Normalization methods

1. In the following question use these letters to denote the three normalization methods we discussed in class by:

A – Mean and variance normalization
B – Quintile (ranking) normalization
C – Invariant genes normalization

Below we present three alternative normalization methods. For each of these additional methods determine which of the methods learned in class make stronger or weaker *assumptions*. If there the assumptions that cannot be directly compared chose 'cannot be compared'. Note that some categories may be empty for some of the new methods.

1. Normalizing by assigning genes in each array to the same predefined Gaussian distribution.

   Methods that make stronger assumptions than this:

   None.

   Methods that make weaker assumptions than this:

   A, B

   Methods that cannot be compared:

   C

When normalizing using a predefined distribution for all arrays we arrive at arrays with the same values (as in B) and the same mean and variance (as in A). In addition, unlike B we do not use any values from the measurements for the global values (just the ranking) and rely on the predefined distribution for the values we end up with. The assumptions made by C are not directly comparable (though we accepted answers that said that C had weaker assumptions as well).

2. Normalizing by only using the spike in controls (which are added to each of the samples and contains genes from a different species).

   Methods that make stronger assumptions than this:

   A, B, C
   Methods that make weaker assumptions than this:

   None

Methods that cannot be compared:

None

Spike controls require very little assumptions regarding the actual similarities between the global or partial values and ranking of the real experimental samples. Unlike A, B and C when using spike controls we rely on the procedure used to carry out the experiments which is controlled by the experimentalist.

3. Normalizing by equating the mean, variance and third moment of the distribution.

Methods that make stronger assumptions than this:

B

Methods that make weaker assumptions than this:

A

Methods that cannot be compared:

C

For B, since all values are the same all moments are equal as well so B is stronger. A is weaker since it only equates the mean and variance. Again, C cannot be directly compared since C looks at specific gene rankings whereas this assumptions looks at global mRNA quantities.

Multiple hypothesis testing.

2. We carried out microarray experiments to compare cancer and healthy cells. For each of the 2000 genes we tested we computed the log likelihood ratio score and used the chi-square distribution to assign p-values for these scores. The table below contains the scores and their corresponding p-values.
Since we tested 2000 genes, we can use the Bonferroni correction to correct for multiple hypothesis testing. Another way is to use randomization tests. We have carried out 1000 such tests (by randomizing the labels of the experiments). In the table below we also present the number of randomization experiments containing at least one gene with a score above each of the three scores:

score	10	20	30
p-value	0.05	0.01	0.001
# of times we found a gene with a lower p-value	300	100	40

a. What is the uncorrected p-value required for a corrected p-value of 0.1 according to the Bonferroni correction?

   0.1 / 2000

b. What is the uncorrected p-value required for a corrected p-value of 0.1 according to the randomization correction method?

   0.01 . As the table above indicates, there were 100 out of 1000 randomization runs that led to a value of 20 or higher for at least one gene. Thus, for a p-value of 0.1 we need the uncorrected p-value that corresponds to a score of 20 which is 0.01.

c. Is there a method (Bonferroni or randomization) that is stricter (leads to lower corrected p-values) for *all* three randomization corrected p-values that can be derived from the table above?

   Yes. In all cases the randomization correction leads to higher (less strict) p-values.

d. Assume we have identified 200 genes with a p-value < 0.01. What is the false discovery rate (FDR)?

   Since we started with 2000 genes we would expect to see 20 genes with this p-value. Thus, the FDR is 20 / 200 = 10%.

e. What is the FDR if we identify 10 genes with the *Bonferroni corrected* p-value for the original p-value of 0.1 (the p-value in your answer to a)?

The Bonferroni corrected p-value for 0.1 is 0.1 / 2000. Thus, the expected number of genes at this p-value is 0.1 (2000 * (0.1 / 2000)). Since we identified 10 genes the FDR is 1%.