A Device for the Smoothing of Time Series

By Walter A. Hendricks

When sample estimates are plotted in time series, part of the irregularity displayed by the plotted points is chargeable to sampling errors in the individual estimates. This paper describes a method for smoothing the series to eliminate effects of the sampling errors. Fluctuations that can be ascribed legitimately to actual changes in the "true" values are retained.

WHEN ESTIMATES on the same item are made from sample data at periodic intervals, there is usually considerable interest in the change taking place from one survey date to the next. Anyone acquainted with the appearance of time series is familiar with the zig-zag pattern displayed by such successive estimates. From some viewpoints, we might be interested in the smoothest possible trend line that could be drawn on the chart. However, the change shown by this kind of trend line between any two successive dates presents a more idealized picture of actual change than is desired in many instances.

Even though the individual estimates are subject to sampling errors, only part of the average departure of the individual observations from the smooth trend line can usually be ascribed to these sampling errors. Some of the difference is caused by departures of the "true" values from the idealized smooth trend line. But how can one determine how much of the variation of the individual observations about the trend line can be ascribed to true deviations and how much to sampling errors?

To answer this question we need some device whereby the time-series chart can be smoothed only to the extent that effects of sampling errors in the estimates are removed. The difference between two successive adjusted estimates can then be interpreted as a valid estimate of the actual change. This problem was brought into prominence recently when a farm population estimate from a well-known national sample survey showed an extremely large decline from the previous year.

Any qualified statistician would realize that sampling errors in the two successive estimates were probably responsible for part of the observed decrease and that the true decrease was likely smaller than was indicated. This paper presents one device by which effects of sampling errors can be eliminated from a time series when the size of the sampling error is known. It also shows how current estimates can be adjusted for sampling errors.

An Illustrative Series

To illustrate the nature of the problem, consider the following average reported prices per pound paid by farmers for coffee in Georgia on four quarterly survey dates in 1956, 1957, and 1958.

Year	Quarter	Price per pound (cents)
1956	1	97. 6
	2	97. 1
	3	103. 0
	4	103. 0
1957	1	108. 0
	2	104. 0
	3	102. 0
	4	99. 3
1958	1	97. 0
	2	93. 2
	3	90. 4
	4	89. 1

The average price for each quarter has a standard error of about plus or minus 1 cent. This error was computed from the variability of the individual reports received in any one survey and, of course, it varies slightly from one quarter to another. However, the average standard error of about 1 cent is adequate for the computations to follow.

The problem at hand is to smooth the time series of average prices in such a way as to eliminate the effects of the sampling errors. If there were no trends in the average price level during the 3-year period, this smoothing could be accomplished by applying the regression equation:

$$Y-\bar{y}=b(Y-\bar{y}) \qquad (1)$$

In this equation, Y is the average reported price for any one date, $\hat{Y}$ is an estimate of the true price, and $\bar{y}$ is the average price for all 12 quarters. The regression coefficient b can be deduced immediately from the observed variability of the 12 values of Y about $\bar{y}$ and the sampling error in any one Y:

$$b=(\sigma_y^2-\sigma^2)/\sigma_y^2 \qquad (2)$$

In this equation σ_y^2 is the variance of Y about $\bar{y}$ and σ^2 is the squared sampling error in Y.

The equation for estimating b takes this form because an observed deviation, $\hat{y}$, consists of a true value, $\hat{y}$, plus an independent random error component. The regression coefficient of $\hat{y}$ on y is given by dividing the expected value of $y\hat{y}$ by the expected value of y^2. It is clear that $E(y^2)=\sigma_y^2$ and $E(y\hat{y})=E[(\hat{y}+e)\hat{y}]=\sigma_{\hat{y}}^2=\sigma_y^2-\sigma^2$.

Unfortunately, the data at hand, and in fact most series of this kind, contain varying trends over time. Equation (1) needs to be modified to read:

$$\hat{Y}-\bar{y}_t=b(Y-\bar{y}_t) \qquad (3)$$

This equation takes account of the change in the level of $\bar{y}$ from one point in time to another. Some modification is also required in the computation of b. Equation (2) needs to be modified so that σ_y^2 is now estimated from the average of squared deviations of the form $(Y-\bar{y}_t)^2$.

To use these relationships, we need estimates of $\bar{y}_t$. The problem can be approached by considering the first three points in the series. The average of the first and third reported prices, $(97.6+103.0)/2=100.3$, can be regarded as an estimate of $\bar{y}_t$ corresponding to the second reported price, 97.1. The squared deviation is $(97.1-100.3)^2=10.24$. An estimate of σ_y^2 from that one degree of freedom is given by $(2)(10.24)/3=6.826$.

This follows directly from the fact that when we have three values of Y, spaced at equal time intervals, the two independent degrees of freedom involved can be represented by Y_3-Y_1, which measures the time trend, and $Y_2-\bar{y}_t=Y_2-(Y_1+Y_3)/2$ which measures the effect of the deviations of the three values of Y from this linear time trend. The mean square associated with this second degree of freedom, as computed in analysis of variance, is $(2/3)\ (Y_2-\bar{y}_t)^2$. This mean square is an estimate of σ_y^2.

If computations of this kind are made by grouping reported prices successively by 3's in an overlapping fashion, we obtain the following 10 deviations, each representing the deviation of a given Y from its corresponding $\bar{y}_t$. These deviations are, $Y-\bar{y}_t=$

−3.20
+2.95
−2.50
+4.50
−1.00
+ .35
− .20
+ .75
− .50
− .75

These deviations are not independent of each other because successive groups of observations overlap. However, two-thirds of the average of the squares of the 10 deviations, 3.199, can be regarded as an estimate of σ_y^2. The regression coefficient then becomes

$$b=(3.199-1.000)/3.199=0.687.$$

The computations described are useful in illustrating the theoretical aspects of the procedure. But for practical work, the computations can be set up more conveniently. Any one deviation of the form $Y-\bar{y}_t$ is nothing more than half the negative of a second difference of Y. The computation of the first and second differences of Y is shown in table 1.

The estimate of σ_y^2 can be obtained by squaring the second differences in table 1, computing the average of these squares, and dividing by 6. The regression coefficient b is computed as before.

TABLE 1.—*Computation of first and second differences of Y*

Year	Quarter	Y	$\Delta(Y)$	$\Delta^2(Y)$
	1	97.6	------	------
1956	2	97.1	−0.5	------
	3	103.0	+5.9	+6.4
	4	103.0	.0	−5.9
	1	108.0	+5.0	+5.0
1957	2	104.0	−4.0	−9.0
	3	102.0	−2.0	+2.0
	4	99.3	−2.7	−.7
	1	97.0	−2.3	+.4
1958	2	93.2	−3.8	−1.5
	3	90.4	−2.8	+1.0
	4	89.1	−1.3	+1.5

Table 2.—*Computation of smoothed average prices*

Year	Quarter	$\Delta^2(\hat{Y})$	$\Delta(\hat{Y})$	$\hat{Y}$
1956	1	-------	-------	99.0
	2	-------	−0.5	98.5
	3	+4.4	+3.9	102.4
	4	−4.1	−.2	102.2
1957	1	+3.4	+3.2	105.4
	2	−6.2	−3.0	102.4
	3	+1.4	−1.6	100.8
	4	−.5	−2.1	98.7
1958	1	+.3	−1.8	96.9
	2	−1.0	−2.8	94.1
	3	+.7	−2.1	92.0
	4	+1.0	−1.1	90.9

The process of smoothing the series of quarterly prices now reduces to replacing the original series by a new series in which the second differences of Y are equal to $b.\Delta^2(Y)$. From these adjusted second differences, we can compute the corresponding first differences and values of $\hat{Y}$.

The adjusted second differences of Y are shown in the third column of table 2.

The first differences in the fourth column of the table are computed by cumulative addition of the second differences, starting with a leading term in the column of first differences.

If we represent this leading term by d, the 11 first differences are:

$$
\begin{array}{l}
d \\
d+4.4 \\
d+\ .3 \\
d+3.7 \\
d-2.5 \\
d-1.1 \\
d-1.6 \\
d-1.3 \\
d-2.3 \\
d-1.6 \\
d-\ .6
\end{array}
$$

The estimate of d is derived by imposing the condition that the sum of these 11 first differences must be equal to the difference between the first and last values of Y in the original data. This gives $11d-2.6=89.1-97.6$ or $d=-0.5$. Starting with this value of d, it is possible to compute the 11 first differences shown in the fourth column of table 2.

The adjusted values of Y shown in the last column of the table are obtained in similar fashion by cumulative addition, starting with a leading term in that column:

$$
\begin{array}{l}
\hat{Y}_o \\
\hat{Y}_o-0.5 \\
\hat{Y}_o+3.4 \\
\hat{Y}_o+3.2 \\
\hat{Y}_o+6.4 \\
\hat{Y}_o+3.4 \\
\hat{Y}_o+1.8 \\
\hat{Y}_o-\ .3 \\
\hat{Y}_o-2.1 \\
\hat{Y}_o-4.9 \\
\hat{Y}_o-7.0 \\
\hat{Y}_o-8.1
\end{array}
$$

The estimate of $\hat{Y}_o$ is obtained by imposing the condition that the sum of the 12 values of $\hat{Y}$ must equal the sum of the 12 values of Y. This gives $12\hat{Y}_o-4.7=1183.7$ or $\hat{Y}_o=99.0$. Starting with this leading term, the values of $\hat{Y}$ in the last column of table 2 are computed by successive addition of the appropriate first differences in column 4.

Application to a Current Estimate

A reported average price for the first quarter of 1959 was not yet available when this paper was written. But for purposes of illustration, suppose it was reported at 85.0 cents.

When this observation is included with the original series of data, the first difference corresponding to that date is:

$$\Delta(Y)=85.0-89.1=-4.1.$$

The second difference is:

$$\Delta^2(Y)=-4.1-(-1.3)=-2.8.$$

The adjusted second difference is:

$$\Delta^2(\hat{Y})=(0.687)(-2.8)=-1.9.$$

The corresponding first difference is given by adding this figure to the cumulative total in the fourth column of table 2:

$$\Delta(\hat{Y})=-1.1-1.9=-3.0.$$

The adjusted average price is then given by adding this first difference to the last adjusted price in the last column of table 2:

$$\hat{Y}=90.9-3.0=87.9.$$

In other words, if the reported average price for the first quarter of 1959 was 85.0 cents, the best estimate of the actual price on that date would be 87.9 cents. In terms of change from the preceding date, the decrease would be 3.0 cents rather than the 4.1 cents observed in the raw data.