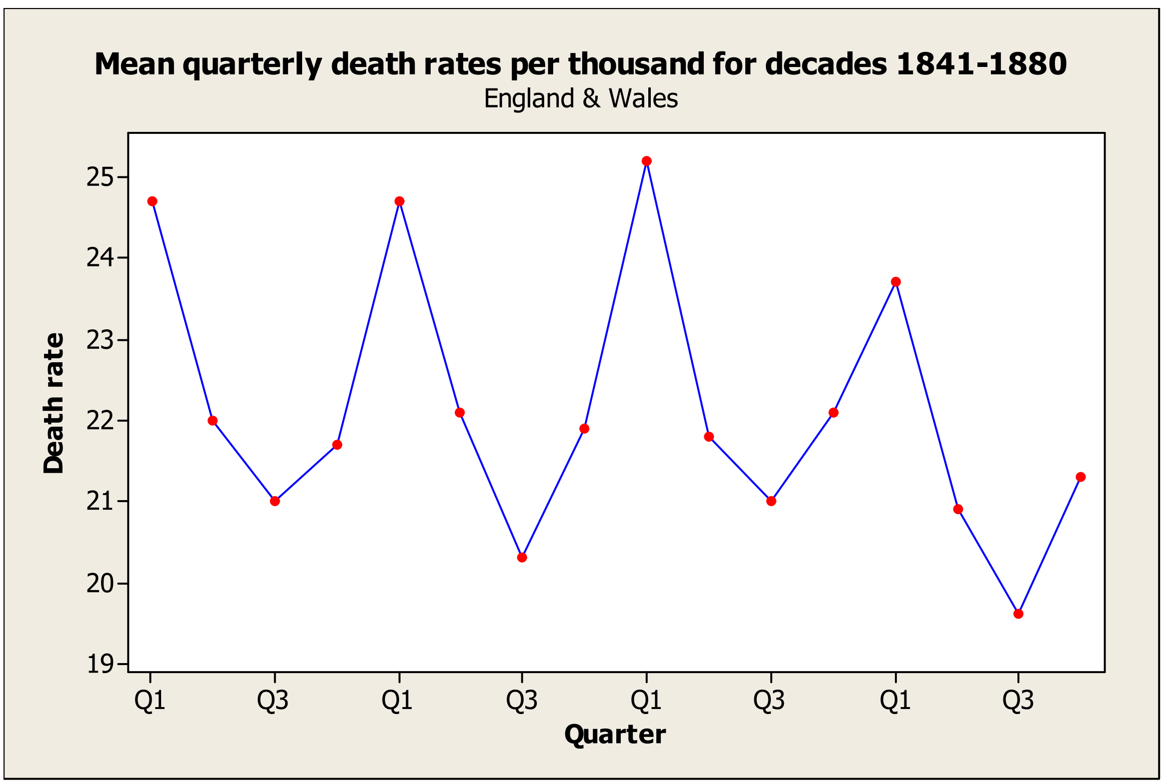
Activity Solution: The Sample Autocorrelation

Recall the "death rate" data gives the death rates, per thousand people, for England and Wales during year-quarters of four decades of the nineteenth century. The figures are given below, along with the plot.

		Quarter			
		1	2	3	4
Decade	1841-50	24.7	22.0	21.0	21.7
	1851-60	24.7	22.1	20.3	21.9
	1861-70	25.2	21.8	21.0	22.1
	1871-80	23.7	20.9	19.6	21.3
		98.3	86.8	81.9	87.0



The sample mean of the series is 22.1 and the standard deviation is 1.6 (in years, both to 1 d.p. which is sufficient accuracy for here).

1. *Clickers question:* Without performing the calculation, provide a guess of the sample autocorrelation function (the sample acf, r_k) at lag 4.

That is, have a guess at r_4.

Looking at the scatter plot one would expect quite a high, positive correlation between values four time units apart. So a guess of $r_4 \approx 0.7$ *would be reasonable. To verify, plot* $x(t+4)$ *against* $x(t)$ *for the data above.*

2. *Clickers question:* Using the formula, find r_4 from the data.
   From the data, we should find

$$\begin{aligned} r_4 &= \frac{\sum_{t=1}^{12} (x(t) - \bar{x})(x(t+4) - \bar{x})}{\sum_{t=1}^{16} (x(t) - \bar{x})^2} \\ &= \frac{\sum_{t=1}^{12} (x(t) - 22.1)(x(t+4) - 22.1)}{15 \times 1.6^2} \\ &= 0.672. \end{aligned}$$

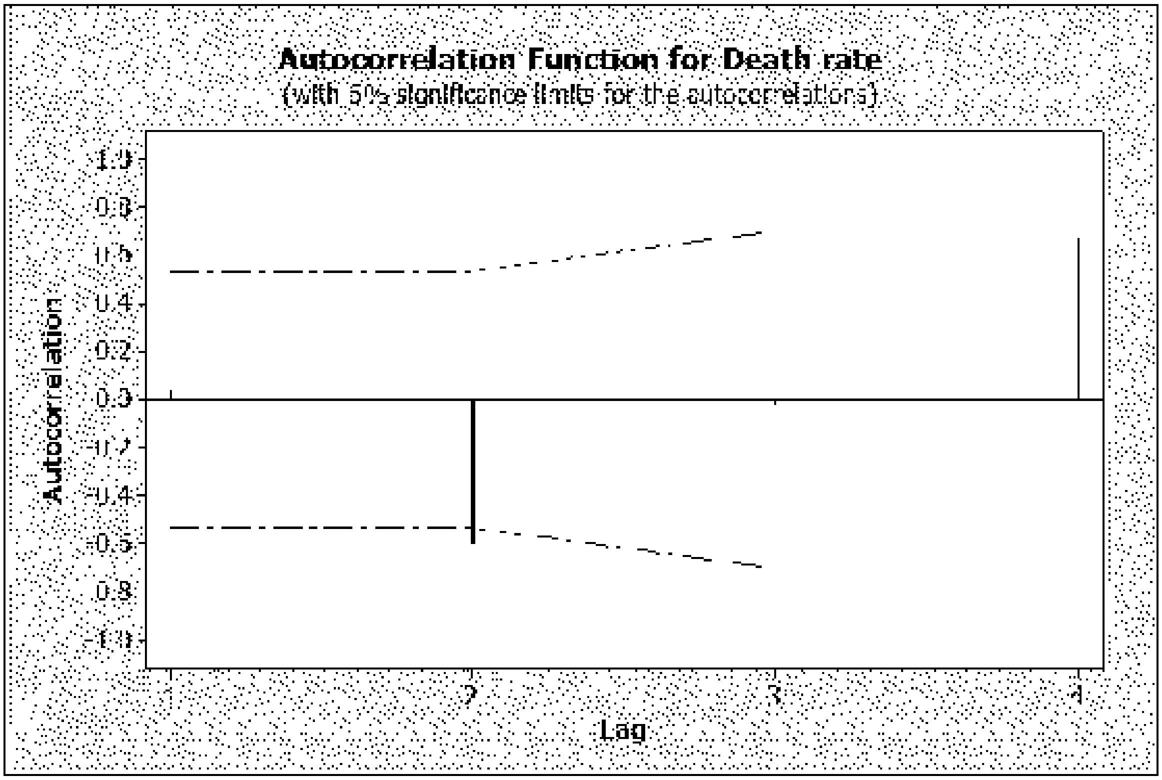
3. *Clickers question:* Without performing the calculation, provide a guess of the sample autocorrelation function (the sample acf, r_k) at lag 2. That is, have a guess at r_2.
   Values two time points apart tend to lie on opposite sides of the mean, so we would expect the sign to be opposite to that of r_4 *but the magnitude to be similar. A guess might be* $r_2 \approx -0.7$. *In fact here* $r_2 = -0.603$.

4. *Clickers question:* Without performing the calculation, provide a guess of the sample autocorrelation function (the sample acf, r_k) at lag 1. That is, have a guess at r_1.
   Values one time point apart tend to be on opposite sides of the mean, and about equally often $x(t+1)$ *exceeds* $x(t)$ *as vice versa. So we might guess* $r_1 \approx 0$. *Using the formula, to find* r_1 *from the data we see*

$$\begin{aligned} r_1 &= \frac{\sum_{t=1}^{15} (x(t) - \bar{x})(x(t+1) - \bar{x})}{\sum_{t=1}^{16} (x(t) - \bar{x})^2} \\ &= \frac{\sum_{t=1}^{15} (x(t) - 22.1)(x(t+1) - 22.1)}{15 \times 1.6^2} \\ &= 0.038. \end{aligned}$$

5. A plot of r_k against the lag k is known as the *correlogram.* Sketch what you think the correlogram would look like here for lags between $k = 0$

and $k = 4$.



6. Re-cap what you have done during this activity. Why do you think you were asked to do this activity? What have you learned?
The idea of autocorrelation has been introduced. The sample autocorrelation measures the correlation between values in a time series k time units apart, for $k = 1, 2, 3, \ldots$. The distance k can be varied, and is known as the lag. The sample autocorrelation can be plotted against the chosen lag, to give a plot known as a correlogram. This is a key tool in exploring time series. In the activity, sample autocorrelations were estimated from the time series plot, then computed using the formula. This gave some intuition to the concept, though in practice software would be used to find autocorrelations.
Learning outcomes encountered in this activity include:

   (a) *Define the sample autocorrelation function and the correlogram.*
   (b) *Describe the behaviour of the correlogram for series that show seasonal fluctuations.*