Assignment 1: CS 215

Due: 18th August before 11:55 pm, 100 points

All members of the group should work on all parts of the assignment. Copying across groups or from other sources is not allowed. We will adopt a zero-tolerance policy against any violation.

Submission instructions:

1. You should type out a report containing all the answers to the written problems in Word (with the equation editor) or using Latex, or write it neatly on paper and scan it. In either case, prepare a single pdf file.
2. The report should contain names and roll numbers of all group members on the first page as a header.
3. Put the pdf file and the code for the programming parts all in one zip file. The pdf should contain the names and ID numbers of all students in the group within the header. The pdf file should also contain instructions for running your code. Name the zip file as follows: A1-IdNumberOfFirstStudent-IdNumberOfSecondStudent-IdNumberofThirdStudent.zip. (If you are doing the assignment alone, the name of the zip file is A1-IdNumber.zip, if there are two students it should be A1-IdNumberOfFirstStudent-IdNumberOfSecondStudent.zip).
4. Upload the file on moodle BEFORE 11:55 pm on the due date. We will nevertheless allow and not penalize any submission until 10:00 am on the following day (i.e. 19th August). No assignments will be accepted thereafter, and your group will get no scores for this assignment.
5. Note that only one student per group should upload their work on moodle, though all group members will receive grades.
6. Please preserve a copy of all your work until the end of the semester.

Questions: *In the following problems, you can use the mean, median and standard deviation functions from MATLAB.*

1. Generate a sine wave in MATLAB of the form $y = 6.5\sin(2.1x + \pi/3)$ where x ranges from -3 to 3 in steps of 0.02. Now randomly select a fraction $f = 30\%$ of the values in the array y (using MATLAB function 'randperm') and corrupt them by adding random values from 100 to 120 using the MATLAB function 'rand'. This will generate a corrupted sine wave which we will denote as z. Now your job is to filter z using the following steps.

   - Create a new array y_{median} to store the filtered sine wave.
   - For a value at index i in z, consider a neighborhood $N(i)$ consisting of $z(i)$, 8 values to its right and 8 values to its left. For indices near the left or right end of the array, you may not have 8 neighbors in one of the directions. In such a case, the neighborhood will contain fewer values.
   - Set $y_{median}(i)$ to the median of all the values in $N(i)$. Repeat this for every i.

   This process is called as 'moving median filtering', and will produce a filtered signal in the end. Repeat the entire procedure described here using the arithmetic mean instead of the median. This is called as 'moving average filtering'. Repeat the entire procedure described here using the first quartile (25 percentile) instead of the median. This is called as 'moving quartile filtering'. Plot the original (i.e. clean) sine wave y, the corrupted sine wave z and the filtered sine wave using each of the three methods on the same figure in

different colors. Introduce a legend on the plot (find out how to do this in MATLAB). Include an image of the plot in your report. Now compute and print the relative mean squared error between each result and the original clean sine wave. The relative mean squared error between y and its estimate $\hat{y}$ (i.e. the filtered signal - by any one of the different methods) is defined as $\frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i y_i^2}$.
Now repeat all the steps above using $f = 60\%$, and include the plot of the sine waves in your report, and write down the relative mean square error values.
Which of these methods (median/quartile/arithmetic mean) produced better relative mean squared error? Why? Explain in your report. [5+5+4+3+3=20 points]

2. Suppose that you have computed the mean, median and standard deviation of a set of n numbers stored in array A where n is very large. Now, you decide to add another number to A. Write a MATLAB function to update the previously computed mean, another MATLAB function to update the previously computed median, and yet another MATLAB function to update the previously computed standard deviation. Note that you are not allowed to simply recompute the mean, median or standard deviation by looping through all the data. You may need to derive formulae for this. Include the formulae and their derivation in your report. Note that your MATLAB functions should be of the following form

```
function newMean = UpdateMean (OldMean, NewDataValue, n),
function newMedian = UpdateMedian (oldMedian, NewDataValue, A, n),
function newStd = UpdateStd (OldMean, OldStd, NewMean, NewDataValue, n).
```

Also explain, how would you update the histogram of A, if you received a new value to be added to A? (Only explain, no need to write code.) **Note:** For updating the median, you may assume that the array A is sorted in ascending order, that the numbers are all unique. For sorted arrays with a even number of elements, MATLAB returns the answer as $(A(N/2) + A(N/2+1))/2$. You may use MATLAB's convention though it is not strictly required. Recall that the standard deviation with n values $A_1, ..., A_n$ is given as $s_n = \sqrt{\sum_{i=1}^{n} (A_i - \bar{A}_n)^2/(n-1)}$ and $\bar{A}_n = \sum_{i=1}^{n} A_i/n$. [5+7+5+3 = 20 points]

3. Consider two events A and B such that $P(A) \geq 1 - q_1$ and $P(B) \geq 1 - q_2$. Show that $P(A, B) \geq 1 - (q_1 + q_2)$. [10 points]

4. Here is a simple example of probability in law. In a certain town, there exist 100 buses out of which 1 is red and 99 are blue. A person XYZ observes a serious accident caused by a bus at night and remembers that the bus was red in color. Hence, the police arrest the driver of the red bus. The driver pleads innocence. Now, a benevolent lawyer decides to defend the distressed bus driver in court. The lawyer ropes in an opthalmologist to test XYZ's ability to differentiate between the colors red and blue, under illumination conditions similar to those that existed that fateful night. The opthalmologist suggests that XYZ sees red objects as red 99% of the time and blue objects as red 2% of the time. What will be the main argument of the defense lawyer? (That is, what is the probability that the bus was really a red one, when XYZ observed it to be red?) **Show clearcut steps for your answer.** [15 points]

5. *In this question and the next one, we will understand the reason why exit polls make some statistical sense.* Consider a village with 100 residents, all of whom participate in an election contested by two candidates A and B. An exit poll is conducted after election day in which three residents (chosen uniformly at random with replacement) are asked as to whom they voted for. The candidate with the majority in the exit poll is declared to be the expected winner by the exit poll agency. If 95 percent of the residents favour candidate A over B and the remaining 5 percent favour B over A, what is the accuracy of this exit poll? (In other words, what is the probability that the exit poll that quizzed 3 voters, declared a majority for A?) Assume that the people give truthful answers in the exit poll. Now, suppose the village has 10,000 residents, all of whom are eligible to vote, and the exit poll again asked only 3 (truthful) voters (chosen uniformly at random with replacement). What is the accuracy of the exit poll now? [15 points]

6. Continuing the previous problem, the question to be asked and answered is what happens if the percentages for A and B are less obvious and unknown! Consider that there are m voters in the village, and there is

a probability $p = k/m$ that the voters prefer A over B. Let us suppose that the exit poll asks a randomly (with replacement) chosen subset $\mathcal{S}$ containing n (truthful) voters. Let us define the quantity $x_i = 1$ if the ith voter voted for A and 0 if he/she voted for B. Let $q(\mathcal{S})$ be the proportion of voters from $\mathcal{S}$ who voted for A out of $n = |\mathcal{S}|$. That is $q(S) = \sum_{i \in \mathcal{S}} x_i/n$. Given this, do as directed:

(a) Prove that $\sum_{\mathcal{S}} \dfrac{q(\mathcal{S})}{m^n} = p$. This is the average of the values of $q(\mathcal{S})$ across all subsets $\mathcal{S}$ of size n.

(b) Prove that $\sum_{\mathcal{S}} \dfrac{q^2(\mathcal{S})}{m^n} = \dfrac{p}{n} + \dfrac{p^2(n-1)}{n}$.

(c) Prove that $\sum_{\mathcal{S}} \dfrac{(q(\mathcal{S}) - p)^2}{m^n} = \dfrac{p(1-p)}{n}$.

(d) Hence argue that the proportion of n-sized subsets $\mathcal{S}$ (out of m^n) for which $|q(S) - p| > \delta$, is less than or equal to $\dfrac{1}{\delta^2}\dfrac{p(1-p)}{n}$. Note that this proportion is quite small. This is also a nice application of one of the inequalities we studied in class (which one?). What is the significance of this result? [5+7+2+(3+3) = 20 points]