Randomization and Permutation

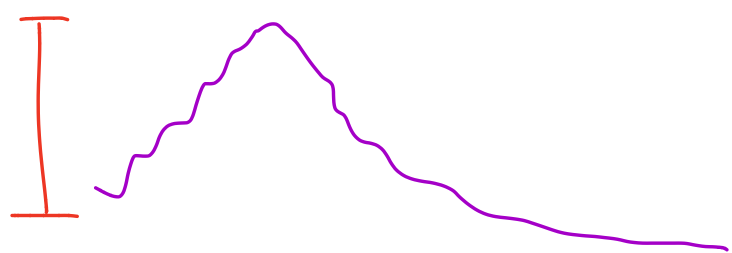
- Compensate for lack of power

- S gene set
- $Z_S = \{Z_i : i \in S\}$ $\quad m = |S|$

$$S(Z_S) = \sum_{i=1}^{S} \frac{S(Z_i)}{m}$$

then we use KS testing on the estimated CDFs.

ES(s)



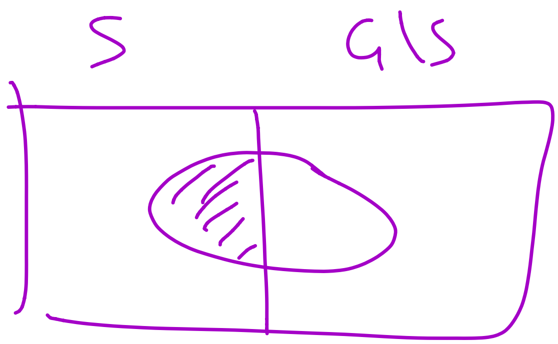
How to asses significance of the enrichment statistic?

$|S| = m \qquad S \subseteq G$

$|G| = N$

$A = \{ i : |z_i| > c \} \quad |A| = a$

$|S \cap A| = k$



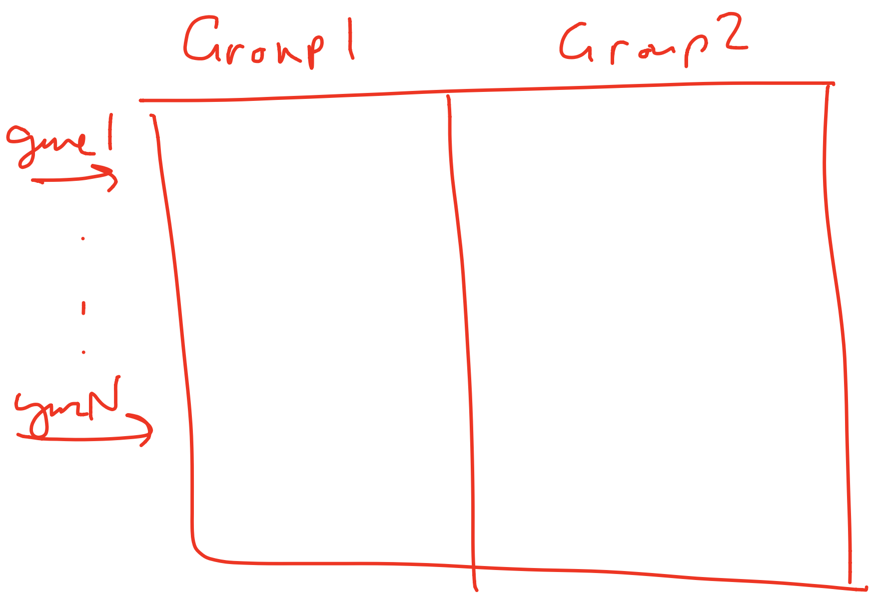
$k \sim \text{Hypergeometric}(m, N-m, a)$

$$P(X=k) = \frac{\binom{m}{k}\binom{N-m}{a-k}}{\binom{N}{a}}$$

So our test statistic is given by

$$\sum_{k \geq \#\text{obs.}} \mathbb{P}(X=k)$$

In general we don't just use counts we use some score.



- Column permutation to calculate

$$z_*^1, \dots, z_*^B$$

$$S_b^* = S(z_b^*)$$

$$\text{p-val} = \#\{S^{*b} > S\}$$

$$\overline{}$$
$$B$$

- Row randomization
  - Select subset of size m

$$Z_{+}^{b} = \{ z_i : i \in I_{+}^{b} \}$$

$$S_{+}^{b} = (S z_{+}^{b})$$

$$p\text{-val} = \frac{\# \{ S_{+}^{b} > S \}}{B}$$

$$S(z_s) = \sum_{z_i \in} \frac{s(z_i)}{m}$$

$$m_s = \frac{1}{N} \sum_{i=1}^{n} s(z_i)$$

$$sd_s = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (s(z_i) - m_s)^2}$$

then from the permutations

$$S_{**}^{b} = m_s + sd_s \left\{ \frac{S_*^{b} - m_s^*}{sd_*} \right\}$$

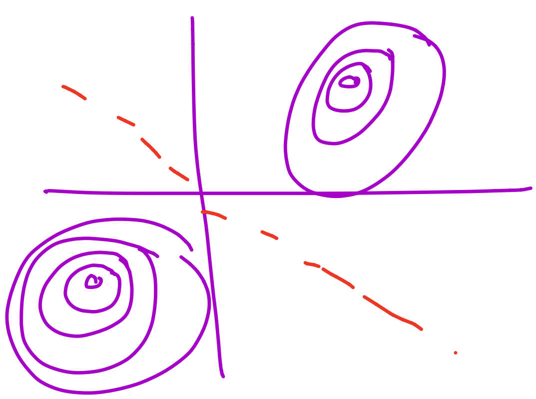
$$P_s^{**} = \frac{\#\{S_{**}^{b} > S\}}{B}$$

Chapter 11 Prediction & Others

Fishers linear discriminant function

$y = 1 \quad x \sim N_N(\delta_1, \Sigma)$

$y = 2 \quad x \sim N_N(\delta_2, \Sigma)$



$$\mathbb{P}(Y=1|X) \propto \mathbb{P}(X|Y=1)\mathbb{P}(Y=1)$$

$$N_N(\delta_1, \Sigma)\,\pi_1$$

$$\log \frac{\mathbb{P}(Y=2|x)}{\mathbb{P}(Y=1|x)} = \log \frac{\mathbb{P}(Y=2)}{\mathbb{P}(Y=1)} + \log \frac{\mathbb{P}(X|Y=1)}{\mathbb{P}(X|Y=2)}$$

$$\log \frac{\exp\left(-\frac{1}{2}(x-\delta_2)^T \Sigma^{-1}(x-\delta_2)\right)}{\exp\left(-\frac{1}{2}(x-\delta_1)^T \Sigma^{-1}(x-\delta_2)\right)}$$

$$\log \exp\left\{(\delta_2-\delta_1)\Sigma^{-1}x - \frac{1}{2}\delta_2^T\Sigma\delta_2 + \frac{1}{2}\delta_1^T\Sigma\delta_1\right\}$$

So the log ratio is

$$\log \frac{\mathbb{P}(y=2|x)}{\mathbb{P}(y=1|x)} = \beta_0 + \beta' x$$

$$\beta^T = (\delta_2 - \delta_1)^T \Sigma^{-1}$$

"Best" for Bayesian stuff

- Bad when $\Sigma_1 \neq \Sigma_2$.
- Bad for high dimensional LDA.

<u>A frequentist model</u>

$$u_i = \frac{x_i - \mu_i}{\sigma_i}$$

$$\vec{u} \sim N\left(\frac{-\delta}{2c_0}, I\right) \text{ for } y=1$$

$$\vec{u} \sim N\left(\frac{\delta}{2c_0}, I\right) \text{ for } y=2$$

$$c_0 = \sqrt{\frac{n_1 n_2}{n}}$$

Construct a prediction function

$$S = \sum \delta_i \cdot u_i \sim N\left(\pm \|\delta\|^2 / 2c_0, \|\delta\|^2\right)$$

Classifier: $S < 0 \quad \hat{y} = 1$

$S > 0 \quad \hat{y} = 2$

$$\mathbb{P}_1(S > 0) = \mathbb{P}_1\left(\frac{S + \|\delta\|^2/2c_0}{\|\delta\|} > \frac{\|\delta\|}{2c_0} \right)$$

$$= \Phi\left(-\frac{\|\delta\|^2}{2c_0}\right)$$

Then to estimate

$$\frac{\overline{X_{i1}} + \overline{X_{i2}}}{2} \qquad \hat{\sigma}_i^2 = \frac{SS_{i1} + SS_{i2}}{n-2}$$

$$\overline{\delta_i} \sim c_0 \frac{\overline{X_{i2}} - \overline{X_{i1}}}{\sigma_i} \sim N(\delta_i, 1)$$

We could also use CV to choose the δ vector.

Bayes & Empirical Bayes

$\delta \sim g \qquad z|\delta \sim N(\delta, 1)$

$$f = \int \phi(z-\delta) g_\delta \, d\delta$$

$$g(\delta|z) = \exp(z\delta - \psi(z)) \exp\left(-\frac{\delta^2}{2} g(\delta)\right)$$

$$\psi(z) = \log\left(\frac{f(z)}{\varphi(z)}\right)$$

in the exp. family.

easy to find mean/variance

$$\mathbb{E}(\delta|z) = z + \ell'(z)$$

$$V(\delta|z) = 1 + \ell''(z)$$