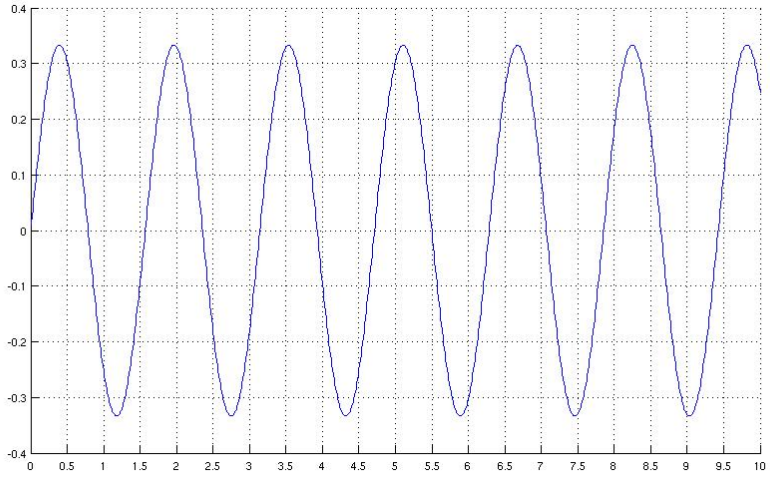
Department of Computer Science University of Bristol

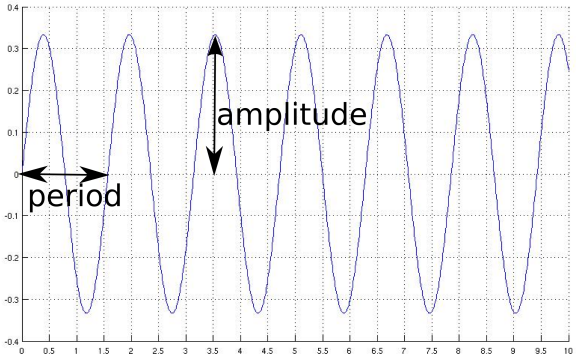
COMS20011: Data-Driven Computer Science

Problem Sheet 1: Data Acquisition

1. On the $sin(x)$ signal below, label the following terms and approximate their values: period, frequency and amplitude



Answer:

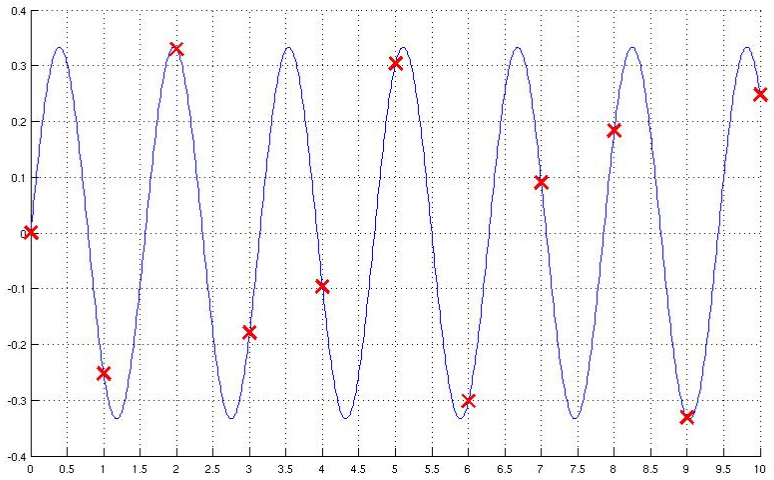


period = 1.57 seconds

frequency = 0.6 [note that we did not label frequency as it is equal to 1/period - more on frequency analysis later in the unit]

amplitude = 0.3

2. For the signal above, convert it into its digital representation using the sampled points. You need to think about the number of bits you would represent each sample as. This is referred to as **Quantisation**. Example, if you need 8 different levels of sound, then 3 bits are sufficient ($2^3 = 8$).



Answer:

Assuming we use 8 levels of sound that correspond to the horizontal grid lines, then each sample is approximated by the binary representation of the closest (rounding) vertical line. The digitisation would thus be:

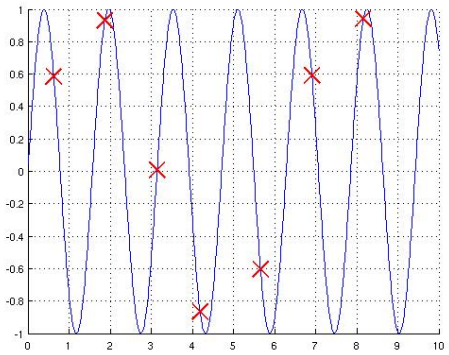
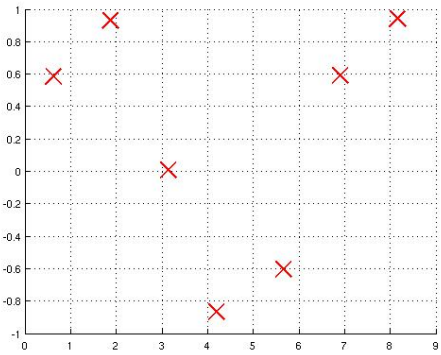
100 001 111 010 011 111 001 100 101 000 110 (11 in total each corresponding to a sampling point as above)

What is the sampling rate in this case??

Answer:

Sampling rate is 1 Hz (1 sample per second)

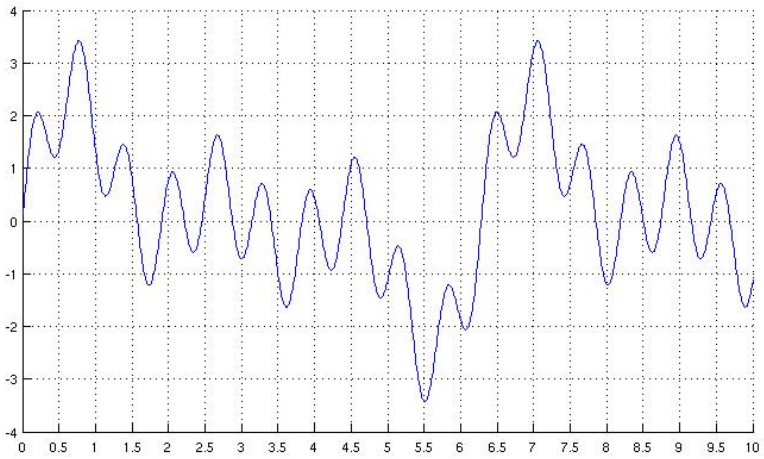
3. Repeat the digitization and reconstruction step for this data below, can you notice any difference?

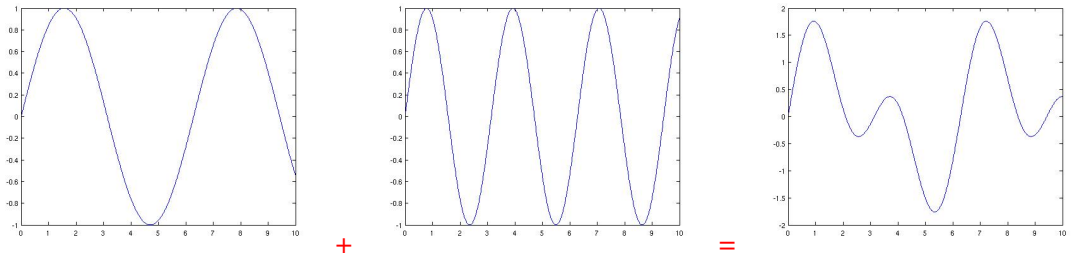
Answer:
Follow the example above again...

4. Based on your understanding of the **Nyquist Sampling Rate** theorem, what is a sufficient sampling rate for the signal below?

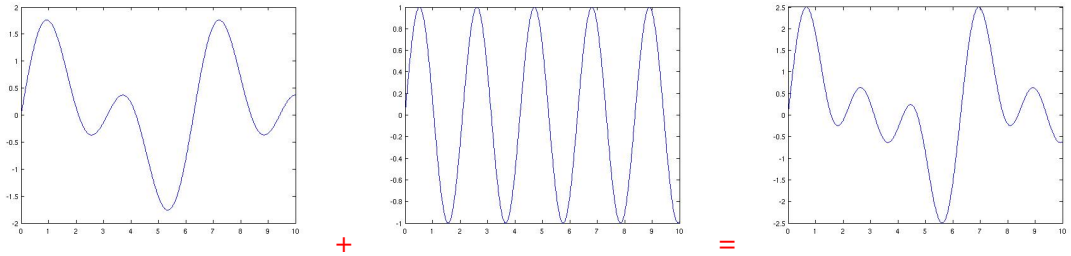
   Note: You might want to look at Fourier Analysis (ahead of our deeper look later in the course) to understand how was this sinusoidal wave constructed.



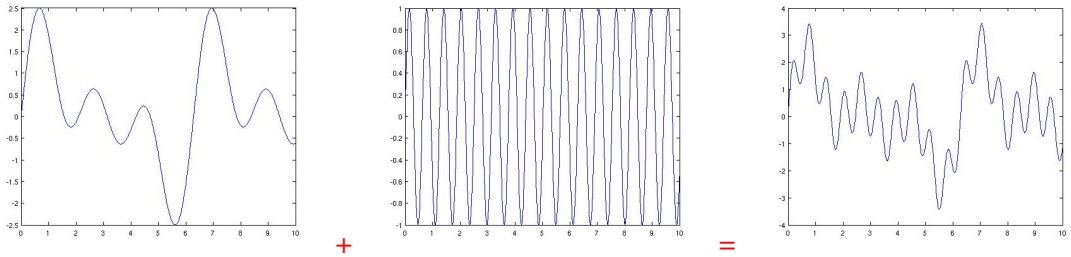
Answer:
One option is to simply get a rough estimate of the highest frequency from the plot above (using the shortest period), which is about 1.6 Hz, from there you can apply the Nyquist theorem, and the sampling rate should be at least twice that, so 3.2 Hz.



Another option is to try to get the different frequency components that were used to build the signal (like in Fourier Analysis). The figures above shows the waves $sin(x)$ and $sin(2x)$ and the sum of the two waves



When adding the summed wave $sin(x) + sin(2x)$ to the higher frequency wave $sin(3x)$ then the wave to the right results



Next we add $sin(x) + sin(2x) + sin(3x)$ to $sin(10x)$ resulting in the frequency required.

The highest frequency in the figure is thus that of the wave $sin(10x)$. The frequency is thus $\frac{10}{2*\pi} = 1.59$ Hz. Following the Nyquist theorem the sampling rate should be at least 3.18 Hz (2×1.59).

5. **Refreshing your memory:**

   For the set of measurements:

   -3, 2, 4, 6, -2, 0, 5

   calculate:

   mean

   median

   variance

   standard deviation

   Answer:

   mean = 1.7

   median = 2

   variance = 12.2

   standard deviation = 3.5

6. **Distance measures:** Calculate the following distance measures for the data provided:

   - D1 = (4,5,6), D2 = (2, -1, 3) - Distance Measure Manhattan Distance L_1
   - D1 = (4,5,6), D2 = (2, -1, 3) - Distance Measure 3-norm L_3
   - D1 = (4, 5, 6), D2 = (2, -1, 3) - Distance Measure Chebyshev Distance L_∞
   - D1 = 'water', D2 = 'further' - Distance Measure Edit Distance
   - D1 = 'weather', D2 = 'further' - Distance Measure Hamming Distance
   - Order, ascendingly, the following words {'tap', 'river', 'liquid', 'ice'} based on their WUP relatedness to: 'water'. Use 1-WUP as the distance measure and the online http://ws4jdemo.appspot.com

   Answer:

   - *11*
   - *6.3*
   - *6*
   - *4*
   - *3*
   - *WUP ('water', ice') = 0.67, WUP('water', tap') = 0.8, WUP('water', 'river') = 0.83, WUP('water', 'liquid') = 0.94*
     D ('water', ice') = 0.33, D('water', tap') = 0.2, D('water', 'river') = 0.17, D('water', 'liquid') = 0.06
     Order = {'ice', 'tap', 'river', 'liquid'}

7. **Distance measures:** Assume you were given a set of whatsapp messages, each with a timestamp (yy-mm-dd hh:mm) and text content (word, word, ...). Propose a distance measure for:

   - calculating whether one message is an exact copy of the other message
   - calculating whether one message was sent before the other message
   - calculating whether one message contains the same set of words as the other message
   - calculating whether one message contains the other message (with potential extras at the start and the end)
   - calculating whether both messages discuss the same topic

   Check your distance measures satisfy: non-negativity, reflexive, symmetric and triangule inequality.

   Answer:

   (a) Calculate the Hamming distance.

   (b) Calculate the difference in the number of minutes relative to a suitable starting time.

   (c) You might wish to propose to use the following measure between message M_1 and M_2

$$D_{NS}(M_1, M_2) = \sum_i \min_j hamming(w_{1i}, w_{2j}) \tag{1}$$

   but this is not symmetric (note that distance measures need to be symmetric). For example, if M_1 = {'a', 'c', 'e'} and M_2 = {'b', 'a'}, $D_{NS}(M_1, M_2) = 2$, but $D_{NS}(M_2, M_1) = 1$. One way to make it symmetric is to

$$D_S(M_1, M_2) = (D_{NS}(M_1, M_2) + D_{NS}(M_2, M_1))/2 \tag{2}$$

   (d) You can use dynamic time warping.

   (e) You can use a similar approach to the one in (c), with a semantic distance measure between words like WUP, where

$$D(M_1, M_2) = (\sum_i \min_j D_{WUP}(w_{1i}, w_{2j}) + \sum_i \min_j D_{WUP}(w_{2i}, w_{1j}))/2 \tag{3}$$