13. Time Series Analysis

13.1 Time Series: Trend

In the previous chapter we considered the dependence of a random variable y on a controlled variable t. As in that case we will assume here that y consists of two parts, the true value of the measured quantity η and a measurement error ε,

$$y_i = \eta_i + \varepsilon_i \quad , \qquad i = 1, 2, \ldots, n \quad . \tag{13.1.1}$$

In Chap. 12 we assumed that η_i was a polynomial in t. The measurement error ε_i was considered to be normally distributed about zero.

We now want to make less restrictive assumptions about η. In this chapter we call the controlled variable t "time", although in many applications it could be something different. The method we want to discuss is called *time series analysis* and is often applied in economic problems. It can always be used where one has little or no knowledge about the functional relationship between η and t. In considering time series problems it is common to observe the y_i at equally spaced points in time,

$$t_i - t_{i-1} = \Delta t = \text{const} \quad , \tag{13.1.2}$$

since this leads to a significant simplification of the formulas.

An example of a time series is shown in Fig. 13.1. If we look first only at the measured points, we notice strong fluctuations from point to point. Nevertheless they clearly follow a certain trend. In the left half of the plot they are mostly positive, and in the right half, mostly negative. One could qualitatively obtain the average time dependence by drawing a smooth curve by hand through the points. Since, however, such curves are not free from personal influences, and are thus not reproducible, we must try to develop an objective method.

S. Brandt, *Data Analysis: Statistical and Computational Methods for Scientists and Engineers*, DOI 10.1007/978-3-319-03762-2_13,

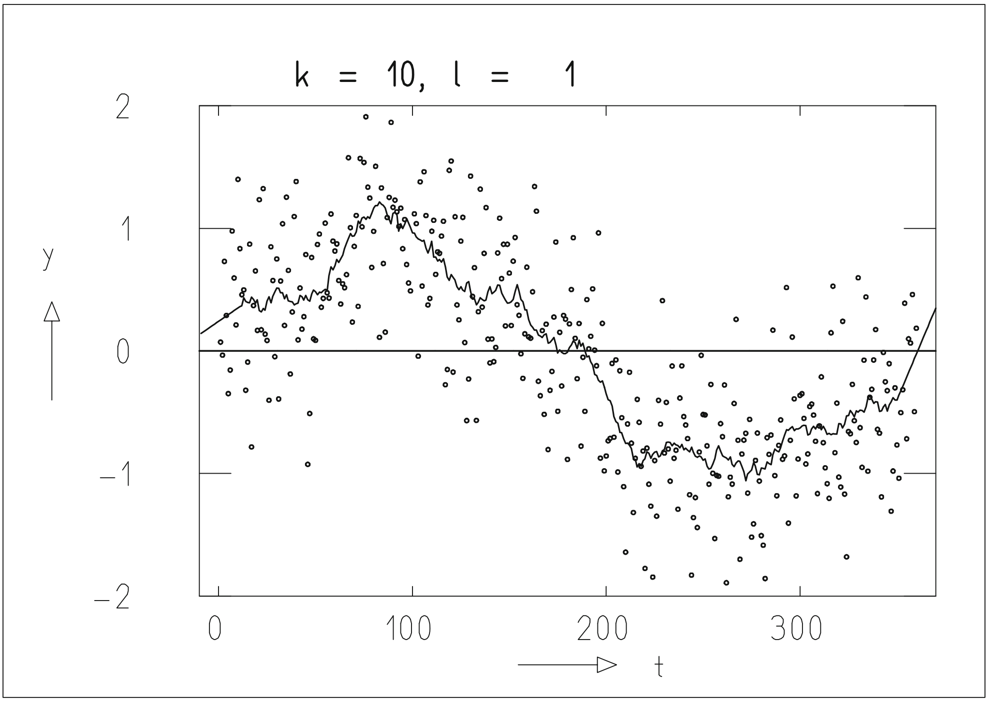

Fig. 13.1: Data points (*circles*) and moving average (joined by line segments).

We use the notation from (13.1.1) and call η_i the *trend* and ε_i the *random component* of the *measurement* y_i. In order to obtain a smoother function of t, one can, for example, construct for every value of y_i the expression

$$u_i = \frac{1}{2k+1} \sum_{j=i-k}^{i+k} y_j \quad , \tag{13.1.3}$$

i.e., the unweighted mean of the measurements for the times

$$t_{i-k},\, t_{i-k+1},\, \ldots,\, t_{i-1},\, t_i,\, t_{i+1},\, \ldots,\, t_{i+k} \quad .$$

The expression (13.1.3) is called a *moving average* of y.

13.2 Moving Averages

Of course the moving average (13.1.3) is a very simple construction. We will show later (in Example 13.1) that the use of a moving average of this form is equivalent to the assumption that η is a linear function of time in the interval considered,

$$\eta_j = \alpha + \beta t_j \quad , \qquad j = -k, -k+1, \ldots, k \quad . \tag{13.2.1}$$

Here α and β are constants. They can be estimated from the data by linear regression.

Instead of restricting ourselves to the linear case, we will assume more generally that η can be a polynomial of order ℓ.

In the averaging interval, t takes on the values

$$t_j = t_i + j\Delta t \quad , \qquad j = -k, -k+1, \ldots, k \quad . \tag{13.2.2}$$

Because η is a polynomial in t,

$$\eta_j = a_1 + a_2 t_j + a_3 t_j^2 + \cdots + a_{\ell+1} t_j^\ell \quad , \tag{13.2.3}$$

it is also a polynomial in j,

$$\eta_j = x_1 + x_2 j + x_3 j^2 + \cdots + x_{\ell+1} j^\ell \quad , \tag{13.2.4}$$

since (13.2.2) describes a linear transformation between t_j and j, i.e., it is merely a change of scale. We now want to obtain the coefficients x_1, x_2, ..., $x_{\ell+1}$ from the data by fitting with least squares. This task has already been treated in Sect. 9.4.1. We assume (in the absence of any better knowledge) that all of the measurements are of the same accuracy. Thus the matrix $G_y = aI$ is simply a multiple of the unit matrix I. According to (9.2.26), the vector of coefficients is thus given by

$$\widetilde{\mathbf{x}} = -(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}\mathbf{y} \quad , \tag{13.2.5}$$

where A is a $(2k+1) \times (\ell+1)$ matrix,

$$A = -\begin{pmatrix} 1 & -k & (-k)^2 & \ldots & (-k)^\ell \\ 1 & -k+1 & (-k+1)^2 & \ldots & (-k+1)^\ell \\ \vdots & & & & \\ 1 & k & k^2 & \ldots & k^\ell \end{pmatrix} \quad . \tag{13.2.6}$$

For the trend $\widetilde{\eta}_0$ at the center of the averaging interval ($j = 0$) we obtain from (13.2.4) the estimate

$$\widetilde{\eta}_0 = \widetilde{x}_1 \quad . \tag{13.2.7}$$

It is equal to the first coefficient of the polynomial. According to (13.2.5), $\widetilde{x}_1$ is obtained by multiplication of the column vector of measurements $\mathbf{y}$ on the left with the row vector

$$\mathbf{a} = (-(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}})_1 \quad , \tag{13.2.8}$$

i.e., with the first row of the matrix $-(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$. We obtain

$$\widetilde{\eta}_0 = \mathbf{a}\mathbf{y} = a_{-k}y_{-k} + a_{-k+1}y_{-k+1} + \cdots + a_0 y_0 + \cdots + a_k y_k \quad . \tag{13.2.9}$$

Table 13.1: Components of the vector **a** for computing moving averages.

$$\mathbf{a} = (a_{-k}, a_{-k+1}, \ldots, a_k) = \frac{1}{A}(\alpha_{-k}, \alpha_{-k+1}, \ldots, \alpha_k)$$

$$\alpha_{-j} = \alpha_j$$

$\ell = 2$ and $\ell = 3$

k	A	α_{-7}	α_{-6}	α_{-5}	α_{-4}	α_{-3}	α_{-2}	α_{-1}	α_0
2	35						−3	12	17
3	21					−2	3	6	7
4	231				−21	14	39	54	59
5	429			−36	9	44	69	84	89
6	143		−11	0	9	16	21	24	25
7	1105	−78	−13	42	87	122	147	162	167

$\ell = 4$ and $\ell = 5$

k	A	α_{-7}	α_{-6}	α_{-5}	α_{-4}	α_{-3}	α_{-2}	α_{-1}	α_0
3	231					5	−30	75	131
4	429				15	−55	30	135	179
5	429			18	−45	−10	60	120	143
6	2431		110	−198	−135	110	390	600	677
7	46 189	2145	−2860	−2937	−165	3755	7500	10 125	11 063

This is a linear function of the measurements within the averaging interval. Here the vector **a** does not depend on the measurements, but rather only on ℓ and k, i.e., on the order of the polynomial and on the length of the interval. Clearly one must choose

$$\ell < 2k + 1 \quad ,$$

since otherwise there would not remain any degrees of freedom for the least-squares fit. The components of **a** for small values of ℓ and k can be obtained from Table 13.1.

Equation (13.2.9) describes the moving average corresponding to the assumed polynomial (13.2.4). Once the vector **a** is determined, the moving averages

$$u_i = \widetilde{\eta}_0(i) = \mathbf{a}\mathbf{y}(i) = a_1 y_{i-k} + a_2 y_{i-k+1} + \cdots + a_{2k+1} y_{i+k} \qquad (13.2.10)$$

can easily be computed for each value of i.

Example 13.1: Moving average with linear trend
In the case of a linear trend function,

$$\eta_j = x_1 + x_2 j \quad ,$$

the matrix A becomes simply

$$A = -\begin{pmatrix} 1 & -k \\ 1 & -k+1 \\ \vdots & \\ 1 & k \end{pmatrix} \quad .$$

One then has

$$\begin{aligned}
A^{\mathrm{T}}A &= \begin{pmatrix} 1 & 1 & \dots & 1 \\ -k & -k+1 & \dots & k \end{pmatrix} \begin{pmatrix} 1 & -k \\ 1 & -k+1 \\ \vdots & \vdots \\ 1 & k \end{pmatrix} \\
&= \begin{pmatrix} 2k+1 & 0 \\ 0 & k(k+1)(2k+1)/3 \end{pmatrix} \quad , \\
(A^{\mathrm{T}}A)^{-1} &= \begin{pmatrix} \dfrac{1}{2k+1} & 0 \\ 0 & \dfrac{3}{k(k+1)(2k+1)} \end{pmatrix} \quad , \\
\mathbf{a} &= (-(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}})_1 = \frac{1}{2k+1}(1, 1, \dots, 1) \quad .
\end{aligned}$$

In this case the moving average is simply the unweighted mean (13.1.3). ■

For more complicated models one can obtain the vectors **a** either by solving (13.2.8) or simply from Table 13.1. Because of the symmetry of A, one can show that polynomials of odd order (i.e., $\ell = 2n$ with n an integer) have the same values of **a** as those of polynomials of the next lower order $\ell = 2n - 1$. One can also easily show that **a** has the symmetry

$$a_j = a_{-j} \qquad , \qquad j = 1, 2, \dots, k \quad . \qquad (13.2.11)$$

13.3 Edge Effects

Of course the moving average (13.2.10) can be used to estimate the trend only for points i that have at least k neighboring points both to the right and left, since the averaging interval covers $2k+1$ points. This means that for the first and last k points of a time series one must use a different estimator. One obtains the most obvious generalization of the estimator by extrapolating the polynomial (13.2.4) rather than using it only at the center of an interval. One then obtains the estimators

$$\begin{aligned}\widetilde{\eta}_i = u_i &= \widetilde{x}_1^{(k+1)} + \widetilde{x}_2^{(k+1)}(i-k-1) + \widetilde{x}_3^{(k+1)}(i-k-1)^2 + \cdots \\ &\quad + \widetilde{x}_{\ell+1}^{(k+1)}(i-k-1)^{\ell} \quad , \qquad i \le k \quad , \\ \widetilde{\eta}_i = u_i &= \widetilde{x}_1^{(n-k)} + \widetilde{x}_2^{(n-k)}(i+k-n) + \widetilde{x}_3^{(n-k)}(i+k-n)^2 + \cdots \\ &\quad + \widetilde{x}_{\ell+1}^{(n-k)}(i+k-n) \quad , \qquad i > n-k \quad . \end{aligned} \tag{13.3.1}$$

Here the notation $\widetilde{x}^{(k+1)}$ and $\widetilde{x}^{(n-k)}$ indicates that the coefficients $\widetilde{x}$ were determined for the first and last intervals of the time series for which the centers are at $(k+1)$ and $(n-k)$.

The estimators are now defined even for $i < 1$ and $i > n$. They thus offer the possibility to continue the time series (e.g., into the future). Such extrapolations must be treated with great care for two reasons:

(i) Usually there is no theoretical justification for the assumption that the trend is described by a polynomial. It merely simplifies the computation of the moving average. Without a theoretical understanding for a trend model, the meaning of extrapolations is quite unclear.

(ii) Even in cases where the trend can rightly be described by a polynomial, the confidence limits quickly diverge from the estimated polynomial in the extrapolated region. The extrapolation becomes very inaccurate.

Whether point (i) is correct must be carefully checked in each individual case. The more general point (ii) is already familiar from the linear regression (cf. Fig. 12.1). We will investigate this in detail in the next section.

13.4 Confidence Intervals

We first consider the confidence interval for the moving average u_i from Eq. (13.2.10). The errors of the measurements y_i are unknown and must therefore first be estimated. From (12.3.2) one obtains for the sample variance of the y_j in the interval of length $2k+1$,

$$s_y^2 = \frac{1}{2k-\ell} \sum_{j=-k}^{k} (y_j - \widetilde{\eta}_j)^2 \quad , \tag{13.4.1}$$

where $\widetilde{\eta}_j$ is given by

$$\widetilde{\eta}_j = \widetilde{x}_1 + \widetilde{x}_2 j + \widetilde{x}_3 j^2 + \cdots + \widetilde{x}_{\ell+1} j^\ell \quad . \tag{13.4.2}$$

The covariance matrix for the measurements can then be estimated by

$$G_y^{-1} \approx s_y^2 I \quad . \tag{13.4.3}$$

The covariance matrix of the coefficients $\mathbf{x}$ is then given by (9.2.27),

$$G_{\widetilde{x}}^{-1} \approx (A^{\mathrm{T}} G_y A)^{-1} = s_y^2 (A^{\mathrm{T}} A)^{-1} \quad . \tag{13.4.4}$$

Since $u_i = \widetilde{\eta}_0 = \widetilde{x}_1$, we thus have for an estimator of the variance of u_i

$$s_{\widetilde{x}_1}^2 = (G_{\widetilde{x}}^{-1})_{11} = s_y^2 ((A^{\mathrm{T}} A)^{-1})_{11} = s_y^2 a_0 \quad . \tag{13.4.5}$$

From (13.2.6), (13.2.7), and (13.2.8) one easily obtains that $(A^{\mathrm{T}} A)_{11}^{-1} = a_0$, since the middle row of A is $-(1, 0, 0, \ldots, 0)$.

Using the same reasoning as in Sect. 12.3 we obtain at a confidence level of $1 - \alpha$

$$\frac{|\widetilde{\eta}_0(i) - \eta_0(i)|}{s_y a_0} \le t_{1-\frac{1}{2}\alpha} \quad . \tag{13.4.6}$$

For a given α we can give the confidence limits as

$$\eta_0^{\pm}(i) = \widetilde{\eta}_0(i) \pm a_0 s_y t_{1-\frac{1}{2}\alpha} \quad . \tag{13.4.7}$$

Here $t_{1-\frac{1}{2}\alpha}$ is a quantile of Student's distribution for $2k - \ell$ degrees of freedom. The true value of the trend lies within these limits with a confidence level of $1 - \alpha$.

Completely analogous, although more difficult computationally, is the determination of confidence limits at the ends of the time series. The moving average is now given by (13.3.1). Labeling the arguments in the expressions (13.3.1) $j = i - k - 1$ and $j = i + k - n$, we obtain

$$\widetilde{\eta} = \mathbf{T}\mathbf{x} \quad . \tag{13.4.8}$$

Here $\mathbf{T}$ is a row vector of length $\ell + 1$,

$$\mathbf{T} = (1, j, j^2, \ldots, j^\ell) \quad . \tag{13.4.9}$$

According to the law of error propagation (3.8.4) we obtain

$$G_{\widetilde{\eta}}^{-1} = T G_x^{-1} T^{\mathrm{T}} \quad . \tag{13.4.10}$$

With (13.4.4) we finally have

$$G_{\widetilde{\eta}}^{-1} \approx s_{\widetilde{\eta}}^2 = s_y^2 T (A^{\mathrm{T}} A)^{-1} T^{\mathrm{T}} \quad , \tag{13.4.11}$$

where s_y^2 is again given by (13.4.1).

The quantity $s_{\widetilde{\eta}}^2$ can now be computed for every value of j, even for values lying outside of the time series itself. Thus we obtain the confidence limits

$$\eta^{\pm}(i) = \widetilde{\eta}(i) \pm s_{\widetilde{\eta}} t_{1-\frac{1}{2}\alpha} \quad . \tag{13.4.12}$$

Caution is always recommended in interpreting the results of a time series analysis. This is particularly true for two reasons:

1. There is usually no a priori justification for the mathematical model on which the time series analysis is based. One has simply chosen a convenient procedure in order to “separate out” statistical fluctuations.

2. The user has considerable freedom in choosing the parameters k and ℓ, which, however, can have a significant influence on the results. The following example gives an impression of the magnitude of such influences.

Example 13.2: Time series analysis of the same set of measurements using different averaging intervals and polynomials of different orders

Figures 13.2 and 13.3 contain time series analyses of the average number of sun spots observed in the 36 months from January 1962 through December 1964. Various values of k and ℓ were used. The individual plots in Fig. 13.2 show $\ell = 1$ (linear averaging) but different interval lengths ($2k+1 = 5$, 7, 9, and 11). One can see that the curve of moving averages becomes smoother and the confidence interval becomes narrower when k increases, but that then the mean deviation of the individual observations from the curve also increases. The extrapolation outside the range of measured points is, of course, a straight line. (For $\ell = 0$ we would have obtained the same moving averages for the inner points. The outer and extrapolated points would lie, however, on a horizontal line, since a polynomial of order zero is a constant.) The plots in Fig. 13.3 correspond to the interval lengths $2k+1 = 7$ and $\ell = 1, 2, 3, 4$. The moving averages lie closer to the data points and the confidence interval becomes larger when the value of ℓ increases. ■

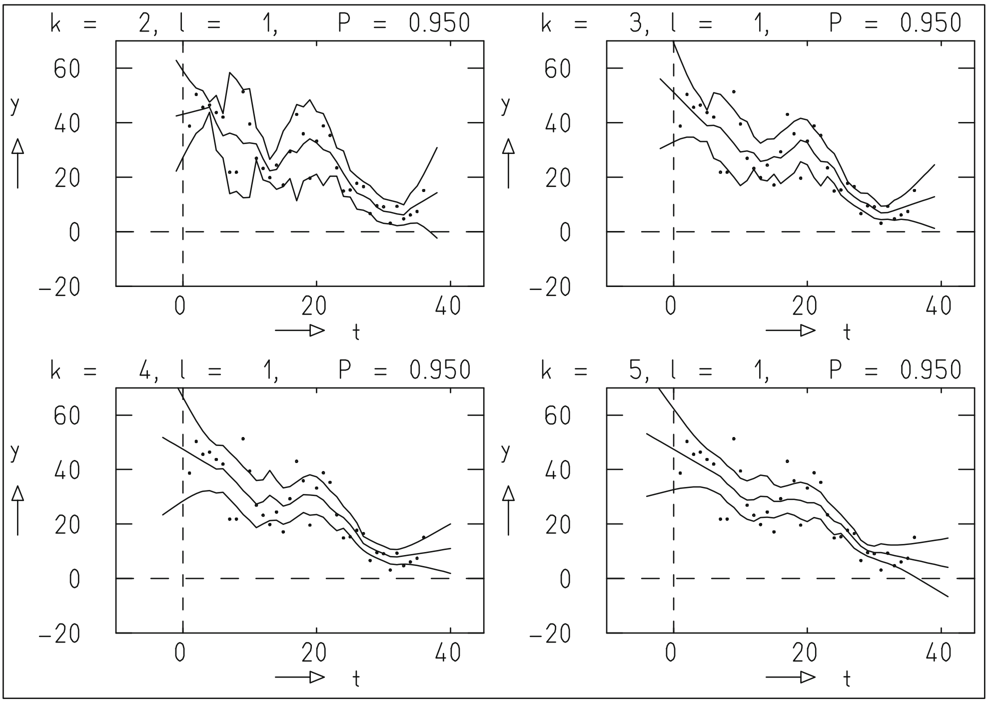

Fig. 13.2: Time series analyses of the same data with fixed ℓ and various values of k.

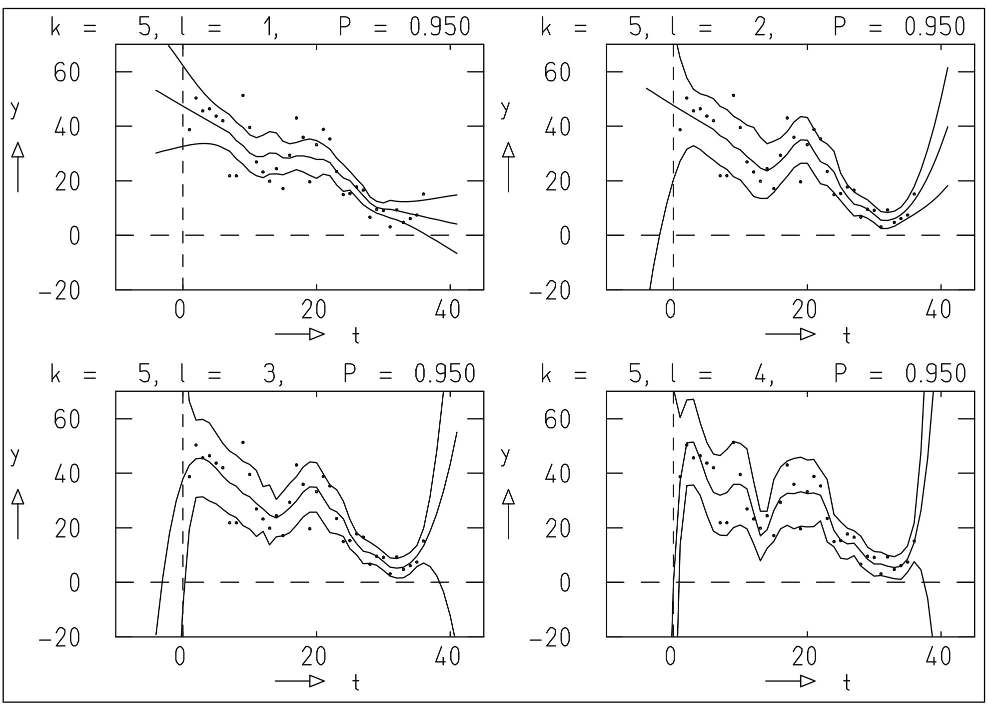

Fig. 13.3: Time series analyses of the same data with fixed k and various values of ℓ.

From these observations we can derive the following qualitative rules:

1. The averaging interval should not be chosen larger than the region where one expects that the data can be well described by a polynomial of the given order. That is, for $\ell = 1$ the interval $2k+1$ should be chosen such that the expected nonlinear effects within the interval remain small.

2. On the other hand, the smoothing effect becomes stronger as the length of the averaging interval increases. As a rule of thumb, the smoothing becomes more effective for increasing $2k+1-\ell$.

3. Caution is required in extrapolation of a time series, especially in nonlinear cases.

The art of time series analysis is, of course, much more highly developed than what we have been able to describe in this short chapter. The interested reader is referred to the specialized literature, where, for example, smoothing functions other than polynomials or multidimensional analyses are treated.

13.5 Java Class and Example Programs

Java Class for Time Series Analysis

`TimeSeries` performs a time series analysis.

Example Program 13.1: The class E1TimSer demonstrates the use of TimesSeries

The uses the data of Example 13.2. After setting some parameters it performs a time series analysis by a call of TimesSeries with $k = 2$, $\ell = 2$. The data, moving averages, and distances to the confidence limits (at a confidence level of 90 %) are output numerically.

Example Program 13.2: The class E2TimSer performs a time series analysis and yields graphical output

The program starts works on the same data as E1TimSer. It allows interactive input for the parameters k and ℓ and for the confidence level P and then performs a time series analysis. Subsequently a plot is produced in which the data are displayed as small circles. The floating averages are shown as a polyline. Polylines in a different color indicate the confidence limits.

Suggestion: Produce the individual plots of Figs. 13.2 and 13.3.