

DISTRIBUTIONS OF WEEKLY AVERAGES OF DIURNAL TEMPERATURE MEANS AND RANGES ABOUT HARMONIC CURVES¹

CHRISTOPHER BINGHAM

The Connecticut Agricultural Experiment Station, New Haven, Conn.
[Manuscript received April 11, 1961]

ABSTRACT

The distributions of weekly averages of diurnal temperature maxima, minima, means, and ranges are found to be non-normal, although the errors in using the normal distribution do not impair the usefulness of derived estimates of probability. The central tendency and the variance are estimated by harmonic regression. This enables the estimation of probabilities for any week from as few as five parameters. A three-term harmonic curve fitted to individual years is adequate to describe the course of temperature.

1. INTRODUCTION

Useful probability statements about the occurrence of diurnal temperature means or ranges cannot be made without knowledge of the form of the distribution of the values and a practicable means of estimating the parameters which specify the distribution. The present paper explores the distributions of the weekly average of the daily mean temperature ($=\frac{1}{2}(\max + \min)$) and the weekly average diurnal range, as well as the related distributions of the weekly average diurnal maxima and minima.

2. HARMONIC REGRESSION

The fundamental tool which will be used in this development is harmonic regression. It is well known that any set of data, x_1, x_2, \dots, x_{2n} at equally spaced times t_1, t_2, \dots, t_{2n} may be exactly fitted by a series of the form

$$y = a_0 + \sum_{p=1}^n A_p \cos(pt - \phi_p), \quad t \text{ measured in degrees.} \quad (1)$$

This is the sum of cosine curves, each with semi-amplitude A_p and time of maximum $t = \phi_p/p$. Equation (1) can also be written in the form used herein,

$$y = a_0 + \sum_{p=1}^n (a_p \cos pt + b_p \sin pt) \quad (2)$$

where

$$a_p = A_p \cos \phi_p; \quad b_p = A_p \sin \phi_p; \quad \text{and} \quad a_p^2 + b_p^2 = A_p^2, \quad p = 1, \dots, n. \quad (3)$$

Such a sum will be called an n -termed Fourier series, and $(a_p \cos pt + b_p \sin pt)$ will be called the p th term. In many

applications it is found that very few terms are needed to give an excellent fit to the data, so that the residuals from the curve are of the same magnitude as the basic errors of observations. For instance, Craddock [6] observed that over much of the Northern Hemisphere, a two-term series provided an adequate fit to the mean monthly temperature. He did not, however, examine any variability between years of the coefficients of the two-term curve which best fit each year. Without this knowledge, a proper error term for significance tests for the reality of given terms is not available. Bliss [4] remedied this deficiency, describing a technique paralleling the standard analysis of variance for the fitting of orthogonal polynomials.

The mathematical model underlying Bliss's analysis is the following. Each observation for the j th unit of time in the i th year, y_{ij} , is considered as a sum

$$y_{ij} = (a_0 + \alpha_{i0}) + (a_1 + \alpha_{i1}) \cos t_j + (b_1 + \beta_{i1}) \sin t_j \\ + (a_2 + \alpha_{i2}) \cos 2t_j + (b_2 + \beta_{i2}) \sin 2t_j \\ + \dots + (a_r + \alpha_{ir}) \cos rt_j + (b_r + \beta_{ir}) \sin rt_j + \epsilon_{ij} \quad (4)$$

where $t_j = j/k \cdot 360^\circ$, $j = 0, 1, 2, \dots, k-1$ (if the units are weeks, $k = 52$). The ϵ_{ij} are independently distributed normal deviates with zero means and common variance σ^2 . The vectors $(\alpha_{i0}, \alpha_{i1}, \dots, \alpha_{ir}, \beta_{i1}, \beta_{i2}, \dots, \beta_{ir})$ are independent observations from a $2r+1$ variate multivariate distribution with zero mean vector. This model covers both the case of a single curve applicable to every year (when the variances $V(\alpha_{i0}) = V(\alpha_{i1}) = \dots = V(\beta_{ir}) = 0$), and the case of random variation of the curve from year to year. In the latter case the variance components of the coefficients enter into the variance of y_{ij} . For details on the analysis of variance see Bliss's [4] bulletin. His analysis is designed to test several hypotheses. The

¹ This work was supported in part by funds of a regional project in agricultural climatology, NE-35.

simplest such hypothesis is that $a_j^2 + b_j^2 = 0$, for particular values of j . One can also test the adequacy of r terms to fit the data, and the reality of the between years variance components associated with the coefficients. Significance tests are made using the F-test for variance ratios.

It should be pointed out that a certain arbitrariness must occur in the above model. This is the choice of a zero point from which we measure the time t . To illustrate this, let us suppose that

$$y_{ij} = (a_0 + \alpha_{i0}) + (a_1 + \alpha_{i1}) \cos t_j + (b_1 + \beta_{i1}) \sin t_j \\ + (a_2 + \alpha_{i2}) \cos 2t_j + (b_2 + \beta_{i2}) \sin 2t_j.$$

Then if $0 \leq \theta < 360^\circ$, y can equally well be expressed by

$$y_{ij} = (a_0 + \alpha_{i0}) + (a_1' + \alpha_{i1}') \cos t_j' + (b_1' + \beta_{i1}') \sin t_j' \\ + (a_2' + \alpha_{i2}') \cos 2t_j' + (b_2' + \beta_{i2}') \sin 2t_j', \quad (5)$$

where $t_j' = t_j + \theta$ and $a_1' = a_1 \cos \theta - b_1 \sin \theta$

$$b_1' = a_1 \sin \theta + b_1 \cos \theta$$

$$a_2' = a_2 \cos 2\theta - b_2 \sin 2\theta$$

$$b_2' = a_2 \sin 2\theta + b_2 \cos 2\theta,$$

with a similar relationship holding for the α 's and the β 's. If one fits a curve of the form (5) to data, exactly the same curve is obtained as when the more conventional form is fitted. It is easily seen that the variance component contributed by α_i will, in general, be different from that contributed by α_i' . However, the sum of the components for a given term will be unchanged. That is, $V(\alpha_i) + V(\beta_i) = V(\alpha_i') + V(\beta_i')$. Further, the semi-amplitudes A_i will be unchanged by the change of origin. Since the analysis of variance does not separate the two components of each term, this is sufficient to ensure that the analysis is invariant under any choice of origin.

Estimates, a_{i0}^* , a_{i1}^* , . . . , b_{ir}^* , of the coefficients, $(a_0 + \alpha_{i0})$, $(a_1 + \alpha_{i1})$, . . . , $(b_r + \beta_{ir})$, for the i th year are easily made as follows.

$$a_{i0}^* = \frac{1}{k} \sum_{j=0}^{k-1} y_{ij} \\ a_{im}^* = \frac{2}{k} \sum_{j=0}^{k-1} y_{ij} \cos mt_j \\ b_{im}^* = \frac{2}{k} \sum_{j=0}^{k-1} y_{ij} \sin mt_j, \quad m=1, 2, \dots, r; \quad t_j = \frac{j}{k} 360^\circ. \quad (6)$$

Estimates for the mean coefficients a_0 , a_i , . . . , b_r are obtained by averaging the above a_{im}^* 's and b_{im}^* 's over years. These are least squares estimates, and hence if the assumptions stated are fulfilled, they are the best unbiased estimates; further, if the error component is normally distributed they are also maximum likelihood estimates. The estimates remain unbiased even when the restriction of independence of residuals is removed.

Using the estimates a_{i0}^* , . . . , a_{ir}^* , b_{i1}^* , . . . , b_{ir}^* one can compute an "expected" value, \hat{y}_{ij} , for the j th week of the i th year by

$$\hat{y}_{ij} = a_{i0}^* + a_{i1}^* \cos t_j + b_{i1}^* \sin t_j + \dots \\ + a_{ir}^* \cos rt_j + b_{ir}^* \sin rt_j. \quad (7)$$

Adequacy of the model is seen in the deviations

$$d_{ij} = y_{ij} - \hat{y}_{ij} \quad (8)$$

of observed values from their "expected" values. These deviations will occasionally be referred to simply as d , without subscripts. Thus the use of this model produces two additional arrays, d_{ij} and \hat{y}_{ij} , similar in form to the original data. Any operations or computations that may be performed on the original data, y_{ij} , may be performed on the d_{ij} and on the \hat{y}_{ij} . For example, to the sample variance for the j th week,

$$s_j^2(y) = \frac{1}{f-1} \left[\sum_{i=1}^f y_{ij}^2 - \frac{1}{f} \left(\sum_{i=1}^f y_{ij} \right)^2 \right] \quad (9)$$

where f is the number of years in the sample, corresponds the variance of the deviations d_{ij} for the j th week.

$$s_j^2(d) = \frac{1}{f-1} \left[\sum_{i=1}^f d_{ij}^2 - \frac{1}{f} \left(\sum_{i=1}^f d_{ij} \right)^2 \right]. \quad (10)$$

Bliss applied this technique to a 14-year record of monthly mean temperatures at New Haven, Conn. His findings confirmed those of Craddock: A two-term series accounted for more than 97 percent of the observed sum of squares. It is well known that the variance of the temperature is higher in winter than in summer. Bliss found that a simple sine curve fitted to the log-variance of monthly mean temperatures adequately described the yearly trend of this variance. Concurrently I applied the technique to the monthly averages of the diurnal temperature range for the same period at New Haven [3]. Both the first and second terms of the regression curve were significant. The trend of the range was quite unexpected with maxima in May and October and absolute and relative minima in January and July, respectively. The distributions of both the mean and the range were normal for all practical purposes, although there were very slight indications of systematic skewness.

3. METHOD AND DATA

To carry out the laborious calculations for the application of the technique to weekly averages of temperature maxima, minima, means, and ranges, electronic data processing equipment was necessary. A program was written for the IBM 650 with the following functions. From cards containing weekly sums or averages of the maximum and minimum temperatures, coefficients a_{ip}^* and b_{ip}^* , $p=1, 2, 3$, of the three-term Fourier curve which

best fitted the data (maximum, minimum, mean, or range, depending on a coded instruction card) of each individual year were computed. Simultaneously sums of powers needed for moments up to the fourth were accumulated. From these "annual" coefficients the expected values \hat{y}_{ij} for each week were computed. In addition, the raw moments of the deviations, $d_{ij}=y_{ij}-\hat{y}_{ij}$, were accumulated, y_{ij} being the observed and \hat{y}_{ij} the expected values for a given week and year (equation (7), (8)). Finally the machine calculated an analysis of variance. Later it was found useful to write a program to convert raw moments to moments about the mean and compute

$$g_1=k_3/k_2^{3/2} \text{ and } g_2=k_4/k_2^2 \quad (11)$$

where $k_i, i=2, 3, 4$, are Fisher's k -statistics [8]. I also found it advantageous to write a program to carry out a harmonic analysis of unreplicated data and to synthesize curves from the estimated coefficients.

The complete set of programs was carried out on the data from only two stations, Storrs, Conn., and Keedysville, Md. both for the years 1926-56. In addition the mean and the range were studied for Uniontown, Pa., Eau Claire, Wis. both for 1926-56, and for Easton, Md. for two independent periods, 1896-1926, and 1926-56. These last will be referred to as Easton I and Easton II, respectively. The climatological year starting March 1 and omitting February 28 and 29 was used throughout.

4. CENTRAL TENDENCY

The mean squares from the analysis of variance for the harmonic regression fitted to the maximum, the minimum, the range, and the mean at Storrs are given in table 1. The sums of squares may be easily calculated using the degrees of freedom. The correct F-tests were applied according to Bliss [4]. The 30-year averages of the mean and the range for Easton II together with the average harmonic curve are shown in figures 1 and 2 respectively.

Maximum, Minimum and Mean.—As was expected, the first (sine curve) term of the Fourier series accounted for approximately 90 percent of the total sums of squares for

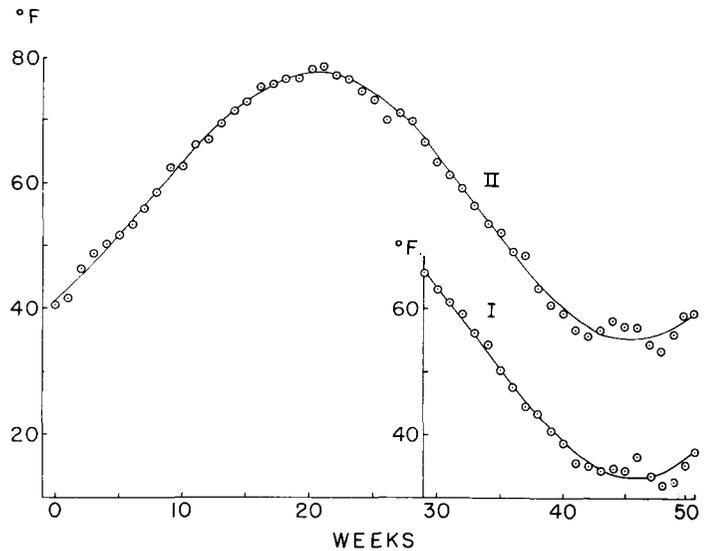


FIGURE 1.—30-year average of mean weekly temperature and the fitted harmonic curve, Easton, Md., 1896-1926 (designated I), and Easton, Md., 1926-1956 (designated II). Week 0 begins March 1.

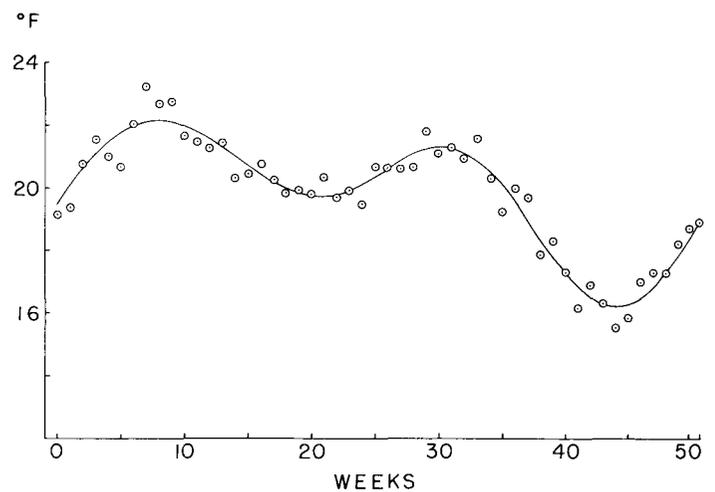


FIGURE 2.—30-year average of weekly average diurnal temperature range and the fitted harmonic curve, Easton II, 1926-1956.

TABLE 1.—Mean squares from analysis of variance of harmonic regression applied to weekly averages of diurnal temperature maxima, minima, ranges, and means, Storrs, Conn., 1926-1956

Row	Source of variation	D.F.	Mean squares			
			Maximum	Minimum	Range	Mean
1	Between years	29	*113.649	*73.631	†65.891	‡80.546
2	1st term	2	‡216080.805	†173223.360	†2631.135	†193994.295
3	2d term	2	‡494.045	‡12.245	‡578.350	96.540
4	3d term	2	14.715	*227.890	‡295.545	47.425
5	Scatter about curve	45	*35.370	†49.283	*15.096	‡32.424
6	1st term x year	58	†72.379	‡51.113	†22.515	‡66.064
7	2d term x year	58	†43.090	‡32.114	†23.743	‡32.081
8	3d term x year	58	*36.092	*29.750	*14.552	‡29.283
9	Residual	1305	25.868	23.206	11.424	21.105

*P<.05
†P<.01.
‡P<.001.

the maximum, the minimum, and the mean. The fit appeared excellent. In a few cases the second or third terms were significant (i.e., larger than one could expect from chance variation under the null hypothesis) but in terms of the percentage of sum of squares accounted for (<<1 percent) they were trivial. However, the interactions of terms by years all were significant and for the second and third terms were more important than the average effect of these terms. This indicates that the shape of individual years cannot in general be adequately described by a simple sine curve, despite the good fit to the averages.

An interesting feature was the consistency over a wide area of the departures of the average values for the period

1926–1956 of the maximum, minimum, and mean from the average Fourier curve. The magnitude and direction of the deviations for all four eastern stations were almost indistinguishable. In addition, at Keedysville and Storrs, the pattern of deviations of the mean was almost perfectly reproduced by the corresponding deviations of the maximum and the minimum. The time of year when the fit was least good and when the bulk of the sum of squares for scatter arose was the period from mid-December to mid-February. The last two weeks of December were considerably below the fitted curve, January above it, and February again below it. Although differing in detail the same general pattern of winter temperatures was observed in both independent 30-year samples from Easton, Md. (see fig. 1). The warm January is reminiscent of the fabled “January thaw.” However, the date of the “thaw” is reportedly well defined at January 20–23 [11] [14]. This does show up clearly in the Easton I record but is not visible in the Easton II record or at other stations analyzed for the later period. This casts doubt on the reality of this “singularity,” especially in view of the occurrence of maximum variability in January. No other such similarities in deviations are apparent in both records from Easton.

Range.—As shown in figure 2, the range followed the pattern uncovered in the preliminary analysis of the monthly data at New Haven [3]. We observed pronounced maxima in early- to mid-May and in mid-September with a summer relative minimum considerably above the winter minimum. The summer dip was least pronounced at Keedysville and Eau Claire but was clearly present in all records, including both independent Easton records. There were few, if any, recognizable similarities in deviations from the fitted curve between the stations. However, at all stations, the first two terms of the Fourier series were highly significant, with the third of lesser importance although still significant. The interaction of terms by years was quite uniformly significant. This, of course, indicates that there is considerable variation between years in the shape of the yearly course of diurnal range. Of considerable interest is close agreement of the phase of the first term (fundamental sine curve) with that of the sun. This apparently reflects the close relationship between the diurnal range and the energy input. That this relationship is not overpowering is, however, clearly shown by the spring and fall maxima. Despite first guesses that these maxima were corollaries of clearer skies in both spring and fall, sunshine and cloudiness records showed that in fact the spring period tended to be cloudier than the summer, although the high daytime gains and nighttime losses of radiation due to clear skies in the fall remain an acceptable explanation for the autumnal maximum. An examination of the daily temperature record for Mt. Carmel, Conn. suggested a possible explanation. It was observed that there was a considerable number of days when the daytime temperature went very far above the nighttime minimum but

returned to previous levels after sundown. In the late spring, the annual course of insolation is considerably ahead of the course of soil temperature, and hence, in general, ahead of the course of nighttime air temperatures. This makes it possible when conditions are right, for the temperature to rise sharply in the daytime and yet return at night, as we observed, to about the same level as the previous night. This behavior is corroborated by a comparison of the variances (between year and within year) of the maximum and the minimum temperatures. During these spring months the variance of the maximum is definitely greater than that of the minimum.

5. DISTRIBUTIONS

The analysis as discussed to this point has provided an efficient method for estimating a central tendency or location statistic for the distributions of temperatures and temperature ranges. Instead of 52 individual means, we have as few as three coefficients (for a sine curve) which give the location of the distribution for every week. However, for statements of probability, knowledge of the shape of the distributions is required. If it can be shown that the distributions are normal, or Gaussian, then the mean and the variance completely specify the distribution. On the other hand, if the distributions are non-normal, higher moments are necessary to specify or to approximate them. In addition, for the analysis of variance to be fully applicable, several assumptions should be met: The deviations of each observation from the theoretical regression curve for the year should be (i) normally distributed, (ii) homoscedastic (i.e., of equal variance), and (iii) independent. To examine the distributions of the observations and to test the first and second of these assumptions, moments were computed and the following procedure carried out. The sample variance for each week, $s_j^2(y)$, $j=0, \dots, 51$, (equation (9)) was computed from the observed values under consideration, to be referred to as y , the maximum, minimum, range, or mean. I also computed $g_1(y)$ and $g_2(y)$ (equation (11)) as measures of skewness and kurtosis. In addition, I calculated the same statistics $s_j^2(d)$, $g_1(d)$, $g_2(d)$ for the distributions of the deviations, d , of the observed values from the best fitting annual curves (see equation (10) for example). This process gave 52 values, one for each week, of the sample statistics just mentioned, for each of the two related distributions.

Variance.—The observed variance of the maximum and the minimum for Storrs and Keedysville changed smoothly over the year, as did the variance of the mean for all stations. Corroborating the previously mentioned tendency, the variance of these variates was considerably higher during the winter months than during the summer, with a reasonably smooth transition between the extremes (see figs. 3, 4, 5). Unfortunately for strict fulfillment of the conditions of the analysis of variance, the variance of the deviations, d , about the individual curves showed the same pattern, although to a reduced degree. In the case

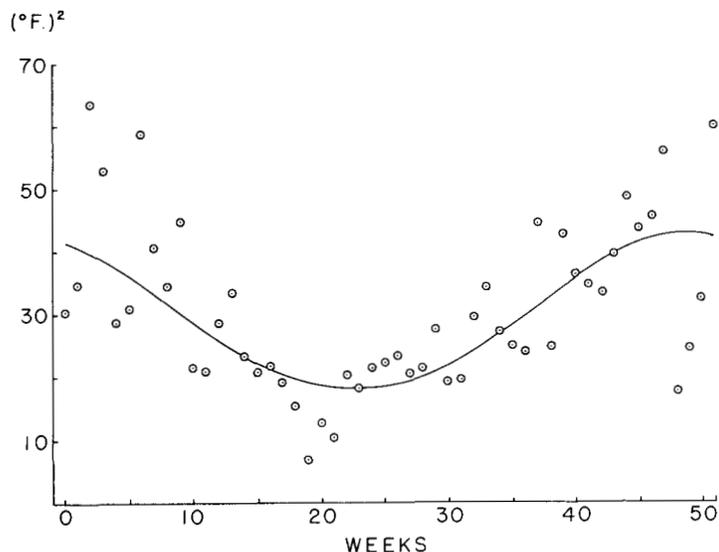


FIGURE 3.—Sample variance of weekly average diurnal maximum temperature and transform of the sine curve fitted to the log variance, Storrs, Conn. 1926-1956.

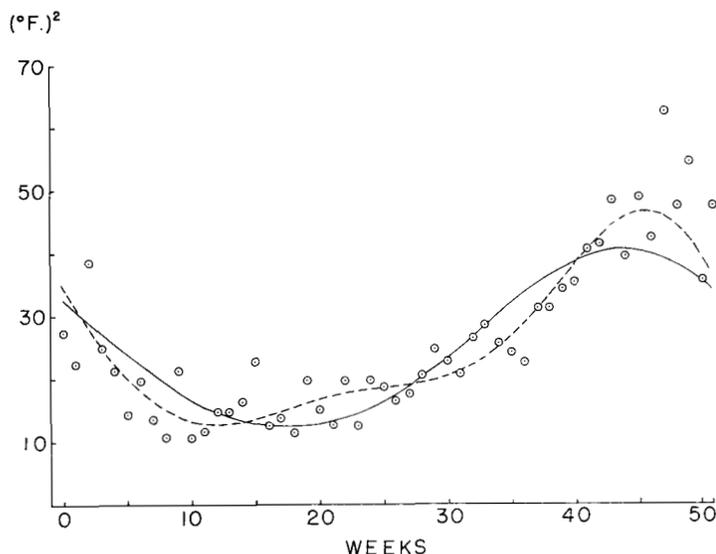


FIGURE 4.—Sample variance of weekly average diurnal minimum temperature and transforms of the sine curve (solid line) and the two-term Fourier curve (dashed line) fitted to the log variance, Storrs, Conn., 1926-1956.

of the range there were indications that the variance followed a double maximum pattern. Hardly visible at Storrs, it was more apparent at Keedysville, and still more so at Uniontown. Furthermore, it could be discerned in both independent Easton records. (The observed variance of the range at Uniontown is shown in figure 6.)

Since the trends in the variance were pronounced, we summarized the pattern by applying to the variance the same techniques of harmonic regression previously used on the temperature variates themselves. In order to minimize the non-normality of the distribution of the variance, it was transformed to its logarithm before the analysis was carried out [2]. As Bliss indicated, the an-

nual course of the log variance of the mean could be approximated by a simple sine curve. The same was true for the log variance of the maximum temperature. In both cases neither the second nor the third term was significantly different from zero. However, the analysis of variance indicated that, for the minimum, the higher terms were significant. The second term component was quite pronounced at Storrs while the third term was important at Keedysville, although at both stations the sine curve was clearly the dominating feature. The range exhibited a markedly different character. The log variance showed a significant second term trend, but little evidence of any single wave. For all four variates, when the deviations, d , about the annual curves were considered, the analysis showed that the F values for all terms decreased although the same relative importance of terms seemed to be the rule.

Correlations Between the Variates.—Of related interest to the course of the variance of the individual variates is

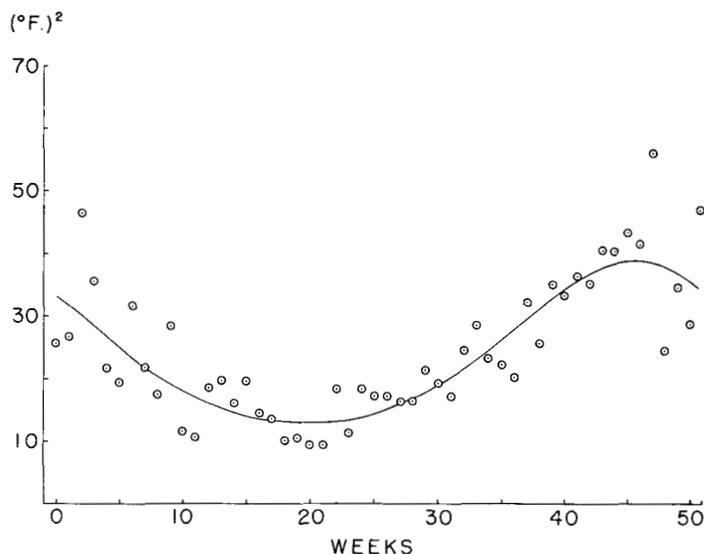


FIGURE 5.—Sample variance of weekly average diurnal mean temperature and transform of the sine curve fitted to the log variance, Storrs, Conn., 1926-1956.

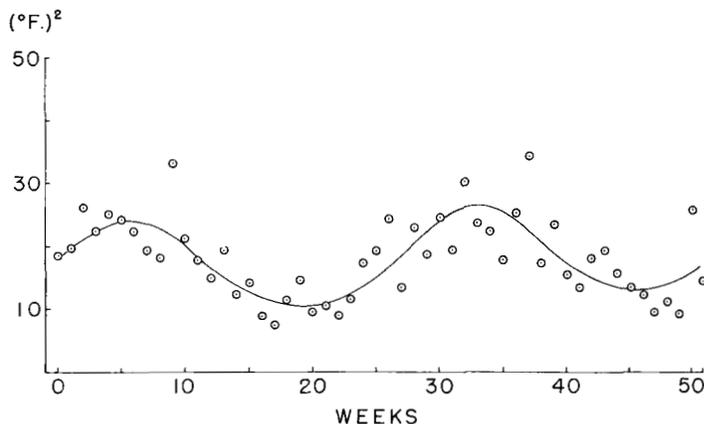


FIGURE 6.—Sample variance of weekly average of diurnal temperature range and transform of the two-term Fourier curve fitted to the log variance, Uniontown, Pa., 1926-1956.

the course of the correlation between them. Because the maximum-minimum and the mean-range coordinate systems are orthogonal, the correlation between the average maximum and the average minimum temperatures contains all the available information on the various correlations. Although not computed directly in our original program, this correlation is easily calculated from the variances of the maximum, minimum, and range as

$$r(\text{max}, \text{min}) = \frac{s_{\text{max}}^2 + s_{\text{min}}^2 - s_{\text{range}}^2}{2s_{\text{max}} \times s_{\text{min}}} \quad (12)$$

This calculation was carried out for the two stations, Keedysville and Storrs, for which all these variances were available. At both stations, the correlation r was always positive, reaching a maximum in winter and a minimum in summer. Like the variance of the mean, it followed a fairly smooth trend between these extremes. To obtain a quantitative description of its seasonal course we transformed r to $z = \tanh^{-1}r$, in order to stabilize its variance and minimize the non-normality of its distribution [9] and carried out a harmonic regression on each station.

The variance about a regression curve fitted to z derived from samples of 30 is theoretically $1/27 = 0.037037$. We can insert this as an added line in the analysis of variance with infinite degrees of freedom. This can be used in place of the ordinary error row in an analysis of variance of replicated data to test the adequacy of the fit and the reality of the regression. For both Storrs and Keedysville the scatter of z about a sine curve, when tested in this manner, was not significantly different from the theoretical variance ($P \geq 0.2$, $F < 1.2$), indicating there was no removable systematic variation in the residuals about the curve. The sine curve was highly significant ($P \leq 0.01$). Thus we may conclude that the transform, z , of the correlation between the weekly averages of maximum and minimum temperatures follows a sine curve. The only apparent regular departure from the sine curve is in April and early May when, at Keedysville, there were eight successive weeks when the observed correlation was less than the "expected" correlation, calculated from the sine curve. At Storrs a similar feature was observed for the same eight weeks, with the exception of one week in which the observed correlation slightly exceeded the "expected." In addition to this, there were other similarities between the two stations in the observed correlations. This is a further reflection of the previously noted similarities in the temperature records at quite widely separated stations (340 miles).

One might hope that, by using the "expected" values of $r(\text{max}, \text{min})$, s_{max}^2 , and s_{min}^2 , computed from the sine curves fitted to their transforms, one could recover the curve describing the course of the variance of the range by the formula,

$$s_{\text{range}}^2 = s_{\text{max}}^2 + s_{\text{min}}^2 - 2rs_{\text{max}} \times s_{\text{min}} \quad (13)$$

I hoped in this way to obtain a purely mathematical explanation for the behavior of s_{range}^2 based on simple assumptions concerning the behavior of the other second moments. The results, however, are not convincing. For Storrs, the curve of s_{range}^2 computed in this way is far too flat, although its only notable feature, a peak at week 10, does reflect a similar peak in the observed variance. At Keedysville, where the observed variance of the range displayed a pronounced double maximum pattern, I found only a single peak at the time of one observed maximum, with a point of inflection at the less important of the two minima.

Homoscedasticity and Independence of Residuals.—The above analysis has made one thing clear: Our assumption of homoscedasticity is not fulfilled. Although the mathematical model postulated implies some sort of a periodic form for the variance of the actual values, y , the variance of the deviations, d , from the annual curves should be constant. In the case of the average diurnal range this is nearly true. However, the other three variables are far from homoscedastic. In any case, the estimates of the regression coefficients are still unbiased estimates of the population values. It is difficult to assess the effect of this upon the tests in the analysis of variance. According to Cochran [5], the F-test for non-regressive designs is sufficiently robust as to be not misleading, although the more general case is not covered. Since, for the mean and the two extremes, the sine curve predominates, we can clearly accept the adequacy of the overall fit, although the tests cannot be considered exact. Since the variability of the variance of the range is far less prominent and regular, the tests should be less affected by heteroscedasticity.

Of equal or greater importance, perhaps, although more difficult to assess, is the possibility of dependent residuals. Direct tests for this are available [1] but laborious to apply and, as they stand, are not readily applicable to our computing scheme. Bliss [4] states that the technique of fitting separate curves to each year will have the effect of removing serial correlation between weeks, and leaving substantially independent residuals. From inspection of individual years, the observed residuals appear to be random. On the assumption of independent residuals, it follows that the moments for each week, as computed from the residuals, are independent and thus also g_1 and g_2 computed from these moments are independent. Serial correlation in these should tend to inflate the significance of trends in these statistics. Thus we would expect that if the residuals are independent, the F values in tests for the reality of regression curves fitted to the annual course of g_1 and g_2 should be lower for the distribution of the residuals than for the distribution of the raw observations. This, as we shall see, was observed and can be considered evidence of independence. However, no conclusive test has been found to clarify this point.

Tests for Normality.—We now examine the assumption of normal distributions for our variates. Being able to work with normal distributions is desirable for three

reasons. Firstly, such normality is assumed in our use of variance ratio tests in the analysis of variance, although it has been shown [5] that departures from normality, if not extreme, have little effect on the F test. Secondly, the normal distribution is known and easy to apply. Thirdly, if a distribution is normal all information about the values of the parameters defining it is subsumed in the sample mean and variance [7]. Furthermore, the powerful central limit theorem, applicable because we are considering averages of several observations in our distributions, tells us that the distributions approach normality.

A further consideration is also important: the use to which the probability statements derived from the distribution will be put. If extreme accuracy at all levels of probability is required, for instance, we must be extremely stringent regarding our tests of the distribution. On the other hand, in practical climatological applications we are not interested in, for example, the difference between a once-in-20 and a once-in-25 event. Even if tests show that almost certainly there is some departure from the normal curve, we may accept a normal approximation if the errors in probability estimated from our approximation do not impair the usefulness of the estimates.

As measures of departures from normality I chose, as previously mentioned, $g_1 = k_3/k_2^{3/2}$ and $g_2 = k_4/k_2^2$ where k_i are Fisher's k statistics. Departure of these from zero is indicative of non-normality. Both statistics are asymptotically normally distributed with mean zero and variance depending, in the null case, only on the sample size. Skewness or asymmetry is measured by g_1 while g_2 measures kurtosis. For ease in machine computation, g_2 was chosen in preference to

$$a = \frac{\sum |x_i - \bar{x}|}{s}$$

recommended by Geary and Pearson [10]. The g_1 's and g_2 's can be tested in two ways. First, compare the observed distributions of the g 's with those expected for samples drawn from independent normal populations. The ordinary significance test using a standard error is of this type. Second, examine the yearly course of g_1 or g_2 for meaningful patterns. An improbably regular pattern is as clear evidence of the presence of non-zero skewness or kurtosis as are high values of g_1 and g_2 .

The chief obstacle to the first type of test is our ignorance of the exact distributions of g_1 and g_2 under the null hypothesis. It is known that for samples as small as 30, such as concern us here, the distribution of g_1 is not far from normal while that of g_2 is strongly positively skewed. When the cumulative sample frequencies for the g 's at Keedysville were plotted on probit paper, the curves reflected this expectation. The distributions of the g_1 's were quite linear indicating approximate normality while the g_2 's exhibited a concave upward curve characteristic of positively skewed distributions.

TABLE 2.—Number of exceedences of the 1st, 5th, 95th, and 99th percentiles in 52 values of $g_1(y)$ and $g_1(d)$ calculated from weekly averages of diurnal temperature maxima, minima, ranges, and means, Keedysville, Md., 1926–1956

	$g_1(y)$			$g_1(d)$		
	>upper 5 percent	<lower 5 percent	Total	>upper 5 percent	<lower 5 percent	Total
Maximum.....	4	1	5	1	1	2
Minimum.....	2	3	5	2	2	4
Range.....	3	1	4	2	3	5
Mean.....	5	1	6	1	0	1
	>upper 1 percent	<lower 1 percent	Total	>upper 1 percent	<lower 1 percent	Total
Maximum.....	1	0	1	0	0	0
Minimum.....	2	2	4	0	0	0
Range.....	0	0	0	0	0	0
Mean.....	1	0	1	0	0	0

Although the exact distribution of g_1 is unknown, Geary and Pearson [10] have given approximate extreme percentage points. A comparison of the observed frequencies of g_1 beyond these points seemed the optimal procedure. Since these percentage points are not available for g_2 for sample size less than 100, this comparison was not possible for g_2 . In table 2 are given the number of times the upper and lower 5 percent and 1 percent levels for g_1 are exceeded for the 52 weekly distributions of the observations, y , and the residuals, d . We see that the agreement is as good as could be expected for the distribution of the deviations, d , about the annual curves although there are improbably few (i.e., no) values beyond the 1 percent point. However, in the distribution of the observations, y , themselves, the most striking feature is the appearance of four $g_1(y)$'s beyond the 1 percent point among the 52 g_1 's derived from the minimum temperature. Further, the upper 5 percent point of $g_1(y)$ for the maximum and the mean is exceeded too often. Thus a certain degree of non-normality is indicated in all the original variates except the range. There is, however, no indication that this is true for the deviation from the annual curves, d . This last result is important since it is assumed in our mathematical model.

Clearly, the above tests lose much of their validity if there is appreciable serial correlation between weekly values since this makes the g_1 's serially correlated. This affects the shape of the observed distribution of the 52 values of $g_1(y)$ for each variate. Thus the second type of test referred to above may be more applicable. This can best be done by fitting a regression curve on time to the computed statistics to uncover any significant pattern over the year. If the regression accounts for a significant part of the variation, it indicates a real departure from the expected values of zero. Because of the general robustness of the F-test, the departure of the distributions of the g_1 's and g_2 's from normality should not invalidate tests of significance of a regression curve. Accordingly, as previously mentioned, a three term harmonic curve was fitted to each of all available sets of g_1 's and g_2 's, with

somewhat mixed results. For the distribution of d there are no more F values significant at the 5 percent level than one would expect, with the exception of g_1 for the maximum temperature. For this, the second term for Storrs is significant and at Keedysville both the second and third terms are significant, although in both cases they are quite small. One may conclude, however, that there was no appreciable systematic skewness or kurtosis in the residual variation, whether one considers the mean, range, maximum, or minimum.

When we examine the observations y , there is a consistent pattern among the $g_1(y)$'s. For four out of the six records analyzed, including Easton II, the second term of the curve fitted to the course of $g_1(y)$ for the mean temperature was significantly different from zero. In the fifth record, Easton I, the amplitude of the second term was more than twice as large as any other, although not quite significant. Eau Claire, on the other hand, demonstrated a strong first term and small second and third terms. This discrepancy may possibly be traced to climatic differences between the Atlantic coastal and the Lake States. The same double maximum pattern was even more apparent in $g_1(y)$ for the minimum temperatures. At both stations analyzed, Storrs and Keedysville, the second term was highly significant, an almost identical pattern emerging. There appeared to be a tendency toward positive skewness in the fall and especially in the spring and negative skewness in the summer and especially in the winter.

We conclude that the distributions of the deviations d from annual curves are Gaussian for all four variates, while the observations y of the minimum and the mean have a skewness that changes seasonally. Since this pattern of skewness is most pronounced in the minimum and since the mean, as defined herein, is in part derived from the minimum, the primary pattern of non-normality is likely to be in the distribution of the minimum temperatures.

Although the tests discussed so far have not indicated any real departures from normality in the distributions of the range and the maximum, one further test placed in doubt the normality of the distributions of these variates, too. By the central limit theorem, the distribution of averages of independent g_1 's or g_2 's should be approximately normally distributed with zero expectation and variance $1/n$ times the variance of a single value. Thus if we consider the mean \bar{g}_i of the 52 values of g_i , $i=1, 2$, we can treat it as a normal deviate with variance $1/52 V(g_i)$. These then can be compared with the percentage points of the normal distribution. Any means outside, say, the 5 percent level would indicate a significant average departure from normality. Since we have no a priori knowledge of the direction in which deviations from the null hypothesis should occur, the proper test to use is the two-tailed comparison. When this test was carried out, no average departures from zero were found for either the $g_1(d)$'s or the $g_2(d)$'s of

the distributions of the deviations, d , for any of the four variates studied. Neither were there significant non-zero average departures in the $g_2(y)$'s. However, both the maximum and the range displayed definite signs of positive average skewness among the y 's. At both Keedysville and Storrs $\bar{g}_1(y)$ surpassed the upper 5 percent level and, in the case of Storrs, the upper 1 percent level. For the range, all except one station, Storrs, with $\bar{g}_1(y) = -0.0041$, showed some positive skewness, with two records, Easton I and Uniontown, surpassing the 1 percent and 5 percent levels respectively.

Conclusions Regarding Distributions.—First, the variation of all four variates about the annual harmonic curves is Gaussian. This is important from a theoretical point of view, since it increases our faith in the underlying model. However, because we cannot predict the shape of the yearly course of the variate, this knowledge is of little use from a practical point of view. Second, we have quite clear indications of skewness but not kurtosis in the distributions of all four variates. In the case of the range and the maximum temperature, the average skewness is consistently and, in some cases, significantly positive. On the other hand, although they display no average skewness, the distributions of the mean and the minimum have skewness that varies systematically over the year. Thus, for completely exact climatological statements of probability, normal assumptions will not be sufficient.

In theory one should try to find the exact nature of the distribution for each of the four variates. In the present instance, this could be more misleading than any assumptions of normality. First, there are insufficient data to establish any distribution as being correct without doubt. Second, the variation in skewness, at least for the mean and minimum temperatures, suggests that the "correct" distribution may vary seasonally. This would greatly increase the difficulty of applying any distribution. An alternative which is more approachable is the use of the third and fourth moments in fitting ad hoc distributions to the variates using either one or several of the Pearson curves or the Edgeworth approximation. Since we have no evidence of the departure of the fourth moment from the value expected for the normal distribution, the third moment, as reflected in $g_1(y)$, should be sufficient. Further, due to the large sampling variation in $g_1(y)$ resulting from our use of small samples, our best estimate of the population skewnesses should be the values calculated from the fitted periodic regression curve, or in the case of the maximum or the range, from the yearly average. Even if such an approximation does not provide an exact fit, the degree of change from the normal approximation should indicate the magnitude of error.

Tables of the cumulative distribution functions of the first Edgeworth approximation for different values of γ_1 , the population measure of skewness, are available [13]. Referring to these I found that for small departures from

symmetry ($|g_1| \leq 0.2$) use of this approximation does not markedly affect probability points given by the normal distribution. Assuming a standard deviation, σ , of 7° F. (near the maximum for any variate) at $g_1 = \pm 0.2$ the upper and lower 1 percent points are displaced approximately 1° F. with lesser changes nearer the median. Thus it seems safe to say that use of the normal distribution for the maximum and the range does not introduce any serious errors since the greatest average $g_1(y)$ (Storrs maximum) is 0.1643. For the minimum temperature at Keedysville which displayed the strongest systematic departures from zero of $g_1(y)$, the "expected" value of $g_1(y)$ ranged from about -0.48 to $+0.52$. It is true that the 1 percent level is quite distorted by this amount of asymmetry. However, the 5 percent level is only about a degree off (again letting $\sigma = 7^\circ \text{ F.}$). Conversely, a deviation which by the normal distribution would be surpassed 5 times in 100 years would, under these circumstances, occur about 6.5 times in 100 years. Similarly the estimated 10 percent point would be exceeded about 11 times every 100 years. The Edgeworth approximation should be fairly accurate when the shape of a distribution is not far removed from the normal form. Since, as mentioned earlier, the central limit theorem assures that the distributions approach normality, the use of this approximation in the present case should be appropriate. Hence, these figures show that the use of the normal curve will be satisfactory on all levels of probability except beyond the 5 and 95 percentiles where one should probably apply some extreme value distributions.

6. PROBABILITY ESTIMATION

The above considerations on the distribution of temperature variates lead to a satisfactory method of estimating probabilities. Since we may assume for our purposes that the variates are normally distributed, specification of the means and variances completely determines the distribution. Our harmonic curves fitted to the variates provide the first of these parameters, while the curves fitted to the log variance provide the second. In the case of the maximum and mean temperatures both curves are essentially sine curves, with higher terms, even when statistically significant, changing the estimates little. The course of the minimum temperature is adequately described by a sine curve. The higher terms of the curve fitted to the variance do not seem to make an appreciable difference. For example, for Storrs the 5 percentile and the 1 percentile are changed a maximum of 0.8° and 1.2° F. by adding a second term. Similarly, the probability assigned to a given departure from the mean seems to be changed by at most 0.02 to 0.03. Since this is less than the level of accuracy ordinarily desired, the sine curve alone can be used. Hence for the maximum, the minimum, and the mean, the estimation of probabilities for any week of the year reduces to the fitting of six constants. I should point out

that although a sine curve is adequate to describe the courses of these temperatures at the stations studied, the possibility is not excluded that at some locations higher terms may be necessary. Each term can, however, be specified by only two additional constants.

For the range, at least two terms, i.e., five parameters, should be used for estimating the central tendency. Since at some stations the second term in the curve describing the course of $\log s^2_{\text{range}}$ is of prime importance, five constants seem to be necessary to estimate the variance. However, in cases where there is no discernible pattern in the variance, it is probably best to use a single average variance derived from the row for scatter about the average curve. Thus for Storrs one would use $15.7 (\text{ }^\circ \text{ F.})^2$ as the variance of the range (table 1, row 5).

Although the above discussion puts forward a relatively simple technique for estimating the mean and variance of temperatures from a small number of constants, there is a further simplification that may produce equivalent results, at least for the mean temperature. I noticed that not only does the log variance of the mean temperature follow a simple sine curve but also the time of the maximum variance is about 180° , or 6 months, out of phase with the temperature. For example, the yearly maximum of the mean temperature at Keedysville, computed from the best fitting simple sine curve, falls $137^\circ 27'$ or 139.4 days after March 4 (the midpoint of week zero) while the maximum of $\log s_{\text{mean}}$ is at $313^\circ 48'$ or 318.2 days after March 4 and $176^\circ 21'$ or 178.8 days after the maximum temperature. This suggests that the log variance and the mean may be linearly related with a negative slope. When \bar{y} was plotted against $\log s$, this was seen to be substantially the case. In fact, the logarithm so little changes the relationships among the standard deviations that an almost equally good fit to a linear relationship is obtained if one plots \bar{y} against s . If one assumes that the phase difference between the log variance and the mean temperature is exactly 180° , which is very close to the observed difference, then one can rather easily compute the slope of the regression of $\log s$ on \bar{y} from a knowledge of the amplitudes of the respective sine curves.

If

$$\begin{aligned} \bar{y} &= a_0 + a_1 \sin t + b_1 \cos t \\ &= a_0 + A_1 \cos (t - \phi_1) \end{aligned} \tag{14}$$

where A_1 is the semi-amplitude and ϕ_1 is the time in degrees of the maximum of the mean as measured from the start of the climatological year, and if

$$\begin{aligned} \widehat{\log s} &= a'_0 + a'_1 \sin t + b'_1 \cos t \\ &= a'_0 + B_1 \cos (t - \phi_1 - 180^\circ), \end{aligned} \tag{15}$$

then

$$\frac{\bar{y} - a_0}{A_1} = \cos (t - \phi_1) = -\cos (t - \phi_1 - 180^\circ) = -\frac{\widehat{\log s} - a'_0}{B_1}$$

Thus

$$\widehat{\log s} = -\frac{B_1}{A_1} \bar{y} + \frac{B_1}{A_1} a_0 + a'_0 \tag{16}$$

Hence we can give the "regression" of $\log s$ on \bar{y} from a knowledge of the sine curves fitted. To test the accuracy of this method, both the ordinary least squares line and the line computed from equation (16) were calculated for the mean temperatures at Storrs and at Easton I and Easton II and are given in table 3 together with the regression line calculated from equation (16) for the other stations studied. We note a remarkable homogeneity among the slopes of the regression lines for the three records which cover two Maryland stations, with the slope at Storrs quite close. However, when one crosses the Alleghenies to Uniontown, the relation changes radically, and is even more changed at Eau Claire. Even if the observed interstation homogeneity in the three easternmost stations is fortuitous, the stability of the regression lines between two independent records at Easton shows that it may be possible to estimate temperature probabilities for any week of the year from only five parameters, the overall mean, the phase angle, and the amplitude of the temperature curve, and the amplitude and mean of the log variance. This relationship does, however, need further study, especially with respect to the geographical stability of the "regression" of $\log s$ (or s) on \bar{y} .

One point in the above discussion may need some explanation. It is a well-known fact that in drawing samples from a single normal population, s^2 , the estimate of the population variance, and \bar{y} , the estimate of its mean, are independently distributed. How then do we have such a marked correlation between the means and the log variances, when we have seen that the distributions are not far from normal? The question is, however, meaningless since, although we have drawn 52 samples from normal or near normal populations, these populations have not been the same. There is neither a priori nor a posteriori evidence that the population of mean temperatures for week one is the same (i.e., has identical characteristics) as the population for any other week. In fact, our investigation has demonstrated that these distributions are definitely different. Thus one should not expect the independence between s^2 and \bar{y} which would result if the populations were the same. A relationship between the true values (not merely the estimates) of the averages and

variances of the mean temperatures, say, is clearly indicated by the data at hand. Its reality can be indicated statistically but its cause must be physical. In a similar vein the standard tests for equality of variance, say, between two different weeks are quite useless. In advance of such a test, we have a well established result that the variances are, in general, unequal.

7. PROBLEMS IN FURTHER APPLICATIONS

To this point we have spoken only about the probabilities of average temperatures or average diurnal ranges for a given week, without reference to the probabilities for derived quantities, or averages over longer periods such as months or seasons. If the mathematical model were completely correct, one could, theoretically, compute quite complicated probabilities concerning degree days, frost dates, etc., from a knowledge of the joint distribution of the coefficients of the curves which describe each year. This knowledge is also needed for efficient use of the parameters in relating them, as being in some sense a description of a yearly temperature regime, with other elements, climatological, agricultural, or physical. Before proceeding further, we should inquire to what degree the estimated coefficients for a given year do contain the salient features of that year. To this end, I tabulated the published departures of the monthly mean temperatures from the long-term normal for the regions or States in which our stations are located. Years were then picked by eye which had various characteristics, e.g., warmer than normal spring combined with colder than normal winter, etc. Then, using the best fitting harmonic curves for these years for the stations involved, the "expected" temperature, \hat{y} , for each week was computed and compared against the 30-year average curve. In every case the unusual pattern for which that year had been chosen was observed in the fitted harmonic curve. Although they lack fine detail, the curves for each year definitely reflect the course of the temperature for that year.

Determination of the distribution of the true coefficients is far more difficult since they cannot be observed directly. We can examine only estimated coefficients. A further complication is the previously discussed dependence of the coefficients upon our arbitrary choice of origin. Thus a full discussion of this problem would involve determining all relationships among the coefficients which are not changed by any choice of origin. However, despite the above objections, I considered it valuable to examine more carefully the estimates of the coefficients for the particular choice of origin we used. To provide a minimal test of normality in the marginal distributions of the separate coefficients, the ranked estimated coefficients for the curves fitted to the mean temperature were plotted against rankits (expected order statistics for the normal distribution [9]) for each of the two Easton samples. Departure from normality of the estimates should be reflected in non-linearity in the plotted points. In every case where any noticeable non-linearity occurred in one

TABLE 3.—Regression equations of $\log s$ on \bar{y} computed by least squares and by equation (6)

Station	Least squares line	Equation (6)
Easton I.....	$\log s = 1.020113 - 0.006454\bar{y}$	$\log s = 1.02479 - 0.006543\bar{y}$
Easton II.....	$\log s = 1.018815 - 0.006057\bar{y}$	$\log s = 1.02260 - 0.006128\bar{y}$
Storrs.....	$\log s = 0.926768 - 0.005239\bar{y}$	$\log s = 0.92902 - 0.005277\bar{y}$
Keedysville.....	(1).....	$\log s = 1.02814 - 0.006067\bar{y}$
Uniontown.....	(1).....	$\log s = 1.18340 - 0.008311\bar{y}$
Eau Claire.....	(1).....	$\log s = 0.89526 - 0.002067\bar{y}$

¹ Not computed.

record, it was not visible in the other. Thus I concluded that there was no evidence for rejecting the hypothesis of normality among the estimates of the coefficients. To the degree that non-normality in the distribution of the actual coefficients would be reflected in non-normality of the estimates, this indicates normality in the marginal distributions of the actual coefficients. Hence, they are probably distributed in a multivariate normal distribution. Further, we computed covariance and correlation matrices for the seven estimated coefficients for each of the two Easton samples, and then combined them to get an estimated covariance matrix based on a 60-year sample. It can easily be shown that for our choice of origin, the winter maximum of the variance of the mean temperature implies that there should be positive correlation between a_0 and a_1 and negative correlation between a_0 and b_1 . Hence any test for the significance of these observed correlations should be single-tailed, and we should also compute the partial correlation $r(a_1, b_1|a_0)$. We found that there were several correlations of statistical significance, i.e., larger than one could expect to occur by chance alone. However, their practical significance is negligible since they are all less than 0.5 in absolute value. The physical or climatological meaning of such correlations is not clear, especially in view of the fact that any one may be reduced to zero by some choice of an origin in fitting the polynomials.

There is one further problem which would limit the application of this method, even were the above problems solved. There appears to be considerable difficulty in making a meaningful correspondence between the parameter space and the space of "year types." In other words, there is no clear way to pick out all those sets of coefficients $a_0, a_1, \dots, a_3, b_1, \dots, b_3$, which, when used in a harmonic curve, would produce a "warm spring." Unless one can accomplish this, knowledge of the distribution of the coefficients cannot be fully applied.

There are many applications, however, of the average curve computed for a location, in which the form of the distribution of the yearly coefficients is of minor importance. Prescott [12] used maps of the phase and amplitude of sine curves fitted to monthly temperatures for stations in Australia to locate suitable homoclims for new crops. Except for the difficulties attached to solving polynomial equations involving trigonometric terms, expression of the yearly course of temperature by a periodic curve provides a means of locating the time when the average temperature is at a maximum or a minimum. If

$$y = a_0 + a_1 \cos t + b_1 \sin t + a_2 \cos 2t + b_2 \sin 2t + \dots,$$

the relative maxima and minima occur at those values for t for which

$$-a_1 \sin t + b_1 \cos t - 2a_2 \sin 2t + 2b_2 \cos 2t - \dots = 0.$$

Similarly, the day on which the average maximum, minimum, or mean temperature passes some critical value can be computed from the coefficients of the average curve. In the case of the mean temperature, of course, this type of problem is generally simple since the temperature can be considered to follow a sine curve. However, for more complex curves, a process of successive approximation seems to be necessary.

REFERENCES

1. R. L. Anderson and T. W. Anderson, "Distribution of the Circular Serial Correlation Coefficient for Residuals from a Fitted Fourier Series," *Annals of Mathematical Statistics*, vol. 21, 1958, pp. 59-81.
2. M. S. Bartlett, "The Use of Transformations," *Biometrics*, vol. 3, No. 3, Mar. 1947, pp. 39-52.
3. C. Bingham, "Analysis by Harmonic Regression of Diurnal Temperature Range" (abstract), *Bulletin of the American Meteorological Society*, vol. 39, No. 8, Aug. 1958, pp. 441-442.
4. C. I. Bliss, "Periodic Regression in Biology and Climatology," *Bulletin* 615, The Connecticut Agricultural Experiment Station, New Haven, Conn., 1958, 55 pp.
5. W. G. Cochran, "Some Consequences When the Assumptions for the Analysis of Variance Are Not Satisfied," *Biometrics*, vol. 3, No. 3, Mar. 1947, pp. 22-38.
6. J. M. Craddock, "The Variation of the Normal Air Temperature Over the Northern Hemisphere During the Year," a paper of the Meteorological Research Committee (London), M.R.P. No. 917. (A copy is available in the Library of the Meteorological Office.)
7. H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946, 575 pp. (pp. 494-495).
8. R. A. Fisher, *Statistical Methods for Research Workers*, 6th edition, Oliver and Boyd, Edinburgh, 1936, 339 pp. (p. 75).
9. R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, 3d edition, Oliver and Boyd, London and Edinburgh, 1948, 112 pp.
10. R. C. Geary and E. S. Pearson, *Tests of Normality*, London, 1938, 15 pp.
11. H. Landsberg, *Physical Climatology*, 2d edition, Gray Printing Co., Inc., Dubois, Pa., 1958, 446 pp. (p. 152).
12. J. A. Prescott, "The Value of Harmonic Analysis in Climatic Studies," *The Australian Journal of Science*, vol. 5, 1943, pp. 117-119.
13. H. R. Tolley, "Frequency Curves of Climatic Phenomena," *Monthly Weather Review*, vol. 44, No. 11, Nov. 1916, pp. 634-642 (see table, pp. 640-641).
14. E. W. Wahl, "The January Thaw in New England (An Example of a Weather Singularity)," *Bulletin of the American Meteorological Society*, vol. 33, No. 9, Nov. 1952, pp. 380-386.