} \sum_{q=1}^{j} C_q \sum_{u,v=1}^{n} (R_{ip} R_q)^{|u-v|}$$

$$= \sum_{i=0}^{s} \sum_{p=1}^{J} C_{ip} \sum_{q=1}^{j} C_q \times \left[\frac{1 + R_{ip} R_q}{1 - R_{ip} R_q} - \frac{2 R_{ip} R_q}{n} \frac{1 - (R_i R_q)^n}{(1 - R_{ip} R_q)^2} \right].$$

We shall later on apply this formula to some simple cases, as e.g. the case when $x_i(t)$ is a harmonic variable.

3. Sampling distributions of statistical parameters evaluated from non-random data samples

The random sampling theories depend on the assumption that the residuals, $\varepsilon(t)$, of eq. (10) are normally distributed, are uncorrelated, and even mutually independent. We shall now try to carry out an analysis of the more general case when successive values of $\varepsilon(t)$ satisfy the following auto-regression equation:

$$\varepsilon(t+j) = a_1 \varepsilon(t+j-1) + \ldots + a_j \varepsilon(t) + \delta_{t+j}, \quad (26)$$

where δ_{t+j} is the residual. Multiplying (26) by $(\mathrm{var}\,\varepsilon)^{-1} \varepsilon(t-k)$ and averaging over the universe, we derive

$$r_{j+k} = a_1 r_{j+k-1} + \ldots + a_j r_k, \quad (27)$$

assuming that

$$\overline{\varepsilon(t-k)\,\delta_{t+j}} = 0, \quad (28)$$

neglecting the regression coefficients $a_{j+1}, a_{j+2}, \ldots, a_{j+k}$ of $\varepsilon(t-1), \varepsilon(t-2), \ldots, \varepsilon(t-k)$ which could have appeared in eq. (26).

The formal solution of (27) is known, see e.g. MILNE (1949), namely

$$r_k = \sum_{m=1}^{j} C_m R_m^k, \quad (29)$$

where the values of R_m are equal to the roots of the algebraic equation

$$z^j = a_1 z^{j-1} + \ldots + a_j. \quad (30)$$

The constants C_m can be established from the values of r_k at 0, 1, 2, ..., $j-1$ lags. As a consequence of relation (28) we have

$$\overline{\delta_{t+j}\,\delta_{t+j-k}} = 0 \quad (31)$$

for all values of $k \neq 0$.

We shall now assume that the residuals δ_{t+j}, $t = 1, 2, \ldots, n$ are independent quantities. We shall also assume that δ_{t+j} is normally distributed. This second assumption is not really necessary, some other prescribed distributions might also be discussed. But the normal assumption is convenient as we may compare our results with those of the classical random sample theory.

If our assumptions of both independence and normal distribution are fulfilled, the combined frequency distribution of n consecutive residuals is proportional to

$$\left\{ \exp\left[-\sum_{t=1}^{n} (\delta_{t+j})^2 \left\{ 2\,\mathrm{var}\,\varepsilon\,(1 - R_\varepsilon^2) \right\}^{-1} \right] \right\} \prod_{t=1}^{n} d(\delta_{t+j}), \quad (32)$$

where $\mathrm{var}\,\varepsilon(1 - R_\varepsilon^2)$ is the universe variance of δ_{t+j} according to (26). Expanding the numerator of the exponent and assuming that $j \ll n$, we have

$$\sum_{t=1}^{n} (\delta_{t+j})^2 \approx (1 + a_1^2 + \ldots + a_j^2) \sum_{t=1}^{n} \varepsilon(t)\,\varepsilon(t) - 2(a_1 - a_1 a_2 - \ldots - a_{j-1} a_j) \sum_{t=1}^{n} \varepsilon(t)\,\varepsilon(t+1) - \ldots - 2 a_j \sum_{t=1}^{n} \varepsilon(t)\,\varepsilon(t+j). \quad (33)$$

Expanding $\sum_{t=1}^{n} \varepsilon(t)\,\varepsilon(t+q)$, $1 \leq q \leq j$, we obtain from eq. (12) that

$$\sum_{t=1}^{n} \varepsilon(t)\,\varepsilon(t+q) = \sum_{t=1}^{n} e(t)\,e(t+q) + \sum_{i=0}^{s} (c_i - \gamma_i)^2 \sum_{t=1}^{n} x_i(t)\,x_i(t+q) + \sum_{i=0}^{s} (c_i - \gamma_i) \sum_{t=1}^{n} \{ e(t)\,x_i(t+q) + e(t+q)\,x_i(t) \} + \sum_{i \neq j} (c_i - \gamma_i)(c_j - \gamma_j) \times \sum_{t=1}^{n} \{ x_i(t)\,x_j(t+q) + x_i(t+q)\,x_j(t) \}$$

$$\sim \sum_{t=1}^{n} e(t)\,e(t+q) + \sum_{i=0}^{s} (c_i - \gamma_i)^2 \sum_{t=1}^{n} x_i(t)\,x_i(t+q) \quad (34)$$

when neglecting the third and the fourth term, which both are assumed to be small compared to the two remaining quadratic terms. We consider this assumption just as a first order approximation, and the resulting sampling distribution should also be interpreted as an approximation of the real distribution.

We shall now introduce the sample "correlation" coefficient $R'_{i,q}$, which is defined as follows

$$R'_{i,q} = \{\sum_{t=1}^{n} x_i(t)\, x_i(t+q)\} \{\sum_{t=1}^{n} x_i(t)\, x_i(t)\}^{-1}. \quad (35)$$

Furthermore, r'_k shall denote the auto-correlation function at k lags for $e(t)$ within the given sample.

Using these definitions and the approximations introduced in derivation of eq. (34), we derive the following relation:

$$\sum_{t=1}^{n} (\delta_{t+j})^2 \approx \varphi(r'_*) \sum_{t=1}^{n} e(t)\, e(t) + \sum_{i=0}^{s} (c_i - \gamma_i)^2\, \varphi(R'_{i*}) \sum_{t=1}^{n} x_i(t)\, x_i(t) \quad (36)$$

where

$$\varphi(R'_{i*}) = (1 + a_1^2 + \ldots + a_j^2) - 2\,(a_1 - a_1 a_2 - \ldots - a_{j-1} a_j)\, R'_{i,1} - 2\,(a_2 - a_1 a_3 - \ldots)\, R'_{i,2} - \ldots - 2\, a_j R'_{i,j}. \quad (37)$$

It is easy to verify that

$$\varphi(r_*) = 1 - R_\varepsilon^2. \quad (38)$$

Introducing spherical coordinates in the n-dimensional space, the volume element $\prod_{t=1}^{n} d(\delta_{t+j})$ will be proportional to

$$\{\sum_{t=1}^{n} (\delta_{t+j})^2\}^{\frac{1}{2}(n-2)}\, d\{\sum_{t=1}^{n} (\delta_{t+j})^2\};$$

compare e.g. the derivations of the χ^2-distribution in KENDALL (1946). Let us define χ_{I}^2 and χ_{II}^2 by the following equations:

$$\chi_{\mathrm{I}}^2 = \frac{\varphi(r'_*)}{\varphi(r_*)} \frac{n\,\overline{e^2}}{\mathrm{var}\ \varepsilon} \quad (39\,\mathrm{a})$$

and

$$\chi_{\mathrm{II}}^2 = \sum_{i=0}^{s} \frac{\varphi(R'_{i*})}{\varphi(r_*)} (c_i - \gamma_i)^2 \frac{n\,\overline{x_i^2}}{\mathrm{var}\ \varepsilon} \equiv \sum_{i=0}^{s} \chi_i^2. \quad (39\,\mathrm{b})$$

The combined frequency distribution of the δ_{t+j}, $t = 1, 2, \ldots, n$ is then proportional to

$$\{\exp -\tfrac{1}{2}(\chi_{\mathrm{I}}^2 + \chi_{\mathrm{II}}^2)\}\ (\chi_{\mathrm{I}}^2 + \chi_{\mathrm{II}}^2)^{\frac{1}{2}(n-2)}\, d(\chi_{\mathrm{I}}^2 + \chi_{\mathrm{II}}^2), \quad (40)$$

i.e. $(\chi_{\mathrm{I}}^2 + \chi_{\mathrm{II}}^2)$ has a χ^2-distribution with n degrees of freedom. Referring to the derivation of eq. (24), the expected degrees of freedom of $\overline{ne^2}$ $(\mathrm{var}\ \varepsilon)^{-1}$ are

$$\nu_{\mathrm{I}} = n - n^{-1} \sum_{i=0}^{s} \sum_{u,v=1}^{n} r_{u-v} R_{i,u-v} \equiv n - \sum_{i=0}^{s} \nu_i = n - \nu_{\mathrm{II}}, \quad (41)$$

where ν_{II} is the expected degrees of freedom taken over by the statistical parameters derived for the sample regression equation.

Relation (40) is fulfilled if:

(1) χ_I^2 has a χ^2-distribution with ν_{I} degrees of freedom and

(2) χ_{II}^2 has another χ^2-distribution (approximately independent of the χ_I^2-distribution) with ν_{II} degrees of freedom, or

(2a) the $\chi_0^2, \chi_1^2, \ldots, \chi_s^2$ do all have χ^2-distributions (approximately independent of the χ_I^2-distribution) which are mutually independent.

Introducing χ_i^{*2} and ν_i^* as the values of χ_i^2 and ν_i when no serial correlations are present, eqs. (39a) and (41) show that

$$\chi_{\mathrm{I}}^2 = \frac{\varphi(r'_*)}{\varphi(r_*)} \chi_{\mathrm{I}}^{*2},$$

$$\nu_{\mathrm{I}} = \frac{n - n^{-1} \sum_{i=0}^{s} \sum_{u,v=1}^{n} r_{u-v} R_{i,u-v}}{n - s - 1} \nu_{\mathrm{I}}^*, \quad (42\,\mathrm{a})$$

$$\chi_{\mathrm{II}}^2 = \frac{\sum_{i=0}^{s} \frac{\varphi(R'_{i*})}{\varphi(r_*)} (c_i - \gamma_i)^2 \overline{x_i^2}}{\sum_{i=0}^{s} (c_i - \gamma_i)^2 \overline{x_i^2}} \chi_{\mathrm{II}}^{*2},$$

$$\nu_{\mathrm{II}} = \frac{n^{-1} \sum_{i=0}^{s} \sum_{u,v=1}^{n} r_{u-v} R_{i,u-v}}{s + 1} \nu_{\mathrm{II}}^*. \quad (42\,\mathrm{b})$$

Positive serial correlation will always lead to the result that $\nu_{\mathrm{I}} < \nu_{\mathrm{I}}^*$ and $\nu_{\mathrm{II}} > \nu_{\mathrm{II}}^*$. As the expectancy of the χ^2-distribution depends on ν, use of ν^* may cause serious errors, especially when there are rather few degrees of freedom.

The factors $\varphi(r'_*)/\varphi(r_*)$ and $\varphi(R'_{i*})/\varphi(r_*)$ re-

TABLE 1

$r_1', R_{i,1}'$	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
$\frac{\varphi(r_*')}{\varphi(r_*)}, \frac{\varphi(R_{i*}')}{\varphi(r_*)}$	1.86	1.69	1.51	1.34	1.17	1.00	0.83	0.66	0.49	0.31

flect the influence due to sampling fluctuations of the auto-correlation functions of $e(t)$ and $x_i(t)$ respectively. To clarify the importance of these two factors, we may consider the simple Markov case where $R_{i,1} = r_1 = 0.75$, i.e.

$$\frac{\varphi(r_*')}{\varphi(r_*)} = \frac{1 + (0.75)^2 - 1.50\, r_1'}{1 - (0.75)^2} = 1 - 3.43\,(r_1' - 0.75),$$

$$\frac{\varphi(R_{i*}')}{\varphi(r_*)} = 1 - 3.43\,(R_{i,1}' - 0.75).$$

Table 1 gives the size of the two factors for some sample values of the coefficients of correlation, when the population correlations $R_{i,1}$ and r_1 are both 0.75. To illustrate relation (42a) we may consider residual variance in two samples where $\nu > 30$, i.e. $z = \sqrt{2\chi_I^2} - \sqrt{2\nu_I} - 1$ is approximately a normal deviate with unit variance. We shall further assume that the two samples have equal size, $n = 50$, and that they consist of data from the same universe. The relations of successive observations correspond to those of the simple Markov case tabulated above. Introducing $s = 0, R_{0,\,u-v} = 1$ (see discussion between eq. (25a) and eq. (25b), $r_{u-v} = 0.75^{|u-v|}$, we derive from eq. (42a) that

$$\nu_I = \frac{50 - 50^{-1} \sum_{u,v=1}^{50} 0.75^{|u-v|}}{49}\, 49 = 50 - 6.52 = 43.48.$$

Suppose now that the first sample has fairly large time scale patterns with $\chi_I^{*2} = 98.00$ and $r_1' = 0.90$. Using Table 1 we notice that $\chi_I^2 = 0.49\ \chi_I^{*2}$. As $z^* = 14.00 - 9.85 = 4.15$ ($P \sim 2 \times 10^{-5}$) and $z = 9.80 - 9.27 = 0.53$ ($P \sim 0.30$), we have demonstrated that the difference $(z^* - z)$ may become relatively large for intercorrelated data. P denotes here the probability of having such a large deviation, or even a larger deviation with the same sign from the zero mean. Let us next consider a sample of data with smaller scale patterns where $\chi_I^{*2} = 24.50$ and $r_1' = 0.55$, i.e. $z^* = 7.00 - 9.85 = -2.85$ ($P \sim 2.2 \times 10^{-3}$) and $z = 9.10 - 9.27 = -0.17$ ($P \sim 0.43$). Consequently we have just considered another case where the difference $(z^* - z)$ becomes too large to be neglected.

These two illustrations are typical cases commonly observed within geophysical data samples. If the investigator however applies the χ_I^{*2}-distribution test to some serially correlated data where $r_k \geqslant 0$ and $R_{i,\,k} \geqslant 0$, he is likely to derive the following results for a great number of applications:

(A) The number of degrees of freedom is too high (cf. relation 24). This may lead to the discovery of more "significant" deviations for low values of χ_I^{*2} than for high values of χ_I^{*2}, as the expectancy of χ_I^{*2} increases with ν^*.

(B) If the number of degrees of freedom is wrong by a fairly small fraction, one is still likely to find an amazingly high frequency of "significant" departures on both tails of the χ_I^{*2}-distribution, as was the case in the two examples described above. The factors $\varphi(r_*')/\varphi(r_*)$ and $\varphi(R_{i*}')/\varphi(r_*)$ may be interpreted as a kind of "scale" factors which adjust the observed χ_I^{*}-distribution of related data back to a real χ_I^2-distribution. These two factors cannot be neglected even in fairly large samples where the degrees of freedom correction is rather small.

4. Application of the χ^2 test procedure to the case of harmonic variables

Harmonic components are orthogonal functions if the sample size is a whole number of periods. The mean is also zero in this case, and it may be of interest to apply our theory to such functions. Suppose that for $i = p$ $(1 \leqslant p \leqslant s)$, the variable $x_p(t)$ is a harmonic component,

$$x_p(t) = \sin(\alpha_p t + \beta_p), \tag{43a}$$

then $$R_{p,\,u-v} = \cos \alpha_p (u - v).$$

The function $\varphi(R_{p*})$ has a very simple form in

the case of $x_p(t)$ being a harmonic component and r_k satisfying relation (29). Making extensive use of the relations between the roots R_m of eq. (30) and the constants $a_1, a_2, \ldots, a_j$, we can show that

$$\begin{aligned}
&\prod_{m=1}^{j} (1 - 2R_m \cos \alpha_p + R_m^2) \\
&= \prod_{m=1}^{j} (\exp \alpha_p \sqrt{-1} - R_m)(\exp -\alpha_p \sqrt{-1} - R_m) \\
&= \{\exp j\alpha_p \sqrt{-1} - [\exp (j-1)\alpha_p \sqrt{-1}] \sum_m R_m \\
&\quad + [\exp (j-2)\alpha_p \sqrt{-1}] \sum_{m \neq m'} R_m R_{m'} - \ldots\} \\
&\quad \times \{\exp -j\alpha_p \sqrt{-1} - [\exp -(j-1)\alpha_p \sqrt{-1}] \\
&\quad \times \sum_m R_m + [\exp -(j-2)\alpha_p \sqrt{-1}] \\
&\quad \times \sum_{m \neq m'} R_m R_{m'} - \ldots\} \\
&= 1 + a_1^2 + \ldots + a_j^2 - 2(a_1 - a_1 a_2 - \ldots) \\
&\quad \times \cos \alpha_p - \ldots - 2a_j \cos j\alpha_p \\
&= \varphi(R_{p*}),
\end{aligned}$$

as $$\sum_m R_m = a_1, \quad \sum_{m \neq m'} R_m R_{m'} = -a_2, \text{ etc.}$$

Consequently, for the harmonic case, we have shown that

$$\varphi(R_{p*}) = \prod_{m=1}^{j} (1 - 2R_m \cos \alpha_p + R_m^2). \quad (43\,b)$$

Introducing the relation (29) for r_{u-v}, we derive for the p'th component of v_{II} in eq. (41):

$$\begin{aligned}
v_p &= n^{-1} \sum_{u,v=1}^{n} R_{p,u-v}\, r_{u-v} \\
&= \sum_{m=1}^{j} C_m \frac{1 - R_m^2}{1 - 2R_m \cos \alpha_p + R_m^2} \\
&\quad - 2n^{-1} \sum_{m=1}^{j} C_m R_m \\
&\quad \times \frac{\cos \alpha_p - 2R_m + R_m^2 \cos \alpha_p - R_m^n \{\cos (n+1) \\
&\quad \times \alpha_p - 2R_m \cos n\alpha_p + R_m^2 \cos (n-1)\alpha_p\}}{(1 - 2R_m \cos \alpha_p + R_m^2)^2}.
\end{aligned} \quad (43\,c)$$

Consequently, the degrees of freedom taken over by the p'th component is in its asymptotic form the following:

$$v_p \approx \sum_{m=1}^{j} C_m \frac{1 - R_m^2}{1 - 2R_m \cos \alpha_p + R_m^2}. \quad (43\,d)$$

The constants C_m of $r_k = \sum_{m=1}^{j} C_m R_m^k$ can be established from the lag correlations r_0 through r_{j-1}. Splitting the denominator above into the product $(1 - R_m \exp \alpha_p \sqrt{-1})(1 - R_m \exp -\alpha_p \sqrt{-1})$, we may write the right hand side of (43 d) as follows:

$$\begin{aligned}
&\sum_{m=1}^{j} \frac{C_m (1 - R_m^2)}{1 - 2R_m \cos \alpha_p + R_m^2} \\
&\quad = -1 + \sum_{m=1}^{j} \frac{C_m}{1 - R_m \exp \alpha_p \sqrt{-1}} \\
&\quad\quad + \sum_{m=1}^{j} \frac{C_m}{1 - R_m \exp -\alpha_p \sqrt{-1}}.
\end{aligned}$$

We can now show that

$$\begin{aligned}
&\prod_{m=1}^{j} (1 - R_m \exp \alpha_p \sqrt{-1}) \sum_{m=1}^{j} \frac{C_m}{(1 - R_m \exp \alpha_p \sqrt{-1})} \\
&\quad = \prod_{m=1}^{j} (1 - R_m \exp \alpha_p \sqrt{-1}) \\
&\quad\quad + \sum_{m=1}^{j} (r_m - \sum_{k=1}^{m-1} a_k r_{m-k}) \exp {}_m\alpha_p \sqrt{-1}).
\end{aligned}$$

Then, by repeated use of the recurrence formula (27), we may demonstrate by elementary, but laborious computations that

$$\begin{aligned}
&\varphi(R_{p*}) \sum_{m=1}^{j} \frac{C_m (1 - R_m^2)}{1 - 2R_m \cos \alpha_p + R_m^2} \\
&\quad = \varphi(R_p) - 2(a_1 - a_1 a_2 - \ldots) r_1 - 2 \\
&\quad\quad \times (a_2 - a_1 a_3 - \ldots) r_2 - \ldots - 2a_j r_j \\
&\quad\quad + 2(a_1 - a_1 a_2 - \ldots) \cos \alpha_p + 2 \\
&\quad\quad \times (a_2 - a_1 a_3 - \ldots) \cos 2\alpha_p \\
&\quad\quad + \ldots + 2a_j \cos j\alpha_p \\
&\quad = \varphi(r_*).
\end{aligned}$$

For the harmonic case, with no sampling of the function $\varphi(R_{p*})$, we have consequently found a relation between the scale function and the asymptotic value of the degrees of freedom,

$$\nu_p \approx \frac{\varphi(r^*)}{\varphi(R_{p*})}. \qquad (43\,e)$$

ν is the mean (or expectancy) of the χ^2-distribution, see e.g. KENDALL (1946). If $\gamma_p = 0$, expectation of c_p^2 is directly proportional to ν_p, and a few cases shall be given as illustrations of eq. (43e). In the first order Markov case $\nu_p = (1 - R_1^2)/(1 - 2R_1 \cos\alpha_p + R_1^2)$. If $R_1 > 0$ ("persistence"), we deduce that $\nu_p > 1$ when $0 \leqslant \alpha_p < \arccos R_1$, likewise $\nu_p < 1$ when $\arccos R_1 < \alpha_p \leqslant \pi$. "Persistence" will therefore tend to diminish amplitudes of short-periodic oscillations, and increase the relative importance of long-periodic oscillations. This result was derived by G. WALKER (1931). In the second order Markov case, the denominator of ν_p is equal to $(1 - 2R_1 \cos\alpha_p + R_1^2) \times (1 - 2R_2 \cos\alpha_p + R_2^2)$. If both roots are positive, the ν_p-distribution is qualitatively the same in the first order Markov case above. But if $R_1 > 0$ and $R_2 < 0$, the ν_p-distribution may have a minimum in the interval $0 < \alpha_p < \pi$, The location of this minimum depends on the relative sizes of R_1 and R_2. R_1 and R_2 may also be complex conjugate roots, and in this case the ν_p-distribution may have a maximum somewhere in the given interval. This discussion may be carried on to still higher order Markov schemes, the roots R_1, R_2,... completely determining the relative shape of the expected amplitude response of a harmonic variable.

If the values of α_p are equally spaced in the interval from 0 to π, we may, to a good approximation, replace a summation by a continuous integration. We can then show that the mean of the asymptotic values of ν_p is approximately one. If we in relation (43d) introduce $z = e^{\alpha_p \sqrt{-1}}$ and integrate with respect to α_p, we derive that

$$\bar{\nu}_p \approx -1 + \sum_{m=1}^{j} \frac{C_m}{\pi} \times \int_0^{\pi} \left(\frac{\exp \alpha_p \sqrt{-1}}{\exp \alpha_p \sqrt{-1} - R_m} + \frac{\exp -\alpha_p \sqrt{-1}}{\exp -\alpha_p \sqrt{-1} - R_m} \right) d\alpha_p,$$

or

$$\bar{\nu}_p \approx -1 + \sum_{m=1}^{j} \frac{C_m}{\pi \sqrt{-1}} \int_c \frac{dz}{z - R_m} = -1 + \sum_{m=1}^{j} 2\,C_m = 1, \qquad (43\,f)$$

according to the theory of residues. This result is rather important, and should always be kept in mind. The ratio

$$\frac{\nu_p(\alpha_p = 0)}{\nu_p(\alpha_p = \pi)} = \prod_{m=1}^{j} \left(\frac{1 + R_m}{1 - R_m} \right)^2$$

will generally differ much from unity, even for quite moderate values of R_m.

As the χ_p^2-values of the various harmonic components may not be statistically independent, it is advisable to apply the χ_I^2 test before focusing too much interest on individual details.

Let us demonstrate relation (43b) for a special case when $j = 3$ and we have the following set of complex conjugate roots,

$$r_k = C_1 R_1^k + C_2 \varrho^k \cos \alpha k. \qquad (44)$$

Introducing this relation for r_k we find that

$$\begin{aligned}
\varphi(R_{p*}) &= (1 - 2R_1 \cos\alpha_p + R_1^2) \\
&\times (1 - 2\varrho \cos\alpha_p \exp \alpha_p \sqrt{-1} + \varrho^2 \exp 2\alpha \sqrt{-1}) \\
&\times (1 - 2\varrho \cos\alpha_p \exp -\alpha \sqrt{-1} + \varrho^2 \exp -2\alpha \sqrt{-1}) \\
&= (1 - 2R_1 \cos\alpha_p + R_1^2) \\
&\times \{1 + 4\varrho^2 \cos^2 \alpha_p + \varrho^4 - 4\varrho(1 + \varrho^2) \\
&\times \cos\alpha_p \cos\alpha + 2\varrho^2 \cos 2\alpha\},
\end{aligned}$$

i.e.,
$$\begin{aligned}
\varphi(R_{p*}) &= (1 - 2R_1 \cos\alpha_p + R_1^2) \\
&\times [(1 - \varrho^2)^2 - 4\varrho(1 + \varrho^2) \cos\alpha_p \cos\alpha \\
&+ 4\varrho^2 (\cos^2 \alpha_p + \cos^2 \alpha)]. \qquad (45)
\end{aligned}$$

In case of trend, $\alpha \approx 0$, $\varrho \approx 1$, we recognize that

$$\varphi(R_{p*}) \approx 4(1 - 2R_1 \cos\alpha_p + R_1^2)(1 - \cos\alpha_p)^2.$$

We shall tabulate $\varphi(R_{p*})$ as a function of R_1 and ϱ when $R_{p,k} = \cos 2\pi k/14.765$, i.e., when considering a semilunar harmonic as one of variables, see BRIER & BRADLEY (1964). We shall further assume that the quantity $y(t)$ which we shall relate to $x_p(t)$, is a meteorological variable that is likely to have trend, i.e. $\cos\alpha \approx 1$.

BRIER & BRADLEY (1964) have recently published a paper on the apparent correlation of a precipitation index for the United States with a semilunar harmonic. The day-to-day persistence correlation of the precipitation index is

TABLE 2. *Tabulation of eq. (45) for the case* $\alpha \approx 0$.

ϱ	0	0.30	0.36	0.50	0.70	1.00
$\varphi(R_{p*})$ when $R_1 = 0$	1	0.294	0.225	0.114	0.046	0.032
– – $R_1 = 0.36$	0.473	0.139	0.106	0.055	0.022	0.015
– – $R_1 = 0.72$	0.207	0.061	0.047	0.024	0.010	0.007

0.36. As they did not consider the influence of a possible trend within the data, the discussion above may be applied to their case. There are 18,250 days in their sample. If we put $\varrho \approx 1$, $R_1 = 0.36$, $\alpha \approx 0$, $n = 18{,}250$, we find that $\varphi(R_{p*}) = 0.015$, and the χ^2-contribution from the p'th harmonic components would be as follows:

$$\chi_p^2 \approx \frac{0.015}{\varphi(r_*)}\, 18{,}250\, \frac{\frac{1}{2}(c_{p1}-\gamma_{p1})^2 + \frac{1}{2}(c_{p2}-\gamma_{p2})^2}{\operatorname{var}\varepsilon} \qquad (46)$$

Subscripts p_1 and p_2 refer to the semilunar sinus and semilunar cosinus respectively. We shall postulate no correlation in the universe, i.e. $\gamma_{p1} = \gamma_{p2} = 0$. If we derive that the square of the "correlation" coefficient $\frac{1}{2}(c_{p1}^2 + c_{p2}^2)(\operatorname{var}\varepsilon)^{-1}$ is equal to 0.001303 (cf. BRIER & BRADLEY (1964)) assuming $\{y(t) - y(t)\}^2 \approx \operatorname{var} y \approx \operatorname{var}\varepsilon$), we find $\chi_p^2 \approx 0.36/\varphi(r_*)$. As we should expect quite a considerable random component in the used precipitation index, $\varphi(r^*)$ is probably not less than, say 0.36. $\chi_p^2 \approx 1$ is according to our theory a rather low, insignificant value even for one degree of freedom. We may therefore conclude that the "correlation" of 0.001303 is not significantly different from zero if there is a trend. Further statistical analysis of this study is not possible, as there is no information about the persistence correlation at more than one lag. The precipitation index can vary from 0 to 40, but no frequency distribution (which may vary during the year) is given.

We shall finally mention that our theory covers the case when we have a multiple-root solution,

$$r_k = (c_1 + c_2 k + \ldots + c_j k^{j-1}) R_1^k.$$

STUMPFF (1936) and HISDAL *et al.* (1956) have demonstrated that the serial correlation of hourly sea-level pressure records is well described by the double-root solution

$$r_k = (c_1 + c_2 k) R_1^k,$$

in Germany as well as in Antarctica.

5. Sampling distributions of spectral bands

We may also apply the formula (43b) to the case with variables each containing a spectral band of harmonics. We shall assume that the bands do not overlap, and define such a variable, $x_B(t)$, by the following equation:

$$x_B(t) = \sum_{p=p_L}^{p_u} G(p) \sin(\alpha_p t + \beta_p), \qquad (48)$$

where $G(p)$ is any (finite) function of p in the spectral interval from p_L to p_u. To avoid minor complications we shall just study samples where all harmonics with argument $[(\alpha_i \pm \alpha_j)t + (\beta_i \pm \beta_j)]$ have an integer number of periods from $t = 1$ to $t = n$.

The serial correlation of $x_B(t)$ at τ lags can then be computed using relation (48). We shall introduce $k_p^2 = \{G(p)\}^2 / \sum_{p=p_L}^{p_u} \{G(p)\}^2$ as a measure of the influence due to the p'th harmonic, i.e.,

$$R_{B,\tau} = \sum_{p=p_L}^{p_u} k_p^2 \cos \alpha_p \tau. \qquad (49)$$

Using relation (43b) we may deduce that

$$\varphi(R_{B*}) = \sum_{p=p_L}^{p_u} k_p^2 \varphi(R_p) = \sum_{p=p_L}^{p_u} k_p^2 \prod_{m=1}^{f} (1 - 2R_m \cos \alpha_p + R_m^2). \qquad (50)$$

The asymptotic value of the degrees of freedom can next be derived from relation (43e).

The case $s = 0$ *(time series)*

It may be useful to repeat some of the derived results for the special case $s = 0$, as this case has obtained much attention in the literature. In the discussion between relations (25a) and (25b) we derived the following simplified form of relation (24):

$$\operatorname{var}\varepsilon \approx \overline{\{e(t)\}^2}\,\frac{n}{n - n^{-1}\sum_{u,v=1}^{n} r^{u-v}}. \qquad (51\,a)$$

This relation corresponds to relation (24) if we put $R_{0,u-v}$ equal to one.

The expected degrees of freedom, ν_I, for the sample variance of $e(t)$, or $y(t)$, is therefore

$$\nu_I = n - n^{-1}\sum_{u,v=1}^{n} r_{u-v} = n - \nu_{II}; \qquad (51\,b)$$

$\nu_{II} = n^{-1}\sum_{u,v=1}^{n} r_{u-v}$ is the expectation of the degrees of freedom taken over by the sample mean $\bar{y} = c_0 x_0 = c_0$.

As

$$\chi_I^2 = \frac{\varphi(r'_*)}{\varphi(r_*)}\,\frac{n\,\overline{\{e(t)\}^2}}{\operatorname{var}\varepsilon} = \frac{\varphi(r'_*)}{\varphi(r_*)}\,\frac{n\,\overline{\{y(t)-\bar{y}(t)\}^2}}{\overline{\{y(t)-\bar{y}(t)\}^2}},$$

it is still very important to have a fairly good approximation to the autoregressive scheme of the universe, cf. relation (26). If we know the universe autocorrelations r_τ, $\phi(r_*)$ is also known.

6. The limiting case of random sampling

Although the following relations are known from the theory of random sampling, see e.g. KENDALL (1946), we wish to demonstrate that similar formula may be derived from our relations.

If $j = 0$ in eq. (26), we find $r_{u-v} = 0$ for $u \neq v$ and $\phi(r_*) = \phi(R_{i*}) = 1$, $R_\varepsilon = 0$. Introducing these values in relations (39a) and (40), we notice that the frequency distribution of $(c_i - \gamma_i)$ is proportional to

$$\left[\exp -\frac{n\overline{x_i^2}}{\operatorname{var}\varepsilon}(c_i-\gamma_i)^2\right] d(c_i-\gamma_i),$$

i.e. $(c_i - \gamma_i)$ is normally distributed with zero mean and variance equal to $\operatorname{var}\varepsilon\,(n\overline{x_i^2})^{-1}$. The denominator of the variance is a sample estimate. We shall use eq. (24) to relate $\operatorname{var}\varepsilon$ to the sample residual variance. It should then be proper to test the significance of $(c_i - \gamma_i)$ by the "Student's" t-test,

$$t = \frac{(c_i-\gamma_i)\sqrt{\overline{x_i^2}\,(n-s-1)}}{\sqrt{\overline{e^2}}}, \qquad (52)$$

with $(n - s - 1)$ degrees of freedom.

7. Serial correlations of data having diurnal and annual trends

If there are various trends in the data, the serial correlations may have some peculiar variations which may be misinterpreted as significant deviations from the general Markov scheme given by eq. (26). If we e.g. study temperature records from stations at middle latitudes, the annual trend is very strong within the months close to the equinoxes. The annual trend is most pronounced for the high temperatures observed from slightly before noon and into the late afternoon. The lowest annual trends are found for the temperatures observed during the night and the early morning.

The way a trend influences a correlation has e.g. been described by NORDØ (1959), discussing linear and harmonic trends in the data. Close to the equinoxes the trend is almost linear within a given month. Let us therefore just study the linear trend, which according to NORDØ (1959, p. 5), gives the following relation for time t_0:

$$r_\tau(t_0) \approx \frac{\varrho_\tau + k_{t_0,t_0+\tau}}{(1+k_{t_0,t_0}^2)^{\frac{1}{2}}\,(1+k_{t_0+\tau,t_0+\tau}^2)^{\frac{1}{2}}} \qquad (53)$$

Here $r_\tau(t_0)$ is the apparent serial correlation at τ lags, and ϱ_τ is the serial correlation when the trend is removed.

The trend factor $k_{i,j}$ is given by the following relation:

$$k_{i,j} \approx \frac{\{T_i(n) - T_i(1)\}\{T_j(n) - T_j(1)\}}{12\left[\{T_i(t) - T_i(t)\}^2\{T_j(t) - T_j(t)\}^2\right]^{\frac{1}{2}}}. \qquad (54)$$

$T_i(t)$ is the mean value at time t of the month, $T_i(n)$ and $T_i(1)$ are the mean values of the last and the first day of the month, respectively. A typical value for the monthly trend just beyond the equinoxes is 5–7°C for hourly temperature observations at noon and in the afternoon, and 1–2°C less for the temperature records late in the night. The maximum values of $k_{i,i}$ should therefore be 0.2–0.4 at the equinoxes, and the highest values are likely to be found in the fall. Table 3 shows the apparent correlation at a lag of one day as a function of $k_{i,i}$ when $\varrho_{24\,h} = 0.70$. Consequently we should expect that the correlation $r_{24\,h}$ shows a marked

TABLE 3. *Influence of a monthly trend when* $\varrho_{24h} = 0.70$.

$k_{i,i}$	0	0.10	0.20	0.30	0.40	0.60	0.80	1.00
r_{24h}	0.70	0.73	0.75	0.78	0.79	0.81	0.83	0.85

diurnal variation during the spring and the fall months, and almost no diurnal variation at the midwinter and midsummer months. Our analysis seems to be verified, at least partly, by some extensive serial correlations carried out by GODSKE (1962, see p. 168). Whether the apparent peaks at 07 a.m. to 10 a.m. in May and August may be caused by similar trend effects is questionable.

The correlograms of the Oslo air temperature at hourly lags may be considered as another demonstration of the trend effects described by relation (53), see GODSKE (1962, p. 172). According to our analysis, see formulae (53) and (54), $r_\tau(t_0)$ should have a 24-hour periodic component equal to $k_{t_0,t_0+\tau}\,(1+k_{t_0,t_0})^{-\frac{1}{2}}\,(1+k_{t_0+\tau,t_0+\tau})^{-\frac{1}{2}}$. The Oslo correlograms referred to above, show indeed a pronounced diurnal variation of $r_\tau(t_0)$ for the month of April.

Next we shall consider a diurnal variation of the variance caused by variations of temperature on time scales less than a day. Variations of cloud cover over a station, showers etc. will e.g. during the summer season cause large temperature fluctuations at the ground in the day time when the solar heating is strong. In the long winter night similar effects are present, as e.g. the infrared radiation loss from the surface which is much dependent on the cloud cover.

Let us denote the temperature variation on the larger than a day time scale by $T'(t)$, and the smaller scale variation by $T''(t)$. Introducing these quantities in the relation (53), we derive ($\tau \geq 1$ day)

$$r_\tau(t_0) \approx \frac{\varrho'_\tau + k'_{t_0,t_0+\tau}}{(1+k'_{t_0,t_0})^{\frac{1}{2}}\,(1+k'_{t_0+\tau,t_0+\tau})^{\frac{1}{2}}},$$

where

$$\varrho'_\tau = \frac{\overline{\{T'(t)-\overline{T}(t)\}\,\{T'(t+\tau)-\overline{T}(t+\tau)\}}}{[\overline{\{T'(t)-\overline{T}(t)\}^2}\;\overline{\{T'(t+\tau)-\overline{T}(t+\tau)\}^2}]^{\frac{1}{2}}}$$

and

$$k'_{ij} \approx \frac{\delta_{ij}[2\,\overline{T'_i(t)\,T''_i(t)} + \overline{\{T''_i(t)\}^2}] + \frac{1}{12}\{\overline{T}_i(n)-\overline{T}_i(1)\}\,\{\overline{T}_j(n)-\overline{T}_j(1)\}}{[\overline{\{T'_i(t)-\overline{T}_i(t)\}^2}\;\overline{\{T'_j(t)-\overline{T}_j(t)\}^2}]^{\frac{1}{2}}}. \qquad (56)$$

$\delta_{ii} = 1$ and $\delta_{ij} = 0$ for $i \neq j$.

The "noise" term $2\overline{T'_i(t)\,T''_i(t)} + \overline{\{T''_i(t)\}^2}$ is generally positive, although the first term may have quite pronounced annual variations and should be made topic of a special study. Consequently we should expect a pronounced maximum of standard deviation of afternoon temperatures in summer. In winter we should expect a minimum in the afternoon, and at the equinoxes a semidiurnal variation with maxima in the late night and in the afternoon. Our considerations are nicely verified by the study of GODSKE (1962, Fig. 1, p. 164), although the curves of the spring and fall months should be corrected for the yearly temperature trend. The combined influence of trend and "noise" should therefore cause an apparent semidiurnal variation of $r_\tau(t_0)$ in spring and fall, see e.g. the April correlograms referred to above.

If we by careful analysis eliminate most of the trend effects and "noise" effects, the interdiurnal variation of $r_\tau(t_0)$ may become smaller and we may perhaps once again, from a statistical point of view, consider the process to be approximately a Markov process without a diurnal component.

When our records cover prolonged intervals of the year, we should take the yearly variation of ϱ'_τ into account, see relation (44).

8. Verification by numerical experiments

Although the preceding analysis seems to explain most of the variations observed in our meteorlogical data samples, another controlled verification may be performed on data generated on a computer. LORENZ[1] has lately carried out such experiments, tabulating the residual variances for $s = 0, 1, 2, 3, 4$. Each of the orthogonal variables $x_1(t)$, $x_2(t)$, $x_3(t)$, $x_4(t)$, and the residual, obey all first order Markov schemes with $r_1 = 0.75$. There are altogether 66 samples of size $n = 45$ for $s = 0, 1, 2, 3, 4$. In Table 4 we shall present an extract of the results obtained. We have standardized some variances in order to simplify the table.

[1] Personal communication concerning unpublished study.

TABLE 4. *Residual variances obtained from numerical models.*

No. of variables ...		$s=0$	$s=1$	$s=2$	$s=3$	$s=4$
Degrees of freedom	ν_I^*	44	43	42	41	40
	ν_I	38.53	35.10	31.66	28.22	24.78
Expected res. variance, $n=45$	Universe	1.000	0.652	0.456	0.370	0.348
	Random sample	0.978	0.623	0.426	0.337	0.309
	Rel. (24)	0.856	0.509	0.321	0.232	0.192
Mean of computed variances	22 cases	0.854	0.487	0.278	0.215	0.193
	42 cases	0.885	0.516	0.297	0.234	0.203
	66 cases	0.866	0.518	0.314	0.239	0.212

We notice that relation (24) gives fairly good estimates and that the random sample estimates are way off.

Inspecting the distributions of computed residual variances for $s=0$, 1, 2, 3, 4, we find in 34–35 % of all cases that χ_I^{*2} is extending beyond the lower (left) 2.5 % limit. 4–5 % of the χ_1^{*2} is beyond the upper 2.5 % limit. If we correct the degrees of freedom according to the ν_I values listed in Table 4, we still find 8 % and 11 % beyond the respective 2.5 % confidence limits. The "scale" factors $\varphi(r')/\varphi(r_*)$ and $\varphi(R'_{i*})/\varphi(r_*)$ were not evaluated in these model experiments, but later experiments[1] with higher order Markov schemes reveal that they are of vital importance, as pointed out elsewhere in this paper.[2]

[1] Personal communication from Mr. Carl Morey, M.I.T.

[2] Dr. Lorenz has lately computed 128 power spectra on data generated by a computer. The data obey a first order Markov scheme with $R_1=0.75$. The experiments show a satisfactory agreement with our theoretical estimates.

Conclusive remarks

Our sampling analysis is based on proper knowledge of the universe serial correlations of the residuals. In practice this assumption may present a serious problem to the investigator, when selecting the data. It becomes vitally important to use samples with very high degrees of freedom when establishing autoregression equations of the form (26).

Acknowledgement

The author is much indebted to Dr. E. N. Lorenz who has carried out the model experiments which have been used in verifying our theories.

The author is also grateful to Dr. Lorenz for his interest and support of this approach. The author wishes to express sincere thanks to Dr. C. L. Godske for offering many valuable comments concerning the manuscript.

REFERENCES

BAUR, F., 1944, Über die grundsätzliche Möglichkeit langfristiger Witterungsvorhersagen. *Annalen der Hydrographie*, 72, pp. 15–25.

BRIER, G., and BRADLEY, D. A., 1964, The lunar synodical period and precipitation in the United States. *J. Atm. Sci.*, 21, pp. 386–395.

DYNKIN, E. B., 1961, *Die Grundlagen der Theorie der Markoffschen Prozesse*. Springer-Verlag, Berlin.

GODSKE, C. L., 1962, Contribution to statistical meteorology. *Geophysica Norvegica*, 24, pp. 161–182.

HISDAL, V., AMBLE, O., and SCHUMACHER, N. J., 1956, Norwegian-British-Swedish Antarctic Expedition, 1949–1952. *Scientific Results*, Vol. 1, Part 2, pp. 17–22.

KENDALL, M. G., 1946–1948, *The Advanced Theory of Statistics*, I and II, Griffin and Co., London.

LORENZ, E. N., 1956, Empirical orthogonal functions and statistical weather prediction. Mass. Inst. of Tech., Sci. Rep. No. 1, Contract AF 19(604) 1566.

MILNE, W. E., 1949, *Numerical Calculus*, pp. 341–345. Princeton University Press.

NORDØ, J., 1953, A statistical discussion of a possible connection between solar activity and sea-level pressure. The Institute of Theoretical Astrophysics, Publication No. 2, pp. 4–13, 35–36.

NORDØ, J., 1959, Expected skill of long-range forecasts when derived from daily forecasts and past weather data. The Norwegian Meteorological Institute, Sci. Rep. No. 4, pp. 1–15.

NORDØ, J., 1960, Significance of regression equations derived from serially correlated data, and a procedure of selecting optimal predictors. The Norwegian Meteorological Institute, Sci. Rep. No. 8, pp. 3–19.

STUMPFF, K., 1936, Über die Zufallswahrscheinlichkeit von Periodizitäten in Beobachtungsreihen. *Veröff. des Meteor. Inst. der Universität Berlin*, **1** (2), p. 53.

WALKER, G., 1931, On periodicity in series of related terms. *Proc. Roy. Soc. A 131*, pp. 518–532.