Hypothesis testing:

1. Assume we are studying cancer vs. healthy individuals using microarrays. We have 20 cancer patients and 10 healthy individuals in our sample set. We are interested in performing randomization tests to determine if the differences in mean values we observe for each gene are significant. Write a formula for the number of randomization tests that can be performed for this data (no need to provide the actual number, only a formula that when evaluated would give this number).
   Answer: $\binom{30}{20}$
2. For another cancer vs. healthy experiment assume that we have performed 500 permutation tests for each gene in our study. We use these to compute p-values for every gene we test. Out of the 1000 genes we tested, we identified 100 as significant. What is the *minimum* FDR for this set?
   Answer: The best we can hope from the randomization tests is a p-value of 1/500 = 0.002. Since we tested 1000 genes, at this p-value we expect to have 2 genes score higher by chance and so the FDR is 2/100 = 2%.
3. We tested n genes and identified 40 with a p-value of 0.005. We are told that the FDR for this set is 20%. What is n?
   a. 8000 b. 4400 c. 1600 d. 800 e. impossible to tell

   Answer: d. Given this FDR we expect 8 genes to pass the p-value by chance. Since the p-value is 0.005 we have n*0.005 = 8 -> n = 1600

4. We tested n genes using a bonforonnie corrected p-value of 0.001 and identified 50 as differentially expressed. What is *uncorrected* p-value we started with?
   a. 0.05 b. 0.01 c. 0.005 d. 0.001 e. impossible to tell

Answer: e. The corrected p-value is a function of the initial p-value and n (number of genes). Since we do not know both of these it is impossible to tell what is p.

Normalization

For each of the following experiments select the answer(s) that would provide a good normalization for the experiment described. You can select multiple answers for each question but you will be panelized for every wrong answer you selected.

1. Over crowding: We are performing experiments in which we compare wild type e. coli. To e. coli's that are overcrowded (that is, in which we are trying to express as many genes as possible to their highest levels).
   a. Quintile b. Same mean / variance c. dChip d. non of the answers is appropriate for this experiment.

   Answer d. In over crowding we are expressing all genes to higher levels than they are in the wild type cells and so none of the methods applies.

2. We are comparing experiments from two different types of arrays where one set has twice as many genes printed on it as the other (randomly selected).
   a. Quintile b. Same mean / variance c. dChip d. non of the answers is appropriate for this experiment.

   Answer: b. and c. Both are appropriate to a random selection of genes. A. is not good since we have a different number of genes so it would not work.

3. We are in a hurry and would like to use the fastest method described in class
   a. Quintile b. Same mean / variance c. dChip d. non of the answers is appropriate for this experiment.

   Answer b. Both a. and c. require sorting (n log n) while b is linear in the number of genes.

4. We are comparing samples from different tissues of the same individual.
   a. Quintile b. Same mean / variance c. dChip d. non of the answers is appropriate for this experiment.

   Answer: All three are appropriate. This is the classical setting.

8.b. Suppose we are using Gaussian Naive Bayes to classify genes into one of two classes: cancer and healthy, based on their expression values in a stress response experiment. Assume we trained a Gaussian Naive Bayes classifier using two different datasets of known cancer and healthy genes. Dataset one contains C1 cancer genes and H1 healthy genes and dataset two contains C2 cancer genes and H2 healthy genes. Surprisingly, we noticed that the classification models obtained for from these two sets (M1 and M2 respectfully) had exactly the same parameters for mean and variance in the two classes. In other words both had the same mean for healthy genes and the same variances for healthy genes and also the same mean and variances for cancer genes (of course, the means and variances differed between healthy and cancer genes in both models).
We next obtained expression values for a new gene G that was not included in either of the training datasets. When classifying G using M1 we determined that it was a cancer gene. However, when we used M2 we determined that it is a healthy gene. Which of the following answers is correct (you may circle more than one, but would be penalized for any wrong circles).

a. C1 > C2 b. H2 > H1 c. C1+H1 > C2+H2 d. C1/C2 > H1/H2 e. None of these has to be correct

Answer: d. For this result to happen we need the prior for cancer to be higher in M1 than it is in M2. This means that $\frac{C1}{C1+H1} > \frac{C2}{C2+H2}$. All values in this inequality are positive so this is the same as:

$$\frac{C2+H2}{C2} > \frac{C1+H1}{C1} \Rightarrow \frac{H2}{C2} + 1 > \frac{H1}{C1} + 1 \Rightarrow \frac{C1}{C2} > \frac{H1}{H2}$$