

**Multiple testing for gene expression data:  
an investigation of null distributions with  
consequences for the permutation test**

**Katie Pollard & Mark van der Laan**

**Division of Biostatistics, U.C. Berkeley**

**[www.stat.berkeley.edu/~laan](http://www.stat.berkeley.edu/~laan)**

## Gene Expression Data

- Observe  $n$  copies of a  $p$ -dimensional vector  $X \sim P$  of gene expression measurements, plus possibly covariates (e.g.: sequence data) and outcomes (e.g.: survival), which can be censored.
- Each gene expression measurement is a ratio, calculated from the intensities of two fluorescently labeled mRNA samples hybridized to an array spotted with known cDNA sequences.
- Data preprocessing may include background subtraction, normalization, log transformation.

Example: Gene expression in tumors vs. healthy tissues of  $n$  cancer patients.

## Multiple Testing Framework

**Parameters:**  $\mu_j(P) \in \mathbb{R}$ ,  $j = 1, \dots, p$ .

**Null Hypotheses:**  $H_{0,j} : \mu_j(P) = \mu_j^0$ ,  $j = 1, \dots, p$ , where  $\mu_j^0$  are hypothesized null values, frequently zero.

**Test Statistics:** Test  $H_{0,j}$ ,  $j = 1, \dots, p$ , with  $T_{jn}$  defined by

$$T_{jn} \equiv \mu_{jn} - \mu_j^0$$

or  $T_{jn} \equiv \frac{\mu_{jn} - \mu_j^0}{\hat{\sigma}(\mu_{jn})}$ .

**Multiplicity Problem:** thousands of hypotheses tested simultaneously, increased chance of false positives.

## Error Control

Given a vector  $c \in \mathbb{R}^p$ , a **multiple testing procedure**  $MT(c)$  is defined by:

Reject  $H_{0,j}$ , if  $|T_{jn}| > c_j$ ,  $j = 1, \dots, p$ .

The cut-off values  $c_j$  are chosen using a test statistic **null distribution** so that the error rate is  $\leq \alpha$  (target level).

Type I **error rates** which are functions of the distribution  $F_{V_n}$  of the number of false positives of  $MT(c)$ :

- $\int x dF_{V_n}(x)/p = E(V_n)/p$  : per-comparison error rate (**PCER**),
- $\int x dF_{V_n}(x) = E(V_n)$  : per-family error rate (**PFER**),
- $1 - F_{V_n}(k - 1) = Pr(V_n \geq k)$  : family-wise error rate (**FWER**).

# Null Distributions

The correct null distribution is the **projection** of the true test statistic distribution onto the space of mean zero distributions.

Estimated by:

1. Tabled Distributions
  - Ignore correlation
  - Same marginal distribution for all genes
2. Permutations
  - Assumes equality of distributions, not just the parameter(s) of interest
3. Null Restricted Bootstrap
  - Defined via a data null distribution
  - Only correct if the test statistic null distribution does not depend on the parameter(s)
4. Ordinary Bootstrap

# Ordinary Bootstrap Estimator

1. Let
  - $\tilde{P}_n$  be an estimator of  $P$ .
  - $\tilde{\mu}_n = \mu(\tilde{P}_n)$  be the parameter estimate under  $\tilde{P}_n$ .
  - $\mu_n^\#$  be  $\mu_n$  applied to  $n$  i.i.d. copies  $X_1^\#, \dots, X_n^\#$  of  $X^\# \sim \tilde{P}_n$ .
2. Use the distribution of  $Z_n^\# = \sqrt{n}(\mu_n^\# - \tilde{\mu}_n)$ , estimated from a large number  $B$  of resamples from  $\tilde{P}_n$ .
3. Choose cut-offs  $c_n$  so that the error rate is  $\alpha$  under this test statistic null distribution.
  - e.g.: use a common quantile of each marginal distribution.
  - The vector  $c_n$  can be fine tuned so that the error rate is exactly  $\alpha$  up to the discreteness of the resampled distribution.

Then,  $MT(c_n)$  is a **bootstrap based multiple testing procedure** controlling the error rate at level  $\alpha$ .

# Applications to the Two Sample Problem

Suppose we have  $n_1$  observations from Population 1 with mean  $\mu_1$  and  $n_2 = n - n_1$  observations from Population 2 with mean  $\mu_2$ .

1. Testing for a difference in means:  $\mu(P) = \mu_2 - \mu_1$

- **Null Hypotheses:**  $H_{0,j} : \mu_j = \mu_{2,j} - \mu_{1,j} = 0, j = 1, \dots, p.$
- **Test Statistics:**

$$\bar{X}_{2,j} - \bar{X}_{1,j}, j = 1, \dots, p$$

$$\text{or } \frac{\bar{X}_{2,j} - \bar{X}_{1,j}}{\sqrt{\hat{\sigma}_{1,j}^2/n_1 + \hat{\sigma}_{2,j}^2/n_2}}, j = 1, \dots, p.$$

2. Testing for an association with group:  $\mu(P) =$  regression coeffs.

- Logistic Regression Model for each gene:  $j = 1, \dots, p$

$$E(\text{group} \mid X_j) = \frac{e^{\beta_{0,j} + \beta_{1,j} * X_j}}{1 + e^{\beta_{0,j} + \beta_{1,j} * X_j}}$$

- **Null hypotheses:**  $H_{0,j} : \beta_{1,j} = 0, j = 1, \dots, p.$
- **Test statistics:**  $\sqrt{n} * \hat{\beta}_{1,j}, j = 1, \dots, p.$

# Comparison of Null Distributions: Variance and Covariance

Let  $COV(X_j, X_{j'})$  be  $\phi_1$  in population 1 and  $\phi_2$  in population 2. Consider  $T_{jn} = \bar{X}_{2,j} - \bar{X}_{1,j}$ .

| Distribution | $Var(T_{jn})$   | $Cov(T_{jn}, T_{j'n})$                    |
|--------------|---|---|
| Permutations | $\frac{\sigma_{1,j}^2}{n_2} + \frac{\sigma_{2,j}^2}{n_1}$ | $\frac{\phi_1}{n_2} + \frac{\phi_2}{n_1}$ |
| Bootstrap    | $\frac{\sigma_{1,j}^2}{n_1} + \frac{\sigma_{2,j}^2}{n_2}$ | $\frac{\phi_1}{n_1} + \frac{\phi_2}{n_2}$ |

Note: If  $T_{jn} = \frac{\bar{X}_{2,j} - \bar{X}_{1,j}}{\sqrt{\hat{\sigma}_{1,j}^2/n_1 + \hat{\sigma}_{2,j}^2/n_2}}$ ,

- $VAR(T_{jn}) = 1$  for both distributions.
- $COV(T_{jn}, T_{j'n})$  is not equivalent unless  $n_1 = n_2$ .

# Equivalence of Multiple Testing and Confidence Regions

Given an estimate  $Q_{0n}$  (e.g.: ordinary bootstrap) of the test statistic null distribution, choose the vector  $c_n$  such that the error rate  $\theta(\cdot)$  is  $\alpha$  under  $Q_{0n}$ . Then,

- $\{\mu : \sqrt{n}(\mu_n - \mu) < c_n\}$  is an asymptotically correct  $\theta$ -specific  $(1 - \alpha)\%$  confidence region for  $\mu(P)$ .
- If  $\theta(\cdot)$  is the FWER, then this region is a  $(1 - \alpha)\%$  *simultaneous* confidence region for  $\mu(P)$ .

The equivalent multiple testing procedure is  $MT(c_n)$ :

Reject  $H_{0,j}$  if  $\mu_j^0$  is outside the interval  $\left[ \mu_{jn} - \frac{c_n}{\sqrt{n}}, \mu_{jn} + \frac{c_n}{\sqrt{n}} \right]$ .

- *i.e.*: construct an error-specific confidence region for the true parameter and check if it contains the hypothesized value.
- This equivalence holds for *any* type I error rate  $\theta$ .

## Conclusions

1. The ordinary bootstrap method provides a test statistic null distribution which asymptotically controls a type I error rate under weak conditions for any data generating distribution  $P$ .
2. Common practice of using a data null distribution (e.g: null restricted bootstrap) only provides asymptotic control if the null distribution does not depend on the parameters of interest.
3. Two Sample Problem: The permutation distribution has the wrong covariance unless  $n_1 = n_2$  or  $\Sigma_1 = \Sigma_2$ .
4. By constructing an error-specific ordinary bootstrap confidence set and checking if it contains the hypothesized  $\mu^0$ , it is possible to do multiple testing without explicitly knowing the test statistic null distribution.