

Gene Expression Analysis

Lecturer: Jingyi Jessica Li

Scribe: Sepideh Mazrouee

1 Introduction: Gene selection by comparative analysis

Table 1: Expression data matrix

Condition	1	2	3	...	N
Gene1	X_{11}	X_{12}	X_{13}	...	X_{1N}
Gene2	X_{21}	X_{22}	X_{23}	...	X_{2N}
.
.
.
GeneM	X_{M1}	X_{M2}	X_{M3}	...	X_{MN}

To identify genes that are differentially expressed between 2 expressional conditions (summarized in Table 2)

Table 2: Two sample comparison

	Condition 1	Condition 2
Replicates	$1, 2, \dots, n_1$	$1, 2, \dots, n_2$
Gene 1		
Gene 2		
.		
Gene M		

Hypothesis test for gene m ($m = 1, \dots, M$)

$H_0 : \mu_{m1} = \mu_{m2}$ (True expression of gene m in condition1)

$H_1 : \mu_{m1} \neq \mu_{m2}$

Note: we need to use the above data points to test this hypothesis

Generally, ignore the gene index m. Observe expression values:

x_1, \dots, x_{n1} (from condition1)

y_1, \dots, y_{n2} (from condition2)

1.1 t test:

to test this hypothesis we can do " t test". The underlying assumptions are as below:

$X_1, \dots, X_{n1} \stackrel{iid}{\sim} N(\mu_1, \sigma^2)$

$Y_1, \dots, Y_{n2} \stackrel{iid}{\sim} N(\mu_2, \sigma^2)$

(this is considered as an extreme case)

Define:

$$\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i \quad , \quad \bar{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i \quad \text{Sample Mean} \quad (1)$$

$$S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 \quad , \quad S_Y^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 \quad \text{Sample Variance} \quad (2)$$

$$S_P^2 = \frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2} \quad \text{Pooled Sample Variance} \quad (3)$$

T statistic:

$$T = \frac{\bar{X} - \bar{Y}}{S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \quad \text{if } H_0 \text{ is true} \quad (4)$$

Given data, we can calculate the observed value of T statistic:

$$t = \frac{\bar{x} - \bar{y}}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (5)$$

Figure1 shows the distribution of t test which looks close to a Normal distribution:

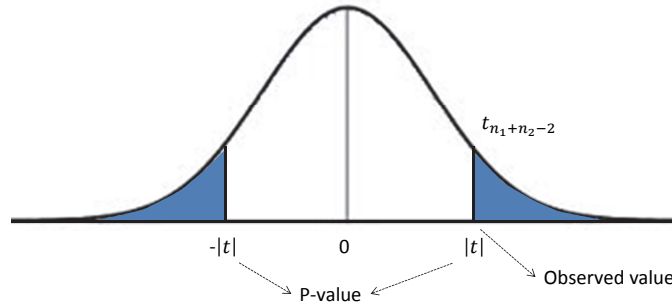


Figure 1: t test Distribution

We first assume it's true and then compare our observation with the distribution. We reject Null Hypothesis at the significance level α (let's say 0.05).

To summarize:

- If $\text{P-value} \leq \alpha$: reject Null Hypothesis \Rightarrow genes are differentially expressed
- If $\text{P-value} > \alpha$: accept Null Hypothesis \Rightarrow genes are NOT differentially expressed

1.2 F -test:

we can also do F -test with the Null hypothesis as the gene expressions have the same mean but different variance under the two conditions (which is considered as another extreme case)

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

$$X_1, \dots, X_{n_1} \sim N(\mu, \sigma_1^2)$$

$$Y_1, \dots, Y_{n_2} \sim N(\mu, \sigma_2^2)$$

$$S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 \quad , \quad S_Y^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 \quad (6)$$

F-Statistic:

$$F = \frac{S_X^2}{S_Y^2} \sim F_{n_1-1, n_2-1} \quad \text{if } H_0 \text{ is true} \quad (7)$$

Let f be the observed value of the F -Statistic. Assuming the null hypothesis is true.

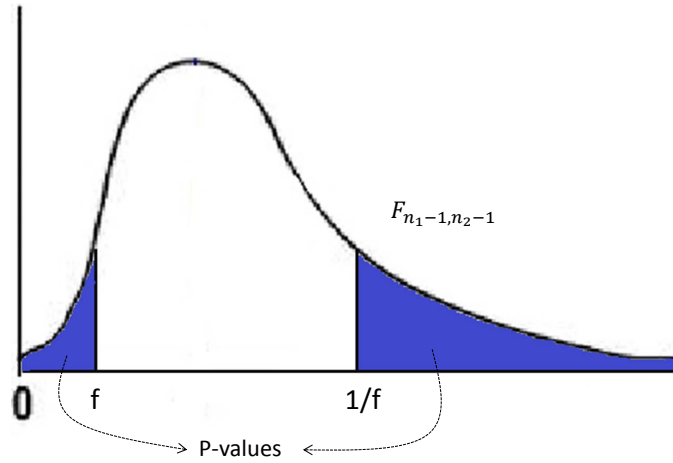


Figure 2: F distribution - two-sided

$$H_1: \sigma_1^2 \neq \sigma_2^2 \quad (\text{two-sided test})$$

Based on different assumptions, we might test one sided as well. Then the alternative hypothesis would change to

$$H_1: \sigma_1^2 < \sigma_2^2 \quad (\text{one-sided test})$$

Note: the second one uses the same data as first one. The only difference is the alternative hypothesis.

1.3 Permutation test:

The two previous cases were both extreme cases. Let's look at a more general case in which we do permutation. Advantages of permutation test could be listed as:

1. Distribution free: it does not apply specific distribution on the data.
2. No need to find the probabilistic distribution of the test statistic.

Let say for previous example if we did not have Normal distribution for data X and Y , we could not be able to apply t test or F test for them.

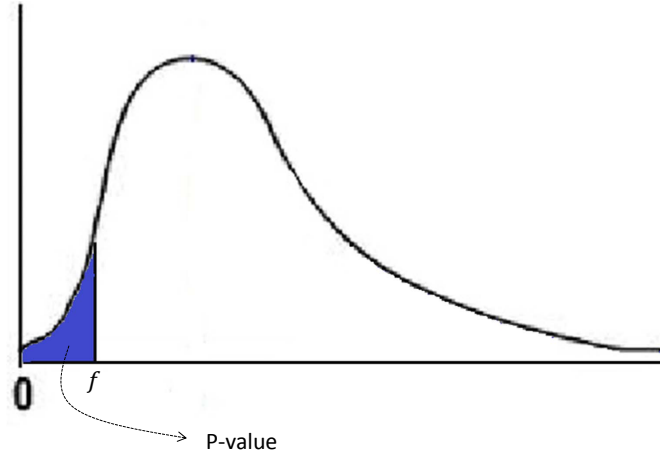


Figure 3: F distribution - one-sided

Procedure: two groups of iid random variables (gene expression values)

$$X_1, \dots, X_{n1}$$

$$Y_1, \dots, Y_{n2}$$

H_0 : distributions of X and Y are the same

H_1 : the distributions of X and Y are the different

So if null hypothesis is true, then $X_1, \dots, X_{n1}, Y_1, \dots, Y_{n2} \stackrel{iid}{\sim}$ common distribution. Then we have $Q = \binom{n_1+n_2}{n_1}$ possible ways to group $X_1, \dots, X_{n1}, Y_1, \dots, Y_{n2}$ into two groups (each of such groupings is a permutation).

Assumption: All the Q permutations have the same probability (iid) if H_0 is true.

If we care about the mean difference, we can still use t statistic:

$$T = \frac{\bar{X} - \bar{Y}}{S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (8)$$

In each permutation say m compute the value of $t_m : (m = 1, \dots, Q) \Rightarrow$ Empirical distribution of t_1, \dots, t_Q is as below

Example: $t_{obs} = \frac{\bar{x} - \bar{\mu}}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 3$

$$n_1 = n_2 = 5 \text{ (5 replicates in each condition)}$$

$$Q = \binom{10}{5} = 252 \Rightarrow \text{we can compute } t_1, \dots, t_{252}$$

If at significance level $\alpha = 0.05$ then $252 * 0.05 \cong 13$. Then if we sort them ascending order we can find the rejection region:

Approximation: if $\binom{n_1+n_2}{n_1} = Q$ is too large;

Draw N (e.g. $N=1000$) random permutations to compute the p-value .

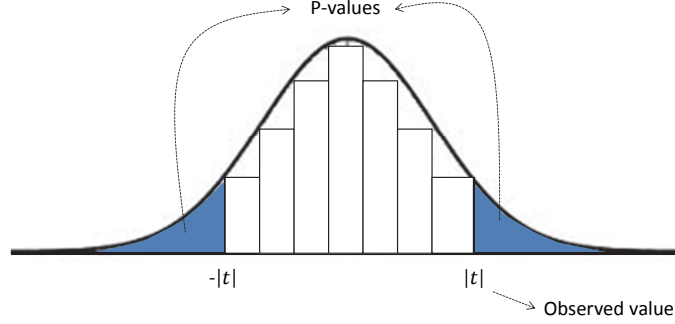


Figure 4: Histogram of t_1, \dots, t_q

Table 3: t observations

t_1
 \cdot
 \cdot
 \cdot
 t_{252}

2 Shrinkage estimator of variance

Back to the 2-sample t -test setting. Given gene g , ($g = 1, \dots, M$)

$$X_{g_1}, \dots, X_{g_n} \stackrel{iid}{\sim} N(\mu_{g_1}, \sigma_g^2)$$

$$Y_{g_1}, \dots, Y_{g_m} \stackrel{iid}{\sim} N(\mu_{g_2}, \sigma_g^2)$$

Often $n + m$ is small, but M is big (small number of reps, but many genes)

Pooled sample variance

$$S_g^2 = \frac{1}{m+n-2} \left(\sum_{i=1}^n (X_{g_i} - \bar{X}_g)^2 + \sum_{j=1}^m (Y_{g_j} - \bar{Y}_g)^2 \right) \quad \text{unstable} \quad (9)$$

$$t_g = \frac{\bar{X}_g - \bar{Y}_g}{\sqrt{(\frac{1}{n} + \frac{1}{m})S_g^2}} \quad \text{unstable too} \quad (10)$$

We will use hierarchical modeling for σ_g^2 , ($g = 1, \dots, M$) (to help stabilize S_g^2)

2.1 Bayesian statistics: a prior for σ_g^2

we know that $\frac{(m+n-2)S_g^2}{\sigma_g^2} \sim \chi_{m+n-2}^2$ (*)

An conjugate prior for χ^2 is inverse- χ^2

prior of $\sigma_g^2 : (g = 1, \dots, M) : Inv - \chi^2(v, s_0^2)$ (**)

from (*): Let $d = m + n - 2$

density: $\rightarrow p(S_g^2 | \sigma_g^2) \propto (\sigma_g^2)^{-\frac{d}{2}} e^{-\frac{dS_g^2}{2\sigma_g^2}}$

From(**) \rightarrow prior: $\pi(\sigma_g^2|v, s_0^2) \propto (\sigma^2)^{-\frac{v}{2}-1} e^{-\frac{vs_0^2}{2\sigma_g^2}}$

By Bayes theorem:

Posterior: $p(\sigma_g^2|S_g^2, v, s_0^2) \propto p(S_g^2|\sigma_g^2)\pi(\sigma_g^2|v, s_0^2) \propto (\sigma_g^2)^{-(\frac{v+d}{2}+1)} \exp[-\frac{vs_0^2+dS_g^2}{2\sigma_g^2}]$

$$\begin{aligned} \sigma_g^2|S_g^2 &\sim \text{Inv} - \chi^2 \left(v + d, \frac{vs_0^2+dS_g^2}{v+d} \right) \\ \Rightarrow \hat{\sigma}_g^2 &= E[\sigma_g^2|S_g^2] = \frac{1}{v+d-2}(vs_0^2 + dS_g^2) \end{aligned}$$

Choose $v \gg d \Rightarrow \frac{vs_0^2+dS_g^2}{v+d}$

Given a pre-specified prior parameter v , we can find the prior parameter s_0^2 by maximizing the joint density $\prod_{g=1}^M p(S_g^2|v, s_0^2) = \prod_{g=1}^M \int p(S_g^2, \sigma_g^2|v, s_0^2)d\sigma_g^2 = \prod_{g=1}^M \int p(S_g^2|\sigma_g^2)\pi(\sigma_g^2|v, s_0^2)d\sigma_g^2$

Then we replace S_g^2 by $\hat{\sigma}_g^2$ in the t statistic