Bioinformatics – Lecture Notes

Announcements

Reference: Microarray Data Analysis and Visualization by Arun Jagota

4 days of classes including today.

Final Projects due May 2 (2 weeks from today).

Class 27 – April 23, 2002

I. Finding genes that are expressed differently in two populations

Assume that we have a population and from that population we have two sub-populations. For example, The population can be the set of people having a specific cancer and the subpopulations can be those who have received treatment and those who have not (treated and control). Another example could be as follows. Consider a population with a certain cancer. If the cancer has two subtypes (Leukemia has two subtypes). The set of people with each subtype forms a subpopulation. We want to determine if genes are expressed differently in the two subpopulations.

A. How to determine if a specified gene is expressed differently in two populations

1) Simple test

One way to determine if gene x is expressed differently in two populations, control vs. treated is to compute the mean levels of gene x in the populations t and c, $\overline{e}_{x,c}$ and $\overline{e}_{x,t}$. A threshold $\boldsymbol{q}$ can be chosen (usually 2 or 2.5) for which if $\frac{\overline{e}_{x,t}}{\overline{e}_{x,c}} \geq \boldsymbol{q}$, x is considered to be significantly up-regulated in the treated group compared to the control. The caveat with this approach is that when the group sizes n_c and n_t are small, genes with low expression levels on average might have a much higher variance than genes with high expression levels. A small $\boldsymbol{q}$ can result in many false positives, while a large $\boldsymbol{q}$ can result in many false negatives.

One way to alleviate this problem is to choose a high threshold $\boldsymbol{q}$ for gene with low expression levels and a low threshold for genes with high expression levels. A better way to solve this problem is to use a standard statistical test.

2) Statistical inference using the normal distributions

In order to use statistical inference, consider $\bar{e}_{x,d} = \bar{e}_{x,c} - \bar{e}_{x,t}$. If $\bar{e}_{x,d}$ is significantly different from 0, the one can conclude that x is expressed differently in the two populations. Assume that the groups contain n_c and n_t samples and that the sample population for the treated and control groups are chosed randomly form the population of treated and control individuals. Also assume the two groups are drawn independently and that the two populations are independent. If n > 30 we can invoke the central limit theorem and assume that $\bar{e}_{x,d}$ is normally distributed with mean $m_{x,c} - m_{x,t}$ and standard deviation

$$\sqrt{s_{x,c}^2/n_c + s_{x,t}^2/n_t}$$

This yields a z statistic

$$z(\bar{e}_{x,d}) = \frac{(e_{x,c} - e_{x,t}) - (m_{x,c} - m_{x,t})}{\sqrt{s_{x,c}^2/n_c + s_{x,t}^2/n_t}}$$

that can be used to answer the question is $z(\bar{e}_{x,d})$ sufficiently different from zero to be able to conclude that gene x is expressed differently in the two populations. The null hypothesis asserts that there is no difference, ie. H_0: $m_{x,c} = m_{x,t}$. The alternative hypothesis is H_A: $m_{x,c} \neq m_{x,t}$. Then

$$z(\bar{e}_{x,d}) = \frac{(e_{x,c} - e_{x,t})}{\sqrt{s_{x,c}^2/n_c + s_{x,t}^2/n_t}}$$

A level of significance α is chosen so that if $|z(\bar{e}_{x,d})| \geq |z_{a/2}|$, we reject the null hypothesis with confidence 1- α. Notice that we used $\alpha/2$ since this is a two sided test.

3) One sided- test

If we wanted to test H_0: $m_{x,c} \geq m_{x,t}$. with H_A: $m_{x,c} < m_{x,t}$, we use a one-sided test with $z(\bar{e}_{x,d}) \leq -z_a$ signifying rejection of the null hypothesis with confidence 1-α.

4) T-test

If we do not know the population variances or the sample is small, and wanted to test H_0: $m_{x,c} = m_{x,t}$ with H_A: $m_{x,c} \neq m_{x,t}$, we need to used the t-test

$$t(\bar{e}_{x,d}) = \frac{(e_{x,c} - e_{x,t})}{\sqrt{s_{x,c}^2/n_c + s_{x,t}^2/n_t}}$$

where $s_{x,c}$ and $s_{x,t}$ are the sample standard deviations. Then if $|t(\bar{e}_{x,d})| \geq |t(\min((n_c, n_t) - 1, \alpha/2|$, we can reject the null hypothesis with confidence 1-α.

5) Equal variance

If we want to test to see if the two populations have different variances we can test this using the F-statistic with H_0: $\sigma_{x,c}^2 = \sigma_{x,t}^2$ with H_A: $\sigma_{x,c}^2 \neq \sigma_{x,t}^2$,

$$F = \frac{s_{x,c}^2}{s_{x,t}^2}$$

which is determined from groups of size n_c and n_t chosen randomly from the control and treated populations, repectively. If $F \geq F(n_c - 1, n_t - 1, \alpha/2)$ or $1/F \geq F(n_t - 1, n_c - 1, \alpha/2)$ we reject the null hypothesis with confidence 1-α.

If the F-test failed to reject the null hypothesis, we have to use a different t-test because the variances might be equal. We use a pooled estimate of the variance

$$s_{x,p}^2 = \frac{(n_c - 1)s_{x.c}^2 + (n_t - 1)s_{x.t}^2}{n_c - n_t - 2}$$

and use the test statistic

$$t_{eq}(\bar{e}_{x,d}) = \frac{\bar{e}_{x,d}}{\sqrt{s_{x,p}^2(1/n_c + 1/n_t)}}$$

If $|t_{eq}(\bar{e}_{x,d}) \geq t(n_c - n_t - 2, \alpha/2)$, we can reject the null hypothesis with confidence with confidence 1-α for the two sided test.

For the one sided test and corresponding hypothesis. If $t_{eq}(\bar{e}_{x,d}) \leq -t(n_c - n_t - 2, \alpha/2)$, we can reject the null hypothesis with confidence with confidence 1-α.

6) Bayesian t-test

If n_c and n_t are small and we have some prior knowledge about the expected distributions of the expression levels of gene x in one or both groups we can base our test on the data and the knowledge. This is done by generating $d_{0,c,x}$ random data examples from a normal distribution with mean $\mu_{0,c,x}$ and variance $s^2_{0,c,x}$ and add these examples to our data. In this way we get a new sample mean

$$\hat{m}_{n_c x} = \frac{d_{0.c.x}}{d_{0,c,x} + n_c} m_{0,c,x} + \frac{n_c}{d_{0,c,x} + n_c} \bar{e}_{x,c}$$

If we generate $e_{0,c,x}+1$ examples with mean $\mu_{0,c,x}$ and variance $s^2_{0,c,x}$, the variance is

$$\hat{s}^2_{n_c c} = \frac{1}{n_c + e_{0,c}} \left(e_{0,c,x} s^2_{0,c,x} + (n_c - 1) s^2_{c,x} + \frac{e_{0,c,x} n_c}{e_{0,c,x} + n_c} (\bar{e}_{x,c} - m_{0,x,c})^2 \right)$$

Similarly for the treated data.

7) Nonparametric tests

If we want to test, based on a comparision of the group means, whether or not there is sufficient evidence that gene x is expressed differently in the two populations, we can use the Mann-Whitney U test. Assume that we have a set G_t of expression levels of gene x in n_t individuals treated for cancer and a set G_c of expression levels of gene x in n_c individuals who form a control group. Assume $\mu_{x,t}$ and $\mu_{x,c}$ are the respective means. We want to test whether $H_0: \mu_{x,t} = \mu_{x,c}$ is true.

The data from the two groups G_t and G_c are pooled in one group G_{ct} . The numbers in G_{ct} are sorted into increasing order and ranked from 1 to n_c+n_t. If some numbers are the same, their ranks are averaged and each is assigned the averaged rank. Let S_t be the sum of the ranks from elements of G_t, and S_c be the sum of the ranks from elements of G_c. The U-score for each group is

$$U_c = n_c n_t + \frac{n_t (n_t + 1)}{2} - S_t$$
$$U_t = n_c n_t + \frac{n_c (n_c + 1)}{2} - S_c$$
$$U = \min(U_c, U_t)$$

If each group has at least 10 members, we can assume that under H_0 U is approximately normally distributed with

$$\mu_U = \frac{n_c n_t}{2}$$

$$\sigma_U = \sqrt{\frac{n_c n_t (n_c + n_t + 1)}{12}}$$

and

$$z(U) = \frac{U - \mu_U}{\sigma_U}$$

For a given α, if $|z(U)| \geq |z(\alpha/2)|$, then reject H_0 with confidence 1- α.

B. How to determine if any gene is expressed differently in two populations

As before use the compound test with the Bonforroni correction with the tests discussed today.

II. Classifying samples from two populations

Assume that we want to classify a new sample into one of two populations. For example, cancerous and non-cancerous, or cancer subtype I and cancer subtype II. To do this we want to develop a method that will choose one or more classifiers from a set of thousands of genes and a marker with which we will classify our data.

If we have gene expression data from our two populations, we need to determine which genes we will use to classify new data into one of the two populations. This is called a *binary classification problem*.

A. Variable selection

We previously discussed methods that would identify genes that could be used to discriminate between the two tissues. The problem with these methods is that the treated each gene individually. We need methods that look at the genes collectively. The method we choose has to be able to 1) evaluate the quality of a subset of the variables and 2) search through all subsets of the data

One method is *sequential backward elimination*. This method starts from the full set of genes and removes genes one at a time on the basis of which gene make the smallest reduction (or largest increase) in the performance of the classifier among all candidates considered for removal

Another method is *greedy selection*. In this, we start with an empty set of genes and add the gene that make the best one gene classifier. Then we add the gene that results in the best two-gene classifier, etc. This is faster if we are interested in obtaining a small number of classifiers.

B. Discriminant analysis
C. Multilayer perceptions
D. Training
E. Overfitting
F. Support vector machines