Coughs and Sneezes Coursework

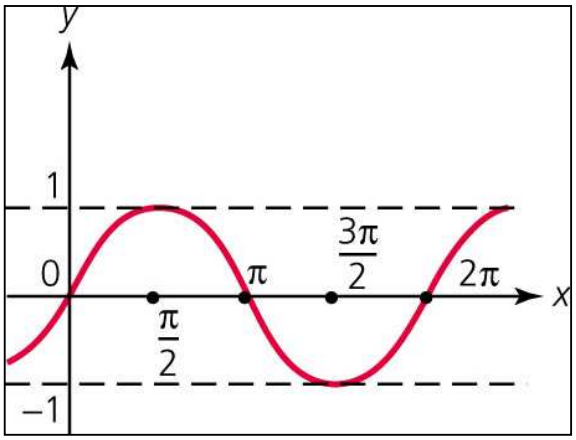
Introduction

At the beginning of a term it is noticed that a large number of University Students who live in a particular hall of residence have a cold. Recording the numbers of students suffering from colds every five days monitors the way in which the cold spreads.
The results are given in the table below where T represents the number of days after monitoring began and S represents the number of students who have a cold.

T	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70
S	25	31	38	43	47	47	45	41	36	30	24	19	14	11	8

In the data set above, there are two 5-day measurements, which have the maximum for the number of students with a cold. Because the number of students with a cold was not measured inbetween 20 and 25 days, I will make the assumption that at 23 days is the maximum with 47 students with a cold because otherwise it makes the modelling difficult because there are two days with the same cases of colds. The maximum value for number of students with a cold will be taken as 48 because sine curves do not have a flat top so there must be a peak, so we assume that the peak of the curve is 48 students.

To find out what the ω value is, you need to do the following sum. The Pi value is not 3.1415 as used for calculating circumferences and areas of a circle in this case but is used for expressing radians. Pi/2 radian refers to 90^{o} on a sine curve. A sine curve looks like this with Pi/2 radians at the bottom of the left-hand curve and Pi radians at the end on the right-hand side. 2Pi Radians are 360^{o}.



In our sine curve, the graph does not go past Pi because with the type of data that we are using, it would not be possible to have a negative number of students having coughs and sneezes. The diagram above is a full sine curve but for our data, we will only be using part of the sine curve.

In this investigation, I will be trying to find:

- Two trigonometrical functions that can be used to model the full data set in the form $S = A \text{ Sin } (\omega T) + C$ and $S = A \text{ Sin } (\omega T+B)+C$
- A polynomial model will be used in the form $S= AT^3+BT^2+CT+D$
- Two equations on the same graph but two different types of graph.
- A 'Trial and Improvement' equation based wholly changing the variables, so the equation is in the form $S = A \text{ Sin } (BT) + C$.

The graph that this data produced is shown at the back. The line of best fit or the regression line is drawn in by hand and goes through all the points. To find the equation of the entire data sets using Microsoft Excel is very tricky because Microsoft Excel does not have a regression line as a Sine wave. It assumes that the data looks like a bell-shaped normal distribution when in fact the graph is a sine curve in a dome shape.

How to make the Models

1. For the First Model, I will be finding an equation in the form of a trigonometrical function $S = A \text{ Sin } (\omega T)+B$. I will be using the maximum value at (T=23, S=48).

- When the '$T = 0$', then '$\omega T = 0$' so 'Sin (ωT) = 0'. Therefore '$S = 0+C$', if you rearrange this equation, you get '25=C'.
- To Find the A value at the beginning of the equation, it is also known as the Amplitude. The Amplitude can be found by taking away the maximum value on the Y-axis (or the S value), which is 48 from the place where the data sets cross the Y-axis, which is at 25. Therefore, the sum is '48-25 = 23'. Therefore, so far, the equation is '$S = 23 \text{Sin } (\omega T)+25$'.

- To find the ω value, you need to: at the maximum along the X-axis, '$\omega T = \text{pi}/2$'. However, we know what T is. $T = 22$. Therefore $\omega*22 = \text{pi}/2\omega$. Substituting that into the equation gives $\omega = \text{pi}/44$, which is equal to 0.07. Therefore, the completed equation is $S = 23 \text{ Sin } (0.07T) + 25$. To see how good this equation, see back of this investigation where all the predicted values that the models predicted using the formulae, are compared to the actual values given from the research. The differences will then be squared to get rid of any minus

differences. These will then be added up and will give a value, the smaller the value, the better the model.

2. For the Second Model, I will be finding an equation in the form of another trigonometrical function S = A Sin (ωT+B)+C. To find this equation, you use the other equation because it gives the Omega value and all the other values are the same. Adding a value into the brackets with the ωT will give a different look to the curve and should improve the accuracy of the model.

- The original equation is as follows but the B value has been added to the brackets as well as the ωT. Y=23Sin(0.07T+B)+25.
- To calculate the B value, I used Data point (40, 13) although any of the data points can be used. Each different data point should give a similar but different B value.
- Using the value for the number of students having a cough and sneeze in the equation. Therefore, the equation becomes 13 = 23Sin(0.07*40)+B)+25.
- Taking away the 25 at the end of the equation, this will hope in trying to get just the B value on one side and a value for it. After taking away the 25 from both sides, the equation becomes 13 -25 = 23Sin(2.8+B).
- Then, you need to take away the 23 from in front of the Sin function. After doing this, the equation becomes -12/23 = Sin (2.8+B).
- To get rid of the Sin function, you do the opposite of Sin, which is Sin^{-1}, which gives the equation to become 2.8 + B = Sin^{-1} -12/23
- Sin^{-1} -12/23 = 2.6
- 2.8+B = 2.6
- B=0.2

Y=23Sin(0.07t+0.2)+25

- For the third equation, I will be trying to find a polynomial equation for the data set. The data will not form a perfectly straight line as is shown in the original data. Unfortunately, a polynomial equation is only going to work if the data is in a perfectly straight line or with a slight curve. A polynomial equation has another disadvantage that no matter what the equation is, the equation will result in results that will assume the data is in a straight line. As the equation is in the form S= AT^3+BT^2+CT+D, there is a certain

chance that the equation goes into enough depth so the loss of the data because of using the polynomial equation as a model.
This equation was not found by myself, but the data was plotted on a graph and Microsoft Excel © was used to calculate the polynomial equation to the third degree. This means there maybe some inaccuracy in the equations produced because Excel occasionally produced odd equations, which are inaccurate, but this equation is a good as it can be.

- For model 4, this model is the two models on the same graph. To do this, you create a graph in Excel for the data set where T = 0 -35 and then you add another data set where T = 40 -70. The first data set will be in the form of a degree 2 polynomial equation and will be used to model the curved bit of the data. The second data set will be a linear equation. Both of these equations was found using Excels 'Add Equation' function.
  To do the table of differences, you type in the first equation for when T = 0-35. Then for T = 40-70 you type the linear equation. This will give a better result as the data set is not a perfect sine curve at all and so having two different models will make the sum of the differences squared closer to 0 than the other models and so make it a better model.

- For model 5, which is the trial and improvement model, the way that this was setup was done using Cell Referencing when you reference the cell but you put a dollar sign so when you copy and paste the formula the cell does not change. The purpose of the dollar sign is to lock the cell reference, so that even when you copy a formula it refers to the same cell. It is therefore called "Absolute Cell Reference". If you wish to lock both the Column and the Row you will need two-dollar signs -one before the column reference (the letter), and one before the row reference (the number). Without the dollar sign it is a "Relative Cell Reference", and you will know that if you copy a formula without a dollar sign down one row it changes the row referred to in the formula.
  This is useful when doing trial and improvement. For reasons not wanting to have a project a hundred pages long, I will only be showing one or two of the equations that I used and what the sum of the differences squared were. This way, I will be able to find what parameters can be changed to get the best set of results. The values started on what they were for the first model being 0.07, 23 and 25. They were changed until a suitable combination had been found where the sum of the differences squared was almost 0.

Results of the Models

1. Model 1, which had an equation of S=23Sin(0.07T)+25 had a sum of the differences to be –42.387279 and a sum of the differences squared to be 370.2480622. This means that the equation had most of the values are below the actual values because of the negative value for the sum of the differences. There was a very high sum of the differences squared, which means that this equation was not very accurate.

2. Model 2, which had an equation of S=23Sin(0.07T+0.2)+25 had a sum of the differences to be 53.201062 and a sum of the differences squared to be 779.3462931. This means that the equation had most of the values are above the actual values because of the positive value for the sum of the differences. There was a much higher sum of the differences squared than Model 1, which means that this equation was even less accurate than Model 1, which was not very accurate.

3. Model 3, which had an equation of $S=0.0005T^3-0.0716T^2+2.4154T+22.551$ had a sum of the differences to be 19 .5625 and a sum of the differences squared to be 82.19396375. This means that the equation had most of the values are above the actual values because of the positive value for the sum of the differences. There was a very small sum of the differences squared compared to Model 1, meaning that so far, Model 3 is the best model.

4. Model 4, which had an equations of $S=-0.0398T^2+1.8988T+23.792$ and $S=-0.9738T+74$ had a sum of the differences to be 0.675 and a sum of the differences squared to be 17.838055. This means that the equation has just about all of the values are above the actual values because of the very small positive value for the sum of the differences. There was a very, very small sum of the differences squared compared to Model 3, meaning that so far, Model 4 is the best model.

5. Model 5, which had an equation of $S=19\sin(0.0644T)+26.5$ had a sum of the differences to be 0.43399 and a sum of the differences squared to be 16.81097. This means that the equation has just about all of the values are above the actual values because of the very small positive value for the sum of the differences. There was a very, very small sum of the differences squared compared to Model 4, meaning that Model 5 is the best model overall.

Conclusion

Model Number	Sum of the differences	Sum of the differences squared
1	–42.387279	370.2480622
2	53.201062	779.3462931
3	19.5625	82.19396375
4	0.675	17.838055
5	0.43399	16.81097

Therefore, looking at the results above, I can tell that where the parameters are not set and can be changed to anything in an attempt to get the sum of the differences squared to equal 0. Realistically, this is not possible because there is not enough variables present and too many will result in confusion. Model 5 is the best model because the sum of the differences squared is the closest number to 0. The sum of the differences does not really tell me anything, only that when that equals 0 either the model is a perfect fit, or more likely that the differences above and below the data cancel each other out.

A trigonometrical model is not the best model type for this data because this type is only really suitable when the data forms a perfect sine wave and in this investigation it doesn't. This can be seen by the high sum of the differences and sum of the differences squared.

A polynomial model is not the best model type for this data because of the same reasons as above. This can be seen by the high sum of the differences and sum of the differences squared.

Two models is quite a good representation of the data and is probably the most accurate way of finding an equation because the 'trial and improvement' method is not very good mathematically because you cant really say how you got those values apart from you had a guess and then changed it.

The 'Trial and Improvement' method is the best but is not mathematically suitable for all the types of data that is available. The Trial and Improvement method will produce better results because the parameters are not just linked to the data.