STAT 218 - 08S2 (C) Semester Two 2008
11 points, 0.0917 EFTS 14/07/2008-12/11/2008

Computational Methods in Statistics
Assignment Cover Sheet

Student ID # :
Surname or Family Name :
First Name or Names :

Course Coordinator: Raazesh Sainudiin

STATEMENT REGARDING DISHONEST PRACTICE

(relating to work submitted for assessment)

The University has a clear interpretation of what constitutes dishonest practice as described in you Calendar. Dishonest practice includes the following:

1. **Plagiarism**, being the presentation of any material (text, data or figures, on any medium including computer files) from any other source without clear and proper acknowledgement of the source of that material. (Guidelines for appropriate acknowledgement will be provided with assignment handouts).
2. **Collusion**, being work performed in whole or in part in conjunction with another person or persons, but submitted as if it had been completed by the named author alone (or joint authors if a group item of work).
3. **Copying**, being the use of material (in any medium, including computer files) produced by another person or persons, with or without their knowledge and approval.
4. **Ghost writing**, being the use of another party (with or without any form of payment) to prepare all or part of an item of work submitted for assessment.

Under the University regulations, evidence of any of these or other forms of dishonest practice by any student(s) represents grounds for disciplinary action and may result in penalties ranging from denial of credit for the item of work in question to exclusion from the University.

- This interpretation of the dishonest practice of collusion is not intended to discourage students from having discussions with each other about how to approach a particular assigned task, and incorporating general ideas coming out of such discussions into their own individual submissions.

DECLARATION:

In signing below, I confirm that I have read and fully understand the statement regarding dishonest practice, as detailed in the University Calendar and briefly outlined above, and hereby certify that this assignment submitted for assessment is entirely my own work.

Signed : .....

Date : .....

ENQUIRIES

Raazesh Sainudiin (r.sainudiin@math.canterbury.ac.nz, phone x7691)
Room 724 Erskine Building
See Course Syllabus for other details.

STAT 218 Assignment **Due: Sep. 25, 2008**

The comments enclosed in square brackets ('[...]') at the end of each question is the problem handing-in protocol. Problems 1, 2 and 3 make $4 + 4 + 4 = 12\%$ of your final grade.

1. Recall the *Laplace*(λ) RV, with the rate parameter $\lambda > 0, \lambda \in \mathbb{R}$ with density function

$$f(x;\lambda) = \frac{\lambda}{2} e^{-\lambda|x|} \ ,$$

DF

$$F(x;\lambda) = \int_{-\infty}^{x} f(y;\lambda)dy = \frac{1}{2}\left(1 + sign(x)\left(1 - e^{-\lambda|x|}\right)\right), \quad where, \quad sign(x) = \begin{cases} 1 & if\ x > 0 \\ 0 & if\ x = 0 \\ -1 & if\ x < 0 \end{cases},$$

and inverse DF

$$F^{-1}(u;\lambda) = -\frac{1}{\lambda}\ sign\left(u - \frac{1}{2}\right)\log\left(1 - 2\left|u - \frac{1}{2}\right|\right) \ , \ u \in [0,1] \ .$$

You should be intimate with this function from the Labworks. Let us use these RVs to make a toy model for pollen dispersal from a source at the origin along a line. In other words, we want a model that would simulate the 'pollen landing points along a line or transect of interest'. Note that the pollen may land on either side of the source that is located at the origin ($x = 0$) in our line of interest. When there is only one type of pollen being produced (by a plant) at the source, we can model the dispersal of the first pollen using the *Laplace*($\lambda = 1.0$) RV X_1. Let the RV X_i model the dispersal of the i-th pollen from the source. To specify the RV X_i, let us make further simplifying assumptions in our model of dispersal of n pollens. Under the **two assumptions** that (1) the dispersal of each pollen is **independent** of the dispersal of any other pollen and that (2) the dispersal of every pollen is **identical** in distribution to that of any other pollen, our model for the dispersal of n pollens from the source is given by the RVs:

$$X_1, X_2, \ldots, X_n \sim F(x; \lambda = 1.0)$$

   (a) Write a `Matlab` **script** (use the `diary` command and edit it to make the final script with generous comments) that would simulate the dispersal of 10 consecutive pollens. The output of this simulation should be 10 real numbers that are stored in a vector named '`x10`'. Using **words** interpret the meaning of these numbers in terms of our model for pollen dispersal. You are required to initialize the fundamental sampler prior to this simulation using `rand('twister',MyStudentID)`, where `MyStudentID` is your integer-valued student ID. [Submit the **script** file with **words** below it – if you use the functions in M-files you created during a previous lab or assignment then this M-file has to be included in your report for full credit].

   (b) Reset the fundamental sampler using '`rand('twister', MyStudentID)`' and simulate the dispersal of 100 pollens by modifying the above script and store the samples in a vector named '`x100`'. Report the five statistics; mean, median, sample standard deviation, minimum and maximum, for the 100 samples and their absolute values:

      i. 100 samples stored in $x100$,

ii. $|x100|$, ie. the absolute values of the 100 samples in $x100$ (‘`abs(x100)`’).

using the built-in `Matlab` commands `mean`, `median`, `std`, `max`, and `min`, respectively. Interpret in **words**, these five statistics for $x100$ and $|x100|$, in terms of our pollen dispersal model. [Submit the final `Matlab` script (use the `diary` command and edit it to make the final script with generous comments) used to carry out the above task of statistical summarization with **words** below it. The two sets of the five summaries should unambiguously appear in the submitted final script. You don't need to submit the modified simulation script from 1(a)].

2. So far we have assumed that there is only one plant, say PLANT located at the origin ($x = 0$). Suppose we have two new plants FAT-PLANT and SKINNY-PLANT that are also located at the origin. Suppose FAT-PLANT produces fatter or heavier pollen and SKINNY-PLANT produces skinnier or lighter pollen when compared with the pollens produced by PLANT. Since a lighter pollen, that is dispersed by wind, would travel farther than a heavier pollen, we may model the dispersal of a pollen from FAT-PLANT according to $Laplace(\lambda = 10)$ RV and that of a pollen from SKINNY-PLANT according to $Laplace(\lambda = 0.1)$ RV. We still assume that each of the three plants independently and identically disperse their pollens. Finally, we assume that there is no differences in the rate of pollen production between the three plants. Thus, the i-th pollen (of any kind) is equally likely to have come from any of the three plants. Such a model of dispersal for a pollen, that may be any one of the three kinds, is called a finite equi-weighted mixture model. Therefore, our model for the dispersal of n pollens (of any kind) from the source is given by the independent and identically distributed RVs:

$$Y_1, Y_2, \ldots, Y_n \sim F(y; \lambda_1 = 1.0, \lambda_2 = 10, \lambda_3 = 0.1)$$

with DF

$$F(y; \lambda_1 = 1.0, \lambda_2 = 10, \lambda_3 = 0.1) = \frac{1}{3}F(x; \lambda_1 = 1.0) + \frac{1}{3}F(x; \lambda_2 = 10) + \frac{1}{3}F(x; \lambda_3 = 0.1)$$

(a) Write a `Matlab` **script** (use the `diary` command and edit it to make the final script with generous comments) that would simulate the dispersal of 10 consecutive pollens of any kind according to the equi-weighted finite mixture model. In other words, draw 10 samples from the RV Y above. The output of this simulation should be 10 real numbers that are stored in a vector named ‘`y10`’. Using **words** interpret the meaning of these numbers in terms of our model for pollen dispersal. You are required to initialize the fundamental sampler prior to this simulation using ‘`rand('twister', MyStudentID)`’ . [Submit the **script** file with **words** below it – if you use the functions in M-files you created during a previous lab or assignment then this M-file has to be included in your report for full credit].

(b) Reset the fundamental sampler using ‘`rand('twister', MyStudentID)`’ and simulate the dispersal of 100 pollens of any kind by modifying the script of 2(a), ie. draw 100 samples from the RV Y above, and store the samples in a vector named ‘`y100`’. Report the five statistics; mean, median, sample standard deviation, minimum and maximum, for the 100 samples and their absolute values:

i. 100 samples stored in $y100$,

ii. $|y100|$.

using the built-in `Matlab` commands `mean`, `median`, `std`, `max`, and `min`, respectively. Interpret in **words**, these five statistics for $y100$ and $|y100|$, in terms of our dispersal model for three kinds of pollen from the same source. Compare and contrast these statistics with their counterparts in (b) above and try to interpret them in **words**. Feel free to increase the sample sizes when computing the five statistics to sharpen your intuitions. [Submit the final `Matlab` script (use the `diary` command and edit it to make the final script with generous comments) used to carry out the above task of statistical summarization with **words** below it. The two sets of the five summaries should unambiguously appear in the submitted final script. You don't need to submit the modified simulation script from 2(a)].

(c) Plot the empirical CDF from the 100 samples of the data array `x100` from part 1. as well as the empirical CDF from the 100 samples in data array `y100` (superimpose the two plots by using the `hold` command and the `ECDF` function from Labworks). By further superimposing the plots of the two DFs $F(x; \lambda = 1.0)$ and $F(y; \lambda_1 = 1.0, \lambda_2 = 10, \lambda_3 = 0.1)$ underlying the simulations, explain how the two empirical CDFs compare with the DFs? [Submit the Figure with the four plots and the discussion below it. The plots should be clearly marked.]

3. You have to generate your own "wavy function" for this problem:

$$\widetilde{f}_c(x) = \exp(-2 + \cos(c_1 x + c_2 x^2) + \sin(c_3 x + c_4 x^2)) \ .$$

First, the following code generates the array $c = (c_1, c_2, c_3, c_4)$ that is specific to your Student ID. If your student ID is 35598968, then the following steps generate your coefficient vector c :

```
>> MyID = 35598968; % write your student ID in the variable MyID
rand('twister',MyID); % Set the seed for rand with MyID
c=rand(1,4)*2.0+2.0 % store the c's in the array called c
c =    2.9276    3.5945    3.9367    3.0110
```

From hereon we assume that such an array `c` that is uniquely tied to your student ID has been declared and available in memory. Next write an M-file called `MyIDPloy4.m` to encode the c-specific $\widetilde{f}_c(x)$ as follows:

MyIDWavy4.m

```
function f = MyIDWavy4(x,c);
% Return: the exponential of a translated sum of cosine and sine of two degree 2
%         polynomials with coefficients specified by the array c
% Input : c is a vector of 4 numbers
% Input : x is the real number(s) at which the function is evaluated
% Output: f(x)
f = exp( -2 + cos( c(1) * x + c(2) * x .^2) + sin( c(3) * x + c(4) * x .^2 ) );
```

Next, get a sense for your wavy function by plotting it over the domain of interest $\mathbb{X} = [-10, 10]$:

```
>> x=-10:0.01:10;
>> plot(x,MyIDWavy4(x,c)) % assuming c has been generated with your ID
```

(a) Implement a von Neumann Rejection Sampler that can produce IID samples from the target shape $\widetilde{f}_c$ over $\mathbb{X} = [-10, 10]$ using proposals from the Uniform$(-10, 10)$ RV. [Submit the M-file for this sampler. You need to find and justify the constant $\widetilde{a}$ such that $\widetilde{a}\frac{1}{10-(-10)} \geq \widetilde{f}_c(x)$ for every $x \in [-10, 10]$ among others.]

(b) Plot the samples from your target density $f_c := \frac{\widetilde{f}_c(x)}{\int_{-10}^{10} \widetilde{f}_c(t)dt}$ on the x-axis to convince yourself that you are indeed simulating from the desired target and summarise the first and second sample moments. [Report the sample mean and sample variance from 1000 samples from f_c over $[-10, 10]$].

(c) What is the average number of proposals before one of your samples gets accepted (you may report an empirical measure here or find the area between $\widetilde{f}_c(x)$ and $\widetilde{a}/20$ using numerical quadrature such as MATLAB's `quad`)? Can you find the optimal constant $\widetilde{a}$ to maximise the efficiency of your rejection sampler based on Uniform$(-10, 10)$ proposals? How else can you go about improving the efficiency of the sampler?