Permutation test

Outline

- Basic Concepts
  - Sample space
  - Exchangeability
- Application in Genomic Studies
  - Pooling results across genes (e.g. SAM)
  - Permutation procedure for pathway analysis

Permutation test vs. standard test

- Standard test (e.g. t test) may be unreliable because the distribution assumptions do not hold.
- Permutation tests offer an alternative testing approach that relies on relatively weak assumptions and yet are quite powerful and simple to apply.

Permutation sample space

Suppose the observed data are $D=\{x_1, x_2, ..., x_n\}$, where $x_i \sim F(x;\theta)$.

The permutation sample space $\Lambda = \{D_1, D_2, D_3, ..., D_M\}$, where each of its element, D_i, is a permutation of the observed data D but has the same amount of information on θ as D.

A Simple Example (I)

- Outcome Y is collected on subjects under two treatment groups:

  $D = \{y_1, y_2, y_3\} \bigcup \{y_4, y_5\},$

  where $y_1, y_2, y_3 \sim f(y;\theta_0); \qquad y_4, y_5 \sim f(y;\theta_1).$

- The likelihood function $L(\theta_0,\theta_1) = \prod_{i=1}^{3} f(y_i;\theta_0) \prod_{i=4}^{5} f(y_i;\theta_1).$

- Under the null $\mathrm{H}_0 : \theta_0 = \theta_1$, $L(\theta_0,\theta_1) = L(\theta_0) = \prod_{i=1}^{5} f(y_i;\theta_0).$

- Then for an arbitrary permutation on D that keeps the same amount of observations by treatment group,

  $\mathrm{D_k} = \{\tilde{y}_1, \tilde{y}_2, \tilde{y}_3\} \bigcup \{\tilde{y}_4, \tilde{y}_5\},$

  the corresponding likelihood function

$$L^*(\theta_0) = \prod_{i=1}^{5} f(\tilde{y}_i;\theta_0) = \prod_{i=1}^{5} f(y_i;\theta_0) = L(\theta_0)$$

A Simple Example (II)

- Therefore, the permutation sample space Λ includes all permutations on D that assigns three in group 0.
- The number of elements in Λ is

$$M=\binom{5}{3}=\frac{5!}{3!2!}=10.$$

- The permutations can be easily exhausted.

- Under H_0, each element of Λ, a permutation of D, has equal chance to be observed.

Permutation test in practice

- Suppose the observed data are $D=\{x_1,x_2,...,x_n\}$, where $x_i \sim F(x;\theta)$.
- $T_D = T(D)$ is a summary statistic.
- Under the null $H_0 : \theta = \theta_0$, the permutation sample space or a random sample of it is $\Lambda = \{D_1, D_2, D_3, ..., D_M\}$. We can generate summary statistics

  $T_m = T(D_m)$, m=1,2,...,M.
- Because under H_0, each D_m has equal chance to be realized as D, $\{T_1, T_2, ..., T_M\}$ form a random sample of T_D.
- $p=\frac{\#\{m : T_m \succ T_D\}}{M}$ is the p-value under this permutation test, where $T_m \succ T_D$ implies that T_m is equally or more extreme than T_D under H_0.

A simple example (III)

Permutation π	Observation 8.6	7.2	5.6	6.0	4.8	Mean Difference (d_π)	$\lvert d_\pi \rvert \geq 1.733$ Yes (Y) or no (N)
1*	2	2	2	1	1	1.733*	Y*
2	2	2	1	2	1	2.067	Y
3	2	2	1	1	2	1.067	N
4	2	1	2	2	1	0.733	N
5	2	1	2	1	2	-0.267	N
6	1	2	2	2	1	-0.433	N
7	1	2	2	1	2	-1.433	N
8	1	2	1	2	2	-1.100	N
9	1	1	2	2	2	-2.433	Y
10	2	1	1	2	2	0.067	N

Exchangeability

- Some observations are exchangeable no matter what (e.g. experimental replicates; observations within the same group). Permutation of these exchangeable observations will not change the findings.
- Some observations are exchangeable only when H_0 is true.
- In a permutation test, one needs to identify the observations that would be exchangeable only when H_0 is true.

A simple example (III)

Permutation π	Observation 8.6	7.2	5.6	6.0	4.8	Mean Difference (d_π)	$\lvert d_\pi \rvert \geq 1.733$ Yes (Y) or no (N)
1*	2	2	2	1	1	1.733*	Y*
2	2	2	1	2	1	2.067	Y
3	2	2	1	1	2	1.067	N
4	2	1	2	2	1	0.733	N
5	2	1	2	1	2	-0.267	N
6	1	2	2	2	1	-0.433	N
7	1	2	2	1	2	-1.433	N
8	1	2	1	2	2	-1.100	N
9	1	1	2	2	2	-2.433	Y
10	2	1	1	2	2	0.067	N

- Total number of permutations under $H_0=(n_1+n_2)!$.
- Number of permutations among the observations that are exchangeable even when H_0 does not hold: $n_1!n_2!$.
- Number of additional permutations contributes by H_0: $A=(n_1+n_2)!/(n_1!n_2!)$.

Randomized block design

- 16 blocks
- Within each block, 6 patients are randomly assigned to 3 treatment groups in pairs.
- Within each block, each pair is exchangeable, thus there are 2!2!2!=8 permutations that will not change the findings.
- Under null, all 6 are exchangeable, thus there are a total of 6!=720 permutations.
- Total number of additional permutation contributed by H_0=720/8=90.
- The 16 blocks are not exchangeable due to block effect.
- Thus the total number of permutations: $A=90^{16}$, a huge number.
- A random sample of all possible permutations would be selected to estimate the p-value.

Pooling Test Results Across Genes

- Minimum p-value=1/A for a permutation test.
- To improve power in microarray analysis, results across genes are pooled together.
- Assumption: under H_0, the permutation statistics for all permutations and all genes are drawn independently from a common null distribution.

Permutation Procedure for SAM

SAM statistic: $d_i = \dfrac{r_i}{s_i + s_0}; i = 1,2,...,p$

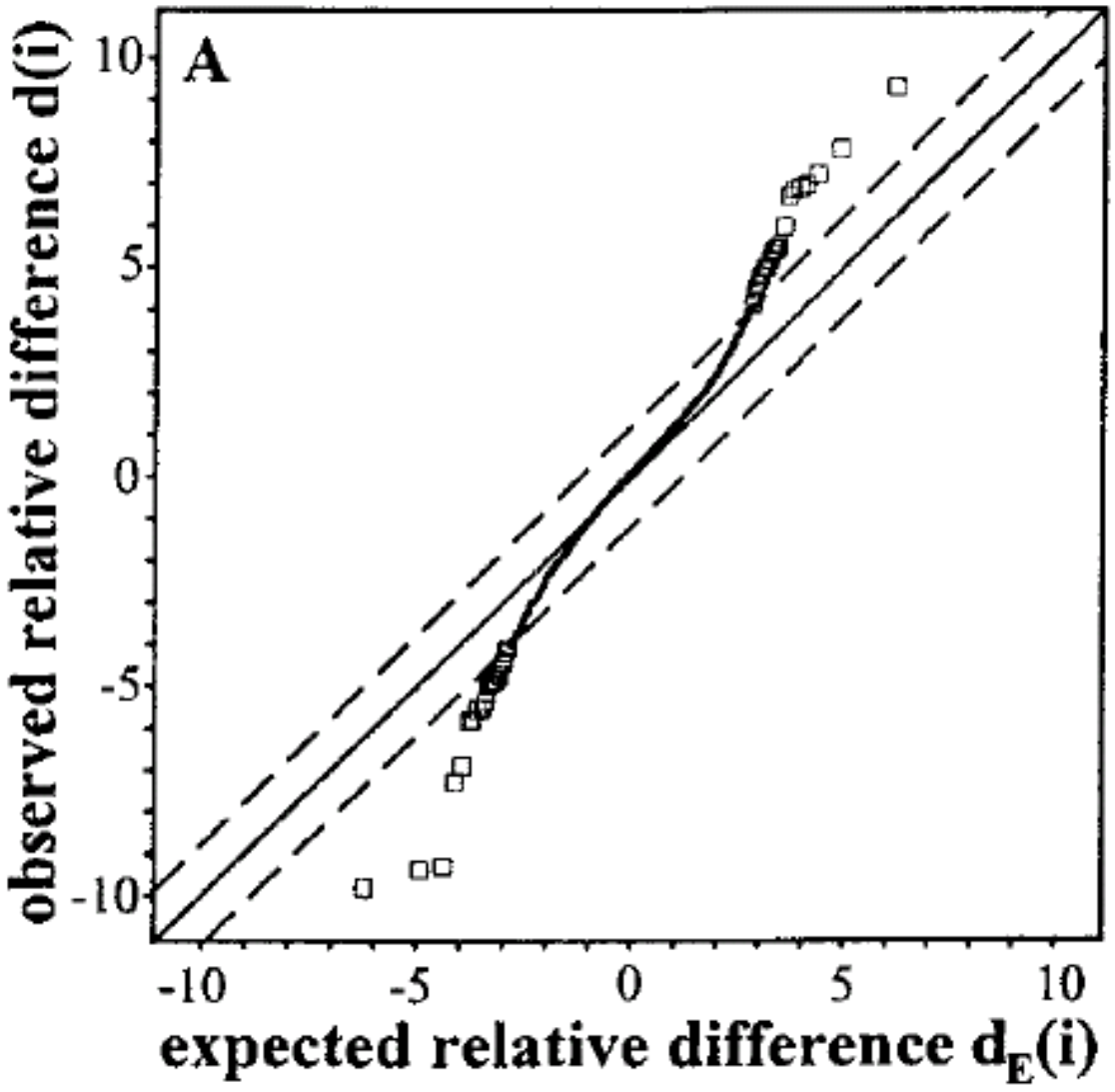
Compute order statistics $d_{(1)} \le d_{(2)} \le ... \le d_{(p)}$

Take B sets of permutation of the response values y_j.

For each permutation b compute the corrsponding order statistics $d^{*b}_{(1)} \le d^{*b}_{(2)} \le ... \le d^{*b}_{(p)}$.

Calculate $\bar{d}_{(i)} = (1/B)\sum_b d^{*b}_{(i)}$.

Permutation Procedure for SAM



Estimate π_0, the proportion of true null

(a) compute $q25$ and $q75$ of the permuted d values

(there are pB such values, $p = \#$ of genes, $B = \#$ of permutations).

(b) compute $\hat{\pi}_0 = \#\{d_i \in (q25, q75)\} / (0.5\,p)$

(d_i are from the original dataset).

(c) $\hat{\pi}_0 = \min(\hat{\pi}_0, 1)$.

For a grid of Δ values, compute the total number of significant genes (from the original dataset) and the median number of falsely called genes (from the B permutations).

$$FDR = \frac{\text{median number of falsely called genes} \times \hat{\pi}_0}{\text{number of genes called significant in the dataset}}.$$

Permutation Procedure in Pathway Analysis

- Pathway analysis: the overall objective is to test whether a group of genes has a coordinated association with a phenotype of interest. (cited from Tian et al. 2005)
- Tian, L., et al., *Discovering statistically significant pathways in expression profiling studies.* Proc Natl Acad Sci U S A, 2005. **102**(38): p. 13544-9.

Two ways to formulate the null

- *Q1: The genes in a gene set show the same pattern* of associations with the phenotype compared with the rest of the genes.

- *Q2: The gene set does not contain any genes* whose expression levels are associated with the phenotype of interest.

- Two different permutation procedure need to take place for these two hypotheses.

The Data

i : index of genes (i=1...B)

j : index of samples (j=1...n)

$\{z_1, ..., z_n\}$ are phenotypes of the subjects (e.g. cancer vs. normal)

$$G_{ki} = \begin{cases} 1 \text{ if the k}^{th} \text{ gene set contains gene i} \\ 0 \text{ if not.} \end{cases}$$

Data matrix

$$\begin{pmatrix} t_1 & t_2 & \cdots & t_B \\ G_{11} & G_{12} & \cdots & G_{1B} \\ \cdots & \cdots & \cdots & \cdots \\ G_{K1} & G_{K2} & \cdots & G_{KB} \end{pmatrix}$$

Hypothesis 1

- *The genes in a gene set show the same pattern* of associations with the phenotype compared with the rest of the genes.
  - Test whether the observed association of genes in a gene set is a random sample from the background distribution of all observed associations.
  - Test statistic: $T_k = \frac{1}{m_k} \sum_{i=1}^{B} G_{ki} t_i$
  - The null distribution of T_k can be obtained by permuting the genes $\{t_1, \ldots, t_B\}$.

Hypothesis 2

- *The gene set does not contain any genes* whose expression levels are associated with the phenotype of interest.
  - Based on the expression levels of genes within the pathway.
  - Test statistic: $E_k = \frac{1}{m_k} \sum_{i=1}^{B} G_{ki} t_i$ has the exact form of the previous statistic.
  - However, the null distribution of E_k is be obtained by permuting the phenotype vector $\{z_1, \ldots, z_n\}$