Northeastern University
College of Science

**Module 6 Homework**

**1. (50 points)** On the Golub et al. (1999) data, consider the "H4/j gene" gene (row 2972) and the "APS Prostate specific antigen" gene (row 2989). Setup the appropriate hypothesis for proving the following claims. Chose and carry out the appropriate tests.

**(a)** The mean "H4/j gene" gene expression value in the ALL group is greater than -0.9.

**(b)** The mean "H4/j gene" gene expression value in ALL group differs from the mean "H4/j gene" gene expression value in the AML group.

**(c)** In the ALL group, the mean expression value for the "H4/j gene" gene is lower than the mean expression value for the "APS Prostate specific antigen" gene.

**(d)** Let $p_{low}$ denote the proportion of patients for whom the "H4/j gene" expression is lower than the "APS Prostate specific antigen" expression. We wish to show that $p_{low}$ in the ALL group is greater than half. Does this test conclusion agree with the conclusion in part (c)?

**(e)** Let $p_{H4j}$ denotes the proportion of patients for whom the "H4/j gene" expression values is greater than -0.6. We wish to show that $p_{H4j}$ in the ALL group is less than 0.5.

**(f)** The proportion $p_{H4j}$ in the ALL group differs from the proportion $p_{H4j}$ in the AML group.

You should state the hypothesis, show the R commands for the tests, show the output of these tests, and state your conclusion based on these outputs.

**2. (10 points)** Suppose that the probability to reject a biological hypothesis by the results of a certain experiment is 0.05. This experiment is repeated 2000 times.
**(a)** How many rejections do you expect?
**(b)** What is the probability of less than 90 rejections?

**3. (10 points)**

For testing $H_0$: $\mu=3$ versus $H_A$: $\mu>3$, we considers a new $\alpha=0.1$ level test which rejects when $t_{obs}=\frac{\bar{X}-3}{s/\sqrt{n}}$ falls between $t_{0.3,n-1}$ and $t_{0.4,n-1}$.

**(a)** Use a Monte Carlo simulation to estimate the Type I error rate of this test when n=20. Do 10,000 simulation runs of data sets from the $N(\mu=3,\sigma=4)$. Please show the R script for the simulation, and the R outputs for running the script. Provide your numerical estimate for the Type I error rate. Is this test valid (that is, is its Type I error rate same as the nominal $\alpha=0.1$ level)?

**(b)** Should we use this new test in practice? Why or why not?

**4. (20 points)**

On the Golub et al. (1999) data set, do Welch two-sample t-tests to compare every gene's expression values in ALL group versus in AML group.

**(a)** Use Bonferroni and FDR adjustments both at 0.05 level. How many genes are differentially expressed according to these two criteria?

**(b)** Find the gene names for the top three strongest differentially expressed genes (i.e., minimum p-values). Hint: the gene names are stored in ***golub.gnames***.

Please submit your R commands together with your answers to each part of the question.

**5. (10 points)** Read the paper "Interval estimation for a binomial proportion" by Lawrence D Brown, T Tony Cai, Anirban DasGupta (2001) Statistical Science pages 101-117. Available at link
http://projecteuclid.org/download/pdf_1/euclid.ss/1009213286

**(a)** Program R functions to calculate the Wald CI, the Wilson CI and the Agresti–Coull CI for binomial proportion. (Formulas are in equations (1), (4) and (5) of the paper.)
**(b)** Run a Monte Carlo simulation to check the coverage of the Wald CI, the Wilson CI and the Agresti–Coull CI for n=40 and p=0.2 at the nominal confidence level of 95%. Do 10,000 simulation runs for calculating the empirical coverages.

Please submit your R functions in part (a). Submit your R script for the simulation in part (b). Also answer part (b) with your numerical estimates of the three coverage probabilities.