Northeastern University
College of Science

Midterm Problems for MATH7340

For this midterm you must answer all questions (7 in total). Provide the numerical answers and plots as requested, and show your derivations. If R is used to arrive at the answer, state so. And please provide an executable R script file with all the R commands used, and clearly label which question the R commands are for.

Problem 1 (10 points)

X follows a distribution with pdf $f_X(x) = 2.469862(xe^{-x^2}), x = 1, 2, 3$; while Y follows a distribution with pdf $f_Y(y) = 2ye^{-y^2}, y > 0$.

a) Find E(X), E(Y), sd(X) and sd(Y).

b) If X and Y are independent, find E(2X-3Y) and sd(2X-3Y).

Problem 2 (10 points)

X follows a standard normal distribution N(mean=0, sd=1), and Y follows a Chi-square distribution with degrees of freedom df=4. Assume that X and Y are independent. Please estimate $E(\frac{X^2}{X^2+Y})$ accurate to two decimal places.

Problem 3 (10 points)

Suppose we decide to use the Monte Carlo method to check coverage of a 95% confidence interval (CI) formula. We generated nsim=1000 data sets from the known distribution, calculate the 95% confidence interval on each data set and check the empirical coverage (that is, the proportion of those 1000 confidence intervals that contains the true parameter). Suppose that the CI formula is wrong, and the true coverage is only 92%. What is the probability that our empirical coverage is greater than 94%?

Problem 4 (10 points)

A random sample from the normal distribution $N(mean = \theta, sd = \theta)$ is provided in the file "normalData.txt". Find the value of MLE $\hat{\theta}$ on this data set.

Instructions on inputting the data set:

You should download the file, put it in the working directory of your R session. Then load it using command

```
y<-as.numeric(t(read.table(file = "normalData.txt", header=T)))
```

Problem 5 (10 points)
On the Golub et al. (1999) data set, complete the following:

a) Use the t-test to test how many genes have mean expression values greater than 0.6. Use a FDR of 10%.

b) Find the gene names of the top five genes with mean expression values greater than 0.6.

Problem 6 (35 points)
On the Golub et al. (1999) data set, compare the "GRO3 GRO3 oncogene" (at row 2715) with the "MYC V-myc avian myelocytomatosis viral oncogene homolog" (at row 2302). I will refer to those two genes as GRO3 gene and MYC gene for short in the following:

a) Draw a histogram of the GRO3 gene expression values.

b) Draw a scatterplot of the GRO3 gene expression values versus MYC gene expression values, labeled with different colors for ALL and AML patients.

c) Use a parametric t-test to check (the alternative hypothesis) if the mean expression value of GRO3 gene is less than the mean expression value of MYC gene.

d) Use a formal diagnostic test to check the parametric assumptions of the t-test. Is the usage of the t-test appropriate here?

e) Use a nonparametric test to check (the alternative hypothesis) if the *median* difference between the expression values of GRO3 gene and the expression values of MYC gene is less than zero.

f) Calculate a nonparametric 95% one-sided upper confidence interval for the *median* difference between the expression values of GRO3 gene and of MYC gene.

g) Calculate a nonparametric bootstrap 95% one-sided upper confidence interval for the *mean* difference between the expression values of GRO3 gene and of MYC gene.

Problem 7 (15 points)

On the Golub et al. (1999) data set, complete the following:

a) Find the row number of the "HPCA Hippocalcin" gene.

b) Find the proportion among ALL patients that the "HPCA Hippocalcin" gene is negatively expressed (expression value<0).

c) We want to show that "HPCA Hippocalcin" gene is negatively expressed in at least half of the *population* of the ALL patients. State the null hypothesis and the alternative hypothesis. Carry out the appropriate test.

d) Find a 95% confidence interval for the difference of proportions in the ALL group versus in the AML group of patients with negatively expressed "HPCA Hippocalcin" gene.