Exercise 9: Univariate analysis

April 25, 2017

Your collaborators from a microbiology lab sent you their data for analysis. They measured the gene expression levels of all 4297 genes in *E. coli* for 12 distinct strains. In each one of the strains, a single gene had been overexpressed. Each measurement was done in 3 independent biological replicates, amounting to 36 samples. The data was sent to you as a MATLAB file named `ex9.mat` which contains the following:

- `genes` – a cellarray with the names of the *E. coli* genes.
- `expression.strainXX` – a matrix containing the gene expression data of strain `XX`, columns correspond repetitions and rows correspond to genes. The data is given in $\log_2$ scale.

You open the file in MATLAB (using the command `load ex9`) and immediately realize that the person who sent you the data forgot to include the names of the strains. You send her an email asking for it but all you get is an out-of-office reply. Since you have a deadline coming up, you decide to try and figure out yourself what each strain represents.

1 First look at the data

1.1 Calculating means and standard deviations

Calculate the mean expression levels and standard deviations per gene (for the samples corresponding to strain01). Use the unbiased estimator for the standard deviation, i.e. $\sigma = \frac{\sum_i^n (x_i-\mu)^2}{n-1}$.

Hint: Compare the results of `mean(X, 1)` and `mean(X, 2)`. Which one gives the correct answer? Similarly try `std(X, 0, 1)` versus `std(X, 0, 2)`. Read the help for `std` and see that you understand what the second argument (i.e. 0) stands for.

1.2 Plot histograms

Plot two histograms of the mean and standard deviation of all gene expression levels in strain01, as two subplots in the same figure.

Hint: Use the standard MATLAB function `hist(Y, 50)`, where Y is an array containing the means or STDs and 50 is the number of bins you would like to see in the histogram.

1.3 Genes with high standard deviations

Four genes have a higher standard deviation than 0.6. Which four genes are they?

Hint: To answer this question, first use the function `ind = find(Y > 1)` to get the set of indices of the genes, and then `genes(ind)` will print out the names of these genes.

1.4 Test significance of fold-change (part I)

A *fold-change* is defined as the ratio between the expression of the same gene in two conditions or strains. Note that it is quite common to present the fold change in a logarithmic scale, specifically $\log_2$. However, in this exercise we will not use the logarithmic scale when discussing fold changes.

Find the mean and standard deviation of the $\log_2$ expression of the gene *dinI* for strain01.

Hint: to find the index of a gene called 'xxx' in the cellarray, use the command `find(ismember(genes, 'xxx'))`

Do the same for the gene *rydC* and the strain04 and fill in the following tables:

Mean $\log_2$ expressions		
strain	gene:*dinI*	gene:*rydC*
01	$\mu = ?$	$\mu = ?$
04	$\mu = ?$	$\mu = ?$

Standard deviations		
strain	gene:*dinI*	gene:*rydC*
01	$\sigma = ?$	$\sigma = ?$
04	$\sigma = ?$	$\sigma = ?$

1.5 Test significance of fold-change (part II)

Use the unpaired two sample student t-test to answer these two questions:

- What is the fold-change* in the expression of *dinI* in strain01 compared to strain04? Is it a significant** change?

- What is the fold-change* in the expression of *rydC* in strain01 compared to strain04? Is it a significant** change?

* provide your answer in the linear scale, i.e. not in $\log_2$ scale like the original data table.
** the standard threshold for significance is 5%.

Hint: The formula for calculating the t statistic is: $t = \frac{|\mu_1 - \mu_2|}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}$, where n is the number of repeats in each condition, i.e. 3. The next step is to use the Student's t distribution to check the probability of having a t statistic of this value. Use the equation `p = 2 - 2*tcdf(t, 4)`, where 4 is the number of degrees of freedom.

2 Volcano plots

2.1 Calculate p-values

Now, perform an unpaired two sample student t-test on all genes between strain01 and strain04. Check that the p-values match the ones you calculated in Question 1.5.

Hint: It is much easier to use the function `PValues = mattest(X, Y, 'VarType', 'equal');` from the bioinformatic toolbox, where X is the expression matrix of the first stain and Y is the matrix for the second strain. Note that we set the parameter `'VarType'` to `'equal'` in order to perform a t-test assuming both samples have equal variances.

2.2 Generate a Volcano plot

Use the Matlab function `mavolcanoplot` to draw a Volcano plot of significance (P-value) versus fold changes between the two strains.

Use the Volcano plot to determine which genes are significantly over- or under-expressed in strain01. Use cutoff values of 2.0 fold-change and p-value 0.05.

Hint: The syntax should look like this: `mavolcanoplot(X, Y, PValues, 'Labels', genes);`

2.3 Familywise Error Rate (FWER)

What should be the cutoff p-value when taking Familywise Error Rate into consideration (use the Bonferroni correction)? Which of the over- or under-expressed genes remain significant after accounting for FWER?

2.4 False Discovery Rate (FDR)

Use Storey's positive FDR method to calculate the q-values. Make a new volcano plot with these q-values instead of the original p-values, and answer question 2.2 again using the new plot.

Hint: Use the `mafdr` function.

2.5 Strain relabelling

Assuming that you know that each one of the strains was prepared by overexpressing a single *E. coli* gene, what do you think is the correct label for strain01 and strain04?

Try to guess which gene was overexpressed in strain02, by comparing it to strain01.

2.6 Bonus question

Why do genes with larger absolute fold-changes also tend to have a more significant p-value, in other words – what is the reason for the Volcano shape of the plot?