- N=102 microarrays and M=6033 gene expression measurements from each

<table>
<thead>
<tr>
<th>m</th>
<th>control group</th>
<th>cancer group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>-0.931 -0.840 ... 3.81</td>
<td>-1.12 1.01 ... -0.001</td>
</tr>
<tr>
<td></td>
<td>-1.07 -0.880 ... -0.477</td>
<td>-0.571 -0.811 ... -0.836</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>-0.754 -0.708 ... -0.011</td>
<td>0.457 0.578 ... -0.162</td>
</tr>
<tr>
<td>6033</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Which genes are related to cancer (differentially expressed)?
- Need to test 6033 null hypotheses!
**Strategy**

DATA

\[ \downarrow \]

\( p \)-Values (6033 of them)

\[ \downarrow \]

Multiple Testing Procedure

\[ \downarrow \]

Decisions (6033 of them)
DATA

\[ p \text{-Values (6033 of them)} \]

\[ \Downarrow \]

Multiple Testing Procedure

\[ \Downarrow \]

Decisions (6033 of them)
Properties of $p$-Values

Multiple testing procedures valid if $p$-values from null hypotheses are

1. uniformly distributed and
2. independent
How to get a $p$-value for testing $H_0$ with data $X$

**In words:**
The $p$-value is “the smallest size allowing for the null to be rejected with the observed data $x$”

**Mathematically:**
1. Define decision function $\delta(x; \eta) \in \{0, 1\}$
   - $\eta$ is size and $X$ is data
2. Definition: $p$-Value is $P(x) = \inf\{\eta : \delta(x; \eta) = 1\}$
We can work with $\delta(X; \eta)$ and use the previous definition to get a $P(X)$.

\[
\begin{bmatrix}
X_{11} & X_{12} & \ldots & X_{1N} \\
X_{21} & X_{22} & \ldots & X_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
X_{M1} & X_{M2} & \ldots & X_{MN}
\end{bmatrix}
\rightarrow
\begin{bmatrix}
X_1 \\
X_2 \\
\vdots \\
X_M
\end{bmatrix}
\]

\[\rightarrow\] means test stat applied to each row and transformed to get \(H_{0m} : X_m \sim N(0, 1)\)

- Ex. Compute \(T_m = T(X_{m1}, X_{m2}, \ldots, X_{mN})\) and take \(X_m = \Phi^{-1}(F_T(T_m))\)
Basic Example

- \( H_{0m} : X_m \sim N(0, 1) \) vs. \( H_{1m} : X_m \sim N(\mu_m, 1), \mu_m \neq 0 \)

\[
\delta_m(x_m; \eta_m) = \begin{cases} 
1 & \text{if } x_m > \Phi^{-1}(1 - \eta_m/2) \\
1 & \text{if } x_m < \Phi^{-1}(\eta_m/2) \\
0 & \text{otherwise}
\end{cases}
\]
Example

- Upper cutoff: $\Phi^{-1}(1 - \frac{1}{2}) = 1.645$
- Lower Cutoff: $\Phi^{-1}(\frac{1}{2}) = -1.645$
- Tail area is $\eta_m = .1!$
Example

- Upper cutoff: $\Phi^{-1}(1 - .1(0.75)) = 1.44$
- Lower Cutoff: $\Phi^{-1}(0.1(0.25)) = -1.96$
- Tail area is $\eta_m = .1$!
Example

Upper cutoff: $\Phi^{-1}(1 - .1(.9)) = 1.34$

Lower Cutoff: $\Phi^{-1}(.1(.1)) = -2.32$

Tail area is $\eta_m = .1!$
Example

- Upper cutoff: $\Phi^{-1}(1 - .1(1)) = 1.28$
- Lower Cutoff: $\Phi^{-1}(.1(0)) = -\infty$
- Tail area is $\eta_m = .1$!
For any $h_m$ in $[0,1]$

$$
\delta_m(x_m; \eta_m) = \begin{cases} 
1 & \text{if } x_m > \Phi^{-1}(1 - \eta_m(1 - h_m)) \\
1 & \text{if } x_m < \Phi^{-1}(\eta_m h_m) \\
0 & \text{otherwise}
\end{cases}
$$

Optimal: $h_m(\mu_m) = I(\mu_m < 0)$

BUT WE DON'T KNOW $\mu_m$!!!
\[
\begin{bmatrix}
X_{11} & X_{12} & \ldots & X_{1N} \\
X_{21} & X_{22} & \ldots & X_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
X_{M1} & X_{M2} & \ldots & X_{MN}
\end{bmatrix}
\rightarrow
\begin{bmatrix}
Y_1 & Z_1 \\
Y_2 & Z_2 \\
\vdots & \vdots \\
Y_M & Z_M
\end{bmatrix}
\]

- \(\delta_1(Y_1, Y_2, \ldots, Y_M, Z_1; \eta_1)\)
- Plug \(h_1(Y_1, Y_2, \ldots, Y_M) = I(\mu_1 < 0)\)
- \(Z_1\) is test data (rather than \(X_1\)
\[
\begin{bmatrix}
X_{11} & X_{12} & \ldots & X_{1N} \\
X_{21} & X_{22} & \ldots & X_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
X_{M1} & X_{M2} & \ldots & X_{MN}
\end{bmatrix}
\rightarrow
\begin{bmatrix}
Y_1 & Z_1 \\
Y_2 & Z_2 \\
\vdots & \vdots \\
Y_M & Z_M
\end{bmatrix}
\]

- \( \delta_2(Y_1, Y_2, \ldots, Y_M, Z_2; \eta_2) \)
- Plug \( h_2(Y_1, Y_2, \ldots, Y_M) = I(\mu_2 < 0) \)
- \( Z_2 \) is test data (rather than \( X_2 \))
\[
\begin{bmatrix}
X_{11} & X_{12} & \ldots & X_{1N} \\
X_{21} & X_{22} & \ldots & X_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
X_{M1} & X_{M2} & \ldots & X_{MN}
\end{bmatrix}
\rightarrow
\begin{bmatrix}
Y_1 & Z_1 \\
Y_2 & Z_2 \\
\vdots & \vdots \\
Y_M & Z_M
\end{bmatrix}
\]

- \( \delta_M(Y_1, Y_2, \ldots, Y_M, Z_M; \eta_1) \)
- Plug \( h_M(Y_1, Y_2, \ldots, Y_M) = I(\mu_M < 0) \)
- \( Z_M \) is test data (rather than \( X_M \))
The compound $p$-value:
- Estimate $I(\mu_m < 0)$ via $h_m(Y)$, where $Y = (Y_1, Y_2, \ldots, Y_M)$

$$P(Y, Z_m) = \min \left\{ \frac{\Phi(Z_m)}{h_m(Y)}, \frac{1 - \Phi(Z_m)}{1 - h_m(Y)} \right\}$$

Simple $p$-value:

$$P(X_m) = 2[1 - \Phi(|X_m|)] = \left\{ \frac{\Phi(X_m)}{1/2}, \frac{1 - \Phi(X_m)}{1 - 1/2} \right\}$$

The simple $p$-value uses all the data ($X_m = Y_m + Z_m$) as test data but uses $h_m = 1/2$
Applied Q-value and BH FDR-controlling procedures to different \( p \)-values.

- **Simple**: \( \times \) - take \( h_m = 0.5 \), use all data as test data
- **Compound**: \( \times, o, \triangle \) - use 4 microarrays to estimate \( h_m \)
Analytical Results

Theorem (simplified): If test data are independent and “correctly modeled” under nulls, these compound p-values are independent and uniformly distributed under nulls.

- “Correctly Modeled” - use $T$-test and $\Phi^{-1}(F_T(T_m))$ if data are normal to get $Z_m$s, use nonparametric test otherwise. See Habiger and Pena (2011), JNS.
- Implies many multiple testing procedures will be valid
We have *borrowed information across tests* to refine our upper/lower tailed cutoffs (via $h_m(Y)$)

- **Gain in power related to “similarity” of data under $H_{1m}$s**
  - *Theorem:* If $\mu_1 = \mu_2 = \ldots$ under $H_{1m}$s, $h_m(Y) \to I(\mu_m < 0)$ as $M \to \infty$
  - As $\mu_m$’s become dispersed, less power is gained.
**Compared to Other Approaches**

*Other Approaches: Efron (2001), JASA; Sun and Cai (2007), JASA, etc.*

1. **We allow each $\delta_m$ to use different cutoffs depending on $f_1(z|\mu_m)$ under $H_{m1}$**
   - vs. using same cutoffs based on $f_1(z) = \int f_1(z|\mu)\pi(\mu)d\mu$
   - Is this better?

2. **We split sample. Others double dip.**
   - efficiency vs. validity

3. **We utilize the $p$-value**