Stat 516 Homework 5 Due Thursday, April 7

1. Prove that any method that strongly controls FDR also weakly controls FWER.

2. The purpose of this problem is to get you thinking about a simple hidden Markov model. Suppose a person has two coins. Coin A has probability 0.5 of landing heads and probability 0.5 of landing tails. Coin B has probability 1 of landing heads and probability 0 of landing tails. One of the coins will be chosen randomly to begin a series of flips. With probability 0.7 coin A will be chosen to start the series. Coin B will be selected to start the series with probability 0.3. Once the series of flips has begun, the coin used for subsequent flips will be determined as follows. If the previous flip was made with coin A, the current flip will be made with coin A with probability 0.9 and with coin B with probability 0.1. If the previous flip was made with coin B, then the current flip will be made with coin B with probability 0.4 and with coin A with probability 0.6.

   (a) Suppose you are told that the first flip in the sequence resulted in heads. Given this information, what is the probability that the second flip will also be heads?

   (b) If the first two flips end up being heads, which coin(s) do you believe were used to generate the sequence? Back up your answer with appropriate probability calculations.

3. Let $f(p;\theta,\delta)$ denote the probability density function of a p-value from a two-sample t-test when the test statistic has a noncentral t distribution with θ degrees of freedom and noncentrality parameter δ. Write an R function that will return $f(p;\theta,\delta)$ for any given $p \in (0,1)$, $\theta > 0$, and $\delta \in \mathbb{R}$. Report $f(p;\theta,\delta)$ for $p = 0.1$, $\theta = 18$, and $\delta = 1.6$.

4. Suppose a microarray has only 10 genes. Suppose the test statistics for those 10 genes are provided in the following table.

Gene	1	2	3	4	5	6	7	8	9	10
t statistic	0.26	-2.12	-0.14	-0.25	0.36	2.23	0.6	-4.42	-0.17	-2.14

   Suppose a gene set S consists of genes 2, 6, 8, and 10.

   (a) Compute $P_{\text{hit}}(S,i)$, $P_{\text{miss}}(S,i)$, and $P_{\text{hit}}(S,i) - P_{\text{miss}}(S,i)$ for each $i = 1,\ldots,10$.

   (b) Find $M(S)$ and $m(S)$.

   (c) Find $ES(S)$.

   (d) Write an R function that will return the enrichment score for any gene set when given the vector of all test statistics corresponding to genes on a microarray and a vector if IDs corresponding to genes in the gene set. For example, if the vector of IDs is $(5, 37, 91, 143)$, then elements 5,37,91, and 143 of the test statistic vector are the test statistics corresponding to the four genes in this example gene set.

   (e) Suppose the observed test statistics in a microarray experiment involving 1000 genes are an independent and identically distributed sample from a standard normal distribution. Conduct a simulation to approximate the distribution of the enrichment score of a random subset of 100 genes. Use a simulation sample size of 10,000 and provide a histogram that illustrates your approximation of the distribution of the enrichment score. In each run of the simulation, generate a new set of independent and identically distributed standard normal test statistics.

   (f) Repeat the simulation in part (e) except suppose that the test statistics for genes in the gene set are always the middle 100 test statistics of the 1000 genes. That is, if you put the statistics in order from smallest to largest, the statistics for genes in the set are ranked 451st through 550th of the 1000 statistics. This will imply that the statistics in the set are concentrated near 0.

(g) Compare histograms from part (e) and (f). Does it seem like the enrichment score for a set whose statistics are in the middle of the test statistic distribution will appear "enriched" relative to a random set?