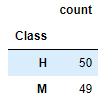
HW4

# Section A

1. The number of genes is 54,675
2. The number of patients is 99
3. The number of patients from each class is M – 49, H – 50



1. The number of genes after cleanup is 54,628

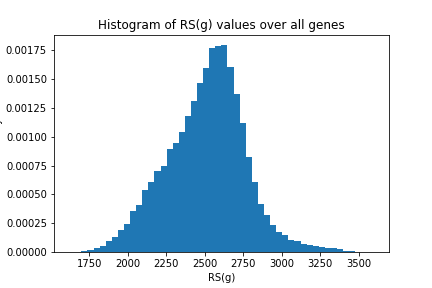
# Section B

The expected sum of ranks for samples labels as M is 2450

1. The minimal value that can take is 1+2+3+…+49=1225.

For questions 3-5 we will calculate the variance

1. We can use the Normal WSRT approximation:   
   similarly we can calculate:  
   Normal WSRT approximation is not accurate enough in the range of the minimal possible values.  
     
   Another way to calculate:  
   There are combination all together. We chose exactly one combination.  
   Under the null assumption: P(
2. There are combination all together. We chose exactly one combination = B! permutations. m+1 can only result from taking the minimal combination m and make one exchange: the sample ranked 49 is exchanged rank with the sample ranked 50. We still had exactly one combination.   
   P(
3. = 2
4. The histogram of values over all genes



# Section C

Since we test 54628 genes, we are in a risk of multiple-test problem of false discovery.

So we are using the Sidack Correction for Familywise error rate correction.

9.3895e-07

We should use this p-value to decide if the null model is contradicted by the result or not.

We found:

WRS

Underexpressed genes: 53

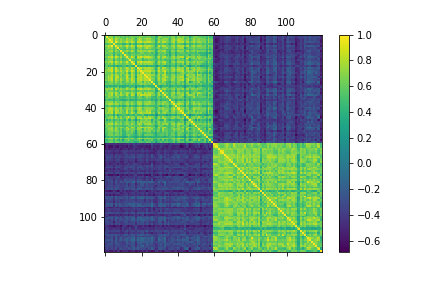
Overexpressed genes: 708

T-Student

Underexpressed genes: 159

Overexpressed genes: 596

# Section D



1. We can report that overexpressed genes are highly correlated with overexpressed genes, and that under expressed genes are highly correlated with under expressed genes. It can also be noted that highly correlated genes have a strong negative correlation with under expressed genes.  
     
   We would like to check 7140 pairs of genes out of the chosen 120.  
   We set alpha to be :

So we are using the Sidack Correction for Familywise error rate correction.

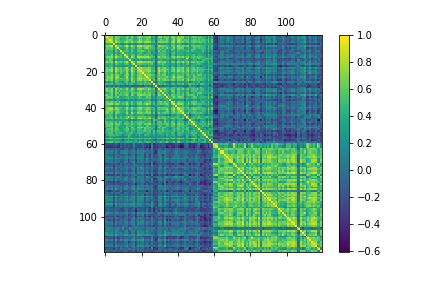
7.1839088e-06

there are 3186 pairs positively correlated.

there are 1019 pairs negatively correlated.

alpha=7.1839089838e-06  
  
this is expected, as we checked the most under-expressed and over-expressed genes, so they are all correlated with the class.

1. Advantages:   
   There can be genes that are correlated but are not included in the set D. A full comparison may find them  
     
   Disadvantages:  
   checking 54,675 genes for correlated pairs means comparisons.  
   1. This can take a very long time .  
   2. If we want to keep the FDR=0.05, we should set an extremely small alpha to compare with the p-value to declare a correlation. This may cause a loss of discoveries.



1. There are 1082 pairs positively correlated.   
   there are 2 pairs negatively correlated.   
   alpha=7.1839089838e-06  
     
   We found less correlation between the genes. This is expected, since checking samples only from class M delete all the correlations resulting from the sensitivity of the genes to the class. If some genes are strongly sensitive to the class, we can expect them to be correlated between them.

# Section E

1.

Since we used the Sidack Correction, and since we set the p-value to meet the need of FDR=0.05, the FDR for of our result is 0.05.

If we wanted to hit 54628 genes with FDR of 0.1, we should use:

= 1.92868884186e-06

If we wanted to hit 54628 genes with FDR of 0.001, we should use:

= 1.83147895738e-08

2.

there are differences between the WRS results and the T-Test result can be due to 2 reasons:

1. T-test is less sensitive to the inner variance within each group. T-Test compares the variance between each group after neglecting the variance of each group. Larger variance can be more harmful to the WSR. In distributions with large variance, T-test can be more reliable.
2. T-test is more sensitive to samples witch are from the “tails”, outliers. So if some of the samples are mistaken (noise) and not resulted form the original distribution, WRS can be more reliable.

3.

Attached qqplots from 3 genes. Each point represents percentages of the samples taken. Blue and orange resales the 2 classes.

