Multiple testing:
False discovery rate and the q-value

Stefanie Scheid

Max-Planck-Institute for Molecular Genetics
Computational Diagnostics

June 19, 2002
Articles

- Storey, J.D. (2001a): The positive False Discovery Rate: A Bayesian Interpretation and the q-value, submitted


The multiple testing problem

- random variables $X_1, \ldots, X_k$ from same family of distribution $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$
- set of $m$ hypotheses for $\theta$
- conclude from one sample
Why not testing separately?

$m$ independent test statistics with level $\alpha$
Why not testing separately?

$m$ independent test statistics with level $\alpha$

\[ \Rightarrow \text{Prob(at least 1 is falsely rejected)} = 1 - (1 - \alpha)^m = \alpha_m \]
Why not testing separately?

\( m \) independent test statistics with level \( \alpha \)

\[ \Rightarrow \text{Prob(at least 1 is falsely rejected)} = 1 - (1 - \alpha)^m = \alpha_m \]

\[ \Rightarrow \text{control } \alpha_m \text{ with multiple test procedure} \]
Two kinds of multiple testing procedures: One-step

- test each null hypothesis “independently” from outcome of others
- e.g. Bonferroni: test each hypothesis to level $\frac{\alpha}{m}$

$\Rightarrow$ multiple level $= \alpha$
Two kinds of multiple testing procedures: Multi-step

- test in sorted order dependent on outcome

- e.g. Bonferroni-Holm: sort according to p-values and test with increasing $\alpha$:
  \[ \frac{\alpha}{m}, \frac{\alpha}{m-1}, \ldots, \alpha \]

$\Rightarrow$ multiple level $= \alpha$
Microarray data

- $k$ samples (e.g. in 2 groups) and $m$ genes
Microarray data

- \( k \) samples (e.g. in 2 groups) and \( m \) genes

- observe gene expression as intensity values:

<table>
<thead>
<tr>
<th></th>
<th>sample 1</th>
<th>sample 2</th>
<th>\ldots</th>
<th>sample ( k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene 1</td>
<td>404</td>
<td>1873</td>
<td>\ldots</td>
<td>151</td>
</tr>
<tr>
<td>gene 2</td>
<td>-2015</td>
<td>-716</td>
<td>\ldots</td>
<td>1227</td>
</tr>
<tr>
<td></td>
<td>\vdots</td>
<td>\vdots</td>
<td>\ddots</td>
<td>\ddots</td>
</tr>
<tr>
<td>gene ( m )</td>
<td>126</td>
<td>42</td>
<td>\ldots</td>
<td>85</td>
</tr>
</tbody>
</table>

Stefanie Scheid - False discovery rate - June 19, 2002
### Number of outcomes

<table>
<thead>
<tr>
<th></th>
<th>Not rejected</th>
<th>Rejected</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null true</td>
<td>$U$</td>
<td>$V$</td>
<td>$m_0$</td>
</tr>
<tr>
<td>Alternative true</td>
<td>$T$</td>
<td>$S$</td>
<td>$m_1$</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>$W$</td>
<td>$R$</td>
<td>$m$</td>
</tr>
</tbody>
</table>
Problems

• multiple test controls $\text{Prob}(V \geq 1) \leq \alpha$
Problems

• multiple test controls \( \text{Prob}(V \geq 1) \leq \alpha \)

• \( m \) is huge \( \Rightarrow \) falsely rejected are likely to occur
Problems

• multiple test controls $\text{Prob}(V \geq 1) \leq \alpha$

• $m$ is huge $\Rightarrow$ falsely rejected are likely to occur

• better control $\frac{|\{\text{falsely rejected}\}|}{|\{\text{rejected in total}\}|}$
Problems

- multiple test controls $\Pr(V \geq 1) \leq \alpha$

- $m$ is huge $\Rightarrow$ falsely rejected are likely to occur

- better control $\frac{\#\{\text{falsely rejected}\}}{\#\{\text{rejected in total}\}}$

- intuitive definition of false discovery rate:

$$FDR = \mathbb{E}\left[\frac{V}{R}\right]$$
Difference between multiple testing and FDR

• **Multiple testing:**
  Fixed error rate $\Rightarrow$ estimated rejection area

• **FDR:**
  Fixed rejection area $\Rightarrow$ estimated error rate
What if $R = 0$?

- Benjamini and Hochberg: $FDR = E \left[ \frac{V}{R} \mid R > 0 \right] \cdot \text{Prob}(R > 0)$

“the rate that false discoveries occur”
What if $R = 0$?

- Benjamini and Hochberg: $FDR = E \left( \frac{V}{R} \mid R > 0 \right) \cdot \text{Prob}(R > 0)$
  
  "the rate that false discoveries occur"

- Storey: $pFDR = E \left( \frac{V}{R} \mid R > 0 \right)$
  
  "the rate that discoveries are false"
FDR in Bayesian terms

**Theorem:** \( m \) identical hypothesis tests are performed with independent statistics \( T_1, \ldots, T_m \) and rejection area \( C \). A null hypothesis is true with a-priori probability \( \pi_0 = \text{Prob}(H = 0) \). Then

\[
pFDR(C) = \frac{\pi_0 \cdot \text{Prob}(T \in C \mid H = 0)}{\text{Prob}(T \in C)} = \text{Prob}(H = 0 \mid T \in C).
\]

Algorithms for calculating \( \widehat{FDR} \) and \( \widehat{pFDR} \) in Storey (2001b).
Simulation study

1. independence between genes
2. loose ("clumpy") dependence
3. general dependence
Simulation study

1. independence between genes

2. loose ("clumpy") dependence

3. general dependence

1. + 2. \( \hat{pFDR} \) is very accurate

3. \( \hat{pFDR} \) is biased upward (overestimation of \( \pi_0 \))
Reasons for “clumpy dependence” in microarrays

- genes work in pathways
  ⇒ small groups of genes interact to produce overall process
Reasons for “clumpy dependence” in microarrays

• genes work in pathways
  ⇒ small groups of genes interact to produce overall process

• cross-hybridization
  ⇒ genes with molecular similarity have evolutionary and/or functional relationship
p-value vs. q-value

• Definition: For a nested set of rejection areas \(\{C\}\) define the \(p\)-value of an observed statistic \(T = t\) to be:

\[
p-value(t) = \min_{\{C : t \in C\}} \text{Prob}(T \in C | H = 0).
\]
p-value vs. q-value

**Definition:** For a nested set of rejection areas \( \{C\} \) define the *p-value* of an observed statistic \( T = t \) to be:

\[
p-value(t) = \min_{\{C : t \in C\}} \text{Prob}(T \in C \mid H = 0).
\]

**Definition:** For an observed statistic \( T = t \) define the *q-value* of \( t \) to be:

\[
q-value(t) = \min_{\{C : t \in C\}} \text{pFDR}(C) = \min_{\{C : t \in C\}} \text{Prob}(H = 0 \mid T \in C).
\]

“posterior Bayesian p-value”
Conclusion

• FDRs are more appropriate in large sets of hypotheses than multiple testing procedures
Conclusion

• FDRs are more appropriate in large sets of hypotheses than multiple testing procedures

• one has to be sure about rejection area
Conclusion

- FDRs are more appropriate in large sets of hypotheses than multiple testing procedures.
- One has to be sure about rejection area.
- Positive FDR ("rate that discoveries are false") is more appropriate than FDR ("rate that false discoveries occur").
Conclusion

• FDRs are more appropriate in large sets of hypotheses than multiple testing procedures

• one has to be sure about rejection area

• positive FDR ("rate that discoveries are false") is more appropriate than FDR ("rate that false discoveries occur")

• q-value can be reported with every statistic ("minimum pFDR over which that statistic can be rejected")