# 1
# Fourier analysis



It is often the case in the physical sciences, and sometimes the social sciences as well, that measurements of a particular variable are collected over a period of time. The collected values form a data set, or *time series*, that may be quite lengthy or otherwise difficult to interpret in its raw form. We then may turn to various types of statistical analyses to aid our identification of important attributes of the time series and their underlying physical origins. Basic statistics such as the mean, median, or total variance of the data set help us succinctly portray the characteristics of the data set as a whole, and, potentially, compare it to other similar data sets.

Further insight regarding the time series, however, can be gained through the use of *Fourier*, *harmonic*, or *periodogram analysis* – three names used to describe a single methodology. The primary aim of such an analysis is to determine how the total variance of the time series is distributed as a function of frequency, expressed either as ordinary frequency in cycles per unit of time, for example, cycles per second, or angular frequency in radians per unit of time. This allows us to quantify, in a way that the basic statistics named above cannot, any *periodic* components present in the data. For example, outside air temperature typically rises and falls with some regularity over the course of a day, a periodic component governed by the rising and setting of the sun as the earth rotates about its axis. Such a periodic component is readily apparent and quantifiable after applying Fourier analysis, but is not described well by the mean, median, or total variance of the data.

In the first two sections of Chapter 1, we will learn some essential terminology of Fourier analysis and the fundamentals of performing Fourier analysis and its inverse, Fourier synthesis. Example data sets and their analyses are presented in Section 1.3 to further aid in understanding the methodology.

As with other types of statistical analyses, statistical significance plays an important role in Fourier analysis. That is, after performing a Fourier analysis, what if we

*Time Series Analysis in Meteorology and Climatology: An Introduction*, First Edition. Claude Duchon and Robert Hale.

find that the variance at one frequency is noticeably larger than at other frequencies? Is this the result of an underlying physical phenomenon that has a periodic nature? Or, is the larger variance simply statistical chance, owing to the random nature of the process? To answer these questions, in Section 1.4 we examine how to ascribe confidence intervals to the results of our Fourier analysis.

In Section 1.5, we take a more detailed look at particular issues that may be encountered when using Fourier analyses. Although not generally requisite to performing a Fourier analysis, the concepts covered are often critical to correct interpretation of the results, and in some cases may increase the efficacy of an analysis. An understanding of these topics will allow an investigator to pursue Fourier analysis with a high degree of confidence.

## 1.1 Overview and terminology

### 1.1.1 Obtaining the Fourier amplitude coefficients

The goal of Fourier analysis is to decompose a data sequence into harmonics (sinusoidal waveforms) such that, when added together, they reproduce the time series. What makes sinusoidal waveforms an appropriate representation of the data is their orthogonality property, their ability to successfully model waves in the atmosphere, oceans, and earth, as well as phenomena resulting from solar forcing, and the fact that the harmonic amplitudes are independent of time origin and time scale (Bloomfield, 1976, p. 7).

Harmonic frequencies are gauged with respect to the *fundamental period*, the shortest record length for which the time series is not repeated. In most practical cases, this is the entire length of the available record, since the record typically does not contain repeated sequences of identical data. The harmonic frequencies include harmonic 1, which corresponds to one cycle over the fundamental period, and higher harmonics that are integer multiples of one cycle. Thus each harmonic is always an integer number of cycles over the length of the fundamental period.

To establish a sense of Fourier analysis, consider a simple example. The heavy line in Figure 1.1 connects the average monthly temperatures at Oklahoma City over the three-year period 2007–2009. By looking at the heavy line only, it is quite evident that there is a strong annual cycle in temperature. It is equally clear that one sinusoid will not exactly fit all the data, so other harmonics are required. The fundamental period, or period of the first harmonic, is the length of the record, three years. The third harmonic has a period one-third the length of the fundamental period, and consequently represents the annual cycle. The thin line in Figure 1.1 shows the third harmonic after it has been added to the mean of all 36 months, that is, the 0-th harmonic. As expected, the third harmonic provides a close fit to the observed time series.

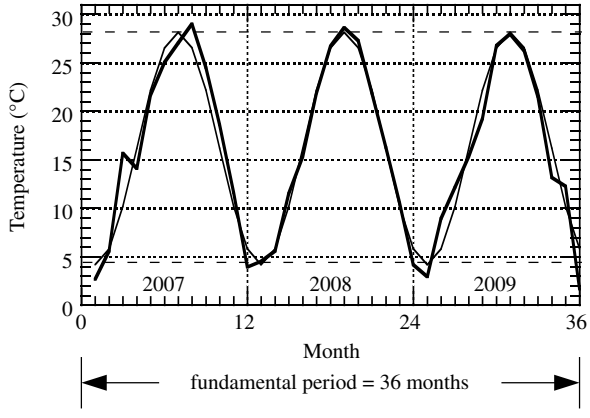



**Figure 1.1** Mean monthly temperatures at Oklahoma City 2007–2009 (heavy line), and harmonic 3 (light line) of the Fourier decomposition.

## 1.1.2 Obtaining the periodogram

The computation of variance arises in elementary statistics as a defined measure of the variability in a data set. When the computation of variance is applied to a time series, it is similarly defined. Now, though, the variance in the data set can be decomposed into individual variances, each one related to the amplitude of a harmonic. Just as adding the sinusoids from all harmonics reproduces the original time series, adding all harmonic variances yields the total variance in the time series. How the decomposition is achieved and how variance is related to harmonic amplitude are discussed in Section 1.2.

A *periodogram* is a plot of the variance associated with each harmonic (usually excluding the 0-th) versus harmonic number and shows the contribution by each harmonic to the total variance in the time series. Henceforth, the term periodogram will be used to refer to the calculation of variance at the harmonic frequencies. The term *Fourier line variance spectrum* is synonymous with periodogram, while the generic term *spectrum* generally means the distribution of some quantity with frequency.

The variance at each harmonic frequency is given by the square of its amplitude divided by two, except at the last harmonic. Figure 1.2 shows the periodogram (truncated to the first 10 harmonics) of the data in Figure 1.1 where we see that harmonic 3 dominates the variability in the data. The small variances at harmonics 6 (period = 6 months) and harmonic 9 (period = 4 months) are easily observed in Figure 1.2, but, in fact, there are nonzero variances at all 18 possible harmonics (excluding the 0-th) and their sum equals the total variance of 75.23 $^{\circ}C^2$ in the 2007–2009 Oklahoma City mean monthly temperature time series.

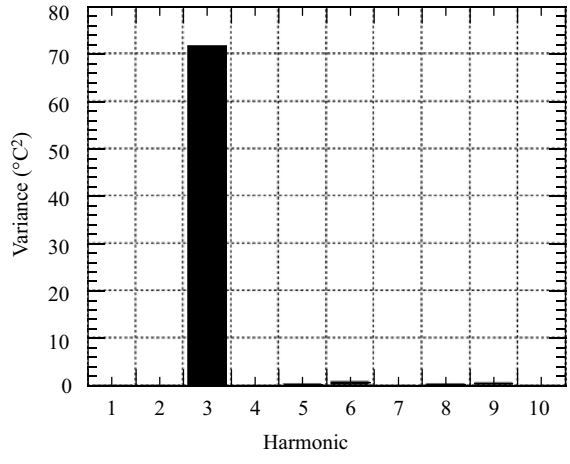



**Figure 1.2** Variance at each harmonic through 10 for the data in Figure 1.1.

The periodogram in Figure 1.2 was computed using the computer program given in Appendix 1.A. This program, written in Fortran 77, performs a 'fast' Fourier analysis of any data set with an even number of data and has been used throughout this chapter to compute the periodograms we discuss.

### 1.1.3 Classification of time series

We can classify time series of data into four distinct types of records. The type of record determines the mathematical procedure to be applied to the data to obtain its spectrum.

The 36 values of temperature $x_n$, in Figure 1.1, connected by straight-line segments for ease in visualization, constitute a *finite digital* record. Digital time series arise in two ways (Box and Jenkins, 1970, p. 23): sampling an analog time series, for example, measuring continuously changing air temperature each hour on the hour; or accumulating or averaging a variable over a period of time, for example, the previous record of monthly mean temperatures at Oklahoma City. With respect to the latter case, if N is the number of months of data and $\Delta t$ the time interval between successive values, the record length in Figure 1.1 is $N\Delta t = 36$ months. In this case, as well as with all finite digital records, all data points can be exactly fitted with a finite number of harmonics. This is in contrast to a *finite analog* record of length T, such as a pen trace on an analog strip chart, for example, a seismograph, for which an infinite number of harmonics may be required to fit the signal.

Figure 1.3 is an example of a finite analog record. Sampling the time series at intervals of $\Delta t$ yields the finite digital record shown in Figure 1.4. The sample values again have been connected by straight-line segments to better visualize the variations

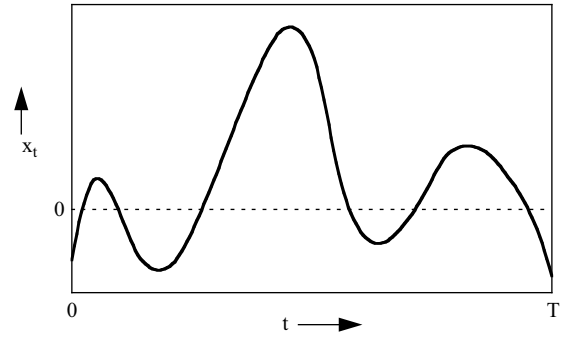



**Figure 1.3** An example of a finite analog data record.

in $x_n$. The sampling interval, $\Delta t$, associated with each datum can be shown on a time series plot to the left or right of, or centered on, each datum – it is a matter of choice. In Figure 1.4, $\Delta t$ is to the right of each datum. One might think that there should be a fifteenth sample point at the very end of the curve in Figure 1.3. However, because of the association of each sampled value with one $\Delta t$, the length of the digital record would be one sample interval longer than the analog record. Conceptually, the fifteenth sample point is the first value of a continuing, but unavailable, analog record.

The concept of an *infinite analog* record is often used in theoretical work. An example would be the trace in Figure 1.3 extended indefinitely in both directions of time. For this case a continuum of harmonics is required to fit the signal, thereby resulting in a *variance density spectrum*. Note, however, that a variance density spectrum can be created also with a finite digital record. How this comes about is

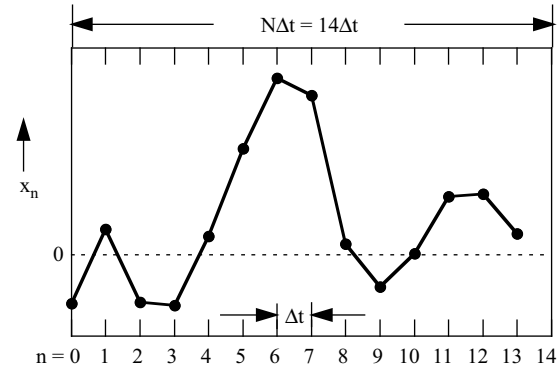



**Figure 1.4** An example of a finite digital data record obtained by sampling the finite analog record in Figure 1.3. There are N = 14 data.

discussed in Chapter 5. An *infinite digital* record would be obtained by sampling the infinite analog record at intervals of $\Delta t$. We will use infinite analog and digital records in Section 3.1.4 (Chapter 3) to determine the effects on the mean value of a time series after it is filtered.

By far the type of record most commonly observed and analyzed in science and technology is the finite digital record. With a few exceptions, this is the type of data record we will deal with in the remainder of Chapter 1, and for which the formulas for computing a periodogram are presented.

## 1.2 Analysis and synthesis

### 1.2.1 Formulas

If one of the data sets collected in your research is a time series of atmospheric pressure, Fourier "analysis" can be used to derive its periodogram and to examine which harmonics dominate the series. Conversely, once the analysis has been done, the original time series of pressure can be reconstructed purely from knowledge of the harmonic amplitudes. Thus Fourier "synthesis" is the inverse process of analysis. Note that the title of this chapter employs the more generic meaning of analysis and includes both the analysis and synthesis terms just described.

The formulas in Table 1.1 are those needed to perform analysis and synthesis. The equations under Fourier Analysis are used to calculate the Fourier coefficients or harmonic amplitudes. The equations under Fourier Synthesis express the time series $x_n$ as the sum of products of cosines and sines with amplitudes $A_m$ and $B_m$, respectively, or, alternatively, the sum of products of cosines only with amplitudes $R_m$ and phase angles $\theta_m$. Notice that the expressions are slightly different depending on whether the time series has an even or an odd number of data. The synthesis equations are equivalent to the forms introduced by Shuster around 1900 (Robinson, 1982).

The arguments of the cosine and sine terms associated with the $A_m$ and $B_m$ coefficients are of the form

$$\frac{2\pi mn\Delta t}{N\Delta t}$$

where m is harmonic number and $n\Delta t$ a point in time along the time axis of total length $N\Delta t$. Thus, $2\pi m$ is the number of radians in the m-th harmonic over the total length of the time series. The product of $2\pi m$ and the ratio $n\Delta t/N\Delta t$ provide location along the sinusoid in radians. Because the time increments ($\Delta t$) cancel, they are not shown in Table 1.1. In Fourier synthesis, the summation is over all harmonics at a given location $n\Delta t$, while in Fourier analysis the summation is over all data locations for a given harmonic m.

**Table 1.1** Formulas used in Fourier synthesis and analysis for an even or odd number of data.

**Fourier Analysis**

$$A_0 = \frac{1}{N}\sum_{n=0}^{N-1} x_n \qquad B_0 = 0$$

$$A_m = \frac{2}{N}\sum_{n=0}^{N-1} x_n \cos\frac{2\pi mn}{N} \qquad B_m = \frac{2}{N}\sum_{n=0}^{N-1} x_n \sin\frac{2\pi mn}{N}$$

$$m = \left[1, \frac{N}{2}-1\right] (N \text{ even}); \quad m = \left[1, \frac{N-1}{2}\right] (N \text{ odd})$$

$$A_{N/2} = \frac{1}{N}\sum_{n=0}^{N-1} x_n \cos(\pi n) \qquad B_{N/2} = 0 \quad (N \text{ even})$$

$$R_m = \sqrt{A_m^2 + B_m^2} \qquad \theta_m = \tan^{-1}\left(\frac{B_m}{A_m}\right)$$

**Fourier Synthesis**

$$x_n = \sum_{m=0}^{N/2}\left(A_m \cos\frac{2\pi mn}{N} + B_m \sin\frac{2\pi mn}{N}\right) = \sum_{m=0}^{N/2} R_m \cos\left(\frac{2\pi mn}{N} - \theta_m\right), \quad n = [0, N-1] \ (N \text{ even})$$

$$x_n = \sum_{m=0}^{\frac{N-1}{2}}\left(A_m \cos\frac{2\pi mn}{N} + B_m \sin\frac{2\pi mn}{N}\right) = \sum_{m=0}^{\frac{N-1}{2}} R_m \cos\left(\frac{2\pi mn}{N} - \theta_m\right), \quad n = [0, N-1] \ (N \text{ odd})$$

**Variance at Harmonic m**

$$S_m^2 = \frac{A_m^2 + B_m^2}{2} \qquad m = \left[1, \frac{N}{2}-1\right] \ (N \text{ even}); \quad m = \left[1, \frac{N-1}{2}\right] \ (N \text{ odd})$$

$$S_{N/2}^2 = A_{N/2}^2 \quad (N \text{ even})$$

**Total Variance**

$$S^2 = \sum_{m=1}^{N/2} S_m^2 \quad (N \text{ even}) \qquad S^2 = \sum_{m=1}^{\frac{N-1}{2}} S_m^2 \quad (N \text{ odd})$$

The variance at each harmonic for even and odd data lengths is given in Table 1.1 under the heading Variance at Harmonic m. Note that the only exception to the general formula for harmonic variance occurs at $m = N/2$ when N is even. The cosine coefficient at N/2 is squared but not divided by two (the sine coefficient is zero). The formulas for the total variance $S^2$ under the heading Total Variance yield the same variance estimates as the formula

$$S^2 = \frac{1}{N}\sum_{n=0}^{N-1} (x_n - \bar{x})^2 \tag{1.1}$$

for computing total variance directly from the data, in which $\bar{x}$ is the series mean. The two formulas in Table 1.1 are nearly the same, the only difference being that the

expression for the upper limit of each summation depends on whether N is even or odd.

### 1.2.2 Fourier coefficients

The method for obtaining the Fourier coefficients is based on the *orthogonality* of cosine and sine functions at harmonic frequencies, where orthogonality means that the sum of the products of two functions over some interval equals zero. The method entails multiplying both sides of a Fourier synthesis equation by one of the cosine or sine harmonic terms, summing over all n, and solving for the coefficient associated with the harmonic term.

For example, consider multiplying both sides of the first Fourier synthesis equation in Table 1.1 (using the $A_m$, $B_m$ form) by $\cos\frac{2\pi kn}{N}$ and summing over all n. The second summation on the right-hand side will have the form and result

$$\sum_{n=0}^{N-1} \sin\frac{2\pi mn}{N} \cos\frac{2\pi kn}{N} = 0 \tag{1.2}$$

where m and k are integers. That this sum is zero can be shown with two examples as well as mathematically. The sine and cosine terms for $m = k = 1$ are shown in Figure 1.5 and for $m = 1$ and $k = 2$ in Figure 1.6. The algebraic signs of the sum of cross products within each quadrant are shown at the base of each figure. Because of symmetry, the absolute magnitude of each sum is the same for each quadrant in

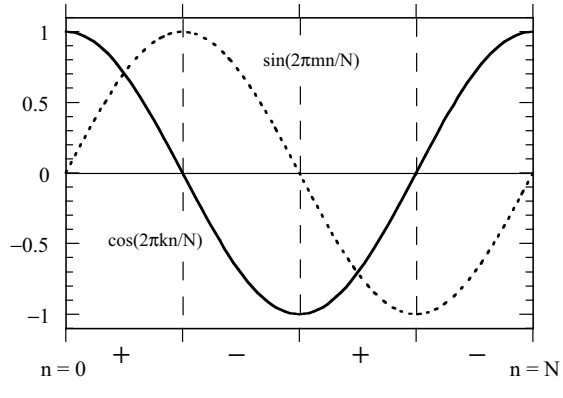



**Figure 1.5** Signs of sums of cross products of cosine and sine terms for $m = k = 1$.

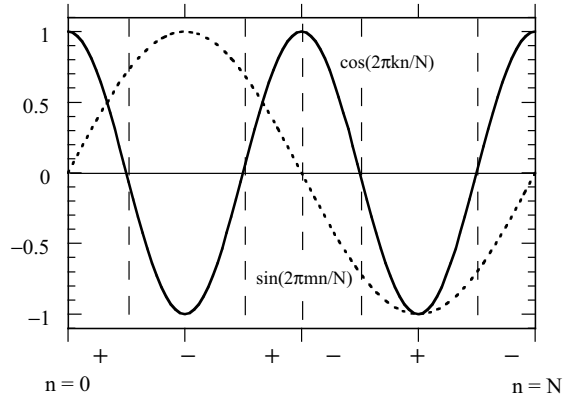



**Figure 1.6** Signs of sums of cross products of cosine and sine terms for m = 1 and k = 2.

Figure 1.5 and similarly for Figure 1.6. Thus the waveforms are orthogonal because the sum of their cross products is zero over the interval 0 to N in each illustration.

It can be surmised from these figures that the sum of the cross products is zero over the fundamental period for any combination of the m and k integers. But how could this be shown mathematically? Firstly, we put the sine and cosine terms in complex exponential form, and then expand the summation above using Euler's formula to obtain

$$\begin{aligned}
&\sum_{n=0}^{N-1} \sin(2\pi mn/N)\cos(2\pi kn/N) \\
&\quad = \sum_{n=0}^{N-1} \frac{1}{2i}\left(e^{i2\pi mn/N} - e^{-i2\pi mn/N}\right)\frac{1}{2}\left(e^{i2\pi kn/N} + e^{-2\pi kn/N}\right) \\
&\quad = \frac{1}{4i}\sum_{n=0}^{N-1}\left(e^{i2\pi(m+k)n/N} + e^{i2\pi(m-k)n/N} - e^{-i2\pi(m-k)n/N} - e^{-i2\pi(m+k)n/N}\right). \qquad (1.3)
\end{aligned}$$

A procedure is developed in Appendix 1.B for finding the sum of complex exponentials. The final two formulas, Equations 1.B.3 and 1.B.4, are very useful for quickly finding the sums of sines and cosines over any range of their arguments. An example of using the first formula follows.

Consider just the first summation on the right-hand side in Equation 1.3. Let

$$Q = \sum_{n=0}^{N-1} e^{i2\pi(m+k)n/N}. \qquad (1.4)$$

Using Equation 1.B.3, Q becomes

$$\begin{aligned} Q &= \frac{1 - e^{i2\pi(m+k)}}{1 - e^{i2\pi(m+k)/N}} \\ &= \frac{1 - \cos[2\pi(m+k)] - i\sin[2\pi(m+k)]}{1 - \cos[2\pi(m+k)/N] - i\sin[2\pi(m+k)/N]} \\ &= 0, \quad m + k \neq 0,\ N. \end{aligned} \tag{1.5}$$

The numerator is zero for all integer values of m and k while the denominator is nonzero except when $(m+k) = 0$ or N, in which cases the denominator is 0 and Equation 1.5 is indeterminate. To evaluate Equation 1.5 for these cases we can apply l'Hopital's rule. The result of taking the first derivative with respect to $(m+k)$ in both the numerator and denominator yields a determinate form with value N. That is

$$\begin{aligned} Q' &= \frac{2\pi\sin[2\pi(m+k)] - i\,2\pi\cos[2\pi(m+k)]}{(2\pi/N)\sin[2\pi(m+k)/N] - i\,(2\pi/N)\cos[2\pi(m+k)/N]} \\ &= N, \quad m + k = 0,\ N. \end{aligned} \tag{1.6}$$

The same result also can be obtained by substituting 0 or N for (m+k) in Equation 1.4. We observe that the first and fourth summations in Equation 1.3 cancel each other for these values.

We can apply the above procedure to the second term in Equation 1.3. The summation will be zero again, except when $(m-k)$ is 0 or N. Employing l'Hopital's rule yields a determinate form with value N for these cases, similar to Equation 1.6. And again, the same results can be obtained from Equation 1.4. Accordingly, when $(m-k) = 0$ or N, the second and third summations in Equation 1.3 cancel. Thus Equation 1.2 is valid for any integer k or m. This includes the possibility that $(k+m)$ is an integer multiple of N.

Now that we have shown that the summed sine–cosine cross product terms akin to Equation 1.2 must be zero, let us consider the sums of sine–sine and cosine–cosine products resulting from multiplying the first Fourier synthesis equation by $\cos\frac{2\pi kn}{N}$ and summing over all n. Following the procedure in Appendix 1.B we find that

$$\sum_{n=0}^{N-1} \sin\frac{2\pi mn}{N} \sin\frac{2\pi kn}{N} = \begin{cases} 0,\ k \neq m \\ \frac{N}{2},\ k = m \neq 0,\ \frac{N}{2}\,(\text{N even});\ k = m \neq 0\,(\text{N odd}) \\ 0,\ k = m = 0,\ \frac{N}{2}\,(\text{N even});\ k = m = 0\,(\text{N odd}) \end{cases} \tag{1.7}$$

and

$$\sum_{n=0}^{N-1}\cos\frac{2\pi mn}{N}\cos\frac{2\pi kn}{N}=\begin{cases}0, & k\neq m\\ \frac{N}{2}, & k=m\neq 0,\ \frac{N}{2}\,(N\text{even});\ k=m\neq 0\,(N\text{odd})\\ N, & k=m=0,\ \frac{N}{2}\,(N\text{even});\ k=m=0\,(N\text{odd}).\end{cases}\tag{1.8}$$

Thus multiplying the synthesis equation for N even by the k-th sine harmonic term and summing yields

$$\sum_{n=0}^{N-1}x_n\sin\frac{2\pi kn}{N}=\sum_{m=0}^{N/2}\left(A_m\sum_{n=0}^{N-1}\sin\frac{2\pi kn}{N}\cos\frac{2\pi mn}{N}+B_m\sum_{n=0}^{N-1}\sin\frac{2\pi kn}{N}\sin\frac{2\pi mn}{N}\right)\tag{1.9}$$

which reduces to

$$\sum_{n=0}^{N-1}x_n\sin\frac{2\pi kn}{N}=B_k N/2,\quad k=\left[1,\frac{N}{2}-1\right]\tag{1.10}$$

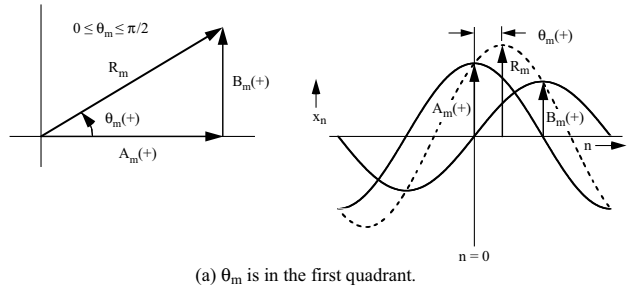
so that

$$B_k=(2/N)\sum_{n=0}^{N-1}x_n\sin\frac{2\pi kn}{N},\quad k=\left[1,\frac{N}{2}-1\right].\tag{1.11}$$

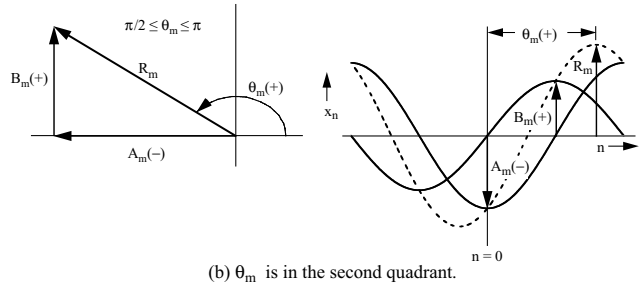
Observe that the sine coefficients for $k=0$, N/2 (N even) are always zero.

The Fourier cosine coefficients, $A_k$, are obtained in a similar manner, but $A_0$ and $A_{N/2}$ are, in general, nonzero. As is evident from Table 1.1, $A_0$ is the mean of the time series. For N odd, an expression similar to Equation 1.9 is used to obtain the Fourier coefficients, the only difference being that the range of harmonics extends from 0 to $(N-1)/2$. Table 1.1 shows the resulting formulas for all Fourier coefficients.

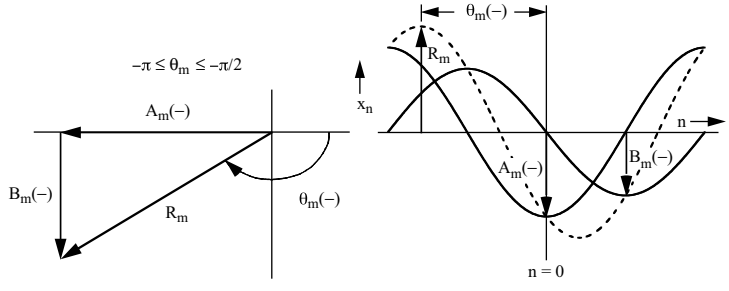
The coefficients $A_m$ and $B_m$ represent the amplitudes of the cosine and sine components, respectively. As shown in the left-hand panel of Figure 1.7a, the cosine coefficient is always along the horizontal axis (positive to the right), and the sine coefficient is always normal to the cosine coefficient (positive upward). In the right-hand panel we see how the cosine and sine vector lengths determine the associated cosine and sine waveforms (ignore the dashed line for the moment). Figures 1.7b–d show various possibilities of waveform relationships depending on the sign of $A_m$ and the sign of $B_m$. More discussion of Figure 1.7 is given in Section 1.2.4.



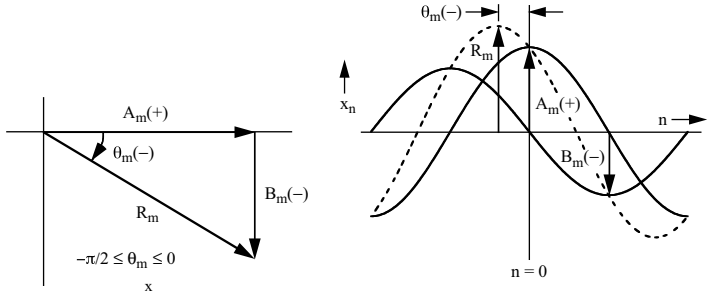

(a) $\theta_m$ is in the first quadrant.




(b) $\theta_m$ is in the second quadrant.




(c) $\theta_m$ is in the third quadrant.




(d) $\theta_m$ is in the fourth quadrant.

**Figure 1.7** (a)–(d) The magnitude and sign of each Fourier coefficient determines the quadrant in which the phase angle lies. Geometric vector lengths in the left hand panels are twice the lengths of the Fourier coefficients in the right hand panels.

An alternative approach can be taken to solve for the Fourier coefficients. As shown by Bloomfield (1976, p. 13), the As and Bs above are identical to the coefficients from a least-squares fit of individual harmonics to the data.

### 1.2.3 Total and harmonic variances

The standard formula for the total variance of a time series of length N

$$S^2 = \frac{1}{N}\sum_{n=0}^{N-1}(x_n - \overline{x})^2 \tag{1.1}$$

was given in section in Section 1.2.1. The total variance is identical to the sum of the variances at the individual harmonics as shown in Table 1.1 for N even and N odd. The variance at an individual harmonic can be derived from Equation 1.1 by first substituting the Fourier synthesis equations for N even or N odd in Table 1.1 into Equation 1.1 for $x_n$ and $\overline{x}$. The substitution for $\overline{x}$ is $A_0$. After expanding the synthesis equation inside the parentheses in Equation 1.1, squaring the result, and performing the required summation, the cross product terms vanish (see Equation 1.2) and, using Equations 1.7 and 1.8, the remaining squared terms will reduce to the equations for variance at any harmonic seen in Table 1.1. With one exception, a harmonic variance is the sum of the squares of the Fourier cosine and sine coefficients divided by two. The exception occurs at harmonic $m = N/2$. The expansion of Equation 1.1 into the sum of harmonic variances is a good exercise in the application of orthogonality in time series analysis.

### 1.2.4 Amplitude and phase representation

Instead of representing a time series by the appropriate sums of ***both*** sines and cosines, an alternative representation is to use ***either*** sines or cosines alone and include phase angles, as seen in the right-hand equations in Table 1.1 under Fourier Synthesis. Because of orthogonality, the cosine term is shifted by 90° or $\pi/2$ radians from the sine term for any harmonic. As a result, a single sinusoid can be represented by two amplitude coefficients ($A_m$ and $B_m$) or, equivalently, by a single amplitude coefficient $R_m$ and a phase angle $\theta_m$. The advantages of the latter are a slightly more compact representation of $x_n$ and only one waveform for each harmonic.

Figure 1.7a illustrates the connection between the two forms of Fourier synthesis. The dashed sinusoid with amplitude $R_m$ in the right-hand panel has been decomposed into a cosine term and a sine term. Their respective amplitudes $A_m$ and $B_m$ depend on the location of the dashed sinusoid relative to the origin $n = 0$, that is, its

phase angle $\theta_m$. As noted earlier, the left-hand panel shows the vector relationship among the three amplitudes and the phase angle. Thus

$$R_m = \sqrt{A_m^2 + B_m^2}$$

$$A_m = R_m \cos\theta_m$$

and

$$B_m = R_m \sin\theta_m$$

so that

$$\theta_m = \tan^{-1}(B_m/A_m), \quad -\pi \leq \theta_m \leq \pi.$$

Substituting the middle two equations above into a cosine–sine synthesis results in the amplitude phase synthesis. We see that phase angle $\theta_m$ is determined by the sign and magnitude of $A_m$ and $B_m$. The sign of each coefficient, not merely the sign of the ratio, determines the quadrant in which the phase angle lies. The left-hand panels in Figures 1.7a–d show the amplitude and phase angle in the quadrant associated with the right-hand panels. We observe in each right-hand panel that, given the dashed line and the origin, we can immediately determine the magnitudes of the cosine and sine coefficients: the cosine coefficient is available at the origin and the sine coefficient 90° to the right.

## 1.3 Example data sets

### 1.3.1 Terrain heights

Table 1.2 contains the data set for this example. The formulas in Table 1.1 are used to perform a Fourier analysis and synthesis. Consider h to be the variation of terrain height above some datum with distance d along a specified direction. Furthermore, let the data in the table represent a finite digital subset of analog periodic data. The data are plotted in Figure 1.8 and connected by straight-line segments. After looking at Figure 1.8 and the tabled data, it should become clear that the waveform repeats itself every 3000 m. Thus one may as well work with 15 values ($n = 0, \ldots, 14$). Or should one use 16 values? Let us determine the difference. Since $\Delta d = 200$ m and the length of the fundamental period $L = 3000$ m, $N = 15$. Every datum must have a space increment $\Delta d$ associated with it. Although 16 points subtend L, the $\Delta d$ associated with the sixteenth point would make the fundamental period 3200, which it clearly is not. In short, the sixteenth point is the first point of the next period and similarly for the thirty-first point in the table.

**Table 1.2** Height (h) versus distance ($d = n\Delta d = n200$ m).

| n | nΔd(m) | h(m) | n | nΔd(m) | h(m) | n | nΔd(m) | h(m) |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 100.00 | 10 | 2000 | 98.25 | 20 | 4000 | 101.75 |
| 1 | 200 | 129.02 | 11 | 2200 | 110.73 | 21 | 4200 | 117.30 |
| 2 | 400 | 127.14 | 12 | 2400 | 97.55 | 22 | 4400 | 110.59 |
| 3 | 600 | 102.45 | 13 | 2600 | 72.86 | 23 | 4600 | 89.41 |
| 4 | 800 | 89.27 | 14 | 2800 | 70.98 | 24 | 4800 | 82.70 |
| 5 | 1000 | 101.75 | 15 | 3000 | 100.00 | 25 | 5000 | 98.25 |
| 6 | 1200 | 117.30 | 16 | 3200 | 129.02 | 26 | 5200 | 110.73 |
| 7 | 1400 | 110.59 | 17 | 3400 | 127.14 | 27 | 5400 | 97.55 |
| 8 | 1600 | 89.41 | 18 | 3600 | 102.45 | 28 | 5600 | 72.86 |
| 9 | 1800 | 82.70 | 19 | 3800 | 89.27 | 29 | 5800 | 70.98 |
| | | | | | | 30 | 6000 | 100.00 |

Further examination of the first 3000 m in Figure 1.8 suggests odd symmetry in the data. That is, if a vertical line were drawn at 1500 m, the heights of any two points equidistant from this line will appear to be a reflection about a horizontal line at 100 m elevation. Consequently, except for the mean, only sine terms will be needed in the Fourier synthesis. Lastly, we notice that the time series exhibits only comparatively slow fluctuations, so that most of its variance should be "explained" (i.e., accounted for) by low harmonic frequencies.

Based on this insight, we first compute the mean and find that $A_0 = 100$ m. Over the first 3000 meters we can easily identify three peaks and three troughs, indicating we should calculate the magnitude of harmonic 3, that is

$$B_3 = (2/15)\sum_{n=0}^{N-1} h_n \sin(2\pi 3n/15) = 20\,\text{m}, \quad N = 15.$$

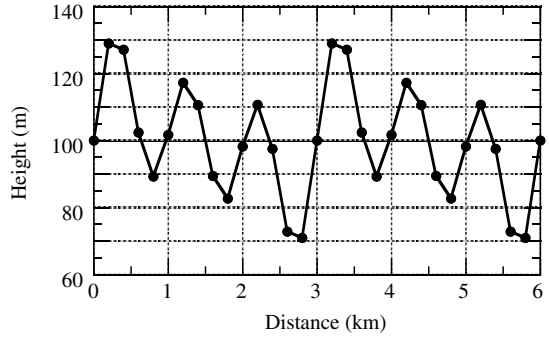



**Figure 1.8** Plot of terrain height data connected by straight-line segments.

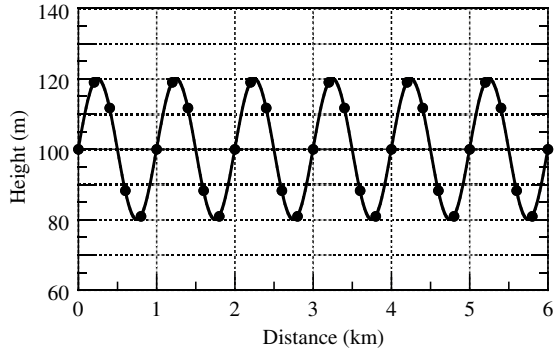



**Figure 1.9** Plot of the mean plus harmonic 3 fitted to the data in Figure 1.8 from 0 to 3 km and repeated from 3 to 6 km.

Thus, its variance is $S_3^2 = B_3^2/2 = 200\ m^2$. The third harmonic added to the mean is plotted in Figure 1.9 from 0 to 3000 m and then repeated to include the entire length of the data set.

Since the average value of the height departures from $A_0$ for the first 1500 m is positive and for the second 1500 m is the same magnitude but negative in sign, harmonic 1 should be nonzero. This is illustrated in Figure 1.10, in which it can be seen that harmonic 1 will have to be a sine wave to account for heights above the mean from 0 to 1500 m and heights below the mean from 1500 m to 3000 m.

Using the formula for $B_1$ we get $B_1 = 10$, so that $S_1^2 = 50\ m^2$. The first harmonic added to the mean is plotted in Figure 1.11, again over the entire time series. The accumulated variance from harmonics 1 and 3 is $250\ m^2$ in comparison to the total variance of $282\ m^2$ computed from Equation 1.1. As there is no apparent high frequency variance in Figure 1.8, we would expect much of the remaining variance to be at a low harmonic frequency. If we try harmonic 2 we find that $B_2 = 8$ m and $S_2^2 = 32\ m^2$. The waveform is shown in Figure 1.12. Since the first three harmonics

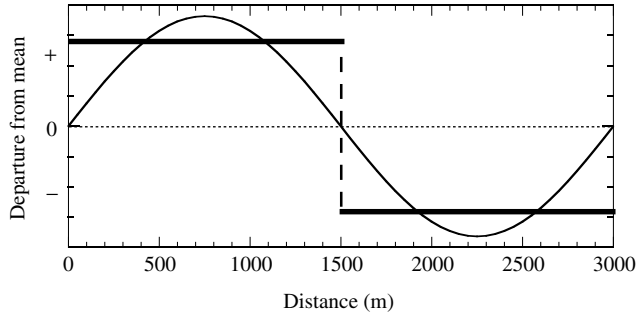



**Figure 1.10** Harmonic 1 results from above average and below average heights as shown.

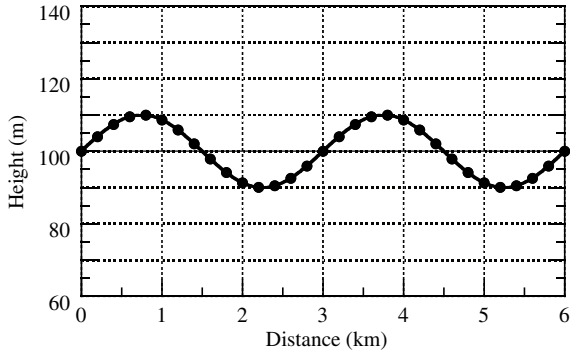



**Figure 1.11** Same as Figure 1.9 except for harmonic 1.

account for 100% of the variance in the data, there is no need to make further calculations (obviously, the data were generated using just the above coefficients!). Figure 1.13 shows the sum of the three harmonics plus the mean, drawn as a smooth curve that passes through all the observations in Figure 1.8.

It is interesting to consider what would happen if a 16-point data length (3200 m) were used, an earlier consideration. Instead of computing the three sine coefficients above to explain 100% of the variance, it would take eight cosine and eight sine nonzero coefficients to account for all the variance. The addition of the one point destroys the symmetry present in the 15-point data length (3000 m).

In Figures 1.9–1.13 the waveforms from 0 to 3000 m were repeated over the interval 3000–6000 m. This is allowed since the data are periodic. By extending the Fourier synthesis using the 15-point record, namely

$$h_n = 100 + 10\sin(2\pi n/15) + 8\sin(4\pi n/15) + 20\sin(6\pi n/15)$$

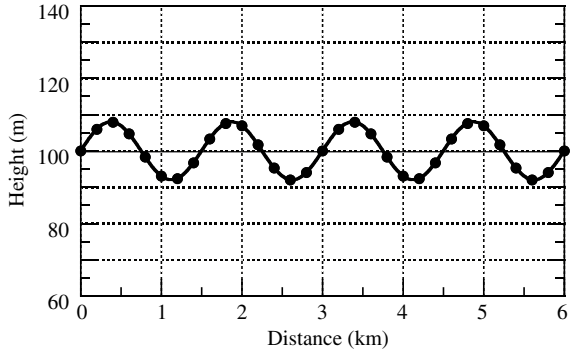



**Figure 1.12** Same as Figure 1.9 except for harmonic 2.

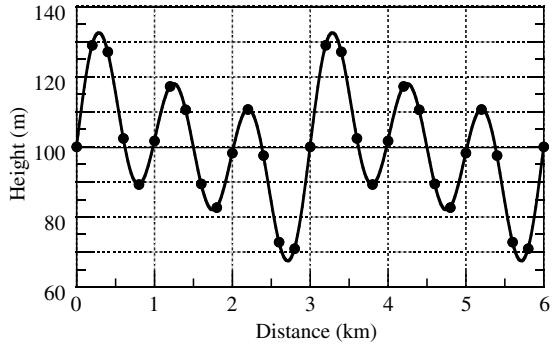



**Figure 1.13** Plot of the mean plus harmonics 1, 2, and 3 fitted to the data in Figure 1.8 from 0 to 3 km and repeated from 3 to 6 km.

the series will exactly match that in Figure 1.8 for $15 \leq n \leq 29$, and, more generally, repeat the observed data for any n replaced by $n \pm 15k$ where k is an integer. With a 16-point record, values of the computed series will be repeated every $\pm 16k$ points but will not match the observed record for $k \neq 0$. For example, the first value in the extended series will be 100 m compared to 129.02 m in the observed series.

Clearly, it is important to determine the correct number of data points when working with periodic data. Many observed variables in meteorology and other physical sciences are externally forced by the sun, so that there are strong diurnal and annual components in the data. These components serve to define the fundamental period.

### 1.3.2 Paradrop days

Table 1.3 shows the mean number of days in January that "paradrop" criteria are met for each hour of the day at Seymour-Johnson Air Force Base, Goldsboro, North Carolina. A paradrop is the insertion of troops or equipment into a site via parachute from an airplane. For a safe paradrop, three meteorological criteria should prevail: ceiling $\geq 2000$ feet (610 m), horizontal visibility $\geq 3$ miles (4.8 km) and surface winds $< 10$ knots (5.1 m/s). Although "ceiling" has a specific definition, it can be taken here to mean there is good vertical visibility between the surface and the height of the ceiling. As an example of paradrop days, from Table 1.3 we see that at 0700, 19.2 days of the 31 days in January meet the safety criteria, on average.

The results of performing a Fourier analysis of the data given in Table 1.3 are shown in Table 1.4. Only the results for the three largest harmonics are presented as they account for 97.8% of the total variance and the remaining variances are all small. Figure 1.14 is a plot of the mean number of days of occurrence versus time, the three largest harmonics (about the mean), and their sum. As expected, their sum provides a good fit to the data.

**Table 1.3** January paradrop days at Seymour-Johnson Air Force Base, North Carolina.

| Hour | Mean number of days | Hour | Mean number of days |
|---|---|---|---|
| 0000 | 21.6 | 1200 | 15.6 |
| 0100 | 21.1 | 1300 | 15.7 |
| 0200 | 21.2 | 1400 | 16.2 |
| 0300 | 20.8 | 1500 | 16.5 |
| 0400 | 20.3 | 1600 | 18.3 |
| 0500 | 20.4 | 1700 | 20.5 |
| 0600 | 20.0 | 1800 | 23.0 |
| 0700 | 19.2 | 1900 | 23.1 |
| 0800 | 19.5 | 2000 | 23.4 |
| 0900 | 18.0 | 2100 | 22.4 |
| 1000 | 17.4 | 2200 | 22.4 |
| 1100 | 17.5 | 2300 | 21.5 |

We can also establish the time or times of the peaks in each harmonic. To do this, we use the formula

$$H = F \times (\text{phase angle } \theta \text{ in degrees})$$

where H is the time in hours after 0000 local time and F is the ratio of the harmonic period to 360°. Thus, from Table 1.4 for the second harmonic

$$H = (12/360) \times (-143.3) = -4.78\,\text{h}.$$

Converting this value to local time, we get 0713 and 1913. Similarly, the peaks for the third harmonic are at 0223, 1023, and 1823 and for harmonic one at 2305.

In Problem 6 at the end of this chapter you are asked to write a Fourier analysis computer program, apply it to the data in Table 1.3, compare your results with those in Table 1.4 and Figure 1.14, and try to ascribe physical meaning to the main harmonics.

**Table 1.4** Results of Fourier analysis of the data in Table 1.3.

| Harmonic | Cosine coefficient | Sine coefficient | Variance | Percentage of total variance | Phase angle in degrees |
|---|---|---|---|---|---|
| 0 | 19.817 | 0 | 0 | 0 | |
| 1 | 2.8188 | −0.6848 | 4.2072 | 74.9 | −13.7 |
| 2 | −1.1965 | −0.8919 | 1.1136 | 19.8 | −143.3 |
| 3 | −0.1743 | 0.5623 | 0.1733 | 3.1 | 107.2 |

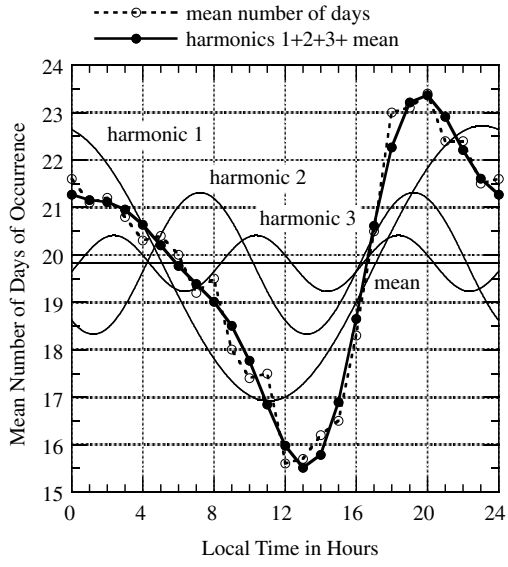



**Figure 1.14** Mean number of days that meet paradrop criteria versus time of day at Seymour-Johnson Air Force Base, Goldsboro, North Carolina for the month of January. Three harmonics explain 98% of the variance in the data.

### 1.3.3 Hourly temperatures

Even if there is a substantial amount of variance at a number of harmonics, one should not believe, in general, that the variance at each harmonic is the consequence of a different physical cause. Instead, often a band of harmonics can be related to a physical phenomenon. Figure 1.15 is an example of a spectrum that shows variance at particular harmonics and at broad bands of frequencies. The data from which the spectrum was computed are hourly temperatures taken at the Norman, Oklahoma Mesonet site (McPherson *et al.* 2007; http://www.mesonet.org) from 1 December 2006 through 31 March 2007 at a height of 1.5 m. Each hourly temperature is a five-minute average at the top of the hour. The logarithmic x-axis is in frequency in cycles/h converted from harmonic number and the y-axis is proportional to the product of variance and frequency. In contrast to the periodogram with line variance in Figure 1.2, the spectrum amplitudes here are connected by straight-line segments, the usual method of presentation. There are two broad frequency bands of interest. One contains periods from about 12 to around 30 days (0.0035–0.0014 cycles/h) and the other from about four to eight days (0.0104–0.0052 cycles/h). The variances in these two bands are due to the passage of long waves in the westerlies (major ridge–trough systems, i.e., Rossby-type waves) and short waves (minor ridge–trough systems and fronts), respectively. Their largely aperiodic nature results in the distribution of variance across a band of

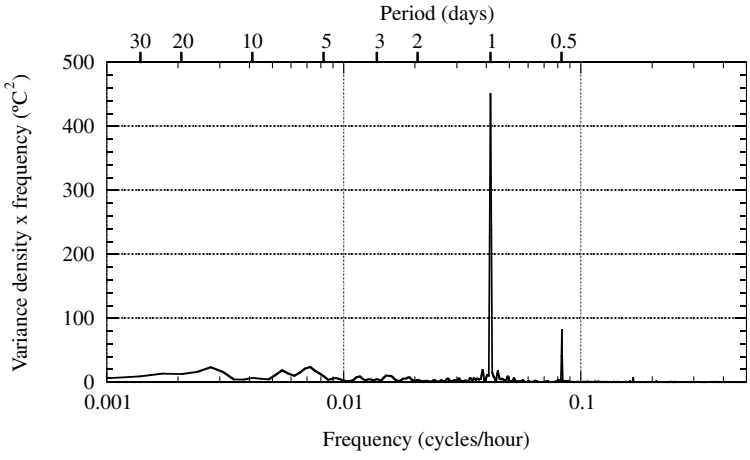



**Figure 1.15** Temperature spectrum at 1.5 m height for December 2006–March 2007 at Norman, Oklahoma.

frequencies, the width of which can vary from year to year. Variance at periods longer than 30 days is not shown because there are too few cycles over the 121 days of data to yield satisfactory estimates.

The two well-defined peaks at periods of 24 and 12 hours are due to the daily cycle of solar heating. Similar to the paradrop data in the previous section, the diurnal temperature variation is a deformed sinusoid such that a semi-diurnal component is also required. In fact, close inspection of Figure 1.15 shows a small amount of variance at a period of six hours (0.1667 cycles/h), the fourth harmonic of the daily cycle of temperature. Thus all three variances are required to explain the variance in the daily cycle.

As a final comment about Figure 1.15, we point out that when the product of variance density and frequency is plotted against the logarithm of frequency, the result is an equal-area representation. Thus this is the plot design to use when the goal is to compare variances from different frequency bands. Although we mentioned in Section 1.1.3 that variance density would be discussed in Chapter 5, here it is only necessary to know that variance and variance density are directly proportional to each other to understand Figure 1.15.

### 1.3.4 Periodogram of a rectangular signal

Fourier analysis necessarily fits sinusoids to a time series; thus it is interesting to observe what happens when data are intrinsically not comprised of sinusoids. The heavy solid line in Figure 1.16a shows a periodic rectangular signal that might represent, for example, whether it is raining or not or the occurrence and non-

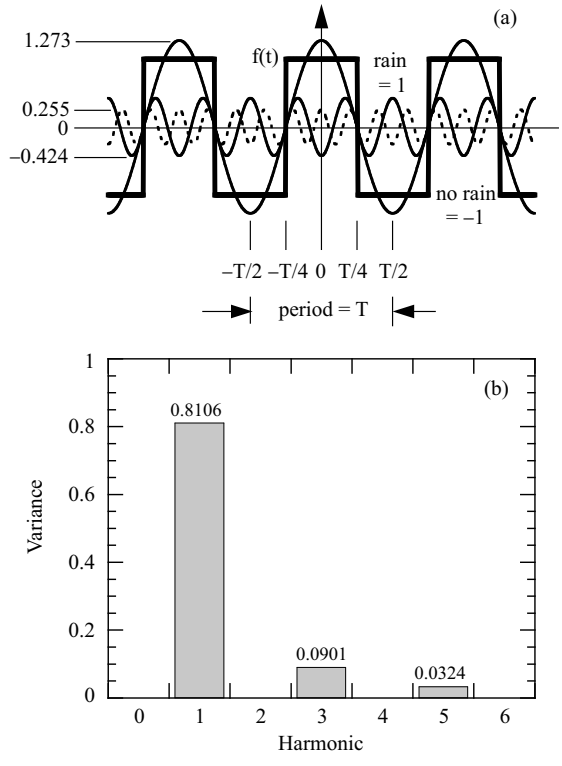



**Figure 1.16** (a) A hypothetical rain – no rain analog signal, f(t), showing the first three nonzero cosine harmonics. (b) Resulting periodogram with fundamental period shown in (a).

occurrence of some periodic phenomenon. Because of the location of the time origin, the signal is an even function and, therefore, only Fourier cosine coefficients will be required. By analogy with Fourier analysis for digital data, the expression for the Fourier cosine coefficients for this periodic analog record is

$$A_m = \frac{2}{T}\int_{-T/2}^{T/2} x(t)\cos(2\pi mt/T)dt, \quad m = 0, 1, 2, \ldots$$

where T is the fundamental period. The periodogram in Figure 1.16b shows that only odd-numbered harmonics are needed to synthesize the signal. The two light solid lines and dashed line in Figure 1.16a are the waveforms of the first, third, and fifth harmonics.

Given the waveform of the first harmonic, the waveforms of the third and fifth harmonics serve to improve the Fourier synthesis. The positive and negative overshoots of the rectangular signal by the first harmonic are compensated by the addition of the associated negative and positive peaks, respectively, in the third

harmonic. However, as can be seen in Figure 1.16a, the addition results in overcompensation that, in turn, is compensated by the fifth harmonic. In this way, adding the waveforms of successive odd harmonics better approximates the flat peaks and troughs in the rectangular signal and sharpens the change in value from 1 to $-1$ and $-1$ to 1. In practice, if you computed a periodogram that showed evidence of decreasing variance at alternate harmonics, you should be wary of the presence of a rectangular wave. Other special signals (for example, triangular and saw-tooth) also show characteristic spectra.

In summary, we are reminded that Fourier analysis fits sinusoids to data regardless of the signal being generated by the physical (or mathematical) process. It is up to the analyst to keep this in mind when interpreting a periodogram.

## 1.4 Statistical properties of the periodogram

Section 1.4.1 provides basic statistical concepts and terminology needed to understand the remainder of Section 1.4. Section 1.4.2 discusses the term *expectation* and shows how it is used to find statistical properties, for example mean, variance, and covariance, of digital and analog data. Expectation is used in Appendix 1.C to derive the formulas for the distribution of variances at the Fourier harmonics. Section 1.4.3 deals with the main result of Appendix 1.C, namely, that the frequency distribution of variance at any harmonic is proportional to a chi-square variable. This conclusion requires that the data we analyze come from a normal white noise process. That is, the data have a normal distribution and the periodogram of the data shows no preference for large or small variance at any harmonic. In practice, we assume the data are at least approximately normally distributed and, if the data are not white noise, the data length is sufficiently long that the properties of the chi-square distribution of variance at a harmonic and their independence from one harmonic to the next are effectively met.

Knowing that the statistical distribution of variances at a harmonic is chi-square opens the window to finding confidence limits for the underlying variance spectrum from which a sample periodogram has been computed as well as testing the null hypothesis that the periodogram came from a white noise process. To put these ideas into practice, we will deal with two data sets: one a 100-year record of autumn temperatures; the other a five-year record of monthly mean temperature, both from central Oklahoma, USA. The theory and application of confidence limits are presented in Sections 1.4.4–1.4.6.

### 1.4.1 Concepts and terminology

The computation of a periodogram is purely an algebraic manipulation of the data. The interpretation of a periodogram depends on how one views the data. If one views a data set as resulting from a physical phenomenon or mathematical process that

produces an identical data set each time the phenomenon or process is initiated, the data set is considered *deterministic*. Each data set produced is identical to every other data set, and likewise for the associated periodograms. If one views a data set as resulting from a physical phenomenon or mathematical process that produces a different data set each time the phenomenon or process is initiated, however small the differences, the data set is *nondeterministic*, or equivalently, the data are *stochastic* or *random*. Because no two data sets are alike, there is, conceptually, a *population* of data sets, with each data set a member or *realization* of the population. The population can be finite or infinite. The periodogram of a nondeterministic data set is also one realization of a population of periodograms. In this concept, each realization, whether a data set or a periodogram, represents an equally valid statistical representation of the physical phenomenon or mathematical process being analyzed. Observed time series in natural science are typically nondeterministic, although deterministic components can exist in the series. In Section 1.4 we focus our attention on nondeterministic data sets and the statistical properties of the resulting periodograms. As part of this effort, we need certain additional terminology.

A *random variable* (rv) is a variable that has associated with it a range of values and either a *probability distribution* (pd) if the variable is digital, or a *probability density function* (pdf) if the variable is analog. For example, random variable (rv) X might take on any integer value from 151 to 250, inclusive. The probability distribution gives the probability of occurrence assigned to each of the 100 possible values, with the sum of the probabilities equal to one. Alternatively, rv X might take on any real value within the range 151 to 250, of which there are infinitely many possibilities. In this case the probability of X *exactly* assuming any particular value (say, exactly 200) is zero. However, there exists a finite probability that X will lie within a range of values (say, 199.99–200.01) that is a subset of the overall range of possibilities. Thus, in the case of analog data, it is necessary to describe probabilities using probability densities, the relative likelihoods of the values within the overall range. The probability density function describes these relative likelihoods and, in parallel with the case of digital data, the integral of the probability density function over the range of rv X is one.

Now we develop the concept of a time series of random variables. Imagine a time series of data from time $t_0$ to time $t_T$ collected from an experiment. Continue to repeat the experiment, thereby forming successive data sets (realizations) of x(t) from $t_0$ to $t_T$. The values of x at, say, time $t'$, where $t_0 \leq t' \leq t_T$ form random variable $X(t')$. This concept is illustrated in Figure 1.17, which shows a selection of realizations stacked one upon the other with $t_0$ and $t_T$ lying beyond the ends of the time axis shown. The intersection of the left-hand vertical dashed line with each realization provides the range of values that comprise rv X at time $t'$. A random variable also can exist at any other point along the time axis (for example, $X(t'')$ at $t''$). In the experiment above, the time axis was finite ($t_0$ to $t_T$). In general, both the time axis and the number of realizations can be infinite. Whether the time axis $X(t'')$

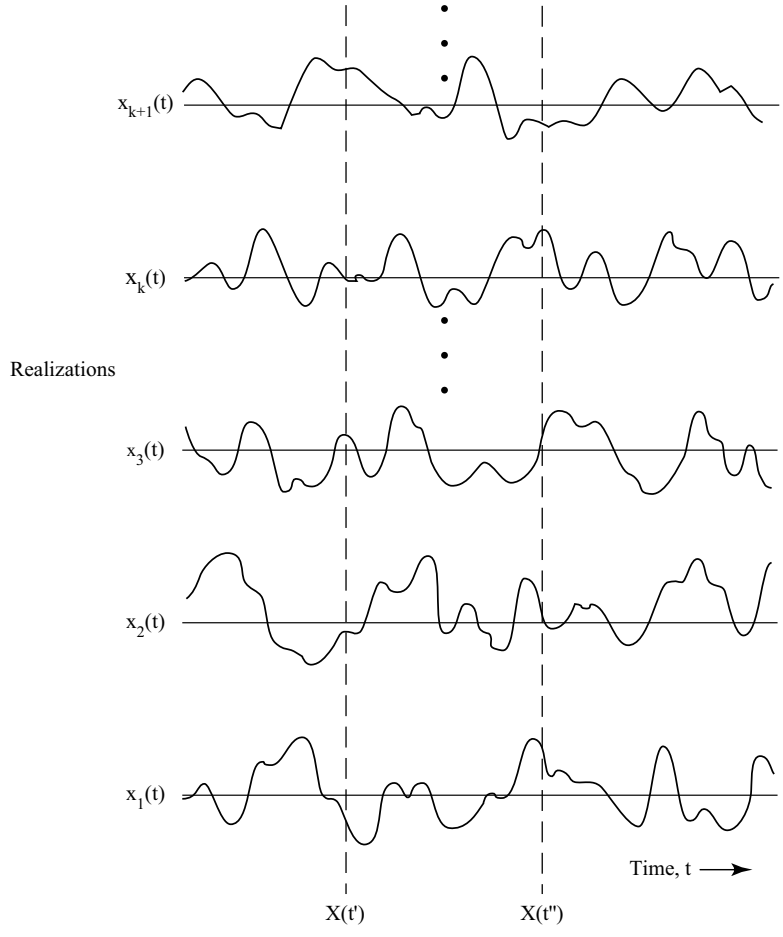



**Figure 1.17** A selection of realizations from a random process. Function X(t′) denotes random variable X at any time t′. The light horizontal lines have the same reference value of X for each realization.

and number of realizations is finite or infinite $t_0$ to $t_T$, the collection of random variables comprise a *random process* or *stochastic process*.

Before continuing with additional concepts we comment on the notation for random variables. Throughout the text we will use only upper case letters to indicate random variables. However, all upper case letters are not random variables; they can be standard mathematical variables, parameters, or constants. We've seen this, for example, with Fourier coefficients in Table 1.1. It is always easy to understand whether or not an upper case letter represents a random variable by the context in which it occurs.

When the data are nondeterministic, we need to consider another attribute: whether the random process is *stationary* or *nonstationary*. If a time series is stationary, the statistical properties of the pd or probability density function do not change with time; a time series is nonstationary if the opposite is true. A simple example of a nonstationary time series is a record of air temperature from winter to summer at a typical middle latitude station. The nonstationarity results from the increasing value of mean daily temperature, that is, a trend. In order to make the series stationary, a low-order polynomial can be fitted to the data set and then subtracted from it. Another common type of nonstationarity occurs when the magnitude of the fluctuations, the population variance, changes with time. An example is the greater variability (gustiness) of the surface wind speed during daytime than nighttime due to the vertical mixing of air as the surface is differentially heated during daylight hours.

If a time series is nonstationary and it is not clear how to remove nonstationary effects, it may be necessary to resort to special analyses using many realizations, divide the data into stationary segments, or apply other methods, such as wavelet analysis (Daubechies, 1992). With the exception of Section 4.1, all mathematical statistics in this text apply to stationary random processes; that is, the population mean and population variance are independent of time. The data sets we analyze can be considered realizations of a stationary random process or can be filtered in such a way to make them stationary or approximately so.

An additional underlying concept needed to derive the statistical properties of a periodogram is a particular random process called *Gaussian* (or *normal*) *white noise*. There are two attributes of this process. With reference to Figure 1.17, the first is that the probability density function of rv $X(t')$ (or $X(t'')$) is Gaussian. The second attribute is that rv $X(t')$ and rv $X(t'')$ are independent of each other. In practice, this means that knowledge of the value of one member of the population at time $t'$ provides no predictability of the value of the same or any other member at any other time. In statistical parlance, the *covariability* or *covariance* between $X(t')$ and $X(t'')$, $t' \neq t''$, is zero, a condition that implies the underlying random process is white noise. The equivalent mathematical statement is derived in the next section.

An examination of the periodogram of any selected time series (a realization) from a Gaussian white noise process would indicate no preference for large or small variances in any part of the spectrum. The average over all possible realizations of the variances at any one harmonic would be identical to that at any other harmonic (with the exception of the highest harmonic for an even number of data, where the variance would be reduced by one-half relative to the other harmonics). That is, the periodogram variances would be uniform with harmonic frequency (less the exception), a condition referred to as "white" by loose analogy with white light wherein no one of the component colors is preferred (there is also acoustic white noise). It is through this connection that the process that produces the uniform variance spectrum is referred to as "white." When we subsequently deal with a white

noise process, it will be assumed to be normal or Gaussian, so that use of either qualifier will be dropped.

## 1.4.2 Expectation

The term *expectation* is of fundamental importance to understanding the statistical properties of the periodogram. The synonym for expectation is average. When expectation is indicated, the average is taken over the entire population, whether it is finite or infinite. The indicator for expectation is the symbol or operator E. When the operator E is applied to a random variable (or a function of a random variable), the question being asked is, "What is the average value of the random variable (or the function of the random variable)." We will see examples of both in subsequent sections.

In the first subsection formulas for the expectation of rv X and general function g(X) for digital data are developed, followed by, in the second subsection, a parallel development for analog data. In the third subsection the expectation of the product of two random variables is developed. The results are formulas for the covariability or covariance between these variables. For those readers familiar with expectation, it may be necessary only to skim through this section to become familiar with the notation.

### *1.4.2.1 Digital data*

Let the sample space S in Figure 1.18 contain a population of N elements, some of which may be the same. Denote distinct elements of S by $x_1, x_2, \ldots, x_K$. In Figure 1.18 $N = 10$ and $K = 6$; two $x_1$ elements are alike, three $x_2$ elements are alike, and two $x_4$ elements are alike.

The symbol for expectation is E and the expectation of random variable X is defined by

$$E[X] = \mu_X = \frac{\sum_{k=1}^{K} x_k n_k}{N} \tag{1.12}$$

where X is a digital random variable, $n_k$ is the number of elements with value $x_k$, and

$$N = \sum_{k=1}^{K} n_k.$$

The expectation operator asks the question – What is the mean value of the quantity in brackets when the entire population is considered? Thus the summation

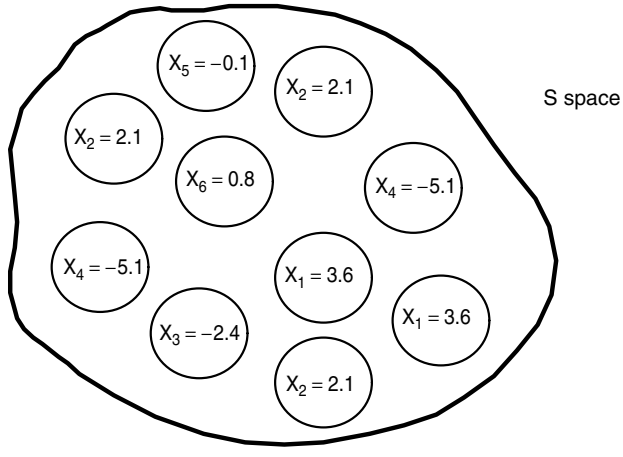



**Figure 1.18** Sample space S with N = 10 elements, some of which are alike.

in Equation 1.12 must include all elements. Since the probability of getting value $x_k$ in a random selection of one of the N elements in S is given by

$$p_{x_k} = \frac{n_k}{N}$$

an equation equivalent to Equation 1.12 is

$$E[X] = \mu_X = \sum_{k=1}^{K} x_k \, p_{x_k}. \tag{1.13}$$

In the example above

$$N = \sum_{k=1}^{6} n_k = 2 + 3 + 1 + 2 + 1 + 1 = 10$$

and

$$\begin{aligned} E[X] = \mu_X &= \sum_{k=1}^{6} x_k \, p_{x_k} \\ &= 3.6 \times \frac{2}{10} + 2.1 \times \frac{3}{10} - 2.4 \times \frac{1}{10} - 5.1 \times \frac{2}{10} - 0.1 \times \frac{1}{10} + 0.8 \times \frac{1}{10} \\ &= 0.16. \end{aligned}$$

Equation 1.13 is good for both finite and infinite populations. In the case of the latter, no empirical determination of the probabilities can be made; they must be known *a priori*.

Now replace digital rv X by a general function of X, namely, g(X). Then,

$$E[g(X)] = \sum_{k=1}^{K} g(x_k)\, p_{x_k}. \tag{1.14}$$

Consider two example functions. Let $g(X) = X^i$ for $i \geq 1$. Then, by analogy with Equation 1.12,

$$E[X^i] = \sum_{k=1}^{K} x_k^i\, p_{x_k} \tag{1.15}$$

is the i-th moment of rv $X^i$ about zero. Of course, $i = 1$ results in the mean $\mu_x$. For the second example function, let $g(X) = (X - E[X])^i$ for $i \geq 1$. The expectation becomes

$$E\left[(X - E[X])^i\right] = \sum_{k=1}^{K} (x_k - \mu_X)^i\, p_{x_k} \tag{1.16}$$

which is the i-th moment about the mean, or the i-th central moment.

A common central moment is the second moment or variance. Accordingly, for $i = 2$ we have

$$E\left[(X - E[X])^2\right] = \sum_{k=1}^{K} (x_k - \mu_X)^2\, p_{x_k} \tag{1.17}$$

or, what is the same,

$$\begin{aligned} \mathrm{Var}(X) = \sigma_X^2 = E\left[(X - E[X])^2\right] &= E[X^2] - 2E[X] \cdot E[X] + (E[X])^2 \\ &= E[X^2] - (E[X])^2. \end{aligned} \tag{1.18}$$

The last form for the variance in Equation 1.18 shows that it is equivalent to the "mean of the squares minus the square of the mean."

For the S space example, substituting Equation 1.13 and Equation 1.15 into Equation 1.18 yields

$$\sigma_X^2 = \frac{[(3.6)^2 \times 2 + (2.1)^2 \times 3 + (-2.4)^2 \times 1 + (-5.1)^2 \times 2 + (-0.1)^2 \times 1 + (0.8)^2 \times 1]}{10} - (0.16)^2$$
$$= 9.732.$$

### 1.4.2.2 Analog data

The expected value of analog random variable X is given by

$$E[X] = \mu_X = \int_{-\infty}^{+\infty} x\, f(x)\, dx \tag{1.19}$$

where f(x) is the *probability density function* of rv X. Integration is involved for an analog variable as opposed to discrete summation for a digital variable. The limits on X extend over the range $-\infty$ to $+\infty$ and include the case in which the probability density function is zero over some portion of this range. For the general analog function g(X),

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x)\, f(x)\, dx. \tag{1.20}$$

Consider two analog example functions following those for digital data. Let $g(X) = X^i$ for $i \geq 1$. Then

$$E[X^i] = \int_{-\infty}^{+\infty} x^i\, f(x)\, dx \tag{1.21}$$

is the i-th moment of rv $X^i$ about zero. Again, when $i = 1$, we obtain the population mean $\mu_X$. Now let $g(X) = (X - E[X])^i$ for $i \geq 1$. Parallel to Equation 1.16,

$$E\left[(X - E[X])^i\right] = \int_{-\infty}^{+\infty} (x - \mu_X)^i\, f(x)\, dx \tag{1.22}$$

is the i-th central moment. The second moment, $i = 2$, is the variance. Thus, for analog data

$$Var(X) = \sigma_X^2 = E\left[(X - E[X])^2\right] = \int_{-\infty}^{+\infty} (x - \mu_X)^2\, f(x)\, dx. \tag{1.23}$$

### 1.4.2.3 Covariance

Consider two analog random variables $X_1$ and $X_2$. We can find the variance of each using Equation 1.23. We now inquire about how these two random variables *covary* in time. That is, do they tend to track each other? When $X_1$ increases, does $X_2$ also tend to increase (or decrease), or does $X_2$ just as likely increase as decrease? The measure of this relationship is called *covariance*. If, when $X_1$ increases (decreases) $X_2$ also tends to increase (decrease), the sign of the covariance (or covariability) will be positive; if, when $X_1$ increases (decreases) $X_2$ tends to decrease (increase), the sign of the covariance (or covariability) will be negative; and, lastly, if, when $X_1$ increases or decreases $X_2$ is just as likely to increase as decrease, the expected covariance is zero and the variables are *independent* of each other. Stated mathematically, we have

$$E[(X_1-\mu_1)(X_2-\mu_2)] = \mathrm{Cov}[X_1,X_2] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x_1-\mu_1)(x_2-\mu_2)f(x_1,x_2)dx_1\,dx_2 \tag{1.24}$$

where $\mathrm{Cov}[X_1,X_2]$ means covariance between random variables $X_1$ and $X_2$ and $f(x_1, x_2)$ is the *joint probability density function* between random variables $X_1$ and $X_2$.

If $X_1$ and $X_2$ are independent, $f(x_1, x_2) = f(x_1) \cdot f(x_2)$; that is, the joint probability density function is equal to the product of the individual probability density functions. With this condition,

$$\begin{aligned}\mathrm{Cov}[X_1,X_2] &= \int_{-\infty}^{+\infty}(x_1-\mu_1)\,f(x_1)\,dx_1\int_{-\infty}^{+\infty}(x_2-\mu_2)\,f(x_2)\,dx_2\\ &= E[X_1-\mu_1]\,E[X_2-\mu_2] \;=\; 0.\end{aligned} \tag{1.25}$$

Because the expectation operator is linear, it can be taken inside the brackets of each term in the product on the right, so that $E[X_1-\mu_1] = \mu_1-\mu_1 = 0$, and similarly for the second term. The expected value of a constant is, of course, the same constant. The formulas for digital data similar to Equations 1.24 and 1.25 are

$$E[(X_1-\mu_1)(X_2-\mu_2)] = \mathrm{Cov}[X_1,X_2] = \sum_{k=1}^{K}\sum_{m=1}^{M}(x_{1k}-\mu_1)\,(x_{2m}-\mu_2)\,p_{x_{1k},x_{2m}} \tag{1.26}$$

where $K = M$ and, for independent variables,

$$\begin{aligned}\mathrm{Cov}[X_1,X_2] &= \sum_{k=1}^{K}(x_{1k}-\mu_1)\,p_{x_{1k}}\sum_{m=1}^{M}(x_{2m}-\mu_2)\,p_{x_{2m}}\\ &= E[X_1-\mu_1]\,E[X_2-\mu_2] = 0.\end{aligned} \tag{1.27}$$

### 1.4.3 Distribution of variance at a harmonic

Let us now shift our focus from comparing variances among harmonics in Section 1.4.1 to examining how variance is distributed at a single harmonic across the population of realizations. The detailed and somewhat lengthy derivation of this distribution is the subject of Appendix 1.C. In this section we present only the results. Our recommendation is that readers complete this section before studying Appendix 1.C.

Basic knowledge of the properties of a chi-square distribution is essential from this point forward. Sufficient background usually can be found in an undergraduate text in statistics. We will expand on this basic knowledge as needed.

In Appendix 1.C it is shown that, for a normal white noise process, the covariance between the sine and/or cosine coefficients at any two harmonics is zero (a result that might have been anticipated from Equations 1.4 and 1.5) and the coefficients are normally distributed. Squaring the coefficients and standardizing them by dividing by their variances yields random variables with a chi-square distribution. Using the additive property of chi-square variable results in harmonic variances that are independent and proportional to a $\chi_2^2$-distribution (a chi-square distribution with two degrees of freedom) except at the frequency origin (harmonic 0) and, for an even number of data N, at harmonic N/2.

In analyzing geophysical data, we are usually concerned with an underlying stochastic process that is other than white noise. For this situation, the sinusoids at the harmonic frequencies are likewise orthogonal (see Equations 1.4 and 1.5) but it is only in the limit as the number of data N in a realization becomes infinite that their variances are independent and have a chi-square distribution. Koopmans (1974, Section 8.2) provides further discussion of the properties of nonwhite noise.

In obtaining the frequency distribution of variance for a general stochastic process, we will assume N is sufficiently large that it is reasonable to apply the results for white noise given in Appendix 1.C. The magnitude of N required to make this assumption reasonable depends on the departure of the random process from white noise. The greater the departure, the larger the value of N, but no specific value can be given. Thus, in accepting a conclusion from statistical analysis of a realization that depends on it being from a normal white process, it is important to express some caution. Assuming that N is sufficiently large, the variances at the interior harmonics are independently distributed according to

$$\frac{C(f_m)}{\Gamma(f_m)} \Rightarrow \frac{\chi_2^2}{2}, \quad \begin{cases} 0 < m < N/2, & N \text{ even} \\ 0 < m \leq (N-1)/2, & N \text{ odd} \end{cases} \tag{1.28a}$$

where random variable $C(f_m)$ is the variance at the m-th harmonic frequency $f_m$, $\Gamma(f_m)$ is the process variance at $f_m$, the arrow indicates "is distributed as," and,

therefore, the above variance ratio is distributed as a chi-square variable with two degrees of freedom divided by two. In general, the value of $\Gamma(f_m)$ is unknown. The next section shows how to determine a confidence interval for $\Gamma(f_m)$. The two degrees of freedom (dof) at each harmonic are a consequence of a sine and a cosine being fitted to the data. There is only one dof at the 0-th harmonic (the mean), regardless of whether N is even or odd. In cases where N is even, there also is only one dof at the N/2-th harmonic. As Table 1.1 shows, the calculations at these harmonics require only a cosine term. That is,

$$\frac{C(f_0)}{\Gamma(f_0)} \Rightarrow \chi_1^2, \qquad \text{N even or odd} \tag{1.28b}$$

and

$$\frac{C(f_{N/2})}{\Gamma(f_{N/2})} \Rightarrow \chi_1^2, \qquad \text{N even.} \tag{1.28c}$$

For N odd, the variance at the highest frequency $[(N-1)/2]$ has two dof as noted in Equation 1.28a. For N even or odd the total number of dof in the periodogram equals the number of data N.

It should be noted that $C(f_m)$ is analogous to $S_m^2$ in Table 1.1. One reason for changing notation is because $C(f_m)$, unlike $S_m^2$, is a random variable. Another reason is that in Section 1.5.6 we will be calculating variance at any frequency, f, and it will be convenient to simply drop the subscript m. For now, our interest remains in dealing with variance at the harmonic frequencies, $f_m$, only.

### 1.4.4 Confidence intervals on periodogram variances

In this and the following section, the underlying stochastic process is unspecified. It may or may not be white noise, but we assume that the number of data, N, is sufficiently large to justify application of independent chi-square distributions derived from a white noise process to the harmonic variances, as discussed in the previous section. In addition, we assume the data follow a normal distribution.

Given that the variance ratio $C(f_m)/\Gamma(f_m)$ follows a chi-square distribution according to Equation 1.28a, we can determine a confidence interval for the ratio using the probability expression

$$\Pr\left\{\frac{\chi_2^2(\alpha/2)}{2} \le \frac{C(f_m)}{\Gamma(f_m)} \le \frac{\chi_2^2(1-\alpha/2)}{2}\right\} = 1-\alpha, \qquad m \neq 0\,(\text{all N}),\ m \neq N/2\,(\text{N even}) \tag{1.29}$$

where $\alpha$ is the level of significance. In this equation observed values of $C(f_m)/\Gamma(f_m)$ vary between confidence limits $\chi^2_2(\alpha/2)/2$ and $\chi^2_2(1-\alpha/2)/2$ in $100(1-\alpha)\%$ of the observations. The term $\chi^2_2(\alpha/2)$ is a particular value of the random variable such that the area beneath its probability density function to the left of this value is $\alpha/2$.

We consider the case in which we have an observed value of $C(f_m)$ and the objective is to find the limits of the confidence interval for the population variance $\Gamma(f_m)$. By rearranging Equation 1.29, the $100(1-\alpha)\%$ confidence interval for $\Gamma(f_m)$ can be obtained from the probability statement

$$\Pr\left\{\frac{C(f_m)}{\chi^2_2(1-\alpha/2)/2} \le \Gamma(f_m) \le \frac{C(f_m)}{\chi^2_2(\alpha/2)/2}\right\} = 1-\alpha, \quad m \neq 0,\ N/2\ (N \text{ even}). \tag{1.30}$$

The interval between $2C(f_m)/\chi^2_2(1-\alpha/2)$ and $2C(f_m)/\chi^2_2(\alpha/2)$ is the $100(1-\alpha)\%$ confidence interval for $\Gamma(f_m)$. By taking the logarithm of the limits of the confidence interval for $\log \Gamma(f_m)$, the lower and upper limits become, respectively,

$$\log C(f_m) + \log(2/\chi^2_2(1-\alpha/2)) \qquad \text{and} \qquad \log C(f_m) + \log(2/\chi^2_2(\alpha/2)).$$

The logarithmic form of expressing the confidence interval is particularly useful in graphical representations of the periodogram. The reason is that the width of the confidence interval will be fixed regardless of frequency when the variances are plotted on a logarithmic axis.

We now apply Equation 1.30 to a set of data. Figure 1.19 is a plot of the data in Table 1.5 covering 100 consecutive years (1906–2005) of average autumn (September, October, November) temperatures from Climate Division 5 in Oklahoma. Climate Division 5 comprises 13 counties in central Oklahoma. Temperatures are in their original units of degrees Fahrenheit.

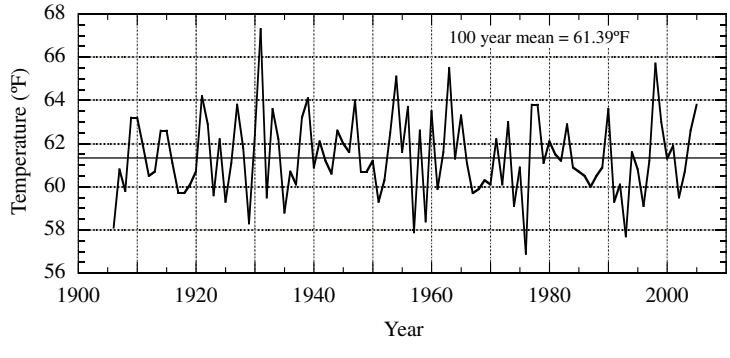



**Figure 1.19** One hundred years of mean autumn temperature (September, October, November) for central Oklahoma (Climate District 5) from 1906 to 2005.

**Table 1.5** One hundred years (1906–2005) of autumn mean temperature (°F) for Oklahoma Climate Division 5 (central part of the state). (*Source*: Oklahoma Climatological Survey.)

| Decade down/ Year across | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1900–1909 | | | | | | | 58.1 | 60.8 | 59.8 | 63.2 |
| 1910–1919 | 63.2 | 61.9 | 60.5 | 60.7 | 62.6 | 62.6 | 61.1 | 59.7 | 59.7 | 60.1 |
| 1920–1929 | 60.7 | 64.2 | 62.9 | 59.6 | 62.2 | 59.3 | 61.1 | 63.8 | 61.8 | 58.3 |
| 1930–1939 | 62.4 | 67.3 | 59.5 | 63.6 | 62.2 | 58.8 | 60.7 | 60.1 | 63.2 | 64.1 |
| 1940–1949 | 60.9 | 62.1 | 61.2 | 60.6 | 62.6 | 62.0 | 61.6 | 64.0 | 60.7 | 60.7 |
| 1950–1959 | 61.2 | 59.3 | 60.3 | 62.5 | 65.1 | 61.6 | 63.7 | 57.9 | 62.6 | 58.4 |
| 1960–1969 | 63.5 | 59.9 | 61.6 | 65.5 | 61.3 | 63.3 | 61.1 | 59.7 | 59.9 | 60.3 |
| 1970–1979 | 60.1 | 62.2 | 60.1 | 63.0 | 59.1 | 60.9 | 56.9 | 63.8 | 63.8 | 61.1 |
| 1980–1989 | 62.1 | 61.5 | 61.2 | 62.9 | 60.9 | 60.7 | 60.5 | 60.0 | 60.5 | 60.9 |
| 1990–1999 | 63.6 | 59.3 | 60.1 | 57.7 | 61.6 | 60.8 | 59.1 | 61.2 | 65.7 | 63.0 |
| 2000–2009 | 61.3 | 61.9 | 59.5 | 60.7 | 62.6 | 63.8 | | | | |

The periodogram is shown by the solid line in Figure 1.20 and was computed using subroutine Foranx in Appendix 1.A. The 95% confidence interval for the population variance is shown on the right-hand side of the figure (solid line) where the dot is to be placed over each sample variance $c(f_m)$ as shown, for example, at harmonic 39. Note that $c(f_m)$ is used to denote a sample value of the rv $C(f_m)$. That the dot with constant width confidence interval around it may be placed at any harmonic is a direct consequence of the logarithmic plot, as described above. Because the periodogram varies wildly, it is not easy to discern bands of small or large variance or a trend in variance with harmonic number. Correspondingly, the 95% confidence interval ($\alpha = 0.05$) for $\Gamma(f_m)$ is very wide. The variability in $c(f_m)$ seen here is typical of periodograms of many kinds of geophysical data and is the chief reason that periodograms of observed data are often smoothed, as discussed in the next section.

### 1.4.5 The smoothed periodogram

To better distinguish bands of large and small variance or a trend in the spectrum, a common practice is to smooth the spectrum by weighting together a number of contiguous variances. The simplest smoothing is the running mean of length n (odd) given by

$$\overline{C}(f_m) = \frac{1}{n} \sum_{j=m-(n-1)/2}^{m+(n-1)/2} C(f_j) \tag{1.31}$$

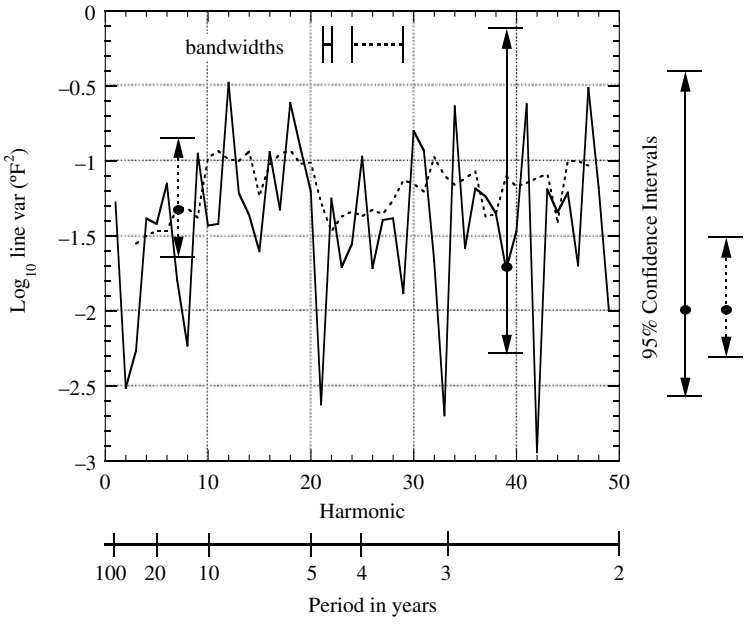



**Figure 1.20** Periodogram of the data in Table 1.5 and Figure 1.19 (solid line). Averaged periodogram using 5-point running mean (dashed line). The respective 95% confidence intervals for the population mean variance at each harmonic are shown to the right and the respective bandwidths at the top of the figure.

in which average variance is calculated only for those harmonics that do not include $f_0$ and $f_{N/2}$ (N even) in the summation. Random variable C is used in Equation 1.31 to indicate we are determining the effects of smoothing on the distribution of variances; in application, however, harmonic variances from a single realization would be smoothed, and lower case variable c would be used as in the previous section. Because of the inability to include the correct number of terms, there will be a loss of $(n-1)/2$ harmonic variances at either end of the smoothed periodogram.

If we assume, as prescribed earlier, that the number of data in a realization is sufficiently large that the variance ratios $C(f_j)/\Gamma(f_j)$ can be approximated by independent $\chi^2$ variables with two dof divided by two and, furthermore, that $\Gamma(f_j)$ is effectively constant over the length n, then, using Equation 1.31, the smoothed variance ratios $\overline{C}(f_m)/\overline{\Gamma}(f_m)$ are approximately $\chi^2$ random variables with 2n dof (Hoel, 1962, p. 268) divided by 2n and are independent every n harmonics.

The dashed line in Figure 1.20 is the result of a five-point running mean and provides an improved picture of the structure of the variance. There are now $2n = 10$ dof associated with each variance ratio. For this realization, periods from 5 to 10 years contain more variance than periods shorter than five years except near the two-year period. This comparison suggests the data should comprise sharp year-to-

year fluctuations superimposed on long-period fluctuations. The plot of the data in Figure 1.19 clearly indicates that there are short period variations; long period variations, though, are less obvious. Thus the need for a periodogram to show the not-so-obvious.

By analogy with the limits of the confidence interval for $\log \Gamma(f_m)$, the limits for $\log \overline{\Gamma}(f_m)$ are

$$\log\overline{C}(f_m) + \log(2n/\chi^2_{2n}(1-\alpha/2)) \quad \text{and} \quad \log\overline{C}(f_m) + \log(2n/\chi^2_{2n}(\alpha/2)).$$

For $n = 5$ and $\alpha = 0.05$, the values for the constant terms above are $-0.31$ and 0.49. The 95% confidence interval is shown by the dashed vertical line on the right of Figure 1.20 and its reduced length relative to that for no smoothing ($n = 1$) reflects its application to an averaged spectrum, namely, to $\log \overline{\Gamma}(f_m)$, which, by our previous assumption, is approximately $\log \Gamma(f_m)$.

To take into account smoothing of the periodogram by other than a running mean, an approximate general formula for the dof r in $\chi^2$ distributions is, for n odd,

$$r = 2\left[\sum_{j=-(n-1)/2}^{(n-1)/2} K^2(f_j)\right]^{-1} \tag{1.32}$$

where $K(f_j)$ is a symmetric weight function centered at frequency $f_0$ such that the sum of the weights is unity (Koopmans, 1974, p. 273). Maintaining unity preserves the total variance in the spectrum. When $K(f_j) = 1/n$, the running mean, $r = 2n$.

Associated with dof is bandwidth $\beta$, the frequency interval between independent adjacent estimates of variance. In the case of a periodogram with no smoothing it is

$$\beta = \frac{1}{N\Delta t} \tag{1.33}$$

which is the frequency difference between harmonics i and $i + 1$. Rewriting Equation 1.33 in the form

$$\beta N\Delta t = 1 \tag{1.34}$$

we see that the product of $\beta$ and $N\Delta t$ is constant. This means that as the length N of a time series increases ($\Delta t$ remains fixed), the bandwidth of each independent periodogram estimate will proportionately decrease and the total number of spectrum estimates will proportionately increase. Because there are two dof associated with each spectrum estimate, increasing the length of a time series in and of itself does not reduce the variability of periodogram estimates.

Equation 1.33 is exact for white noise and approximate for nonwhite noise when N is large. For a five-point running mean the bandwidth would be five times as wide. The bandwidths are shown in Figure 1.20 for their associated spectra. An approximate general formula for the bandwidth is (Koopmans, 1974, p. 277)

$$\beta = \frac{r}{2N\Delta t} = \left[ N\Delta t \sum_{j=-(n-1)/2}^{(n-1)/2} K^2(f_j) \right]^{-1} \tag{1.35}$$

which reduces to $\beta = n/N\Delta t$ for a running mean of length n.

As an example of a simple nonrunning mean filter, consider a three-point smoother (a triangular filter) whose weights are $^1/_4$, ½, $^1/_4$ (sum of weights = 1). From Equation 1.32 the dof will be $5^1/_3$ whereas the number of dof for a three-point running mean is six. From Equation 1.35 the bandwidth of the former is 8/9 as wide as that of the latter. It is of interest to know that the periodogram used to produce the spectrum of hourly temperatures in Figure 1.15 was smoothed with this triangular filter prior to creating the product of variance density and frequency. The purpose was to magnify the two broad frequency bands that were discussed relative to the main peak of the daily cycle of temperature.

### 1.4.6 Testing the white noise null hypothesis

In this section we examine the problem of testing the null hypothesis that a sample of data comes from a random process that is white noise. This is equivalent to the null hypothesis that the expected values of the spectrum variances are uniform with frequency. A white noise test can be an important tool in analyzing spectra of geophysical data. If we observe in a given spectrum a single variance, multiple variances, or a band of variance that seems to be unexpectedly large, the question arises whether these features are a consequence of an underlying physical process or whether they occurred by chance. If the white noise null hypothesis applied to the spectrum cannot be rejected, there is then doubt that the observed large variance or variances are anything more than natural fluctuations in a realization from a white noise process.

In the previous section a method was developed to place a confidence interval for the population variance surrounding each sample variance (variance at a harmonic from a single realization). In contrast, in this section we will place confidence intervals for the sample variances about the estimated population variance, hypothesized to be uniform with frequency. Here, however, it is necessary to consider two types of confidence intervals. These will be demonstrated with two examples, the first of which employs the 100-year record of central Oklahoma temperature data seen in the previous two sections. How these confidence intervals are used in making a white noise test requires some background knowledge, to which we now turn.

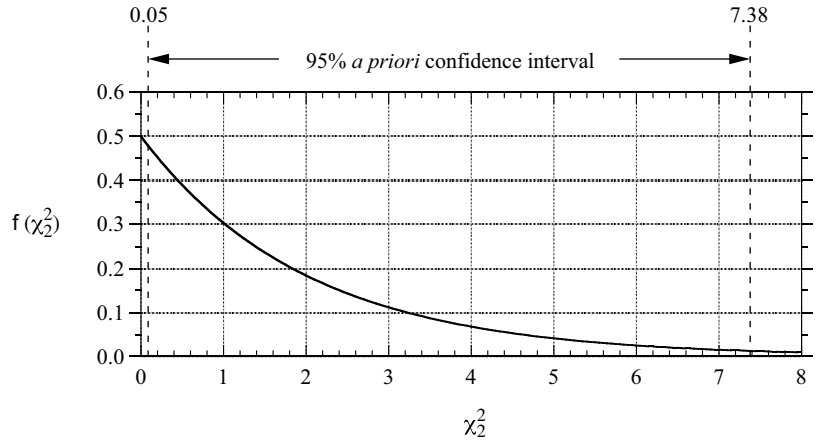



**Figure 1.21** The probability density function (pdf) of a random variable that has a chi-square distribution with two degrees of freedom. Confidence limits for the 95% *a priori* confidence interval are shown by the vertical dashed lines. The area under the curve has unit value.

Figure 1.21 shows the probability density function (or, equivalently, the frequency distribution) of a $\chi^2$ random variable with two dof. The probability density function is given by:

$$f(\chi_2^2) = \frac{\exp\left(-\frac{\chi_2^2}{2}\right)}{2}. \tag{1.36}$$

The two vertical dashed lines encompass what is called the *a priori* 95% confidence interval. Figure 1.21 is the typical way of presenting a probability density function or frequency distribution with confidence limits. When a confidence interval is applied to a spectrum, the width of this interval is oriented in the vertical, as we did in Figure 1.20.

Let us imagine successively withdrawing 29 samples from a population that has the distribution shown in Figure 1.21. Prior to the first withdrawal, the probability that its value will lie outside the interval (0.05, 7.38) is 0.05. Prior to the second withdrawal, the probability that its value will lie outside the same interval is also 0.05. Repeat this 27 more times. Because each withdrawal is independent of any other, the probability is 0.05 that any sample value of $\chi_2^2$ will lie outside the interval (0.05, 7.38).

Now arrange the sample values of $\chi_2^2$ as shown in Figure 1.22, except that each value withdrawn is divided by two (with this adjustment we can apply the results of this section directly to the distribution of periodogram variance ratios). Before even looking at the sample values of $\chi_2^2/2$, we would not be surprised to find one or two lying outside the confidence interval. This follows from the calculation

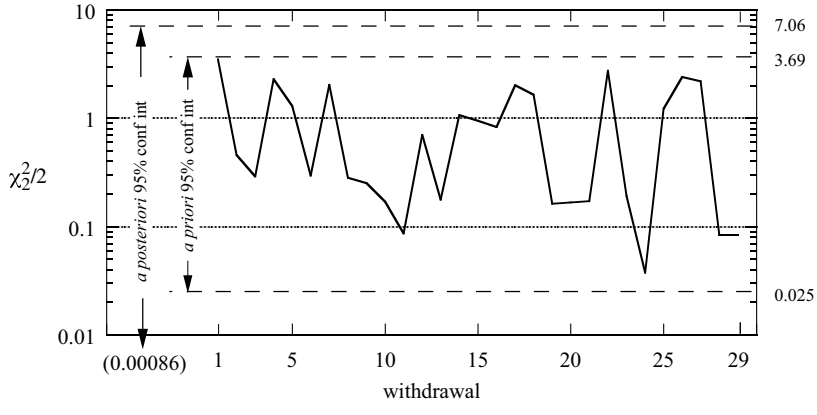



**Figure 1.22** A plot of 29 random withdrawals from a chi-square distribution with 2 degrees of freedom after dividing the value of each withdrawal by 2. The chi-square distribution function is shown in Figure 1.21. 95% *a priori* and *a posteriori* confidence intervals are also shown.

$29 \times 0.05 = 1.45$, where 29 is the number of withdrawals and 0.05 is the level of significance or the probability of rv $\chi^2_2/2$ being greater than 3.69 (7.38/2 from Figure 1.21) or less than 0.025 (0.05/2) per withdrawal. (If we had 100 such data sets we would expect 145 of the 2900 values to lie outside the confidence interval.) In fact, Figure 1.22 shows that one value or point lies very close to the upper confidence limit (withdrawal 1 is 3.501) and the value of withdrawal 24 (0.038) is slightly above the lower confidence limit.

To show the probability, $\alpha$, of observing one or more values from a $\chi^2_2/2$ distribution outside the *a priori* confidence interval, we make use of the binomial distribution

$$\frac{M!}{Z!(M-Z)!} \times p^Z(1-p)^{M-Z}$$

where: p = probability that the value of a randomly selected point will lie outside the *a priori* confidence interval (0.05 for a 95% confidence interval); M = total number of points (29 in this example); and Z = the number of the M points that lie outside the *a priori* confidence limits.

The probability of one or more points lying outside the confidence interval is one minus the probability of no points lying outside the confidence interval, or, in general, $\alpha = 1 - (1-p)^M$. Thus, if $p = 0.05$ and $M = 29$, then $\alpha = 1 - (0.95)^{29} = 1 - 0.2259 = 0.7741$. Instead of having a 5% chance of finding at least one value of $\chi^2_2/2$ outside the confidence interval, we actually have a 77% chance when considering the group of 29 values or points.

In practice, we are sometimes faced with the following dilemmas. In a given data set the number of values that lie outside the *a priori* confidence interval is about as expected, but one of the values is very large. Is the very large value significantly greater than expected? In another case, a few more values than expected lie outside the confidence interval. Is the difference between the expected number and observed number significant?

The solution to both dilemmas is as follows. We really want $\alpha$ to be 0.05. That is, when we consider all the points in the group (29 in this example), we want to find the two particular values of $\chi_2^2/2$ such that there is only a 5% chance that *any one or more* points will lie outside the associated interval. This is called the *a posteriori* confidence interval. With the meaning of p the same as that given previously, except that it now applies to the confidence interval for the group of points, the determination of the limits of this interval follows.

From above, and using the binomial theorem, for $p \ll 1$

$$\alpha = Mp,$$

so that

$$p = \alpha/M.$$

With

$$\alpha = 0.05, \quad M = 29,$$

then

$$p = 0.00172.$$

If we now integrate the probability density function (Equation 1.36) between 0 and $\chi_2^{2*}$ and between $\chi_2^{2**}$ and $\infty$ where $^*$ and $^{**}$ indicate particular values of $\chi_2^2$, and equate both results to p/2, we obtain $\chi_2^{2*}/2 = 0.00086$ and $\chi_2^{2**}/2 = 7.06$. These are the lower and upper limits for the *a posteriori* 95% confidence interval and are plotted in Figure 1.22 (however, the lower limit is off the graph). The *a posteriori* confidence interval deals with all 29 values at one time and the *a priori* confidence interval deals with one value at a time. There is only one chance in 20 that any one or more of the 29 values would lie outside the 95% *a posteriori* confidence interval, and as Figure 1.22 shows none do. This result is in accord with our withdrawal of 29 random samples from a $\chi^2$ distribution with two dof divided by two.

In periodogram analysis we typically use *a posteriori* confidence limits because we want to observe the *entire spectrum* of harmonic variances *after the fact* of calculating

the spectrum. If we wanted to know whether the variance at a *particular harmonic* exceeded the confidence limits *before the fact* of observing the variance at the harmonic in question, we would use the *a priori* confidence interval. Interest in the latter approach is uncommon. Nevertheless, obtaining *a priori* confidence limits is always a natural first step because if none of the spectrum variances exceed these limits, there is no need to proceed to the next step of computing *a posteriori* confidence limits.

To better understand white noise testing we examine two applications to real data.

### *1.4.6.1 White noise test: Example 1*

We revisit the 100-year record of mean autumn temperatures for central Oklahoma (Climate Division 5) given in Table 1.5 and plotted in Figure 1.19. Figure 1.20 showed confidence intervals for the population variance $\Gamma(f_m)$ and smoothed population variance $\overline{\Gamma}(f_m)$ at each harmonic given samples of $C(f_m)$ and $\overline{C}(f_m)$, respectively. The underlying random process was unspecified, but the number of data was assumed sufficiently large to justify using independent chi-square distributions of the harmonic variances derived for a normal white noise process. In this example we will use the same data to find a different kind of confidence interval; that is, we will find confidence limits for observations of rv $C(f_m)$ common to all harmonic frequencies given an estimate $\hat{\Gamma}(f_m)$ of the population variance under the white noise hypothesis. Because a white noise process is hypothesized, there is no restriction on the size of a data set.

The variance of the data set in Table 1.5 is $3.4201^\circ F^2$. As a consequence, our estimate of the population variance at each of the 49 interior harmonics ($m = 0$ and $m = 50$ excluded) in the periodogram under the white noise hypothesis is $\hat{\Gamma}(f_m) = 3.4201^\circ F^2/49.5 = 0.0691^\circ F^2$. The symbol $\wedge$ means "estimate of" and the reason we make this distinction is that variance of the realization ($3.4201^\circ F^2$) is just an estimate of the population variance. In general, we do not test the variance at the highest harmonic for N even, here $m = 50$, because its variance, under a white noise hypothesis, is one-half the interior variances; it is a unique harmonic. The reason for its uniqueness is that its bandwidth is one-half the bandwidth associated with each of the interior harmonics. That the divisor is 49.5 instead of 49 is because the variance of the data set included the variance at $m = 50$. Thus, for the general case of an even number of data, N, the white noise variance at the interior harmonics is determined from the total variance in the data set divided by $(N/2 - 1) + 0.5 = (N/2) - 0.5$. In the general case of an odd number of data, N, the white noise variance at all the harmonics (except $m = 0$) is the total variance in the data set divided by $(N - 1)/2$. The highest harmonic has full bandwidth.

To find the *a priori* confidence interval for the sample variances about their estimated expected value, we rewrite Equation 1.29 to obtain the form

$$\Pr\left\{\frac{\Gamma(f_m)\,\chi_2^2(\alpha/2)}{2} \le C(f_m) \le \frac{\Gamma(f_m)\,\chi_2^2(1-\alpha/2)}{2}\right\}$$

$$= 1-\alpha, \quad m \neq 0\ (\text{all N}); \quad m \neq N/2\ (\text{N even}). \tag{1.37}$$

To be consistent with Figure 1.20, we take logarithms of the limits of the confidence interval and obtain

$$\log\left[\Gamma(f_m)\,\chi_2^2(\alpha/2)/2\right] \quad \text{and} \quad \log\left[\Gamma(f_m)\,\chi_2^2(1-\alpha/2)/2\right]$$

which, for $\alpha = 0.05$ and $\hat{\Gamma}(f_m) = 0.0691°F^2$, are $-2.76$ and $-0.59$, respectively. We expect, on average, 2½ ($0.05 \times 50$) variances to exceed these limits; in Figure 1.23 we note that the variances at three harmonics (12, 42, and 47) fall outside either the upper or lower limit. Should the white noise null hypothesis be rejected? This is a good example in which the answer can be found by calculating the *a posteriori* confidence limits.

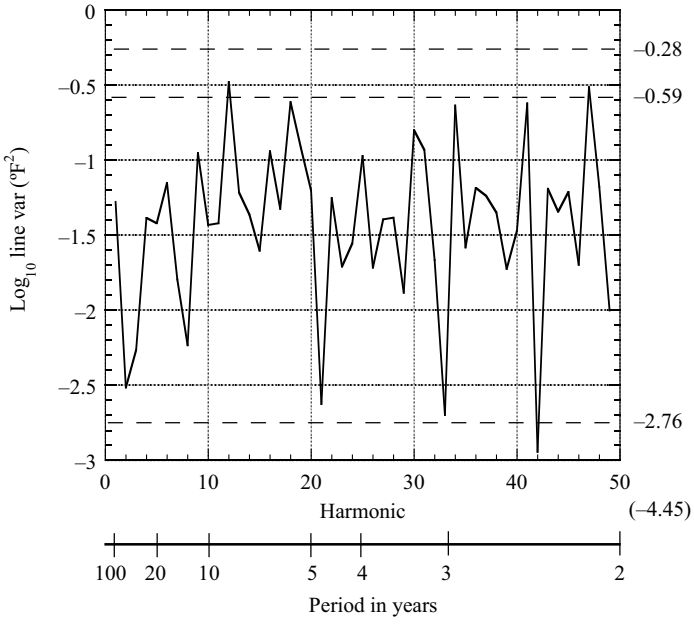



**Figure 1.23** Periodogram of the data in Table 1.5 and Figure 1.19 (see also Figure 1.20). The inner two dashed lines are the 95% *a priori* confidence limits; the upper dashed line is the upper 95% *a posteriori* confidence limit. The lower 95% *a posteriori* confidence limit is located below the graph and has value $-4.45$.

The *a posteriori* confidence interval is determined by replacing $\alpha/2$ in Equation 1.37 and the expressions for the limits of the confidence interval by $p/2$ where $p \approx \alpha/M$ and $M = 49$. Therefore, the parallel equations for *a posteriori* confidence limits are

$$\Pr\left\{\frac{\Gamma(f_m)\,\chi_2^2(p/2)}{2} \le C(f_m) \le \frac{\Gamma(f_m)\,\chi_2^2(1-p/2)}{2}\right\}$$

$$= 1 - p, \quad m \neq 0\ (\text{all N});\ m \neq N/2\ (\text{N even}) \tag{1.38}$$

and logarithms of the limits of the confidence interval are

$$\log\left[\Gamma(f_m)\,\chi_2^2(p/2)/2\right] \quad \text{and} \quad \log\left[\Gamma(f_m)\,\chi_2^2(1-p/2)/2\right].$$

Equation 1.36 can be integrated, as in the previous example, to obtain the values of $\chi_2^2$ for area $p/2 = 0.0005102$ at the left and right extremes of the chi-square distribution (refer to Figure 1.21). The results are $\chi_2^2(p/2) = 0.0010207$ and $\chi_2^2(1-p/2) = 15.161$, so that the logarithms of the lower and upper limits of the 95% *a posteriori* confidence interval are $-4.45$ and $-0.28$. We observe in Figure 1.23 that no variance lies outside this range and, therefore, we cannot reject the null hypothesis that the data are a realization from a white noise process. Thus there appears to be no useful statistical predictability of mean autumn temperature at Oklahoma City other than using its long-term mean as the predictor.

### *1.4.6.2 White noise test: Example 2*

For our second example, we investigate five years of mean monthly temperature at Oklahoma City from 2003–2007. The data are given in Table 1.6 and plotted in Figure 1.24a. As expected, the time series shows a strong annual solar influence. We can consider the annual solar cycle for each year to vary in a different way about a long-term mean annual temperature cycle. Ideally, it is the long-term cycle we would like to remove from the time series before we apply a white noise test. The best we can do, however, is to estimate this cycle using the five years of data available to us. Except for the solar cycle, the approach to obtain confidence limits is similar to that in the first example.

We can estimate the long-term annual cycle by averaging the five years of data month-by-month and then subtracting the appropriate five-year average from each observed monthly mean. The set of residuals, also given in Table 1.6, form a sequence of 60 values from January 2003 through December 2007 and comprise the time series for which the white noise null hypothesis will be tested. The time series is shown in Figure 1.24b.

**Table 1.6** (a) Monthly mean temperatures (°C) at Oklahoma City Will Rogers Airport from 2003 to 2007. The bottom row shows monthly means averaged over the five-year period. (b) Monthly mean residuals, i.e., the appropriate five-year monthly average has been subtracted from each monthly mean. (Note: All monthly means in (a) have been converted to Celsius from the original monthly means in Fahrenheit. *Source*: National Climatic Data Center, Asheville, NC.)

(a)

| Month/ Year | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2003 | 2.67 | 3.17 | 9.72 | 15.78 | 20.61 | 23.28 | 29.06 | 28.22 | 20.78 | 17.56 | 10.28 | 6.17 |
| 2004 | 4.39 | 4.39 | 12.94 | 16.22 | 22.17 | 24.11 | 26.06 | 24.78 | 23.94 | 18.11 | 10.33 | 6.39 |
| 2005 | 4.22 | 8.22 | 10.94 | 16.28 | 20.67 | 25.56 | 26.89 | 27.22 | 25.06 | 17.50 | 12.11 | 3.83 |
| 2005 | 8.72 | 5.39 | 12.83 | 19.61 | 22.56 | 26.67 | 30.11 | 29.94 | 21.83 | 17.11 | 11.61 | 6.39 |
| 2007 | 2.67 | 5.61 | 15.67 | 14.11 | 21.67 | 25.06 | 27.06 | 29.00 | 24.50 | 18.61 | 11.61 | 3.94 |
| Mean | 4.53 | 5.36 | 12.42 | 16.40 | 21.53 | 24.93 | 27.83 | 27.83 | 23.22 | 17.78 | 11.19 | 5.34 |

(b)

| Month/ Year | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2003 | −1.87 | −2.19 | −2.70 | −0.62 | −0.92 | −1.66 | 1.22 | 0.39 | −2.44 | −0.22 | −0.91 | 0.82 |
| 2004 | −0.14 | −0.97 | 0.52 | −0.18 | 0.63 | −0.82 | −1.78 | −3.06 | 0.72 | 0.33 | −0.86 | 1.04 |
| 2005 | −0.31 | 2.87 | −1.48 | −0.12 | −0.87 | 0.62 | −0.94 | −0.61 | 1.83 | −0.28 | 0.92 | −1.51 |
| 2006 | 4.19 | 0.03 | 0.41 | 3.21 | 1.02 | 1.73 | 2.28 | 2.11 | −1.39 | −0.67 | 0.42 | 1.04 |
| 2007 | −1.87 | 0.26 | 3.24 | −2.29 | 0.13 | 0.12 | −0.78 | 1.17 | 1.28 | 0.83 | 0.42 | −1.40 |

Before applying the test, a few comments are in order concerning the method of removing the annual cycle. The total variance of the annual cycle is the sum of variances from harmonics with periods of 12, 6, 4, 3, 2.4, and 2 months. When the periodogram of the residuals is computed (an exercise we recommend), one will discover the variance is zero at these six harmonics. The reason is that we have removed the variances at all harmonics of the annual cycle from the original time series. In fact, had we computed a periodogram of the original data, it would have included the identical variances at the harmonics corresponding to the periods of the estimated annual cycle. In this example, we will replace the zero variances at periods of 12, 6, 4, 3, and 2.4 months by the average of the two adjacent variances. While these replacement values are artificial and are not part of the white noise test, for the sake of appearance they will provide a smoothly varying periodogram in the vicinity of the harmonics of the annual cycle. From Equation 1.28c, the distribution of the variance ratio at the harmonic corresponding to a period of two months is $\chi_1^2$. As in Example 1, the white noise test is applied only to the interior harmonics.

The variance of the residual data shown in Figure 1.24b is 2.3160 °C$^2$. Under the null hypothesis that this data set is a realization of a white noise process, we can

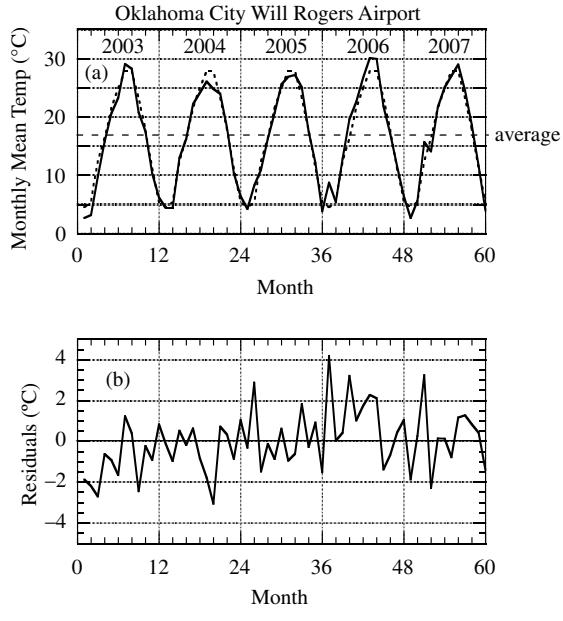



**Figure 1.24** (a) Mean monthly temperatures at Oklahoma City Will Rogers Airport from 2003 to 2007 (solid line) and average mean monthly temperatures (dashed line). (b) Residual mean monthly temperatures (actual – average).

estimate the population variance $\Gamma(f_m)$ at each of the 29 interior harmonics (harmonics $m = 0$ and $m = N/2$ are excluded) in the Fourier spectrum using the equation $\hat{\Gamma}(f_m) = (2.3160/24)\,^\circ C^2 = 0.0965\,^\circ C^2$. The reason for dividing by 24 is that the residual variance does not include any variance from the six harmonic frequencies previously discussed. Since there are a total of 30 harmonic frequencies (60 samples of data), and the variance has been removed from the six harmonics associated with the annual cycle, the residual variance is distributed equally among the remaining 24 harmonics to estimate the population variance.

Upon replacing $\Gamma(f_m)$ by its estimate $\hat{\Gamma}(f_m)$, we conclude from Equation 1.28a that the variance ratio $C(f_m)/\hat{\Gamma}(f_m)$ varies approximately as $\chi^2_2/2$. Figure 1.25 shows the sample variance ratios versus harmonic number where the ratios at harmonics 5, 10, 15, 20, and 25 are the averages of adjacent ratios. Since no ratio lies outside the *a posteriori* confidence interval, the null hypothesis that the sample data come from a random process that is white noise cannot be rejected at the 5% level of significance. Stated another way, this realization can be viewed as a member of a population of similar random time series, the totality of which comprises a white noise random process. In this particular example, computing the *a posteriori* confidence interval was not necessary since none of the variance ratios lie outside the *a priori* confidence interval. The goal of this example was to derive the 95% *a posteriori* confidence limits for variance ratios as opposed to variances in the first example.

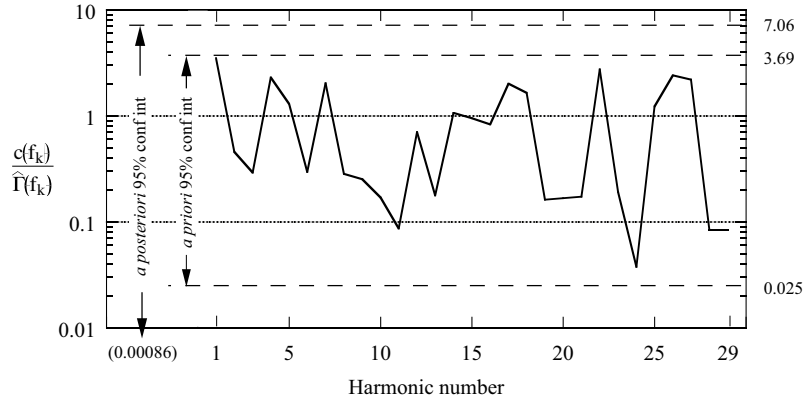



**Figure 1.25** Observed variance ratio versus harmonic frequency for the residuals in Figure 1.24b. The population variance $\Gamma(f_m)$ is estimated from the sample variance. Harmonic frequency $f_m$ has been converted to harmonic number. 95% *a priori* and *a posteriori* confidence intervals are also shown.

A keen observer will recognize that the plot in Figure 1.22 is identical to that in Figure 1.25. In fact, the same data set was used to produce the plot in Figure 1.22. Thus withdrawing 29 values from a chi-square distribution in the discussion in Section 1.4.6 was a little "white" lie! But whether one literally withdrew samples from a chi-square distribution was immaterial to developing an understanding of the mechanics of a white noise test.

Practically speaking, if the data set in this example is representative of other five-year intervals at Oklahoma City, then there is no skill in attempting to forecast mean monthly temperature beyond what can be accomplished by employing the average annual cycle. Had the white noise hypothesis been rejected, there would have been potentially useful skill in mean monthly temperature forecasts apart from the average annual cycle. In conclusion, if there is interest on the part of an investigator to make statistical forecasts of any variable represented by a time series, a good first step is to perform a white noise test of the original data or, if appropriate, the residual data, that is, the original data less the deterministic components.

## 1.5 Further important topics in Fourier analysis

At this juncture, we are able to (i) compute the Fourier coefficients of a data set, (ii) calculate its spectrum or periodogram, (iii) determine a confidence interval for the population variance at each harmonic frequency, and (iv) perform *a priori* and *a posteriori* white noise tests. Now we consider selected topics that will extend our understanding of Fourier analysis. As we have already seen in Table 1.1, the number of harmonics at which variance is computed is directly related to the number of data.

Section 1.5.1 explains why. The second topic, covered in Section 1.5.2, shows, mathematically, why variance calculated at a given harmonic frequency includes not only the variance at that harmonic but also variance from frequencies between nearby harmonics. Thus variance in one part of a spectrum can "bleed" or "leak" to another part of the spectrum. In short, we always view a spectrum through a "window."

Sometimes we are faced with a signal and noise problem. For example, we might be suspicious that there is a 60 Hz signal, say, from a power source, corrupting a data set we collected. That is, a deterministic signal may be embedded in otherwise random data. In Section 1.5.3 we investigate how averaging spectra from a number of data records, each of which contains the deterministic signal, smooths the averaged spectrum so the deterministic signal is more easily discernible. Another approach is discussed in Section 1.5.4, where we examine the effect that increasing the length of a time series has on discriminating a sinusoid from random components. The fifth topic shows how to convert the formulas in Table 1.1 for Fourier synthesis and analysis to complex form; this is developed in Section 1.5.5. Because of trigonometric symmetry, a complex representation is very compact. Using complex forms makes it easy to compute variance at frequencies between harmonics. This is the subject of Section 1.5.6. One interesting result is that the variance at a nonharmonic frequency is uniquely related to the variances at all the harmonic frequencies. The seventh and last topic, in Section 1.5.7, is concerned with adding zeroes to a data set, why we might do that, and how to interpret the resulting spectrum

### 1.5.1 Aliasing, spectrum folding, and the Nyquist frequency

*Aliasing* is a direct consequence of digitally sampling an analog signal. Aliasing has no relevance to purely analog data records. To show how aliasing works, consider the three cases in Figure 1.26. In example (1) an analog sinusoidal wave with frequency 10 Hz is sampled at intervals of 0.1 s as indicated by the arrowheads. The dashed line connects the sample values. Based on just the sample values, we would likely (mistakenly) conclude the underlying signal has constant value. In example (2) there is a 9 Hz sinusoid sampled every 0.1 s. After fitting the sample values with a smooth line, we would likely (mistakenly) conclude that the underlying signal is a 1 Hz sinusoid. Example (3) indicates that for 0.1 s sampling, a 6 Hz sinusoid could just as well be interpreted as a 4 Hz sinusoid. These examples show that in digital sampling there is an inherent ambiguity in the frequency at which the true fluctuations are occurring. This is reflected in their line spectra shown in Figure 1.27. In example (1), all the variance in the true spectrum (solid bar) is at 10 Hz, but the observed spectrum indicates a nonvarying signal, that is, no variance at all. In example (2), the variance in the true spectrum is at 9 Hz while the observed spectrum (open bar) shows variance at 1 Hz. The true and observed spectra in

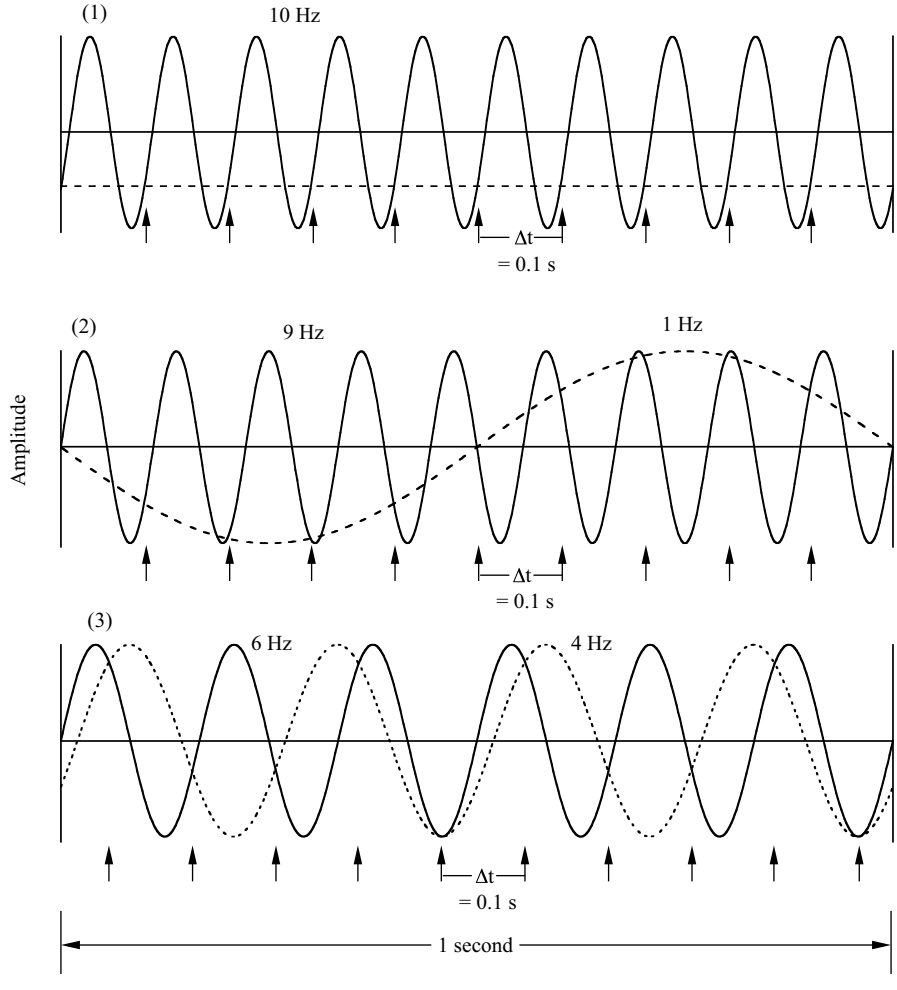



**Figure 1.26** Three examples of aliasing indicated by dashed lines. (1) A 10 Hz sinusoid is sampled as a constant signal. (2) A 9 Hz sinusoid is sampled as a 1 Hz sinusoid. (3) A 6 Hz sinusoid is sampled as a 4 Hz sinusoid.

example (3) follow the same pattern as above. Another way to look at aliasing is that *more than two observations per cycle are required to unambiguously define a sinusoid.* Otherwise, it can be interpreted also as a sinusoid of lower frequency.

The picture that emerges from Figure 1.27 is that the calculated value of variance is folded about 5 Hz. This frequency is called the *folding* or *Nyquist frequency*, $f_\upsilon$, the latter named after Harry Nyquist who did pioneering work in signal analysis (Nyquist, 1928). In general, the Nyquist frequency is determined by the sampling interval $\Delta t$, that is, $f_\upsilon = 1/(2\Delta t)$; in the example just discussed, $f_\upsilon = 5$ Hz.

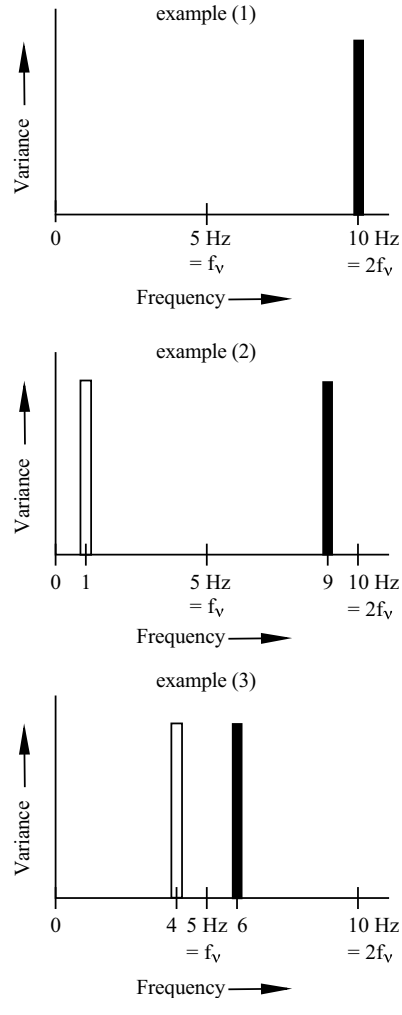



**Figure 1.27** The true and observed spectra corresponding to the three examples in Figure 1.26. In each example above, the true spectrum is indicated by a solid bar and the observed spectrum by an open bar.

Furthermore, it is easy to conclude that the spectrum will repeat itself at frequency intervals of $\pm i/\Delta t$, $i = 1, 2, \ldots$ . As an illustration, consider time series 1 given by

$$x_{1n} = \cos(2\pi f_m n\Delta t + \phi) \tag{1.39}$$

in which the series represents digital sampling of a sinusoid with frequency $f_m$, data point number n, and phase shift $\phi$ at intervals of $\Delta t$. Then define time series 2 as

$$x_{2n} = \cos[2\pi(f_m \pm i/\Delta t)\, n\Delta t + \phi]. \tag{1.40}$$

That is, time series 2 is digitally sampled in the same way as time series 1, except the frequency of the signal being sampled has been increased or decreased by integer multiples of twice the Nyquist frequency. Time series 2 can be expanded as a standard trigonometric angle-sum relation and then, because $\cos(2\pi in) = 1$ and $\sin(2\pi in) = 0$, reduced to

$$x_{2n} = \cos(2\pi f_m n\Delta t + \phi) \tag{1.41}$$

the same formula as for time series 1. As a consequence of digital sampling, the $x_{1n}$ and $x_{2n}$ time series are identical, despite the fact that the underlying signals being sampled are different. Thus the same variance will be computed at $f_m$, $f_m \pm 1/\Delta t$, $f_m \pm 2/\Delta t$, and so on. We notice that negative frequencies are allowed. This is purely a mathematical convenience. While employing a spectrum that has both positive and negative frequencies is especially helpful in understanding aliasing, the "two-sided" spectrum concept also will be used later in Sections 1.5.5–1.5.7. In these sections we will find that mathematical formulas for spectra are more compact and easier to interpret when they include variance at both positive and negative frequencies.

Figure 1.28a summarizes aliasing from a schematic viewpoint. The *aliased spectrum* extends across *all* negative and positive frequencies with the spectrum repeated at intervals of $2f_v = 1/\Delta t$. In the jargon of spectrum analysis, the band of

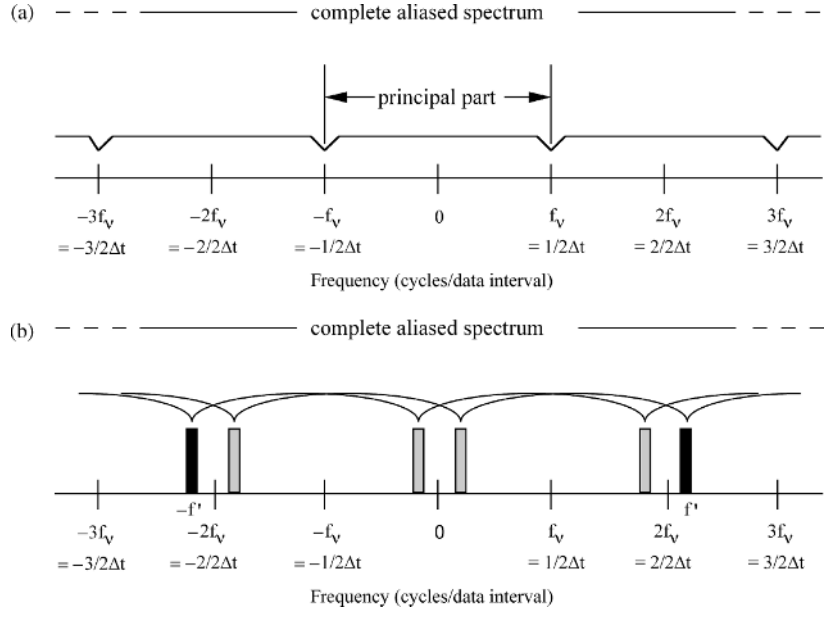



**Figure 1.28** (a) The complete aliased spectrum and its principal part. $f_v$ is the Nyquist frequency. (b) In a two-sided spectrum, one-half the variance appears at $f'$ and one-half at $-f'$ and each is aliased to frequencies $\pm i/\Delta t$ from $\pm f'$ where i is an integer.

frequencies between $-f_\upsilon$ and $+f_\upsilon$ is called the *principal part* of the aliased spectrum. In practice, only the principal part is needed because the spectrum is repeated every $2f_\upsilon$ or $1/\Delta t$; that is, the principal part contains all the variance in the time series. Further insight into aliasing can be obtained by considering an input sinusoid with frequency greater than $f_\upsilon$. Let us use the same frequency scale in Figure 1.28b and place the variance at $f'$ between $2f_\upsilon$ and $3f_\upsilon$. Because we are using both positive and negative frequencies, the total variance at $f'$ is split so that one-half the variance of the sinusoid is at $f'$ and one-half at $-f'$. Figure 1.28b shows the solid vertical bars; their sum is the total variance. From Equations 1.40 and 1.41 the variances will be distributed to the open bars at frequencies $\pm i/\Delta t$ relative to $\pm f'$ as shown by the lines and pointers. If the input frequency happens to be a multiple of $f_\upsilon$, no variance will appear at any frequency in accord with example (1) in Figure 1.27.

The repetition of the variance distribution in the principal part of the aliased spectrum in the remainder of the aliased spectrum is evident. If your preference is to deal only with variances at positive frequencies from 0 to $f_\upsilon$, simply fold the spectrum from 0 to $-f_\upsilon$ around the origin from 0 to $f_\upsilon$ and add the variances.

It should be clear by now that it is important to know whether you are working with a two-sided ($-f_\upsilon$ to $f_\upsilon$) spectrum or a one-sided spectrum (0 to $f_\upsilon$) to get the correct total variance of the time series. In the former the total variance resides between $-f_\upsilon$ and $f_\upsilon$ while in the latter between 0 and $f_\upsilon$. The variances at positive and negative frequencies in the former spectrum are one-half those in the latter. In the periodogram or Fourier analysis in previous sections, including Table 1.1, the harmonic variances were calculated at positive frequencies only, that is, from 0 to $f_\upsilon$.

Let us return to example (2) in Figure 1.26 to create a new analog time series that is the sum of the original 9 Hz sinusoid and the 1 Hz aliased sinusoid (dashed line) and sample it at $\Delta t = 0.1$ s as shown. The value at each sample point will be twice as large (negative or positive) as in the original 9 Hz sinusoid. As a consequence, the observed variance at 1 Hz in the principal part of the aliased spectrum will be four times larger than that with only the original 9 Hz sinusoid. With a 180° phase shift of either wave (flip either sinusoid about the horizontal axis), the value at each $\Delta t$ of the sum waveform is zero, and thus the variance is zero. In a word, this is why we have to be concerned with the effects of aliasing; variance at frequencies greater than $f_\upsilon$ will alter the true variance present at frequencies less than $f_\upsilon$ and produce an erroneous picture of variance. The seriousness of aliasing is in proportion to the ratio of the variance at frequencies outside the principal part to the true variance in the principal part.

Consideration of the potential for aliasing is critical to effective experiment design and proper analysis of results. To minimize aliasing, the sample rate should be such that practically all the variance will be at frequencies less than $1/(2\Delta t)$. If the general structure of the spectrum is unknown before sampling, experimentation may be required with different sampling rates to observe spectrum changes. If, for a given

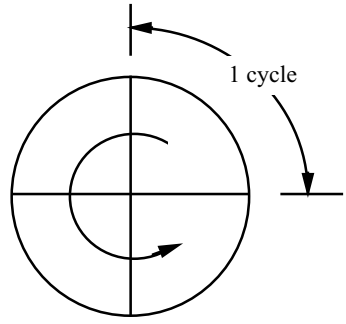



**Figure 1.29** A four-spoke wagon wheel.

sampling rate, the potential exists for serious aliasing and the sampling rate cannot be increased, then one must filter the variance at frequencies $>1/(2\Delta t)$ before sampling. There is no effect on aliased variance if filtering is performed after digitizing. That is, the analog signal must be filtered.

A visual example of aliasing as seen in Western cowboy movies is the familiar changing of the direction of rotation of a wagon wheel as the wagon increases its speed from rest. The digital sampling is done by the camera shutter opening and closing 24 times each second.

Consider the four-spoke wheel in Figure 1.29. One cycle means rotation of the wheel $^1/_4$ revolution. When the wheel turns slowly, we see a continued forward rotation of the set of four spokes because there are many samples (shutter openings and closings) for the small angular rotation. As the wheel rate of rotation increases, the angular separation between successive samples also increases until the separation reaches $\alpha = 45°$, or ½ cycle or $f_v = 0.5$ cycle/$\Delta t$, where $\Delta t = (1/24)$ s. This is the maximum observable frequency or rate of rotation of the wheel and is shown in Figure 1.30a. As the rate of rotation or actual frequency f increases beyond $f_v$, the observed frequency will be negative. This can be understood by referring to Figure 1.30b, keeping in mind that the sampling rate is fixed. Since $\alpha > 45°$, it is apparently easier for our brain to think the wheel has rotated not through angle $\alpha$

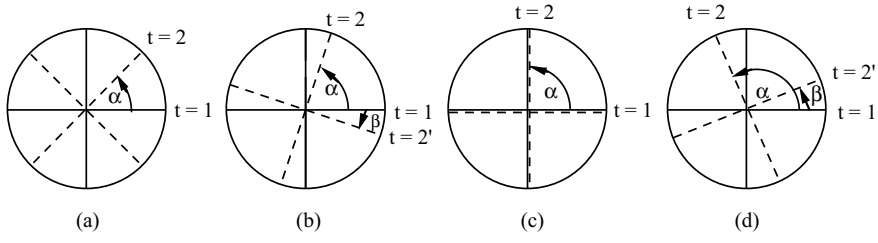



**Figure 1.30** Successive positions of the four-spoke wagon wheel at times $t = 1$ and $t = 2$ for an increasingly higher rate of rotation from case (a) to case (d).

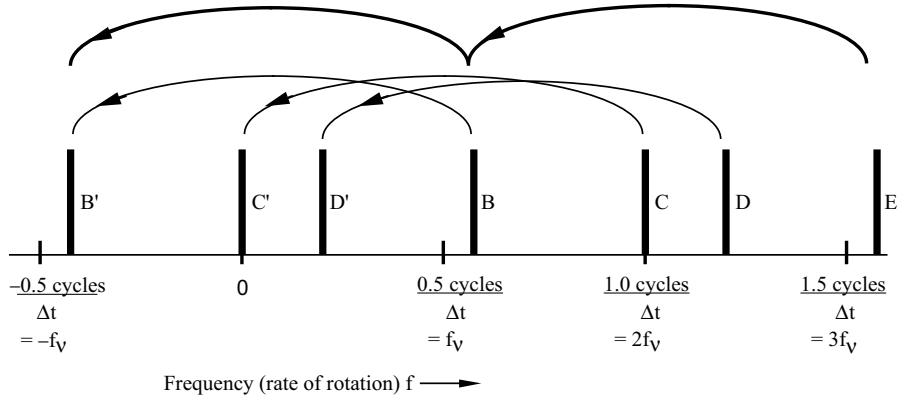



**Figure 1.31** The aliased spectrum for frequency of rotation greater than the Nyquist frequency.

from position t = 1 to position t = 2, but through smaller angle β from position t = 1 to position t = 2′. (Of course, the appearance of the wheel is identical at t = 2 and t = 2′.) In the spectrum in Figure 1.31 this corresponds to the variance at frequency B aliased to frequency B′. As the wheel rotates faster, $\alpha = 90^\circ$ and it appears that there is no motion (Figure 1.30c). Each spoke advances $^1/_4$ revolution or one cycle each time the shutter opens. In the spectrum this corresponds to variance at frequency C aliased to frequency C′ (the origin). From B′ to C′ it appears that the wheel rotation rate is decreasing, that is, becoming less negative.

An increasing forward rate of rotation occurs for $\alpha > 90^\circ$. For example, the variance at frequency D in Figure 1.31 is aliased to D′. As seen in Figure 1.30d, it is, again, apparently easier for our brain to accept rotation of the wheel through angle β at apparent time t = 2′ rather than larger angle α at real time t = 2. As the wheel rotates still faster, say at rate E in the spectrum, the rotation rate will be aliased back to frequency E′ = B′, as shown by the heavy lines. Thus, as the rate of rotation of the wheel increases from zero it reaches a maximum forward rate, which instantaneously becomes the maximum backward rate, which then decreases to zero and the cycle starts all over again – all because of digitally sampling an analog signal.

## 1.5.2 Spectrum windows

Consider the "idealized time series" of 36 consecutive values of temperature at Phoenix, Arizona, during fair weather shown in Figure 1.32a. Because Phoenix is located in the southwestern desert of the United States, we expect a strong diurnal variation in air temperature. It is idealized because other harmonics that would normally contribute to the daily temperature variation, such as the semi-diurnal

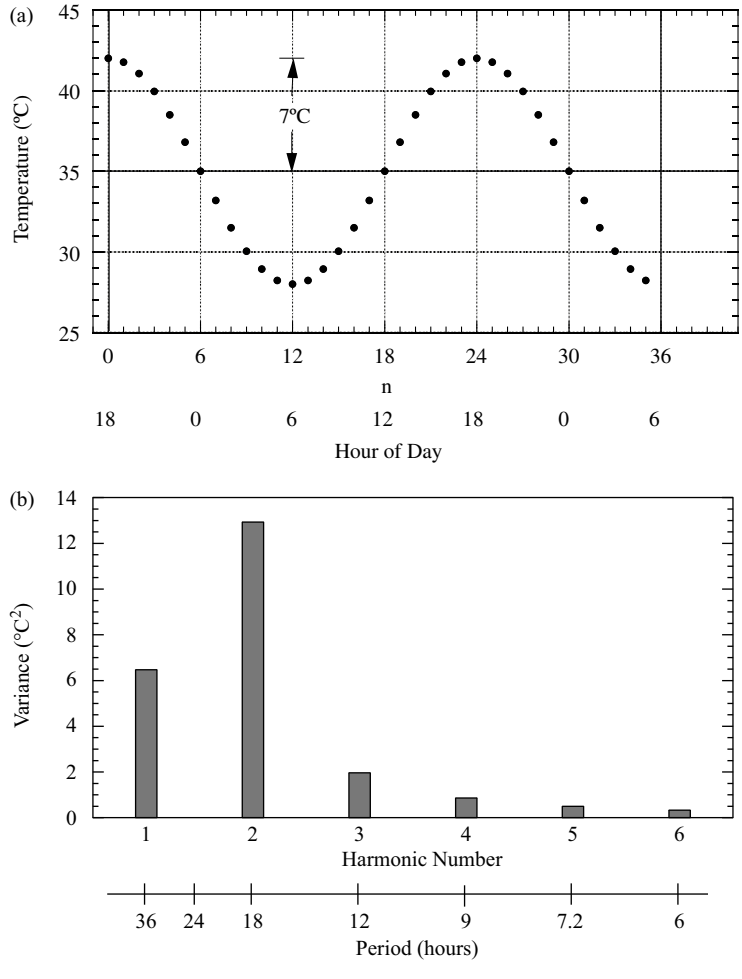



**Figure 1.32** (a) Idealized time series of temperature at Phoenix, Arizona, during fair weather in July. (b) Periodogram to harmonic 6 of idealized time series of temperature in (a).

component seen in Figure 1.15, have been ignored. In a periodogram of this time series, the diurnal variation would occur at a nonharmonic frequency midway between harmonic 1 (fundamental period = 36 hours) and harmonic 2 (period = 18 hours). The purpose of this section is to show how the input variance at a nonharmonic frequency (the daily cycle here) gets distributed to the harmonic frequencies.

Apart from the constant offset value of 35 °C, the time series in Figure 1.32a is given by the sinusoid

$$x_n = a\cos(\omega n\Delta t - \phi), \quad n = 0, 1, \ldots, N-1$$

where x represents temperature, $\Delta t = 1$ hour, amplitude $a = 7\,^{\circ}C$, data length $N = 36$ hours, phase angle $\phi = 0^{\circ}$, and angular frequency $\omega = 2\pi \times 1.5/N$. That is, there are 1.5 cycles over the 36-hour record. Changing to angular frequency is merely a convenience to reduce the number of symbols in each equation. Figure 1.32b is the resulting periodogram in the form of a line spectrum out to harmonic 6. Harmonics 1 and 2, which are adjacent to the frequency of the input wave, account for about 80% of the variance of $x_n$; the higher harmonics account for the remaining variance. The big question is: How did the variance from the input wave get distributed to the various harmonics?

To find the answer, we first substitute $x_n$ above into the equations for $A_m$ and $B_m$ (N even) in Table 1.1. Carrying out the summations is a tedious exercise in trigonometry, and the general procedure is shown in Appendix 1.D. We are really interested in the variance at harmonics, so the Fourier coefficients need to be squared according to $S_m^2 = (A_m^2 + B_m^2)/2$. This step is also given in Appendix 1.D, with the result that

$$S_m^2(\omega) = \frac{a^2}{2}\left\{ \frac{\sin^2[N(\omega+\omega_m)/2]}{N^2\sin^2[(\omega+\omega_m)/2]} + \frac{\sin^2[N(\omega-\omega_m)/2]}{N^2\sin^2[(\omega-\omega_m)/2]} + 2\cos[(N-1)\omega - 2\phi]\frac{\sin[N(\omega+\omega_m)/2]}{N\sin[(\omega+\omega_m)/2]} \times \frac{\sin[N(\omega-\omega_m)/2]}{N\sin[(\omega-\omega_m)/2]} \right\},$$

$$m \neq 0,\ N/2. \tag{1.42}$$

It is not necessary to work through Appendix 1.D at this time. It is important, though, to be able to properly interpret Equation 1.42, and there are two ways. In the first way, $S_m^2(\omega)$ gives the variance at harmonic numbers m, where $0 < m < N/2$, due to an input sinusoid of amplitude a at angular frequency $\omega$. Figure 1.32b is an example. In short, the equation shows how input variance $a^2/2$ is distributed among the harmonic frequencies.

The second way to interpret Equation 1.42 is to consider fixing m successively at 1, 2, 3, . . ., where $\omega_m = 2\pi m/N$, and, then, for each m, allow the input frequency $\omega$ to vary continuously over the range of frequencies in the spectrum. A plot of the ratio $S_m^2(\omega)/(a^2/2)$ for each m provides the "window" through which the spectrum is viewed at that harmonic for input variance at any $\omega$. Figure 1.33 shows the spectrum window, that is, the part of Equation 1.42 in braces, for harmonics $m = 1$, 2, and 3 for a cosine input ($\phi = 0^{\circ}$). The heavy solid line shows the location of the input wave at harmonic 1.5. To get the variance at harmonic 2, we multiply the variance of an integer number of cycles of the input wave, $a^2/2 = 24.5\,^{\circ}C^2$, by 0.5277, the amplitude of the window associated with harmonic 2 (i.e., center curve) at the input frequency. The product is the variance at harmonic 2 in Figure 1.32b. The product of 0.2641,

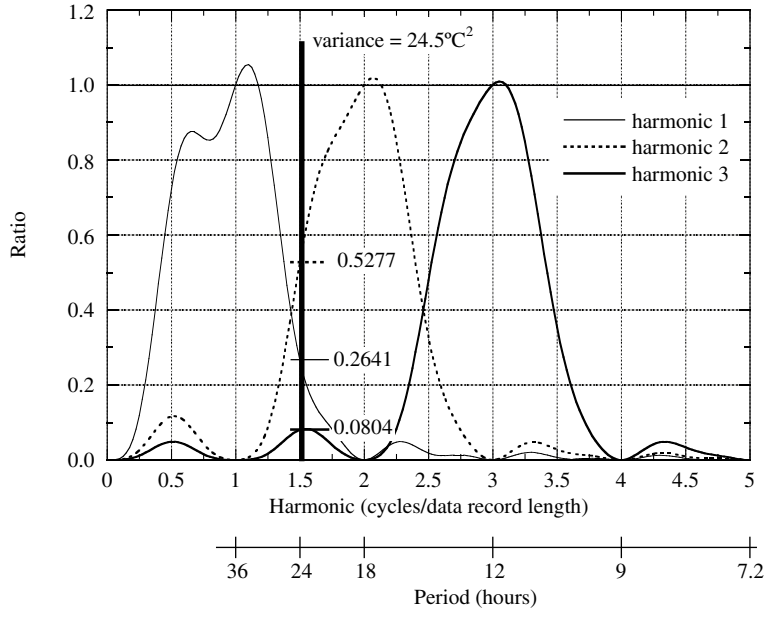



**Figure 1.33** Spectrum windows from Equation 1.42 centered at harmonics 1, 2, and 3 for a cosine wave input. The product of the variance (24.5 °C$^2$) in Figure 1.32a and the intersection of the spectrum windows yields the observed variance at harmonics 1, 2, and 3 in Figure 1.32b.

the amplitude of the window associated with harmonic 1 (i.e., left-hand curve) at the input frequency, and 24.5 °C$^2$ is the value of variance at harmonic 1 and, similarly, the variance at harmonic 3 is the product of 24.5 °C$^2$ and 0.0804 (right-hand curve). The windows for sine wave inputs ($\phi = 90°$) are shown in Figure 1.34; their individual shapes tend to be a reverse image of those in Figure 1.33. We conclude that the spectrum window depends on harmonic number and phase angle for a given N.

The spectrum window for the mean squared value standardized by the input variance is, from Equation 1.D.4,

$$A_0^2/(a^2/2) \;=\; 2\cos^2[(N-1)(\omega/2) - \phi]\frac{\sin^2(N\omega/2)}{N^2\sin^2(\omega/2)}. \qquad (1.43)$$

Figure 1.35 shows the spectrum windows for a cosine input ($\phi = 0°$) and a sine input ($\phi = 90°$). For the Phoenix example ($\phi = 0°$) the value of the window at harmonic 1.5 is 0.001543, so that with $a = 7$ °C, $A_0 = 0.1944$ °C (to get the mean of the time series in Figure 1.32a, add back 35 °C). That $A_0$, the mean of the time series, is not zero is because the Phoenix time series does not have an integer number of cycles. In fact, if $A_0$ is multiplied by $N = 36$, the number of data, the

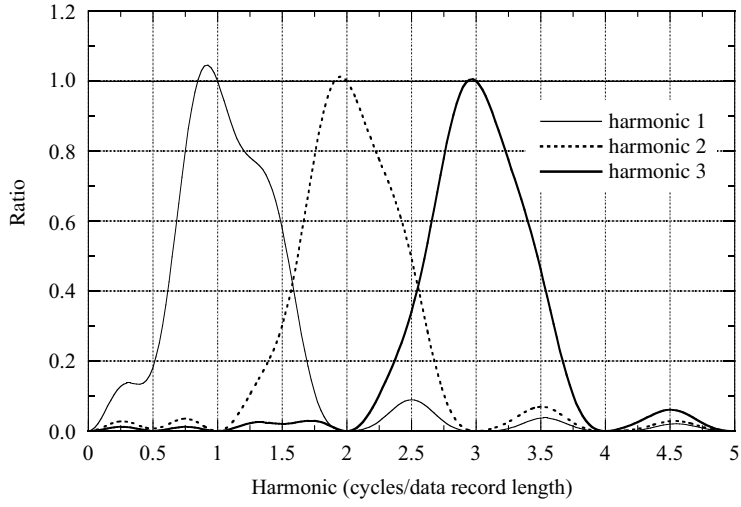



**Figure 1.34** Spectrum windows (Equation 1.42) centered at harmonics 1, 2, and 3 for a sine wave input.

result is 7 °C, the amplitude of the sinusoid. By matching positive departures from 35 °C with negative departures, we see that only one of the two maximum positive values of temperature has a negative equivalent. At the Nyquist frequency, where $m = N/2$, $S^2_{N/2}(\omega)$ can be obtained directly from Equation 1.42 by dividing the right-hand side by two.

In general, the window shape is dependent on harmonic number, the number of data, and the phase angle of the input. When the number of data in a sample is

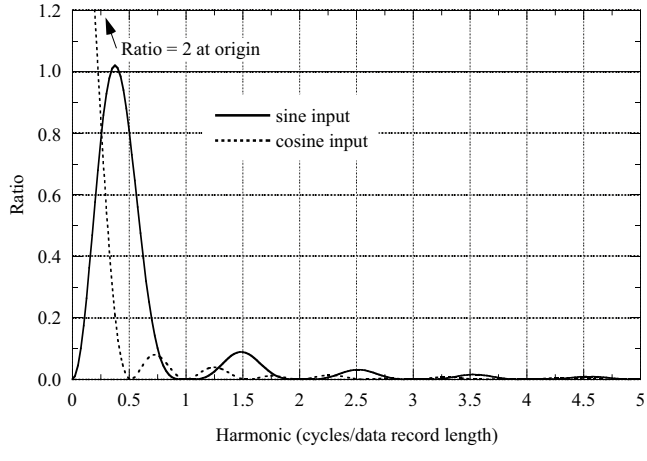



**Figure 1.35** Spectrum windows (Equation 1.43) at the 0-th harmonic for sine and cosine inputs.

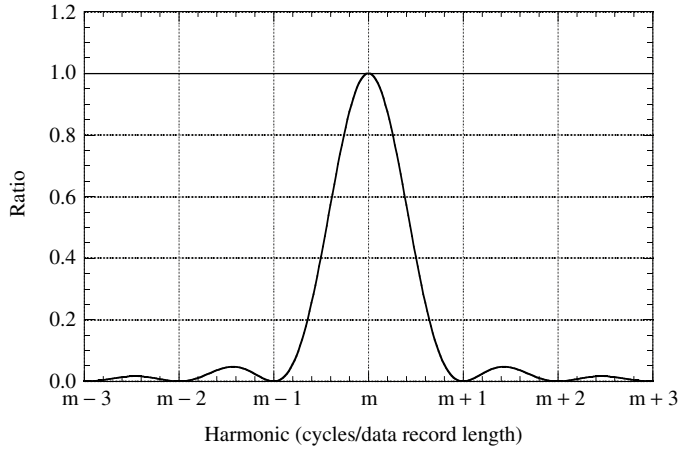



**Figure 1.36** The spectrum window (Equation 1.44) at general harmonic m, when m is away from the low and high frequency ends of the periodogram.

$N \approx 100$ or higher, the main and adjacent lobes at the interior harmonics can be quite accurately modeled by simplifying Equation 1.42 to

$$S_m^2(\omega) = \frac{a^2}{2} \frac{\sin^2[N(\omega - \omega_m)/2]}{[N(\omega - \omega_m)/2]^2} \tag{1.44}$$

which is dependent only on the number of data and the difference between the input frequency and the harmonic where the calculation is made. The maximum error in using Equation 1.44 in place of Equation 1.42 for interior harmonics is about $\pm 3\%$ for $N = 100$. This formula is the square of the familiar "diffraction function" common in optics and is plotted in Figure 1.36.

Assuming N is sufficiently large, we can think of the variance computed at a given harmonic frequency as the integral over the frequency range in the spectrum of a weight function (Equation 1.44) centered at that harmonic times an underlying, but unknown, spectrum. This process is repeated at all harmonics and results in variance "leaking" from one part of the spectrum to other parts of the spectrum. The variance observed at a particular harmonic does not necessarily mean that the data contain a pure tone at that harmonic. To find the variance of the Phoenix diurnal temperature cycle in a periodogram, a record length that is a multiple of 24 hours should be selected.

### 1.5.3 Detecting a periodic signal by averaging spectra

If we were to average together periodograms of equal length realizations from the same random process, harmonic by harmonic, we expect the averaged periodogram would be smoother than any individual periodogram. If a deterministic signal is

present, its magnitude should not be affected by averaging. In this section we use the idea of averaging to investigate the particular problem of detecting a sinusoid embedded in white noise when multiple realizations are available.

The average of the two periodogram random variables $C_1(f_m)$ and $C_2(f_m)$ from realizations of equal length N (even) of a white noise process is

$$\overline{C}(f_m) = \frac{C_1(f_m) + C_2(f_m)}{2}$$

so that

$$\frac{\overline{C}(f_m)}{\Gamma(f_m)} = \frac{1}{4}\left(\chi_2^2 + \chi_2^2\right) = \frac{\chi_4^2}{4}$$

or, in general, averaging u spectra, $u = 1, 2, \ldots$, yields

$$\frac{\overline{C}_u(f_m)}{\Gamma(f_m)} = \frac{\chi_{2u}^2}{2u} \tag{1.45a}$$

for the interior harmonics, and

$$\frac{\overline{C}_u(f_m)}{\Gamma(f_m)} = \frac{\chi_u^2}{u} \tag{1.45b}$$

for the 0-th and Nyquist frequencies, that is, $f_0$ and $f_\nu = f_{N/2}$, respectively.

In parallel with Equation 1.30 we can use Equation 1.45a to determine the confidence interval for the population variance at the interior harmonics given the sample variance. Thus,

$$\Pr\left\{\frac{\overline{C}_u(f_m)}{\chi_{2u}^2(1-\alpha/2)/2u} \le \Gamma(f_m) \le \frac{\overline{C}_u(f_m)}{\chi_{2u}^2(\alpha/2)/2u}\right\} = 1 - \alpha, \quad f_m \ne f_0,\ f_\nu$$

where $\alpha$ is the significance level. The interval between

$$2u\frac{\overline{C}_u(f_m)}{\chi_{2u}^2(1-\alpha/2)} \quad \text{and} \quad 2u\frac{\overline{C}_u(f_m)}{\chi_{2u}^2(\alpha/2)}$$

is the $100(1-\alpha)\%$ confidence interval for $\Gamma(f_m)$. By taking logarithms of the lower and upper limits of the confidence interval for $\log \Gamma(f_m)$, they become, respectively,

$$\log\overline{C}_u(f_m) + \log\left[\frac{2u}{\chi_{2u}^2(1-\alpha/2)}\right] \quad \text{and} \quad \log\overline{C}_u(f_m) + \log\left[\frac{2u}{\chi_{2u}^2(\alpha/2)}\right].$$

As noted in Section 1.4.4, the width of the confidence interval will remain constant regardless of frequency. For example, consider averaging three Fourier spectra so that $u = 3$. The 95% ($\alpha = 0.05$) confidence interval for $\log \Gamma(f_m)$, $f_m \neq f_0, f_v$, extends from $\log \overline{C}(f_m) + \log\ (0.415)$ to $\log \overline{C}(f_m) + \log\ (4.85)$. In a similar manner, Equation 1.45b can be used to find the confidence interval for the population variance at the exterior harmonics.

If there are deterministic components in the spectrum, they will remain unchanged by spectrum averaging. Looking at this method in another way, averaging spectra can be used to detect deterministic components.

Consider the following computer simulation of

$$x_n = b \sin(2\pi f n - \phi) + \varepsilon_n, \quad n = 1, 2, \ldots, N$$

where $b = \sqrt{2}$, $N = 32$, $f = 6.25/N$, $\phi$ is phase angle ($0 \leq \phi < 360^\circ$), and $\varepsilon_n$ is white noise with population variance $\sigma^2 = 5$. If the signal-to-noise variance ratio (SNR) is defined to be

$$\text{SNR} = \frac{\dfrac{b^2}{2}}{\dfrac{\sigma^2}{(N-1)/2}} \tag{1.46}$$

the ratio of the variance of the sinusoid to the white noise variance at an interior harmonic frequency, its value is 3.1.

Each realization of length 32 comprises computer generated normal white noise, $\varepsilon_n$, added to the sinusoid with a different value of $\phi$. If we make the null hypothesis that the variance spectrum comes from a white noise process and rewrite the two-sided equation for the confidence interval for the population variance in the form for a one-sided test only, namely,

$$\Pr\left\{0 \leq \overline{C}_u(f_m) \leq \frac{\chi^2_{2u}(1-\alpha)\Gamma(f_m)}{2u}\right\} = 1 - \alpha \tag{1.47}$$

we can use this formula to obtain the *a priori* upper confidence limit for the distribution of the observed harmonic variances. That we are dealing with only the upper confidence limit is because we are interested in the possible existence of a sinusoid, the indication of which is a peak in the spectrum. Figure 1.37 shows the spectra for six realizations of 32 data each for harmonics 3–9. The lower dashed line in each realization is the average variance for the interior harmonics and the arrow indicates the input frequency of the sinusoid. The upper dashed line shows the 95%

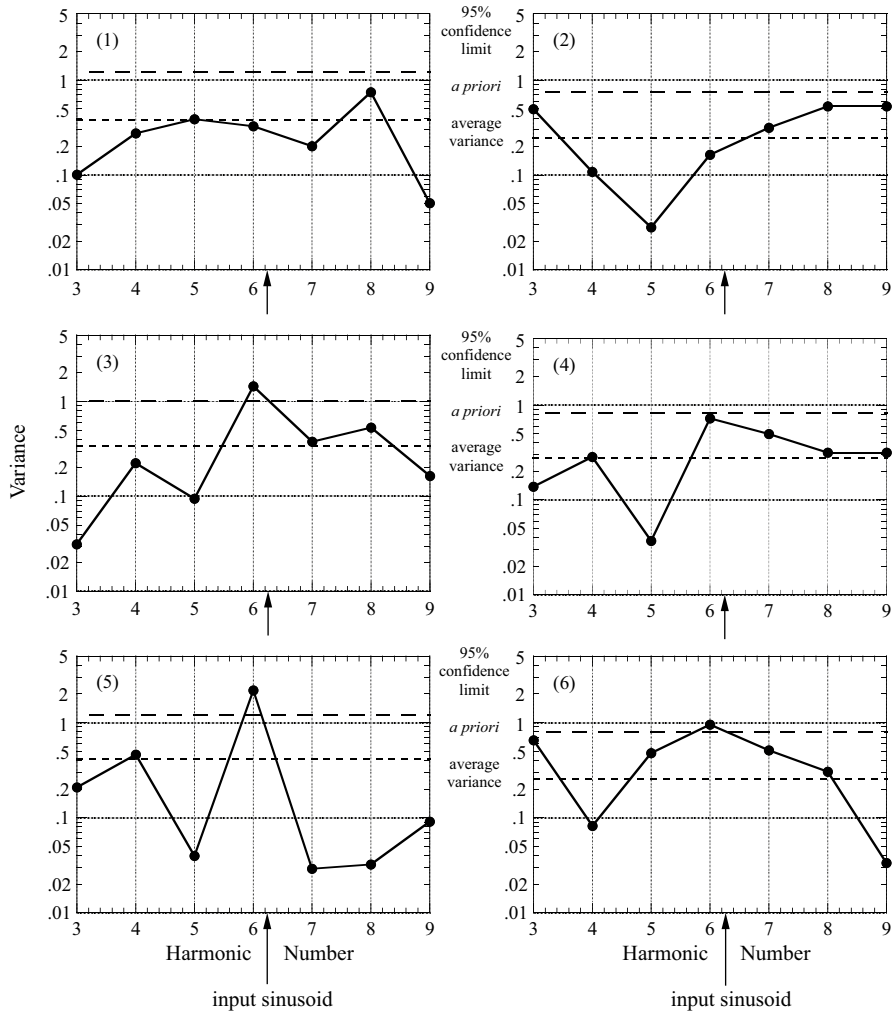



**Figure 1.37** Periodograms of six realizations of a sinusoid plus white noise with signal-to-noise ratio (SNR) = 3.1 (solid line). The input sinusoid is 6.25 cycles over data length N = 32. The upper dashed line is the 95% *a priori* confidence limit. The lower dashed line is the average variance of the 15 interior harmonics for N = 32.

upper confidence limit for the observed variance of each individual realization (u = 1) and is computed from:

$$\frac{\chi_2^2(0.95)}{2}\Gamma(f_m)$$

where $\Gamma(f_m)$ is estimated by summing all of the harmonic variances (signal plus noise) of a realization and dividing by (N − 1)/2. Realizations (1) through (6) of

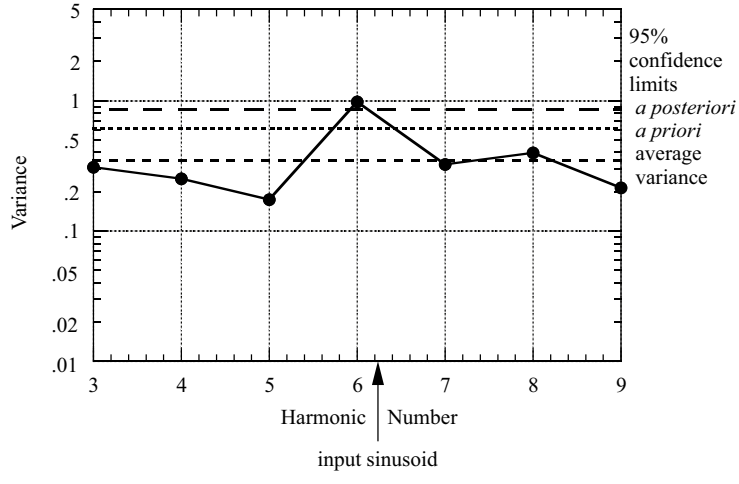



**Figure 1.38** Average of the six periodgrams in Figure 1.37 (solid line). The upper two dashed lines are the *a priori* and *a posteriori* 95% confidence limits, the bottom dashed line is the average variance across all interior harmonics.

white noise yielded variances 6.72, 4.12, 5.60, 4.92, 6.77, and 4.25, respectively, compared to the population variance $\sigma^2 = 5$.

Figure 1.37 shows that the spectra vary considerably from one sample to the next, as expected, and that with a signal-to-noise ratio slightly greater than three, one may very well not detect the sinusoid using a single realization. Figure 1.38 shows the results of averaging the six spectra in Figure 1.37. The *a priori* 95% upper confidence limit is computed from $\chi^2_{12}(0.95)\Gamma(f_m)/12 = 0.61$, where the estimate of $\Gamma(f_m) = 0.348$ is obtained by averaging the estimates from all six realizations. The upper confidence limit occurs at a lower value of variance and closer to the mean than for any single periodogram. The result is the variance at harmonic 6 now clearly stands out.

Let us assume that the six realizations are actual data. Whether we would conclude that there is a significant oscillation at or near harmonic 6 depends on what we know from physical considerations may be occurring there and the likelihood the peak could have occurred by chance. With regard to the latter, we expect to observe, on average, 1 in 20 harmonic variances that exceed the *a priori* upper confidence limit. It is appropriate then to find the 95% *a posteriori* upper confidence limit, so that there is only a 5% chance that any one or more of the 15 harmonic variances will exceed this limit.

Following the procedure in Section 1.4.6, we divide $\alpha = 0.05$ by 15, the number of interior harmonics, the result being 0.0033. Next we find (estimate) from a chi-square table the abscissa of a $\chi^2_{12}$ distribution such that the area to the left is 0.9967. Thus, $\chi^2_{12}(0.9967)/12 = 2.46$. Given the above estimate of $\Gamma(f_m)$, the 95% *a posteriori* upper confidence limit is 0.86. Assume further that there is physical

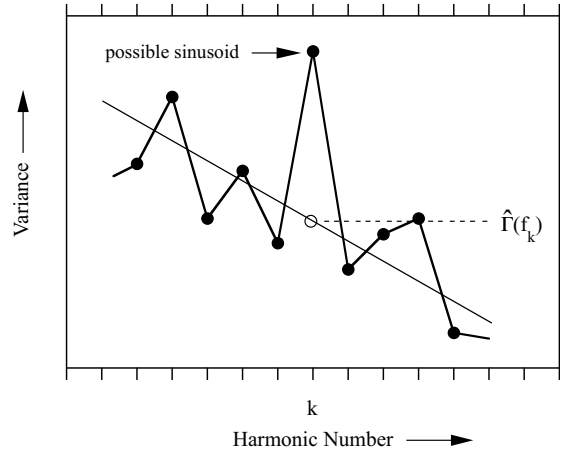



**Figure 1.39** When the white noise null hypothesis is inadequate, it may be advantageous to fit a smooth curve or, as shown above, a straight line to the harmonic variances in the neighborhood of the possible sinusoid. A white noise test can be applied to the departures from the fitted curve.

evidence for a sinusoid close to harmonic 6. Since the white noise null hypothesis has been rejected, the variance in the sinusoid should have been removed from the estimate of $\Gamma(f_m)$, the consequence being the correct *a posteriori* upper confidence limit would be even lower. Therefore, the test is conservative in that if the null hypothesis is rejected at some value of $\alpha$ using the preceding method, the actual value of $\alpha$ is even less.

When the spectrum of random data is not white noise, the estimation of $\Gamma(f_m)$ in Equation 1.47 must be made at the frequency at which a sinusoid is suspected. One way to do this is to apply a straight-line fit to the surrounding periodogram values and use the value of the straight line at $f_m$ as the estimate of $\Gamma(f_m)$, as illustrated in Figure 1.39. Confidence limits for the observed variance then can be computed if the departures from the straight line are suitably white. Another way is to model the underlying stochastic process using an appropriately smooth function and determine the white noise confidence limits with respect to the departures from the model. Crowley, Duchon, and Rhi (1986) show an example of the latter in which they searched for potential solar cycles in annual varve data.

We conclude this section by saying that if one were fortunate enough to have six or more realizations from a random process in which there is a deterministic sinusoid with the signal-to-noise ratio of three or greater, there is a good chance of detecting the sinusoid in the averaged periodogram. The determination of the minimum SNR required for detection and statistical confirmation of a sinusoid, in general, is complicated because the outcome depends on its proximity to the nearest harmonic and the spectrum of the noise. For example, if the sinusoid is located between

harmonics, the spectrum window will distribute its variance to a number of harmonics (Section 1.5.2). If the noise spectrum falls or rises sharply where it is located, we expect that the SNR would have to be very large in order to detect a pure sinusoid.

### 1.5.4 Effect of data length on detecting a periodic component

As noted in Section 1.4.5, the Fourier spectrum of random data can be viewed as an "unstable" spectrum because increasing the data length does not reduce the variability of variance computed at any harmonic from one realization to the next. Rather, an increase in data length results in an increase in frequency resolution; if the data length is doubled, the bandwidth or frequency separation between adjacent harmonic frequencies is halved. The dof for each approximately independent variance estimate is still two; as a result, the statistical distribution of each variance ratio $C(f_m)/\Gamma(f_m)$ remains $\chi_2^2/2$.

As in the previous section, consider a sinusoidal signal to which is added white noise. In this case let the signal have an integer number m cycles. The variance at the harmonic of the sinusoid is the sum of three terms: the variance of the sinusoid, the variance of the random component, and the covariance between the two harmonics – one from the sinusoid, the other from the noise. This can be understood by considering the sample variance of the sum of two sinusoids $x_{1n} = a_1 \cos(2\pi mn/N + \phi_1)$ and $x_{2n} = a_2 \cos(2\pi mn/N + \phi_2)$, where $x_{1n}$ is the sinusoid and $x_{2n}$ the Fourier component of random noise at the frequency of the sinusoid. It can be shown that the variance of the sum

$$S^2(x_{1n} + x_{2n}) = (1/N) \sum_{n=1}^{N} (x_{1n} + x_{2n})^2$$

reduces to

$$S^2(x_{1n} + x_{2n}) = \left(a_1^2 + a_2^2\right)/2 + a_1 a_2 \cos(\phi_1 - \phi_2). \tag{1.48}$$

A convenient way to prove Equation 1.48 is to express $x_{1n}$ and $x_{2n}$ in terms of complex exponentials using Euler's formula and then apply the summation procedure in Appendix 1.B – similar to the way it was applied it in Section 1.2.2. We see from Equation 1.48 that, depending on the magnitudes of $a_2$ and $\phi_2$ in a particular realization of noise, the variance at the harmonic of the sinusoidal signal could be larger or smaller than the variance of the sinusoid itself. In an expected sense there is no preference for the covariance term in Equation 1.48 to be either positive or negative because $a_2$ has no sign preference and is uncorrelated with $\phi_2$.

The important point to remember is that if the data length is doubled, the white noise variance will be distributed over twice as many frequencies and, on average, reduced by a factor of two at the frequency of the sinusoidal signal. The variance of the sinusoid will remain unchanged but occurs at twice the original harmonic number. Of course, there will be a proportional reduction in white noise at any harmonic for other integer multiple increases in data length.

If the sinusoid is not at a harmonic frequency, the most likely case in practice, the results are more complex, but in the expected sense can be qualitatively inferred from multiplying the spectrum window with the sinusoidal input variance, as discussed in Section 1.5.2. For example, if the frequency of the sinusoid lies midway between adjacent harmonics in the periodogram, the variance at the same frequency after doubling the data length will contain all the variance of the sinusoid. Considering only the variance of the sinusoid, its value will be more than twice the values at the adjacent harmonics in the original periodogram because the spectrum window spreads the variance of the sinusoid to *all* harmonics, not just the two adjacent harmonics. If the sinusoid is nearer to one or the other adjacent harmonics in the original periodogram, the variance will be mostly contained in the two harmonic frequencies that surround it in the spectrum for the case of twice the original data length. In summary, increasing the length of a time series that is stationary and contains a deterministic component results in improved ability to distinguish the variance of the deterministic component from the surrounding harmonic variances. At an appropriate stage, one can test for statistical significance of a possible sinusoid using one of the approaches given in Section 1.4.6.

As an example, consider the simulated time series $b\sin(2\pi ft - \pi/4)$, where $b = \sqrt{2}$ and $t = 1, 2, \ldots, N$, to which white noise is added. The white noise has a population variance $\sigma^2 = 5.0$. In Figure 1.40, curve (a) shows the distribution of percentage of total variance for harmonic numbers three through nine for the sinusoid plus a realization of white noise for $N = 32$ and $f = 6/N$. The signal-to-noise ratio, as defined by Equation 1.46, is 3.1. Percentage of total variance is used as the ordinate so that comparisons between realizations are not affected by varying amounts of total sample variance. Curve (a) indicates that the periodogram estimates vary considerably and that there would be no reason to expect an input sinusoid at harmonic 6. The periodogram of the white noise by itself (not presented) shows that, by chance, the value of variance at harmonic 6 is small compared to the adjacent variances and the phase angle between the component of white noise at harmonic 6 and the sinusoidal signal is about 30°. The combination of these two factors yields the value shown at harmonic 6 as dictated by Equation 1.48.

Curve (b) in Figure 1.40 results from extending the sinusoid and the realization of white noise associated with curve (a) to double their lengths. That is, the seed for the white noise random number generator was the same for curves (a) and (b). As expected, the periodogram values are generally reduced in magnitude as the

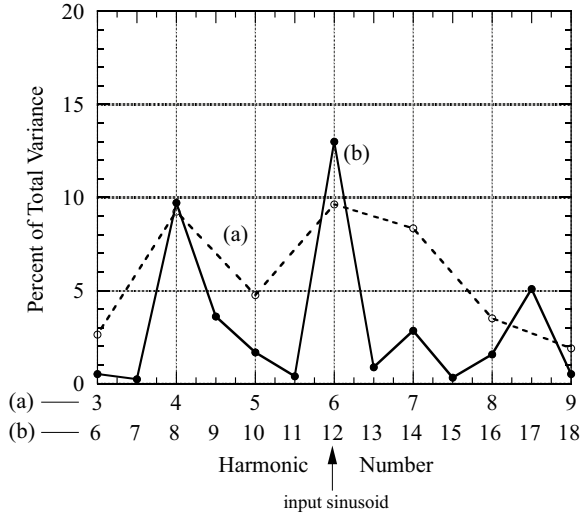



**Figure 1.40** Partial periodograms for sinusoidal input plus white noise. (a) For signal-to-noise ratio (SNR) = 3.1 and data length N = 32. (b) For SNR = 6.2 and N = 64.

white noise variance is now distributed across twice the number of harmonics in (a). The presence of an input sinusoid is more in evidence with a SNR = 6.2, double that in (a) (ignoring the –1 contribution in the denominator of Equation 1.46). Even with this higher SNR, there is still considerable variance at harmonic 8.

Figure 1.41 is similar to Figure 1.40 with two exceptions: the input sinusoid is at harmonic 6.25 in the 32-point data set and there is a second doubling of the initial data length to yield a 128-point data set. The highest peak in curve (a) occurs at harmonic 7 with a comparatively small value at harmonic 6. By chance, the component of white noise at harmonic 6 is nearly out of phase ($\sim 170^\circ$) with the component of the input sinusoid at harmonic 6 (negative covariance term). At the same time, there is only about a $70^\circ$ phase difference between the component of white noise at harmonic 7 and the component of the input sinusoid at harmonic 7 (positive covariance term). The result is that the contribution of the input sinusoid to the variance at harmonic 7 is about 1.5 times greater than that at harmonic 6. This, coupled with the much larger variance of the random component at harmonic 7 than at harmonic 6, yields the magnitudes shown. Curve (b) shows the spectrum when the data are extended to twice the original length so that the SNR is 6.2. Generally, the periodogram values are less than those in (a), as anticipated. The frequency of the input sinusoid now lies midway between adjacent harmonics. The phase differences between the components of white noise and the input sinusoid at harmonics 12 and 13 and the magnitudes of the components account for the similar percentages of total variance at these harmonics. Other large percentages of total variance occur at harmonics 8, 14, and 17, due mainly to the strength of the

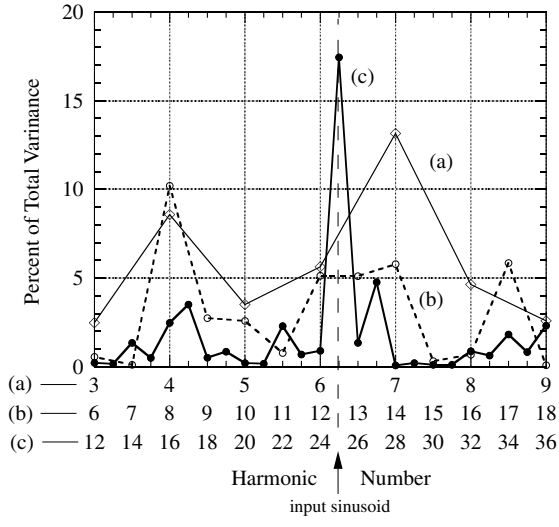



**Figure 1.41** Partial periodograms for sinusoidal input plus white noise. (a) For signal-to-noise ratio (SNR) = 3.1 and data length N = 32. (b) For SNR = 6.2 and N = 64. (c) For SNR = 12.4 and N = 128.

component of white noise as leakage of variance from the input sinusoid diminishes with distance from harmonic 12.5.

In curve (c) in Figure 1.41, the data have been extended to four times the original length, the SNR thus being 12.4. From Table 1.7, which applies to curve (c), the phase difference between the input sinusoid and the sinusoid of noise at harmonic 25 is about 95°. In the "worst case" that could have arisen, the phase angle difference would be 180°, with the result that the ordinate would have been 9.4%. This figure is not too different from the values of 7.3% and 8.6% that were found at harmonics 2 and 10 (not shown), in which situation there would be little evidence for the deterministic component at harmonic 25. The figure of 9.4% can be calculated from values in the total variance and variance columns of Table 1.7 and Equation 1.48. The calculation is a good exercise to demonstrate understanding of how periodogram variances can change due to phase angle differences when two sinusoids are summed.

Now we can adapt Equation 1.47 to the ordinate in Figure 1.41 in order to find the $100(1-\alpha)\%$ *a priori* confidence interval. Dividing each term by the total variance, such that the confidence interval is expressed as a percentage, yields

$$\Pr\left\{0 \le \frac{C(f_m)}{5.882} \le \frac{\chi_2^2(1-\alpha)\Gamma(f_m)}{2 \times 5.882}\right\} = 1-\alpha. \tag{1.49}$$

The white noise estimate for $\Gamma(f_m)$ is $5.882/((N-1)/2) = 2 \times 5.882/127 = 0.0926$. The resulting upper limit of the 95% *a priori* confidence interval is 4.7%. Among the

**Table 1.7** Statistical properties of signal $\sqrt{2}\sin(2\pi\, 25\, n/128 - \pi/4)$ and a realization of a time series of 128 values of white noise.

| Time series | Total variance | Variance at harmonic 25 | Percentage of total variance | Phase angle (degrees) |
|---|---|---|---|---|
| signal | 1.000 | 1.000 | 100.00 | 135.0 |
| noise | 4.936 | 0.084 | 1.70 | 39.6 |
| signal + noise | 5.882 | 1.030 | 17.51 | 118.5 |

63 interior harmonics there are four with values outside this limit. These include harmonic 25 for both the observed (95°) and "worst case" (180°) phase differences. On average, one would expect about three ($63 \times 0.05$) rejects under the white noise null hypothesis. Using only the *a priori* confidence limit, it is unclear whether to reject or not reject the white noise null hypothesis that the data set comes from a white noise process. Accordingly, we calculate the 95% *a posteriori* confidence interval from Equation 1.49 after replacing the argument of $\chi_2^2$ by $1 - \alpha/63 = 0.99921$. Integrating Equation 1.36 between 0 and the upper limit of integration results in $\chi_2^2(0.99921) = 14.278$. The upper limit of the *a posteriori* confidence interval is, therefore, 11.2%. The only harmonic whose percentage of total variance exceeds this limit is that at harmonic 25 for the 95° phase angle difference case. There are no harmonics with values exceeding this limit for the 180° "worst case."

We conclude that even with a SNR as large as 12, it can be difficult, in general, to not only detect a sinusoidal signal in the presence of white noise in a given realization, but to show also that it is statistically significant. Factors that contribute to this difficulty are (i) the occurrence of the sinusoid between harmonics and the attendant spectrum window effects and (ii) the chance occurrence of a combination of amplitude and phase angle of random noise at the same harmonic as the signal that significantly cancels the signal variance.

To briefly summarize Sections 1.5.3 and 1.5.4, we can state that the detection of a sinusoidal signal embedded in noise will be enhanced by either increasing the data length (with resultant increase in the SNR) or averaging a number of periodograms (with resultant narrowing of the spectrum confidence interval). Increasing the data length forces the periodic component to be closer to a harmonic frequency.

### 1.5.5 Complex representation of Fourier series

The most compact expression of a Fourier synthesis is that written in complex exponential form. The purpose of this section is to develop complex exponential forms for Fourier synthesis and analysis producing what is called a *Fourier transform pair*. The periodogram is then expressed in terms of complex coefficients. We shall

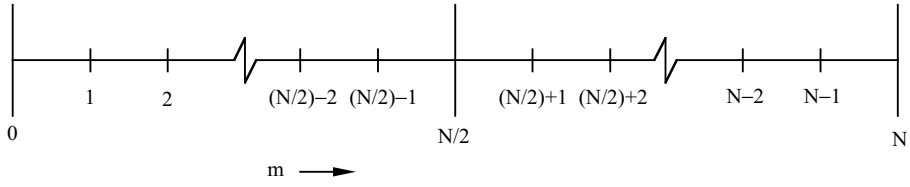



**Figure 1.42** Fourier coefficents $A_m$ and $B_m$ computed at harmonics $(N/2)+1$ to $N-1$ can be exactly matched to Fourier coefficients computed at harmonics 1 to $(N/2)-1$ with appropriate change in sign. N is even.

see that one of the features of the periodogram in complex form is that, with a couple exceptions, the Fourier coefficients are one-half the magnitude given in Table 1.1. The exceptions occur at the zero harmonic (N even or odd) and the Nyquist frequency (N even), in which cases there is no change of magnitude. The compactness also can result in an amplitude spectrum that includes both positive and negative harmonics (or frequencies), as described in Section 1.5.1 and Figure 1.28.

Figure 1.42 shows the locations of the harmonic coefficients $A_m$ and $B_m$ along a harmonic axis that has been extended to twice its usual length. If the range of m in the synthesis formula

$$x_n = A_0 + \sum_{m=1}^{\frac{N}{2}-1} \left( A_m \cos\frac{2\pi mn}{N} + B_m \sin\frac{2\pi mn}{N} \right) + A_{N/2} \cos\pi n \qquad (1.50)$$

from Table 1.1 for N even were to be extended beyond N/2, what would happen to the values of $A_m$, $B_m$, $\cos\frac{2\pi mn}{N}$, and $\sin\frac{2\pi mn}{N}$? Using trigonometric identities for the sum and difference of two angles, the results for the sine terms will be

$$\sin\frac{2\pi\left(\frac{N}{2}+m\right)n}{N} = \sin(\pi n)\cos(2\pi mn/N) + \cos(\pi n)\sin(2\pi mn/N) \qquad (1.51)$$

and

$$\sin\frac{2\pi\left(\frac{N}{2}-m\right)n}{N} = \sin(\pi n)\cos(2\pi mn/N) - \cos(\pi n)\sin(2\pi mn/N). \qquad (1.52)$$

Because the first term on the right-hand side in both equations is always zero and the second term is the same except for sign, it follows that

$$\sin\frac{2\pi\left(\frac{N}{2}+m\right)n}{N} = -\sin\frac{2\pi\left(\frac{N}{2}-m\right)n}{N}. \qquad (1.53)$$

In a similar manner it can be shown that

$$\cos\frac{2\pi\left(\frac{N}{2}+m\right)n}{N} = \cos\frac{2\pi\left(\frac{N}{2}-m\right)n}{N}. \tag{1.54}$$

Thus, the sine terms show odd symmetry about harmonic N/2 and the cosine terms show even symmetry. For N odd, the location of the Nyquist frequency is between adjacent harmonics that surround the Nyquist frequency. Nevertheless, the same pattern of symmetry about the Nyquist frequency holds for N odd as for N even.

Continuing with N even, since $A_m$ and $B_m$ both involve the cosine and sine terms above,

$$A_{\frac{N}{2}+m} = A_{\frac{N}{2}-m} \quad \text{and} \quad B_{\frac{N}{2}+m} = -B_{\frac{N}{2}-m}. \tag{1.55}$$

Thus, noting the even and odd symmetry of the Fourier cosine and sine coefficients, respectively, $x_n$ can be written

$$x_n = \sum_{m=0}^{N-1}\left(A'_m \cos\frac{2\pi mn}{N} + B'_m \sin\frac{2\pi mn}{N}\right) \tag{1.56}$$

where $A'_m = A_m/2$ and $B'_m = B_m/2$, except $A'_0 = A_0$ (N even or odd) and $A'_{N/2} = A_{N/2}$ (N even). Fourier coefficients $A_m$ and $B_m$ are the original coefficients defined in Table 1.1. From this point forward, whenever a primed Fourier coefficient is observed, it means that its value is one-half the value of an unprimed coefficient, except as noted. Primed coefficients have been used already in Appendix 1.D.

Before rewriting the expression above in terms of complex numbers, let us briefly review what we mean by a complex number. A complex number is given by

$$z = x + iy$$

and its complex conjugate by

$$\overset{*}{z} = x - iy$$

where x and y are real numbers and i the *imaginary unit* defined by

$$i = \sqrt{-1}.$$

The real number x is called the *real part* of z (or its conjugate) and the real number y is called the *imaginary part* of z (or its conjugate). (Note that the x notation here is distinct from notation $x_n$ for the time series above.) The complex number z can be

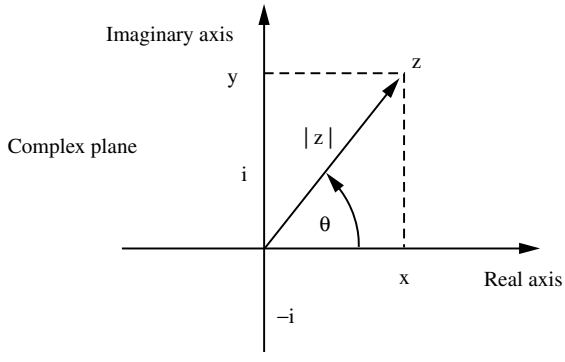



**Figure 1.43** The complex plane.

easily interpreted as a vector in the complex plane shown in Figure 1.43 extending from the origin to the intersection of x and y.

The length of vector z is given by

$$|z| = \sqrt{x^2 + y^2}$$

and its direction by

$$\theta = \tan^{-1}\left(\frac{y}{x}\right).$$

Since

$$x = |z|\cos\theta \quad \text{and} \quad y = |z|\sin\theta$$

it is apparent that $z = x + iy$ can be written in the equivalent form

$$z = |z|\cos\theta + i|z|\sin\theta = |z|(\cos\theta + i\sin\theta)$$

which, from Euler's formula, can be written

$$z = |z|\, e^{i\theta}.$$

This is called the *polar* or *trigonometric* form of a complex number.

As for now, we represent the Fourier coefficients using complex numbers in order to rewrite the synthesis formula in complex exponential form. That is,

$$x_n = \sum_{m=0}^{N-1} (A'_m - iB'_m)\left(\cos\frac{2\pi mn}{N} + i\sin\frac{2\pi mn}{N}\right). \tag{1.57}$$

The cross product terms will vanish in the summation because of their odd symmetry about $m = N/2$. For example, the product of $A'_m$ with $\sin(2\pi mn/N)$ for $0 < m < N/2$

will be identical to the same product at harmonic N − m, but of opposite sign since $A'_m$ is an even function about N/2 and sin(2πmn/N) is an odd function.

If we let $S'_m = A'_m - iB'_m$, then, from Table 1.1 and the Fourier coefficients there divided by two, as required earlier,

$$\begin{aligned} S'_m &= \frac{1}{N}\sum_{n=0}^{N-1} x_n \cos\frac{2\pi mn}{N} - i\frac{1}{N}\sum_{n=0}^{N-1} x_n \sin\frac{2\pi mn}{N} \\ &= \frac{1}{N}\sum_{n=0}^{N-1} x_n \left(\cos\frac{2\pi mn}{N} - i\sin\frac{2\pi mn}{N}\right), \quad 0 \le m \le (N-1),\ N \text{ even.} \end{aligned} \tag{1.58}$$

$S'_m$ is the complex Fourier coefficient at the m-th harmonic frequency.

Using Euler's formula, expressions for $x_n$ and $S'_m$ can be expressed very compactly and symmetrically as

$$S'_m = \frac{1}{N}\sum_{n=0}^{N-1} x_n \exp(-i2\pi mn/N), \quad m = 0, 1, \ldots, N-1 \tag{1.59}$$

and

$$x_n = \sum_{m=0}^{N-1} S'_m \exp(i2\pi mn/N), \quad n = 0, 1, \ldots, N-1. \tag{1.60}$$

These equations constitute a *digital Fourier transform pair*, are valid whether N is even or odd, and could be written also

$$\begin{aligned} S'_m = \frac{1}{N}\sum_{n=0}^{N-1} x_n \exp(-i2\pi mn/N), \quad & m = -[(N-1)/2], \ldots, 0, \ldots, [N/2] \\ & \text{or } m = -[N/2], \ldots, 0, \ldots, [(N-1)/2] \end{aligned} \tag{1.61}$$

and

$$x_n = \sum_{m=-[(N-1)/2]}^{[N/2]} S'_m \exp(i2\pi mn/N), \quad n = 0, 1, \ldots, N-1$$

or

$$x_n = \sum_{m=-[N/2]}^{[(N-1)/2]} S'_m \exp(i2\pi mn/N), \quad n = 0, 1, \ldots, N-1 \tag{1.62}$$

where [q] means truncation of q. The new limits on m follow from the easily proved relation $S'_{m \pm kN} = S'_m$, where k is an integer. Equations 1.60 and 1.62 are referred to as *inverse* Fourier transforms of Equations 1.59 and 1.61, respectively. Whenever there is a Fourier transform pair, the equation for the time or space function in terms of the frequency function is considered the inverse Fourier transform.

If the variance in the periodogram is denoted by $C'_m$, then

$$C'_m = S'_m \times \overset{*}{S}{}'_m = (A'_m - iB'_m)(A'_m + iB'_m) = A'^2_m + B'^2_m,$$
$$m = -[(N-1)/2], \ldots, 0, \ldots, [N/2] \qquad (1.63)$$

where the asterisk again indicates complex conjugate. Ordinarily, variance is not computed at $m = 0$. Notice the periodogram here is two-sided; that is, there are variances at both negative and positive harmonics. Their use is a mathematical convenience. Recall that the primed Fourier coefficients are one-half the values given in Table 1.1 except at $m = 0$ and $m = N/2$ (N even). To match the one-sided spectrum in Table 1.1 for N even or odd, the variances have to be doubled according to

$$S^2_m = C_m = C'_m + C'_{-m} = 2C'_m, \quad m \neq 0;\ m \neq N/2 \text{ (N even)}$$
and (1.64)
$$S^2_{N/2} = C_{N/2} = C'_{N/2}, \quad \text{(N even)}.$$

## 1.5.6 The spectrum at nonharmonic frequencies

It was pointed out in Section 1.2 that the total variance in a data set can be shown to be the sum of the variances at the harmonic frequencies. In terms of accounting for total variance, there is no need to examine the spectrum at a frequency resolution higher than the spacing between adjacent harmonics. Nevertheless, a value of variance can be computed at any frequency by changing the cosine and sine arguments in the spectrum formulations from $2\pi mn\Delta t/(N\Delta t)$ to $2\pi fn\Delta t$ where f is frequency (in cycles per time interval between samples). In this section we derive a formula from which we conclude that the variance spectrum $C'(f)$, available at a continuum of frequencies, is uniquely related to the variances at the harmonic frequencies $C'_m$.

Define $S'(f)$ to be the complex amplitude coefficient at frequency f. Then, by analogy with Equation 1.61,

$$S'(f) = \frac{1}{N}\sum_{n=0}^{N-1} x_n \exp(-i2\pi fn\Delta t), \quad -1/(2\Delta t) \leq f \leq 1/(2\Delta t). \qquad (1.65)$$

Now substitute the first form of Equation 1.62 for $x_n$ to get

$$S'(f) = \sum_{m=-[(N-1)/2]}^{[N/2]} S'_m \sum_{n=0}^{N-1} \left(\frac{1}{N}\right) \exp\left[i2\pi\left(\frac{m}{N} - f\Delta t\right)n\right] \tag{1.66}$$

where [q], as earlier, means the truncated value of q in the summation limits.

Using Equation 1.B.4 to obtain the second summation yields

$$S'(f) = \sum_{m=-[(N-1)/2]}^{[N/2]} S'_m \times E_m(f), \quad -1/(2\Delta t) \le f \le 1/(2\Delta t) \tag{1.67}$$

where

$$E_m(f) = \exp\left[i(N-1)\left(\frac{m}{N} - f\Delta t\right)\pi\right] \times \frac{\sin\left[\left(\frac{m}{N} - f\Delta t\right)N\pi\right]}{N\sin\left[\left(\frac{m}{N} - f\Delta t\right)\pi\right]}. \tag{1.68}$$

Equation 1.67 tells us that each complex Fourier coefficient at a frequency between two adjacent harmonics is a weighted sum of the $S'_m$ harmonic coefficients. This means that calculating Fourier coefficients at nonharmonic frequencies yields no additional insight into the variance structure of the data; all of the variance information is revealed by the harmonic coefficients.

If $S'_m$ and $E_m(f)$ are replaced by their conjugates, the conjugate companion of Equation 1.67 will result and the weighted sum relation will apply to $\overset{*}{S}'(f)$. Because the periodogram is the product of $S'_m$ and $\overset{*}{S}'_m$, we conclude that at any frequency, f, the variance spectrum

$$C'(f) = S'(f) \times \overset{*}{S}'(f), \quad -1/(2\Delta t) \le f \le 1/(2\Delta t) \tag{1.69}$$

is a weighted sum of the variances at the harmonic frequencies. When f is at a harmonic frequency, the weight function is zero at all other harmonics except the one under consideration.

Let us re-examine Equation 1.68 for the case when $m = 0$. Then,

$$E_0(f) = \exp[-i(N-1)\pi f\Delta t]\,\frac{\sin(\pi N f\Delta t)}{N\sin(\pi f\Delta t)}. \tag{1.70}$$

Note that the limit of $E_0(f)$ is one as f tends to zero. If $S'_0$, the mean of the series, is large, the product of $E_0(f)$ and $S'_0$ will provide a large contribution to $S'(f)$ when f is in the neighborhood of the origin. The variance $C'(f)$ will then include a large contribution from the mean. Since it is the second moment about the mean that is desired, this

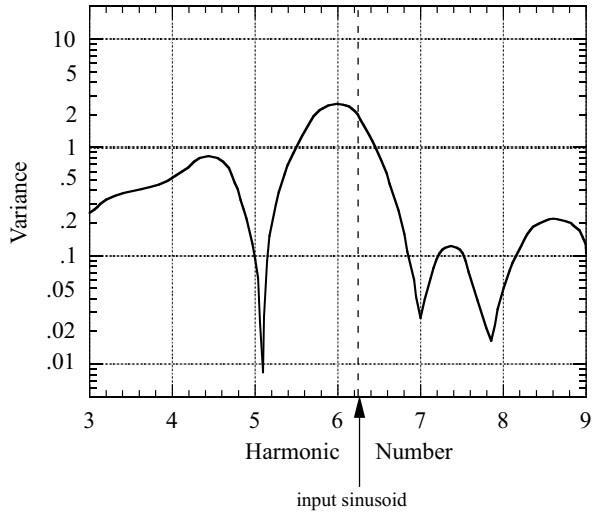



**Figure 1.44** Spectrum of sinusoid with frequency 6.25 cycles over data length N = 32 with added white noise. The ratio of the signal variance to the white noise variance at an internal harmonic (SNR) is 3.1 (see Equation 1.46).

contribution must be deleted. Therefore, in applying Equation 1.69 it is to be understood that the sample mean has been removed before computing $S'(f)$ or $C'(f)$.

One might expect that $C'(f)$ could be used to find the exact frequency of a deterministic signal embedded in noise. Unfortunately, this is not true, and an illustration of this fact is shown in Figure 1.44. The spectrum here is the same as spectrum (5) in Figure 1.37, except that the variances were calculated at frequency increments corresponding to 1/20 the harmonic spacing using Appendix 1.A. Due mainly to the leakage of noise variance from surrounding harmonics, the peak in the spectrum occurs not at the frequency of the input sinusoid (harmonic 6.25) but slightly to the left of harmonic 6.

If the noise is reduced to zero, however, the input frequency can be accurately determined, as shown in Figure 1.45. Why is this? In the first place there is no noise to contend with and, therefore, no noise leakage. In the second place, for interior frequencies and sufficiently large N, the frequency at which the peak in the spectral window in Equation 1.42 occurs when $2\pi f_i$ is substituted for $\omega_m$ is close to the frequency of the input sinusoid $f_i$. Recall from Figures 1.33 and 1.34 that the peak in the spectrum window at harmonic 1 is displaced from harmonic 1 for a pure cosine or sine input. It was found from simulations that when $N > 100$ a reasonably accurate estimate of the frequency of an input sinusoid away from the ends of the spectrum can be made because the window is nearly symmetric and its peak is close to the center of the window. This parallels the earlier finding in Section 1.5.2 that the squared diffraction function in Equation 1.44 provides a good approximation to the window function in Equation 1.42 away from the spectrum ends for $N > 100$. Of course, the larger the value of N, the greater the accuracy. From Figures 1.44

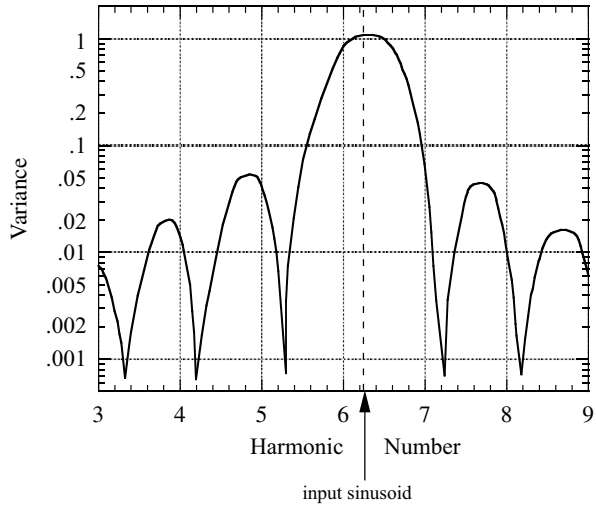



**Figure 1.45** Spectrum of sinusoid with frequency 6.25 cycles over data data length N = 32.

and 1.45 we conclude that the higher the ratio of the deterministic signal variance to the noise variance, the more accurately one can estimate the frequency of the signal.

The various equations developed in this and the previous sections typically are not used in computing a periodogram. Instead, we use a straightforward algorithm employing the formulas in Table 1.1 or a fast algorithm as in the computer program in Appendix 1.A. The latter algorithm permits us to evaluate the Fourier coefficients and variances at as high a resolution in frequency as we wish. When we do this, we now know that the variance computed at an off-harmonic frequency is a weighted sum of all harmonic variances, and the closer they are to a given off-harmonic frequency, the greater their influence.

## 1.5.7 Padding data with zeroes

In this section we investigate a topic of practical interest wherein a time series with zero mean is modified by appending zeroes to it in order to obtain a desired length. The procedure is called "padding data with zeroes" and a common purpose is to match the length of a record with that required when using an FFT (fast Fourier transform) algorithm to analyze many and/or long data sets. As an example, if we had an 83 point sequence and were using a simple FFT requiring $2^k$ points, we could add 45 zeroes to obtain $2^7 = 128$. We will show that the periodogram of the padded data is identical to the variance spectrum (Equation 1.63) of the original data computed at intervals in frequency of 1/128, except for a multiplicative constant. It may seem odd that one would add a sequence of zeroes to a time series (after removal of the mean) and compute a periodogram that has any meaning. The interesting aspect is that in calculating the Fourier coefficients, the padded series can be partitioned

into the original series and the sequence of zeroes, the latter contributing nothing to the coefficients. The result is a spectrum with higher resolution than the periodogram of the original data.

Let us begin by considering a time series of data and subtract its mean from each datum to get data set A. Next consider data set B. It is the same as data set A except that zeroes have been appended in order to apply an FFT. The means of data sets A and B are zero. The variances of both are the same if, for set B, the coefficient of the sum in the expression for variance is the same as that for set A – a condition that we now investigate.

The formula for the Fourier coefficients in data set A is, following Equations 1.58 and 1.59,

$$S'_m = A'_m - iB'_m = \frac{1}{N}\sum_{n=0}^{N-1} x_n \exp(-i2\pi mn/N), \quad m = -[(N-1)/2], \ldots, 0, \ldots, [N/2] \tag{1.71}$$

where m is harmonic number, $x_n$ is the n-th datum, N is the number of data and [q] indicates truncated value, as before.

Of course, we know from the previous section that we can calculate Fourier coefficients at higher resolution in frequency than the harmonic frequencies (Equation 1.65) and they are completely dependent on those at the harmonic frequencies (Equation 1.67). Consider increasing the number of complex coefficients N in set A by a factor R, such that RN is the number of data needed by an FFT. Then the new formula for the high resolution coefficients in data set A is

$$\begin{aligned} S'_r \;&=\; A'_r - iB'_r \;=\; \frac{1}{N}\sum_{n=0}^{N-1} x_n \exp(-i2\pi rn/(RN)), \\ r \;&=\; -[(RN-1)/2], \ldots, 0, \ldots, [RN/2]. \end{aligned} \tag{1.72}$$

To distinguish padded data set B from data set A, we will use bold notation, for example, $\mathbf{x}_n$. Data set B has RN data and, by analogy with Equation 1.71, noting that $\mathbf{x}_n = 0$ beyond $n = N-1$, its Fourier coefficients are

$$\begin{aligned} \mathbf{S}'_r = \mathbf{A}'_r - i\mathbf{B}'_r \;&=\; \frac{1}{RN}\sum_{n=0}^{RN-1} \mathbf{x}_n \exp[-i2\pi rn/RN] \\ &= \frac{1}{RN}\sum_{n=0}^{N-1} \mathbf{x}_n \exp(-i2\pi rn/RN) + \frac{1}{RN}\sum_{n=N}^{RN-1} \mathbf{x}_n \exp(-i2\pi rn/RN) \\ &= \frac{1}{RN}\sum_{n=0}^{N-1} \mathbf{x}_n \exp(-i2\pi rn/RN), \quad r \;=\; -[(RN-1)/2], \ldots, 0, \ldots, [RN/2]. \end{aligned} \tag{1.73}$$

Except for the coefficient of the summation, Equation 1.73 is the same as Equation 1.72. If it were desired that the Fourier coefficients of padded data set B be the same as the high resolution coefficients of data set A, the former coefficients need to be multiplied by R.

Whether one uses padding (Equation 1.73) or direct calculation (Equation 1.72), the total variance derived from the Fourier coefficients must equal the variance in the data. The variance at a Fourier harmonic frequency can be obtained by forming the product $S_m \times \overset{*}{S}_m$, in which the asterisk means complex conjugate. Accordingly, from Equations 1.61 and 1.63,

$$S'_m \times \overset{*}{S}'_m = A'^2_m + B'^2_m = \frac{1}{N^2}\left|\sum_{n=0}^{N-1} x_n \exp(-i2\pi mn/N)\right|^2,$$

$$m = -[(N-1)/2], \ldots, 0, \ldots, [N/2] \tag{1.74}$$

the total variance of which is

$$\sum_{m=-[(N-1)/2]}^{[N/2]} \left(A'^2_m + B'^2_m\right). \tag{1.75}$$

When variances are computed at a higher resolution in frequency than that associated with just the harmonic frequencies, they must be scaled by 1/R. The reason for scaling is that the variances computed at a resolution in frequency greater than the harmonic resolution are not independent (as are the harmonic variances). The dependence is taken into account by reducing the bandwidth associated with each high resolution spectrum variance by 1/R.

Using padding for FFT purposes, with the consequent increase in spectral resolution, the expression for the periodogram variance is

$$\mathbf{S}'_r \times \overset{*}{\mathbf{S}}'_r = \mathbf{A}'^2_r + \mathbf{B}'^2_r = \frac{1}{(RN)^2}\left|\sum_{n=0}^{N-1} \mathbf{x}_n \exp(-i2\pi rn/RN)\right|^2,$$

$$r = -[(RN-1)/2], \ldots, 0, \ldots, [RN/2]. \tag{1.76}$$

However, as mentioned earlier, to match the variances associated with high resolution data set A, the Fourier coefficients in data set B have to be multiplied by R, or, what is the same, the variances in Equation 1.76 have to be multiplied by $R^2$. Thus,

$$A'^2_r + B'^2_r = R^2\mathbf{A}'^2_r + R^2\mathbf{B}'^2_r. \tag{1.77}$$

It follows that for high-resolution data set A, the total variance is given by

$$\frac{1}{R}\sum_{r=-[(RN-1)/2]}^{[RN/2]}(A_r'^2 + B_r'^2) \tag{1.78}$$

and, for data set B, by

$$\frac{1}{R}\sum_{r=-[(RN-1)/2]}^{[RN/2]} R^2\left(\mathbf{A}_r'^2 + \mathbf{B}_r'^2\right) = R\sum_{r=-[(RN-1)/2]}^{[RN/2]}(\mathbf{A}_r'^2 + \mathbf{B}_r'^2). \tag{1.79}$$

If the variances at the nonharmonic frequencies in Equation 1.77 are treated as random variables, they have the same asymptotic (as N tends to infinity) mean, variance, and distribution as those at the harmonic frequencies (Koopmans, 1974, pp. 261–265).

The consequences of padding a time series with zeroes to accommodate analysis with an FFT can be illustrated with a typical example in which the number of data does not match the requirements of the FFT that is to be used. Consider spatially averaged sea surface temperature (SST) in an area bounded by 10° south latitude on its southern edge, the equator on its northern edge, and 80 and 90° west longitude on its eastern and western edges, respectively. This area comprises Niño Regions 1 and 2, which are often used as indicator regions of the current state of the El Niño Southern Oscillation (ENSO). Figure 1.46 shows Niño Region 1 + 2 monthly SST anomalies for the 30-year period of 1981–2010. These values were derived by subtracting the 1981–2010 monthly means from each month's actual SST in a manner identical to that used in Section 1.4.6.2. The original monthly values of SST were obtained from

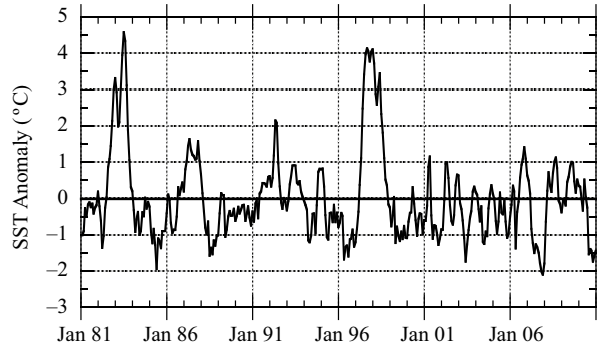



**Figure 1.46** Mean monthly SST (sea surface temperature) anomalies in Niño Region 1 + 2 for the 30-year period of 1981–2010.

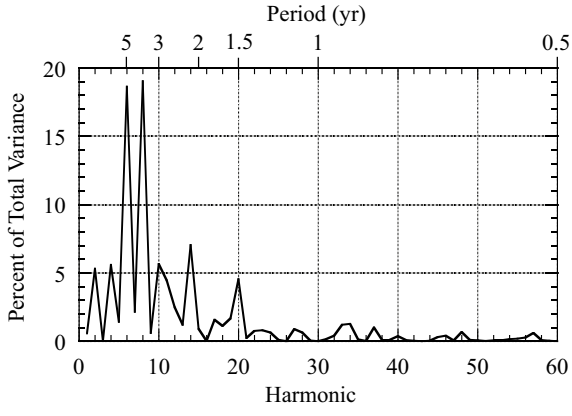



**Figure 1.47** Periodogram to harmonic 60 of SST anomalies in Figure 1.46. Total anomaly variance is 1.440 °C$^2$.

the National Weather Service Climate Prediction Center (http://www.cpc.ncep.noaa.gov/data/indices/).

If one were not concerned with computational speed, the SST anomaly data could be analyzed without implementation of an FFT by using standard Fourier analysis techniques, as in Appendix 1.A, where the number of data is $N = 360$. The first 60 harmonics of the periodogram resulting from such an analysis are presented in Figure 1.47, where these harmonics explain over 96% of the total variance in the time series. Notice that harmonics 30 and 60 necessarily have zero variance, since the 30-year mean was removed from the data. There is a concentration of variance at harmonics 20 and lower, corresponding to periods of 1.5 years and longer, and of particular interest are the peak variances that occur at periods of 3.75 and 5.0 years (harmonics 8 and 6, respectively). The periodogram provides an informative depiction of the cyclic nature of El Niño and La Niña.

Suppose, however, that many thousands of such analyses needed to be performed as rapidly as possible. The computational efficiency of FFTs then becomes necessary, and it may be the case that the FFT requires $N = 2^k$ data points as previously described. By augmenting the 30 years of monthly data with 152 zeroes, we obtain a time series that is 512 ($2^9$) data points in length. An FFT was used to compute the folded version (positive harmonics only) of the Fourier coefficients in Equation 1.73 and the folded version of the periodogram variances in Equation 1.76 where $RN = 512$. The ratios (in percentage) of the individual variances to their sum are shown in Figure 1.48 for the first 85 harmonics to account for the inherently narrower bandwidth of the analysis.

In comparing Figures 1.47 and 1.48 to each other, we can make the general statement that they are different for two reasons. One is that the period (in years)

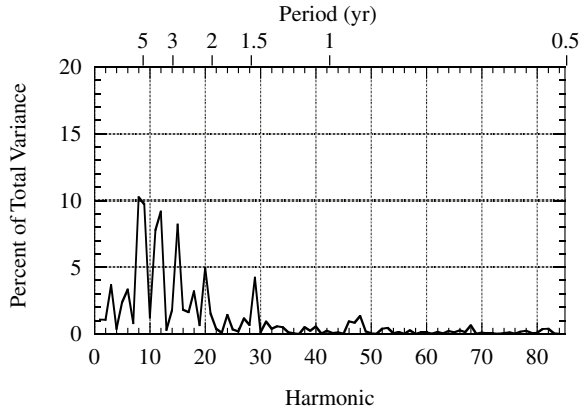



**Figure 1.48** Periodogram to harmonic 85 of SST anomalies in Figure 1.46 after padding with 152 zeroes. Total anomaly variance is 1.440 $°C^2$.

versus harmonic relation is different in each figure; the other is that the bandwidth or width of the spectrum window associated with the variance estimates in each figure is different. To explain these differences, consider an example. As mentioned earlier, Figure 1.47 shows two strong peaks: the first at a period of 5 years (harmonic 6) and the second at a period of 3.75 years (harmonic 8). In Figure 1.48, the periodogram of the padded data, the variance in the first peak lies between harmonic 8, corresponding to a period of 5.333 years, and harmonic 9, corresponding to a period of 4.741 years. The spectrum windows centered at harmonics 8 and 9 transfer most of the variance in the first peak in Figure 1.47 to these two harmonics. Because the location of the first peak is approximately midway between the surrounding harmonics 8 and 9 in Figure 1.48, the variances there are quite similar.

An analogous situation occurs with the second peak in variance in Figure 1.47. This peak, at a period of 3.75 years (harmonic 8), lies between harmonics 11 (period of 3.879 years) and 12 (period of 3.556 years). Again, the spectrum windows centered at these harmonics redistribute the variance that lies between them in Figure 1.47 to these harmonics, as shown in Figure 1.48.

The opposite situation occurs in Figure 1.48 at harmonic 15 and period 2.844 years. The peak at this harmonic falls between harmonics 10 (period of 3 years) and 11 (period of 2.727 years) in Figure 1.47. In this case, a redistribution of a peak in variance in the padded spectrum occurs in the unpadded spectrum. In any periodogram, variance that is intrinsically located between harmonics is distributed to surrounding harmonics by a spectrum window centered at each harmonic, as we learned in Section 1.5.2. In addition, the two periodograms of the SST anomaly data illustrate the effect of data length on finding periodicities, a subject discussed in Section 1.5.4.

# Appendix 1.A Subroutine foranx

```
subroutine foranx (s, n, var, nf, tvar, fr1, fr2, iprnt)
dimension s(1), a(400), b(400), pvar(400), phi(400), var(400),
freq(400)
c*************************************************************
c
c     This subroutine performs a fast Fourier analysis of an even
      number of data points at as
c     many frequencies as desired. The frequency span is between 0.0
      and 0.5 cy/data interval,
c     inclusive. The algorithm used is that given at the end of
      Chapter 9 of Spectral Analysis
c     (Jenkins & Watts, 1968, Holden-Day, San Francisco, 525 pp.)
c       * Input *
c  s  input data array.
c  n  length of s. n is an even number.
c  nf  number of frequencies (including zero) at which variance
      is to be computed.
c     nf.gt.n/2 and is an odd number.
c     nf = n/2 + 1 for standard periodogram.
c  fr1 frequency at which printing begins.
c  fr2 frequency at which printing ends.
c      fr1 and fr2 are less than or equal to 0.5 and fr1 < fr2.
c          * Output *
c     var  the array of spectrum variances at the nf frequencies.
c     tvar the total variance in the data.
c           * Other *
c     iprnt user supplied output device unit number.
c
c     the synthesis form is x(t) = a*cos(wt) + b*sin(wt)
c         = c*cos(wt - phi) phi = phase angle in program
c
c*************************************************************
c     ** set constants, variance scale factor, and frequency array **
c
      data pi /3.1415926536/
      anf = nf
      an = n
      rddg = 180.0/pi
      frfac = an/(2.0*anf - 2.0)
      bb = 0.5/(anf - 1.0)
      do 10 i = 1, nf
      aa = i
 10   freq(i) = (aa - 1.0)*bb
```

```
c
c     ** get Fourier coefficients at 0 and Nyquist frequencies **
c
      a(1) = 0.0
      b(1) = 0.0
      a(nf) = 0.0
      b(nf) = 0.0
      e = n
      tvar = 0.0
      do 20 i = 1, n
      c = i - 1
      r = s(i)/e
      a(1) = a(1) + r
 20   a(nf) = a(nf) + r*cos(c*pi)
c
c     ** get variance in data, start accumulation of variance in
      spectrum **
c
      do 30 i = 1, n
      s(i) = s(i) - a(1)
 30   tvar = tvar + s(i)**2/e
      var(nf) = a(nf)**2
      pvar(nf) = var(nf)*100.0*frfac/tvar
      var(1) = 0.0
      pvar(1) = 0.0
      wvar = var(nf)*frfac
      g = n/2
      phi(1) = 0.0
      phi(nf) = 0.0
      do 40 i = 1, n
 40   s(i) = s(i)/g
c
c     ** J & W algorithm **
c
      nfm1 = nf - 1
      do 50 j = 2, nfm1
      ang = 2.0*pi*freq(j)
      co = cos(ang)
      si = sin(ang)
      v0 = 0.0
      v1 = 0.0
      z0 = 0.0
      z1 = 0.0
      do 60 i = 2, n
      ii = n - i + 2
```

```
      v2 = 2.0*co*v1 - v0 + s(ii)
      z2 = 2.0*co*z1 - z0 + s(ii)
      v0 = v1
      v1 = v2
      z0 = z1
      z1 = z2
 60   continue
      a(j) = s(1) + v1*co - vo
      b(j) = z1*si
      var(j) = (a(j)**2 + b(j)**2)/2.0
      pvar(j) = var(j)*frfac*100.0/tvar
      wvar = wvar + var(j)*frfac
 50   phi(j) = atan2(b(j), a(j))*rddg
c
c     **print results **
c
      kj = 0
      ih1f = 1
      ih2f = nf
      if(fr1.lt.0.0.or.fr1.gt.0.5) go to 111
      if(fr2.lt.fr1.or.fr2.gt.0.5) go to 111
      if(fr1.eq.0.0.and.fr2.eq.0.5) go to 109
      ih1f = 2.0*fr1*(anf - 1.0) + 1.01
      ih2f = 2.0*fr2*(anf - 1.0) + 1.01
c
 109  do 70 j = ih1f, ih2f
      wn = freq(j)*an
      kj = kj + 1
      if(((kj - 1)/25)*25.eq.(kj - 1)) write(iprnt, 102)
 70   write(iprnt,103) wn, freq(j), a(j), b(j), var(j), pvar(j),
      phi(j)
      write(iprnt,104) tvar, wvar
 102  format(//9x, 'har-', 6x, 'freq', 7x, 'cosine', 9x, 'sine',
      *10x, 'line', 7x, 'percent of', 5x, 'phase', /8x, 'monic',
      *5x, 'cy/di', 2(3x,'coefficient'), 5x, 'variance', 6x,
      'total var',*5x, 'angle', /)
 103  format(5x, f8.3, 4x, f6.3, 4(2x,g12.5), 4x, f6.1)
 104  format(///21x, 'variance in data set', g12.5//8x,
      *'variance explained by periodogram', g12.5)
      go to 99
 111   write(iprnt, 112)
 112  format(//, 10x, 'fr1 or fr2 or both out of range')
 99   return

      end
```

## Appendix 1.B Sum of complex exponentials

Let

$$Q = \sum_{n=a}^{b} \exp(i\omega n) = e^{i\omega a} + e^{i\omega(a+1)} + \cdots + e^{i\omega b} \tag{1.B.1}$$

where a and b are integers and $b > a$. Multiply Equation 1.B.1 by $e^{i\omega}$ to get

$$e^{i\omega}Q = e^{i\omega(a+1)} + e^{i\omega(a+2)} + \ldots + e^{i\omega(b+1)}. \tag{1.B.2}$$

Subtract Equation 1.B.2 from Equation 1.B.1 to obtain

$$Q = \frac{\left(e^{i\omega a} - e^{i\omega(b+1)}\right)}{(1 - e^{i\omega})}. \tag{1.B.3}$$

Now multiply the numerator and denominator of Equation 1.B.3 by $\exp(-i\omega/2)$. Then successively withdraw $\exp(ia\omega/2)$ and $\exp(ib\omega/2)$. The result is

$$\begin{aligned} Q &= \sum_{n=a}^{b} \exp(i\omega n) \\ &= \exp[i(a+b)\omega/2] \left[\frac{e^{i\omega(b-a+1)/2} - e^{-i\omega(b-a+1)/2}}{e^{i\omega/2} - e^{-i\omega/2}}\right] \end{aligned}$$

which, using Euler's formula, reduces to

$$Q = \exp[i(a+b)\omega/2] \frac{\sin[\omega(b-a+1)/2]}{\sin(\omega/2)}. \tag{1.B.4}$$

In application of Equations 1.B.3 and 1.B.4 it is important to test $\sin(\omega/2)$ to verify that it is not zero for any values of the argument. If $\sin(\omega/2)$ is zero, then l'Hopital's rule can be applied to these equations to obtain a determinate form. Equation 1.B.1 can be used, also.

## Appendix 1.C Distribution of harmonic variances

The purpose of this appendix is to develop relationships for the statistical distribution of the harmonic variances. Because the chi-square ($\chi^2$) distribution plays a prominent role in the development that follows, it is important to be

familiar with its properties. We begin with the forms for the Fourier cosine and sine amplitudes given in Table 1.1, now treated as random variables, for an even number of data N, namely

$$A_m = \frac{2}{N}\sum_{n=0}^{N-1} X_n \cos\frac{2\pi mn}{N} \tag{1.C.1a}$$

and

$$B_m = \frac{2}{N}\sum_{n=0}^{N-1} X_n \sin\frac{2\pi mn}{N}, \quad m = \left[0, \frac{N}{2}\right]. \tag{1.C.1b}$$

Making use of linear expectation operator E and assuming a purely random process (white noise) represented by random variable $X_n$ with $E[X_n] = 0$ so that $E[A_m] = E[B_m] = 0$, we obtain:

$$\begin{aligned}
\mathrm{Var}[A_m] &= E[A_m^2] \\
&= \frac{4}{N^2}\left\{ E[X_0X_0]\cos\frac{2\pi m0}{N}\cos\frac{2\pi m0}{N} + E[X_0X_1]\cos\frac{2\pi m0}{N}\cos\frac{2\pi m1}{N}\right. \\
&\quad + \cdots + E[X_0X_{N-1}]\cos\frac{2\pi m0}{N}\cos\frac{2\pi m(N-1)}{N} + E[X_1X_0]\cos\frac{2\pi m1}{N}\cos\frac{2\pi m0}{N} \\
&\quad + E[X_1X_1]\cos\frac{2\pi m1}{N}\cos\frac{2\pi m1}{N} + \cdots + E[X_1X_{N-1}]\cos\frac{2\pi m1}{N}\cos\frac{2\pi m(N-1)}{N} \\
&\quad + \cdots + E[X_{N-1}X_0]\cos\frac{2\pi m(N-1)}{N}\cos\frac{2\pi m0}{N} \\
&\quad + E[X_{N-1}X_1]\cos\frac{2\pi m(N-1)}{N}\cos\frac{2\pi m1}{N} \\
&\quad \left. + \cdots + E[X_{N-1}X_{N-1}]\cos\frac{2\pi m(N-1)}{N}\cos\frac{2\pi m(N-1)}{N}\right\}.
\end{aligned} \tag{1.C.2}$$

The expectation $E[X_iX_j] = 0$ for $i \neq j$ because the random variables are uncorrelated; similarly, $E[X_iX_j] = \sigma_X^2$ for $i = j$ because the random variables are completely correlated. The latter relation follows from Equation 1.18 and noting from above that $E[X_n] = 0$. Therefore, Equation 1.C.2 becomes

$$\mathrm{Var}[A_m] = \begin{cases} \dfrac{4}{N^2}\sigma_X^2 \displaystyle\sum_{n=0}^{N-1} \cos^2\frac{2\pi mn}{N} = \frac{4}{N^2}\sigma_X^2\frac{N}{2} = \frac{2}{N}\sigma_X^2, & m = \left[1, \dfrac{N}{2} - 1\right] \\ \dfrac{1}{N^2}\sigma_X^2 \displaystyle\sum_{n=0}^{N-1} \cos^2\frac{2\pi mn}{N} = \frac{1}{N^2}\sigma_X^2 N = \frac{1}{N}\sigma_X^2, & m = 0, \dfrac{N}{2} \end{cases} \tag{1.C.3}$$

and for the sine coefficients

$$\mathrm{Var}[B_m] = \begin{cases} \dfrac{4}{N^2}\sigma_X^2 \displaystyle\sum_{n=0}^{N-1} \sin^2\frac{2\pi mn}{N} = \frac{4}{N^2}\sigma_X^2\frac{N}{2} = \frac{2}{N}\sigma_X^2, & m = \left[1, \dfrac{N}{2} - 1\right] \\ \dfrac{1}{N^2}\sigma_X^2 \displaystyle\sum_{n=0}^{N-1} \sin^2\frac{2\pi mn}{N} = 0, & m = 0, \dfrac{N}{2}. \end{cases} \tag{1.C.4}$$

The sums of the cosine-squared and sine-squared terms can be determined from Equations 1.4 and 1.5.

The covariance between the coefficients at different harmonics may be calculated in a similar manner. For $m \neq k$,

$$\begin{aligned} \mathrm{Cov}[A_m, A_k] &= \frac{4}{N^2}\sigma_X^2 \sum_{n=0}^{N-1} \cos\frac{2\pi mn}{N} \cos\frac{2\pi kn}{N} \\ &= 0 \end{aligned} \tag{1.C.5}$$

since the cosine terms are orthogonal to each other, as demonstrated in the derivation of Equation 1.8. Similarly, for the sine coefficients

$$\mathrm{Cov}[B_m, B_k] = 0. \tag{1.C.6}$$

Lastly, for all m, k,

$$\mathrm{Cov}[A_m, B_k] = 0 \tag{1.C.7}$$

because over their length, any integer number of sine waves is orthogonal to any integer number of cosine waves.

Now assume that each rv $X_n$ from our white noise process has a normal distribution with population mean zero and population variance $\sigma_X^2$. Since random variables $A_m$ and $B_m$ are linear functions of normal random variables from Equation 1.C.1, they also are normally distributed. From statistical theory, the square of a normal random variable with zero mean and unit variance (i.e., a standard normal variable) is distributed as a chi-square variable with one degree of

freedom. Thus, if we square the Fourier coefficients and standardize them by dividing by their variance, we have

$$\frac{A_m^2}{\mathrm{Var}[A_m]} = \frac{A_m^2}{\sigma_X^2/(N/2)} \Rightarrow \chi_1^2, \quad m = \left[1, \frac{N}{2} - 1\right] \tag{1.C.8}$$

$$\frac{B_m^2}{\mathrm{Var}[B_m]} = \frac{B_m^2}{\sigma_X^2/(N/2)} \Rightarrow \chi_1^2, \quad m = \left[1, \frac{N}{2} - 1\right] \tag{1.C.9}$$

and

$$\frac{A_m^2}{\mathrm{Var}[A_m]} = \frac{A_m^2}{\sigma_X^2/N} \Rightarrow \chi_1^2, \quad m = 0, \frac{N}{2} \tag{1.C.10}$$

in which the arrow indicates "is distributed as." Notice that no equation comparable to Equation 1.C.10 is given for $B_m$ when $m = 0, N/2$; the reason is that $B_m$ is always zero for these two values of m. There is another relevant relationship involving $\chi^2$ variables: the sum of any number of mutually independent $\chi^2$ variables whose degrees of freedom sum to $\nu$ is itself a $\chi^2$ variable with $\nu$ degrees of freedom; that is,

$$\chi_{\nu_1}^2 + \chi_{\nu_2}^2 + \cdots + \chi_{\nu_k}^2 = \chi_\nu^2 \tag{1.C.11}$$

where $\nu = \nu_1 + \nu_2 + \cdots + \nu_k$. Thus, dividing Equations 1.C.8 and 1.C.9 by two and then summing yields

$$\frac{\frac{A_m^2 + B_m^2}{2}}{\sigma_X^2/(N/2)} \Rightarrow \frac{\chi_2^2}{2}, \quad m = \left[1, \frac{N}{2} - 1\right]. \tag{1.C.12}$$

The reason for dividing by two is to match the expression for variance at a harmonic given in Table 1.1. The denominators in Equations 1.C.12 and 1.C.10 distribute the population variance $\sigma_X^2$ among the harmonic frequencies in such a way that the variance at the interior harmonics is uniform and twice the value at the frequency origin ($m = 0$) and the highest frequency ($m = N/2$). The variance at $m = 0$ is the variance of the sample mean (i.e., the mean of a realization) about the population mean, the latter value of which is zero in this development.

Now simplify the notation by letting

$$C(f_m) = \frac{A_m^2 + B_m^2}{2}, \quad m = \left[1, \frac{N}{2} - 1\right] \tag{1.C.13a}$$

and

$$C(f_m) = A_m^2, \quad m = 0, \frac{N}{2} \tag{1.C.13b}$$

where $f_m = m/N\Delta t$ is the harmonic frequency for harmonic m. Next, replace the white noise variance $\sigma_X^2/(N/2)$ at the interior harmonics and $\sigma_X^2/N$ at the two exterior harmonics, by $\Gamma(f_m)$ in Equations 1.C.12 and 1.C.10. Taking their expectations, and noting that $E[\chi_\nu^2] = \nu$, yields

$$E\left[\frac{C(f_m)}{\Gamma(f_m)}\right] = E\left[\frac{\chi_2^2}{2}\right] = 1, \quad m = \left[1, \frac{N}{2} - 1\right] \tag{1.C.14}$$

and

$$E\left[\frac{C(f_m)}{\Gamma(f_m)}\right] = E\left[\chi_1^2\right] = 1, \quad m = 0, \frac{N}{2} \tag{1.C.15}$$

with the result that

$$E[C(f_m)] = \Gamma(f_m), \quad m = \left[0, \frac{N}{2}\right]. \tag{1.C.16}$$

We now introduce the term *estimator*. An estimator is a random variable used to estimate a population parameter. For example, in Equation 1.C.13, spectrum estimator $C(f_m)$, as an appropriate function of the Fourier coefficients, is used to estimate the population variance at frequency $f_m$. Equation 1.C.16 shows that $C(f_m)$ is an unbiased estimator of the white noise variance at the harmonic frequencies because its expected value is equal to the population variance $\Gamma(f_m)$. If the expected value were something other than $\Gamma(f_m)$, $C(f_m)$ would be a biased estimator. It is usually desirable that an estimator be unbiased. However, if the calculation of an unbiased estimator requires information that is otherwise unavailable, or if repeated calculations are needed that consume significant computation time, it may be more advantageous to employ a biased estimator.

Since $\mathrm{Var}[\chi_\nu^2] = 2\nu$, we have, following Equations 1.C.14 and 1.C.15,

$$\mathrm{Var}[C(f_m)] = \Gamma^2(f_m), \quad m = \left[1, \frac{N}{2} - 1\right] \tag{1.C.17}$$

and

$$\mathrm{Var}[C(f_m)] = 2\Gamma^2(f_m), \quad m = 0, \frac{N}{2} \tag{1.C.18}$$

showing that the variance of the estimator is uniform at the interior harmonics and twice that value at the exterior harmonics (note the definitions of $\Gamma(f_m)$). That we are dealing with the variance of harmonic variances simply means that each harmonic variance $C(f_m)$ is itself a random variable and thus has a probability distribution function based on an infinity of realizations. Equations 1.C.16–1.C.18 are the expressions for the mean and variance of the probability distribution function.

It is significant that the variances of the harmonic variances are independent of sample size. The collection of additional data does not increase the stability of the estimator. That this is the case is not unexpected because as the length, N, of the time series increases, the number of Fourier harmonics increases accordingly and the separation between them, that is, the bandwidth or frequency averaging distance associated with each harmonic, decreases. The number of data (degrees of freedom for white noise) consumed in a variance estimate remains the same. To effect increased stability of the spectrum estimator requires some form of spectrum averaging.

In the case of N odd, the analysis parallels that above, except that the highest harmonic is $(N-1)/2$. To get $\Gamma(f_m)$ at all harmonics except the frequency origin divide the population variance by N/2; at $m=0$ divide $\sigma_X^2$ by N.

The above derivations have been done under the assumption that the population mean is known, and in this case equal to zero. The derivation could have been done with a known nonzero mean, but the procedure is more tedious. More generally, the population mean is unknown and the total variance in a given time series is taken with respect to the sample mean. If the time series is hypothesized to be a realization of white noise (with mean unknown), the total variance is similarly distributed as above but without any variance contribution at $m=0$. This is because the total variance must be perforce computed about the sample mean.

For the case of an even number of data, the estimate of the total variance, $\hat{\sigma}_X^2$, is divided by $(N-1)/2$ to obtain white noise variances at the interior harmonics and by $N-1$ to obtain the white noise variance at the highest harmonic, N/2. There is no contribution of variance at $m=0$. In the case of an odd number of data, the total variance is divided by $(N-1)/2$ to obtain estimates of the harmonic white noise variances and, again, there is no contribution of variance at $m=0$.

In summary, for N even and the mean of the white noise process known, variances at the interior harmonics have a distribution proportional to $\chi_2^2/2$. Variances at the two exterior harmonics (0 and N/2) have a distribution proportional to $\chi_1^2/1$. For N odd and the population mean known, the distributions of variance at all harmonics are proportional to $\chi_2^2/2$, except at the 0-th harmonic where the distribution is proportional to $\chi_1^2/1$. When the population mean is unknown, the variances have similar distributions except that no variance is generated at the origin.

## Appendix 1.D Derivation of Equation 1.42

The problem is to find the variance at any harmonic frequency when the input is at a nonharmonic frequency. Consider the general input sinusoid $a\cos(\omega n - \phi)$ and take its Fourier transform. From Equation 1.63 for a two-sided spectrum,

$$S'_m = A'_m - iB'_m = \frac{1}{N}\sum_{n=0}^{N-1} a\cos(\omega n - \phi)\exp(-i\omega_m n)$$
$$m = -[(N-1)/2], \ldots, 0, \ldots, [N/2] \qquad (1.D.1)$$

where m is harmonic number and $\omega_m = 2\pi m/N$ is angular frequency. It is assumed that the time step $\Delta t = 1$.

Using Euler's formula, the input sinusoid can be put in complex exponential form such that:

$$(A'_m - iB'_m)\frac{2N}{a} = \exp(-i\phi)\sum_{n=0}^{N-1}\exp[i(\omega-\omega_m)n] + \exp(i\phi)\sum_{n=0}^{N-1}\exp[-i(\omega+\omega_m)n]. \qquad (1.D.2)$$

From Equation 1.B.4,

$$(A'_m - iB'_m)\frac{2N}{a} = \exp\{i[(N-1)(\omega-\omega_m)/2 - \phi]\}\frac{\sin[N(\omega-\omega_m)/2]}{\sin[(\omega-\omega_m)/2]}$$
$$+\exp\{-i[(N-1)(\omega+\omega_m)/2 - \phi]\}\frac{\sin[N(\omega+\omega_m)/2]}{\sin[(\omega+\omega_m)/2]}. \qquad (1.D.3)$$

We can make use of Euler's formula, again, to rewrite the exponential terms of Equation 1.D.3. Equating the real portions of the resulting equation allows us to solve for $A'_m$, and, similarly, equating the imaginary portions yields $B'_m$, so that

$$A'_m = \frac{a}{2}\left\{\cos\left[(N-1)\left(\frac{\omega+\omega_m}{2}\right) - \phi\right]\frac{\sin N[(\omega+\omega_m)/2]}{N\sin[(\omega+\omega_m)/2]}\right.$$
$$\left.+\cos\left[(N-1)\left(\frac{\omega-\omega_m}{2}\right) - \phi\right]\frac{\sin[N(\omega-\omega_m)/2]}{N\sin[(\omega-\omega_m)/2]}\right\} \qquad (1.D.4)$$

and

$$B'_m = \frac{a}{2}\left\{\sin\left[(N-1)\left(\frac{\omega+\omega_m}{2}\right)-\phi\right]\frac{\sin[N(\omega+\omega_m)/2]}{N\sin[(\omega+\omega_m)/2]}\right.$$

$$\left.-\sin\left[(N-1)\left(\frac{\omega-\omega_m}{2}\right)-\phi\right]\frac{\sin N[(\omega-\omega_m)/2]}{N\sin[(\omega-\omega_m)/2]}\right\}. \qquad (1.D.5)$$

These results apply for N even or odd and to a two-sided spectrum. In reference to Equation 1.42, where N is even and the periodogram is one sided, the $A'_m$ and $B'_m$ above have to be doubled except at $m = 0$, N/2. Thus the variance at positive harmonic m is

$$S_m^2(\omega) = \left[(2A_m)^2 + (2B_m)^2\right]/2 = 2A_m^2 + 2B_m^2$$

$$= \frac{a^2}{2}\left\{\frac{\sin^2[N(\omega+\omega_m)/2]}{N^2\sin^2[(\omega+\omega_m)/2]} + \frac{\sin^2[N(\omega-\omega_m)/2]}{N^2\sin^2[(\omega-\omega_m)/2]}\right.$$

$$+ 2\cos[(N-1)\omega - 2\phi]\frac{\sin[N(\omega+\omega_m)/2]}{N\sin[(\omega+\omega_m)/2]}$$

$$\left.\times\frac{\sin[N(\omega-\omega_m)/2]}{N\sin[(\omega-\omega_m)/2]}\right\}, \qquad m \neq 0, \frac{N}{2} \qquad (1.D.6)$$

which is Equation 1.42.

## Problems

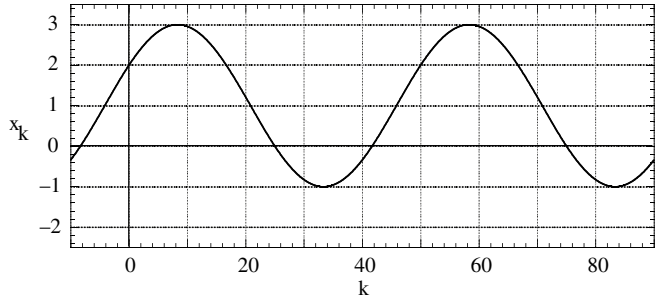
**1** On graph paper, carefully sketch at least one complete cycle of the sinusoid given by

$$y(t) = 1 - 2\cos(0.5\pi t + \pi/4)$$

starting at $t = -1$. (Suggestion: First find the period and location of the maximum or minimum of the cosine term alone without the phase angle. Then adjust the plot to take into account phase and vertical displacement.)

**2** The figure below shows a sinusoid that is digitally sampled according to

$$x_k = a + c\cos\left(\frac{2\pi mk}{N} - \phi_m\right), \quad k = 0, \pm 1, \pm 2, \ldots$$




From the above figure determine:

(a) $a =$ _____

(b) $c =$ _____

(c) an appropriate $m =$ ____ for an appropriate $N =$ ____

(d) $\phi_m =$ ____ degrees

**3** Use Appendix 1.B to show that

$$\sum_{n=0}^{N-1} \sin^2\left(\frac{2\pi kn}{N}\right) = \frac{N}{2}$$

where N is an even integer and $0 \le k \le N/2$.

**4** A time series of length $N\Delta t$ where $N = 50$ is obtained. It then is discovered that the last half of the series, $25\Delta t$, is a repeat of the first $25\Delta t$. How does the variance of the time series of length $50\Delta t$ compare with the variance of the time series of length $25\Delta t$?

**5** Manual Fourier Analysis: use only paper, pencil, and a nonprogrammable hand-held calculator.

The data below are 30-year normal monthly precipitation values for 1971–2000 at San Francisco International Airport (SFO AP), California (37.62 N, 122.40 W) and Oklahoma City Will Rogers Airport (OKC AP), Oklahoma (35.38 N, 97.60 W).

(a) Plot the data on separate graphs and comment on differences you observe between the time series. Can you provide a meteorological explanation for the differences in annual total precipitation regimes?

(b) Choose one of the time series and perform a Fourier analysis of the data sufficient to detect both amplitudes and phase angles of the significant harmonics present (i.e., find enough harmonics to explain at least 95% of the variance in the data).

(c) On a separate graph, plot the significant waves in (b) in the form of the amplitude and phase representation discussed in Section 1.2.4.

(d) Plot the sum of the significant waves (plus the mean) on the time series graph in (a).

(e) What percentage of the variance of the observed series does each harmonic explain?

(f) Compare the observed variance (that of the data set itself) with the explained variance to obtain residual variance.

| Month | San Francisco International Airport Precipitation (mm) | Oklahoma City Will Rogers Airport Precipitation (mm) |
|---|---|---|
| January | 113.0 | 32.5 |
| February | 101.9 | 39.6 |
| March | 82.8 | 73.7 |
| April | 30.0 | 76.2 |
| May | 9.7 | 138.2 |
| June | 2.8 | 117.6 |
| July | 0.8 | 74.7 |
| August | 1.8 | 63.0 |
| September | 5.1 | 101.1 |
| October | 26.4 | 92.5 |
| November | 63.2 | 53.6 |
| December | 73.4 | 48.0 |

**6** Fourier Analysis Using a Computer Program

In this problem we use the paradrop days data in Table 1.3 that were discussed in Section 1.3.2.

(a) Write a computer program that will find the cosine amplitudes, sine amplitudes and phase angles for the largest harmonics that explain at least 95% of the variance.

(b) Convert the phase angles into actual times of the maximum amplitude for the various harmonics.

(c) Plot (1) the original data, (2) each of the largest harmonics, the sum of which explains at least 95% of the variance, and (3) the sum of these harmonics, all on one graph. Compare your results with Figure 1.14.

(d) Can you attach any physical meaning to the individual harmonics? You might consider the typical cycle of daily wind, for example. Does it have a sinusoidal shape?

**7** Recall that the variance of rv X is given by

$$\mathrm{Var}[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx.$$

Let rv X have a uniform probability density function f(x) between a and b and zero elsewhere. If $b - a = 1$, find the variance of rv X for this rectangular distribution.

**8** The observed variance in a periodogram at harmonic k is $8\ ^\circ C^2$. The goal is to find the limits of the 95% *a priori* confidence interval for the population variance $\Gamma(f_k)$ at harmonic k. Assume that

$$\frac{C(f_k)}{\Gamma(f_k)} = \chi_2^2/2$$

(a) Write down the appropriate probability statement(s) of the form $\Pr\{\_\_\} = \_\_$ for the confidence limits on the population variance

(b) What are the upper and lower limits of the 95% *a priori* confidence interval? Recall that

$$f_{\chi_2^2}(x) = \frac{1}{2}e^{-x/2}$$

**9** The observed variance in a periodogram of a time series with $N = 41$ data is found to be $12\ m^2$. The null hypothesis is made that the sample of data comes from a white noise process. Find the limits of the 95% *a posteriori* confidence interval for the observed variances at the harmonic frequencies.

**10** Consider a time series comprising $N = 51$ data with variance $= 40\ Pa^2$. The null hypothesis $H_0$ is made that the realization is from a data population that is white noise. A periodogram of the time series is calculated and the largest value in the periodogram is $10.45\ Pa^2$ and the smallest is $0.0065\ Pa^2$. Show whether $H_o$ will be rejected or not rejected.

**11** If a signal can be described by $x_n = A \sin(2\pi f n \Delta t)$ in which $\Delta t = 0.1$ s and $f = 12$ Hz, at which frequencies in the principal part of the complete aliased spectrum will the variance be observed and what will be the variance at each frequency?

**12** Suppose you have a data set that comprises 100 values of wind speed in which the sampling interval is two seconds. Unbeknownst to you, there was a strong sinusoid with period 1.6 seconds introduced into the analog signal (i.e., before digitization) because of a defective electronic component. A periodogram analysis of the data set is performed.

(a) What is the Nyquist frequency in Hz?

(b) At what positive frequency (in Hz) in the principal part of the aliased spectrum will the erroneous variance occur?

(c) What is the corresponding harmonic number for the frequency found in (b)?

(d) What can be done or what should have been done to eliminate the unwanted signal from appearing in the periodogram? Explain.

**13** An analog temperature signal is sampled once every second. The number of data collected is 40. Unfortunately, a nearby transmitter has added an unwanted frequency of 1.125 Hz.

(a) At what *frequencies* (Hz) in the principal part of the (two-sided) aliased spectrum will the unwanted variance appear?

(b) What are the corresponding *harmonics* in the principal part of the aliased spectrum at which the variances occur?

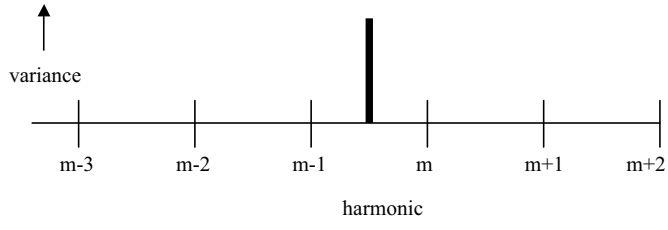
**14** Consider a stagecoach scene in a motion picture (e.g., *How the West Was Won*). The wheels of the stagecoach have a radius $r = 0.6$ m and each has eight spokes. Assume the camera shutter speed is 24 frames per second.

Plot the perceived (which may be the actual) angular speed (radians/second) of one of the wheels versus the speed of the stagecoach as it increases from 0 m/s to the speed at which the wheels are perceived to be stationary, that is, not rotating. (Hint: Sketch an eight-spoke wheel, write down the equation for the stagecoach speed in terms of the angular speed of a wheel, then adapt it to the conditions of the problem.)

**15** Under certain conditions the spectrum window function of the form $[(\sin x)/x]^2$ can be used to estimate the variance at harmonic frequencies due to variance in the data at nonharmonic frequencies.

(a) What are the two primary conditions?

(b) Assume the conditions of (a) are met. Sketch at least one spectrum window centered at a harmonic and calculate the variance at harmonics m − 1, m, and m + 1 based on the figure below. The one-sided variance input shown in the figure is $10\,m^2\,s^{-2}$ located midway between harmonics m −1 and m.




**16** The objectives of this problem are to compare periodograms of hourly temperature for January and July, 2009 at Oklahoma City, OK, and determine whether the hourly temperatures in these months can be modeled as a white noise process after removal of the daily cycle.

***Data***

The data are available on the website http://www.wiley.com/go/duchon/timeseriesanalysis. The filenames are OKC_200901_hrly_temp.xls and OKC_200907_hrly_temp.xls. The data are hourly temperatures in degrees Celsius for January and July 2009. The first column is the sequential hour count, the second column is the date, the third column is the time the temperature was observed in Central Standard Time, and the fourth column is the temperature. The only data needed to work this problem are the hourly temperatures in the fourth column.

(a) Plot the times series of hourly temperature for each month on separate sheets of paper, using the same size for all your plots. Show on each plot frontal passages, cloudy days, clear days, and any other meteorological events that you believe to be present.

(b) Use the Fourier Analysis computer program you designed in problem 6 or subroutine FORANX in Appendix 1.A to compute the periodogram of the 744 points for each month. Plot the $\log_{10}$ variance (or variance on a $\log_{10}$ axis) versus frequency, period, or harmonic for all harmonics. On each plot show the total variance and bandwidth associated with each plot. Place the plots on separate pages.

(c) Compute and plot the *average daily cycle* of temperature for each month. Briefly discuss the principal differences between the two months and their causes.

(d) Remove the daily cycle from the original hourly data for each month to form two new time series (the "uncontaminated" data). Plot the two time series of "uncontaminated" data. Comment on the presence or absence of the daily cycle of temperature.

(e) Plot the periodograms of the uncontaminated hourly data, replacing the variance estimates at the harmonic frequencies of the daily cycle with the average of surrounding variances. On each plot show the total variance and the bandwidth associated with each estimate.

(f) Apply a white noise test to each periodogram in (e). Compute the *a priori* confidence limits and *a posteriori* confidence limits. Place them on the periodograms of variance in which the vertical axis is $\log_{10}$ variance. Do you accept or reject the white noise null hypothesis? If you reject the hypothesis that the sample comes from a population of white noise, what physical phenomenon or phenomena do you think led to its rejection?

## References


Bloomfield, P. (1976) *Fourier Analysis of Time Series: An Introduction*. John Wiley & Sons, Inc., New York.

Box, G.E.P., and Jenkins, G.M. (1970) *Time Series Analysis Forecasting and Control*. Holden-Day, San Francisco, CA.

Crowley, K.D., Duchon, C.E., and Rhi, J. (1986) Climate record in varved sediments of the eocene Green River formation. *J. Geophys. Res.*, **91**(D8), 8637–8647.

Daubechies, I. (1992) *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series on Applied Mathematics, No. 61. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Hoel, P.G. (1962) *Introduction to Mathematical Statistics*. John Wiley & Sons, Inc., New York.

Jenkins, G.M., and Watts, D.G. (1968) *Spectral Analysis and its Applications*. Holden-Day, San Francisco, CA.

Koopmans, L.H. (1974) *The Spectral Analysis of Time Series*. Academic Press, New York.

McPherson, R.A., Fiebrich, C.A., Crawford, K.C. *et al.* (2007) Statewide monitoring of the mesoscale environment: a technical update on the Oklahoma Mesonet. *J. Atmos. Oceanic Technol.*, **24**, 301–321.

Nyquist, H. (1928) Certain topics in telegraph transmission theory. *Trans. Am. Inst. Electr. Eng.*, **47**, 617–644.

Robinson, E. A. (1982) A historical perspective on spectrum estimation. *Proc. IEEE*, **70**(9), 885–907.