StatsM254 Statistical Methods in Computational Biology Lecture 3/4 - April 6, 8 2015

Gene Expression Analysis

Lecturer: Jingyi Jessica Li *Scribe: Douglas Arneson, Adriana Sperlea*

1 GENE EXPRESSION ANALYSIS

Our data are represented by the matrix: $X = (X_{ij})_{nxm}$ where:

- Rows → represent different **genes** (n)
- Columns → represent different **samples** (m)
- X_{ij} → expression level of **gene** i in **sample** j

The following is a representation of data seen in gene expression analysis:

$$\begin{array}{cc} Condition1 & Condition2 \end{array}$$

Samples

$$\text{Genes}\begin{pmatrix} 1,1 & 2,1 & \cdots & m_1,1 \\ 1,2 & 2,2 & \cdots & m_1,2 \\ \vdots & \vdots & \ddots & \vdots \\ 1,n & 2,n & \cdots & m_1,n \end{pmatrix}\begin{pmatrix} 1,1 & 2,1 & \cdots & m_2,1 \\ 1,2 & 2,2 & \cdots & m_2,2 \\ \vdots & \vdots & \ddots & \vdots \\ 1,n & 2,n & \cdots & m_2,n \end{pmatrix}$$

$\nwarrow \nearrow$

The conditions are compared to find differentially expressed genes

Hypothesis - test for every gene. Say we fixed gene 1, and its expression levels are:
$X_1,\cdots,X_{m_1}$ μ_1 where μ is the **population mean**
$Y_1,\cdots,Y_{m_2}$ μ_2

Null Hypothesis - H_0: $\mu_1=\mu_2$
If we reject H_0, gene 1 is called **differentially expressed** (DE)
If we accept H_0, gene 1 is **not** DE

Simple Solution **- t test**

Assumption: $X_1,\cdots,X_{m_1} \sim N(\mu_1,\sigma^2)$
$Y_1,\cdots,Y_{m_2} \sim N(\mu_2,\sigma^2)$

We assume a normal Gaussian distribution (however, this can be relaxed when m_1 and m_2 are large, by Central Limit Theorem)
The main assumption is that **the variance (σ^2) is the same**

Sample Mean: $\bar{X} = \frac{1}{m_1}\sum_{i=1}^{m_1} x_i$ $\bar{Y} = \frac{1}{m_2}\sum_{i=1}^{m_2} y_i$

Sample Variance: ${S_x}^2 = \frac{1}{m_1-1}\sum_{i=1}^{m_1}(x_i - \bar{x})^2$ ${S_y}^2 = \frac{1}{m_2-1}\sum_{i=1}^{m_2}(y_i - \bar{x})^2$

The denominator (i.e. $m_1 - 1$) indicates that the sample variance is **unbiased**

Pooled Sample Variance: ${S_p}^2 = \frac{(m_1-1){S_x}^2+(m_2-1){S_y}^2}{m_1+m_2-2}$

t Statistic: $T = \frac{\bar{x}-\bar{y}}{S_p\sqrt{\frac{1}{m_1}+\frac{1}{m_2}}}$

Under $H_0 : T \sim t_{m_1+m_2-2}$

2 QUESTION

Assuming the two means are the same, are the variances different?

$\mathrm{X}_1,\cdots,\mathrm{X}_{\mathrm{m}_1} \sim \mathrm{N}(\mu,{\sigma_1}^2)$ $\mathrm{Y}_1,\cdots,\mathrm{Y}_{\mathrm{m}_2} \sim \mathrm{N}(\mu,{\sigma_2}^2)$
$H_0 : {\sigma_1}^2 = {\sigma_2}^2$

F-statistic: $\mathrm{F} = \frac{{S_x}^2}{{S_y}^2}$

Under the Normal assumption and H_0:
$\mathrm{F} \sim F_{m_1-1,m_2-1}$

Small sample problem: **often** $m_1 = m_2 = 3$

In t-test: only 6 data points to calculate ${S_p}^2$, which is to estimate σ^2

- This is unstable
- So the t statistic will be unstable

Use Bayesian to help stabilize the estimate

3 NUISSANCE PARAMTER

$\mathrm{X}_1,\cdots,\mathrm{X}_{\mathrm{m}} \sim \mathrm{N}(\mu,\sigma^2)$
μ - **parameter of interest**, it is unknown
σ^2 - **nuissance parameter**, unknown but we don't care about it

$\mathrm{L}(\mu, \sigma^2|X_1, \cdots, X_n) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$

Frequentist: to find μ^* to maximize $L(\mu, \sigma^2|X_1, \cdots, X_n)$ consider σ^2 as fixed
$\hat{\mu}_{MLE} = f(\hat{\sigma}_{MLE})$

Bayesian: prior of σ^2, e.g. inverse-chi-square $\rightarrow$ conjugate of $\mathrm{N}(\mu, \sigma^2)$ to maximize:
$\int_{\sigma^2} L(\mu, \sigma^2|X_1, \cdots, X_n) \cdot p(\mu) \cdot p(\sigma^2) d\sigma^2$
$\propto \int_{\sigma^2} p(\mu, \sigma^2|X_1, \cdots, X_n) d\sigma^2 = p(\mu|X_1, \cdots, X_n)$ to find $\hat{\mu}_{Bayesian}$

4 GENE EXPRESSION ANALYSIS CONTINUED

Gene expression data matrix: $X = (X_{ij})_{m\times(n_1+n_2)}$

- m genes
- n_1 samples in condition 1
- n_2 samples in condition 2

Looking at i^{th} gene:

When n_1 and n_2 are **very small**, the pooled sample variance is:

$S_p{}^2 = \frac{(n_1-1){S_1}^2+(n_2-1){S_2}^2}{n_1+n_2-2}$, where
$S_x{}^2 = \frac{\sum_{i=1}^{n_1}(X_{ji}-\bar{X}_j)^2}{n_1-1}$, $S_y{}^2 = \frac{\sum_{i=1}^{n_2}(Y_{ji}-\bar{Y}_j)^2}{n_2-1}$
$\bar{X}_j = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{ji}$, $\bar{Y}_j = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_{ji}$
$S_p{}^2$ is an **unstable estimate** of σ^2.

To stabilize the estimate of σ^2, we can borrow information from the **prior** of σ^2.

For convenience, we use the **conjugate prior inverse-chi-square distribution**.

Fact: $\frac{(n_1+n_2-2){S_p}^2}{\sigma^2} \sim X^2_{n_1+n_2-2}$

Likelihood: $d \triangleq n_1 + n_2 - 2$
$L(\sigma^2|{S_p}^2) = p(\frac{d{S_p}^2}{\sigma^2}|\sigma^2) \propto (\frac{d{S_p}^2}{\sigma^2})^{\frac{d}{2}-1}e^{-\frac{d{S_p}^2}{2\sigma^2}} \Rightarrow p({S_p}^2|\sigma^2) \propto (\sigma^2)^{-\frac{d}{2}}e^{-\frac{d{S_p}^2}{2\sigma^2}\cdot({S_p}^2)^{\frac{d}{2}-1}}$
$\sigma^2 \sim$ Inverse - $X^2(v, {s_0}^2)$
$p(\sigma^2) \propto (\sigma^2)^{-\frac{v}{2}-1}e^{-\frac{v{s_0}^2}{2\sigma^2}}$

$\Rightarrow$ **Posterior** $p(\sigma^2|{S_p}^2) \propto p({S_p}^2|\sigma^2)\cdot p(\sigma^2)$
$\propto (\sigma^2)^{-\frac{v+d}{2}-1}\cdot e^{-\frac{v{s_0}^2+d{S_p}^2}{2\sigma^2}}$

$\Rightarrow$ **A common approach:**
$\hat{\sigma^2} = E[\sigma^2|{S_p}^2] = \frac{v{s_0}^2+d{S_p}^2}{v+d-2}$

If we set $v >> d$, then:
$\hat{\sigma}^2 \approx \frac{v{s_0}^2+d{S_p}^2}{v+d} = (\frac{v}{v+d}){s_0}^2 + (\frac{d}{v+d}){S_p}^2$

We can fix $v = v^*$ and find ${s_0}^2$ by maximizing the join density of ${S_p}^2$ given v^* and ${s_0}^2$.
$p({S_p}^2|v^*, {s_0}^2) = \int p({S_p}^2, \sigma^2|v^*, {s_0}^2)d\sigma^2 = \int p({S_p}^2|\sigma^2)\cdot p(\sigma^2|v^*\cdot {s_0}^2)d\sigma^2$

Then find ${s_0}^2$ as:

$$({s_0}^2)^* = \arg\max_{s_0^2} p(S_p^2|v = v^*, {s_0}^2)$$

Lastly, plug in v^* and $({s_0}^2)^*$ into
$\hat{\sigma}^2 = (\frac{v^*}{v^*+d})({s_0}^2)^* + (\frac{d}{v^*+d}){S_p}^2$

$\rightarrow$ **t-test**