Separating the Wheat from the Chaff: Statistical Methods for False Discovery Control

Radu Craiu
University of Toronto

joint work with

Shelley Bull, Andrew Paterson and Lei Sun
University of Toronto

London, May 30, 2006

Contents

- Multiple Comparisons.
- False Discovery Rate (FDR) and Non-Discovery Rate (NDR).
- Stratified False Discovery Control.
- “ROC” - based comparisons.

Multiple Comparisons

➠ Summary of events for multiple hypothesis testing:

	Declared non-significant	Declared significant	Total counts
Truth: H_0	U (= True Negatives)	V (= FP/type I errors)	m_0
Truth: H_1	T (= FN/type II errors)	S (= True Positives)	m_1
Total	$m - R$	R	m

- Observed: m, R.
- Unobserved: m_0, m_1, U, V, T and S.

Measures of Type I Error Rate

- Family-Wise Error Rate: FWER $= \Pr(V \geq 1)$.
  - Stringent criterion: e.g. $\alpha \approx 10^{-5}$ required for genome-wide linkage analyses of complex diseases using an Affected Sib-Pair (ASP) design.
  - Diminished power: often few or no discoveries.
- False Discovery Rate: FDR $= \mathrm{E}[V/R]$
  (Benjamini and Hochberg, 1995).
  - Control FDR $\leq \alpha \Longrightarrow$ corresponding FWER $\geq \alpha$.
  - Alternative definitions:

$$\mathrm{FDR} = \mathrm{E}[\mathbf{V}/\mathbf{R}|\mathbf{R} > 0] \Pr(\mathbf{R} > 0) \text{ (BH, 1995)},$$

$$\mathrm{pFDR} = \mathrm{E}[\mathbf{V}/\mathbf{R}|\mathbf{R} > 0] \text{ (Storey, 2002)}.$$

  - In practice: $\Pr(\mathbf{R} > 0) \approx 1$ (Storey and Tibshirani, 2003).

Two Frameworks for FDR control

- Fixed FDR framework: pre-specify FDR at level γ, then find a rejection procedure that rejects as many tests as possible while control FDR at γ.
  - The FDR-adjusted p-value method (Yekutieli, Benjamini, 1999) and the q-value approach (Storey, 2002).
  - Control FDR at γ level $\Longleftrightarrow$ Reject all tests with q-values $\leq \gamma$.

$$p_{(1)} \leq \ldots \leq p_{(m)},$$

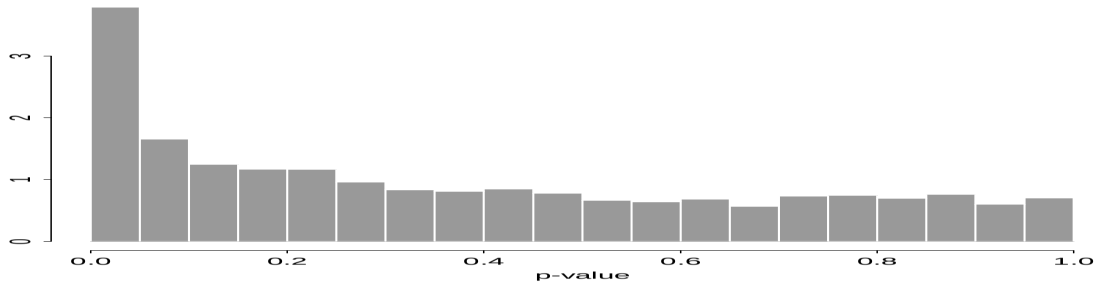
$$\hat{q}_{(m)} = \hat{\pi}_0 \, p_{(m)}, \; \hat{q}_{(i)} = \min \left\{ \frac{\hat{\pi}_0 \, m \, p_{(i)}}{i}, \hat{q}_{(i+1)} \right\}.$$

- Fixed rejection region framework: reject all tests with (unadjusted) p-values $\leq \alpha$ level (pre-specified), then estimate FDR among all positives.

$$\widehat{\text{FDR}}(\alpha) = min\left\{\frac{m\,\hat{\pi}_0\,\alpha}{\mathbf{R} = \{\#\,p_i \leq \alpha\}}, 1\right\}.$$

➠ An estimator for $\pi_0 = m_0/m$:

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{(1-\lambda)\,m}, \text{ with } \lambda = 1/2.$$



Motivating Example I

	Control FDR at 5%		Control FDR at 10%		
	Declared non-significant	Declared significant	Declared non-significant	Declared significant	Total
Truth: H_0	899	**1**	892	**8**	900
Truth: H_1	**81**	**19**	**28**	**72**	**100**
Total	980	**20**	920	**80**	1000

- Control FDR at 5%: miss 81 true signals and identify 19 true signals.
- Control FDR at 10%: miss 28 true signals and identify 72 true signals.

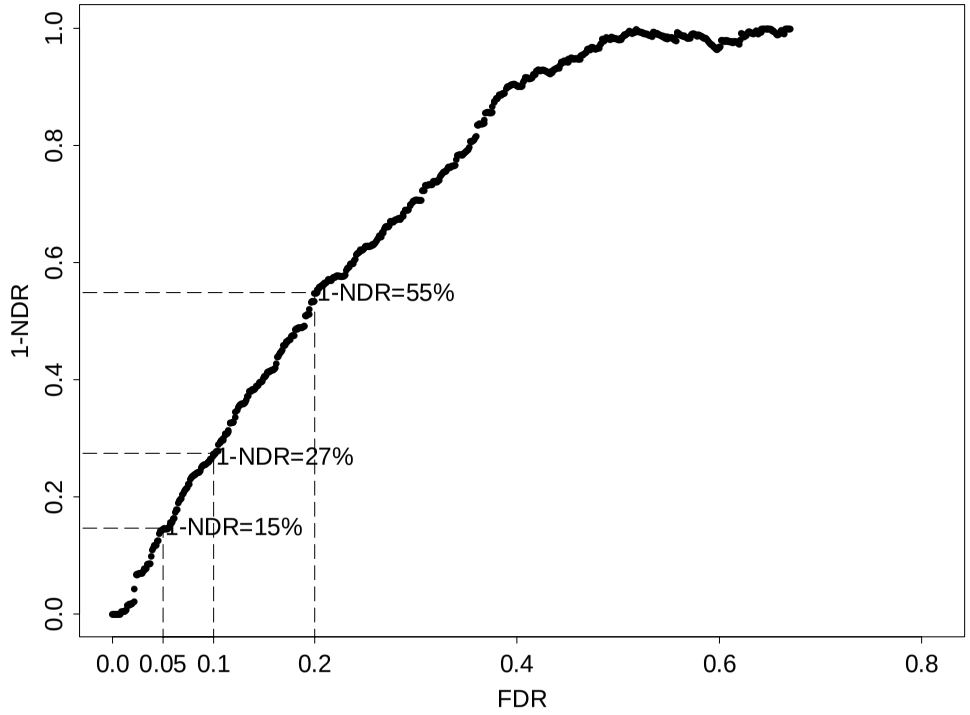
Which FDR level? Measures of type II error rate and power?

Non-Discovery Rate (NDR)

- Definition: $\text{NDR} = \text{E}[\mathbf{T}]/m_1 = 1 - \text{E}[\mathbf{S}]/m_1$.
- Estimation: $\widehat{\text{NDR}} = 1 - \{\mathbf{R}\,(1 - \text{FDR})\}/\{m\,(1 - \hat{\pi}_0)\}$.
  - Accurate estimation of π_0.
- Interpretation: Fixed region framework with threshold α, $\text{NDR} = \overline{\beta(\alpha)}$; $1 - \text{NDR} = \overline{\text{Power}(\alpha)}$.
- Utility: trade-off between FDR and 1 - NDR.

Application - Microarray

- Storey and Tibshirani (2003): $m = 3,170$, $\hat{\pi}_0 = 0.67$.



Stratified False Discovery Control

- Inherent stratification in many genetics studies:
  - high priority markers selected from candidate genes or linkage regions vs. secondary markers included to cover the genome,
  - SNPs vs. microsatellites,
  - each marker tested for association with each of K phenotypes of interest,
  - tests conducted assuming K different genetic models,
  - ...
- Available auxiliary information/variable: stratum indicator.
  - Effective way to incorporate the auxiliary information?
  - Any gain by adjusting for multiple comparisons within stratum?

Motivating Example II

- In a GWA study, assume a map with 105K SNPs, and
  - 5K SNPs were selected from favored regions (stratum 1), among which 100 are truly associated,
  - 100K SNPs were chosen to cover the genome (stratum 2), among which 50 are truly associated.
- Fixed rejection: $\alpha = 0.001$, and $1 - \beta(\alpha) = 70\%$:

	Aggregated	Stratum 1	Stratum 2
m = # SNPs	105,000	5,000	100,000
m_1 = # associated SNPs	150	100	50
E[**V**] = E[# false positives]	105	5	100
E[**S**] = E[# true positives]	105	70	35
E[**R**] = E[# positives]	210	75	135
FDR = E[**V**/**R**]	50%	7%	74%

➠ Fixed FDR: $\gamma = 10\%$, and power follows a normal model, $1 - \beta(\alpha; \mu) = \Phi(\Phi^{-1}(\alpha) + \sqrt{n}\mu/\sigma)$ with $n = 100$,$\mu = 1.8$, $\sigma = 5$:

	Aggregated	Stratum 1	Stratum 2
m = # SNPs	105,000	5,000	100,000
m_1 = # associated SNPs	150	100	50
α used	0.00006	0.0016	0.000016
$1 - \beta(\alpha)$	40%	74%	29%
E[**V**] = E[# false positives]	6	8	1.6
E[**S**] = E[# true positives]	60	74	14.4
E[**R**] = E[# positives]	66	83	16

$$\mathrm{E}[\mathbf{S}] = \mathbf{60} < \mathbf{74 + 14.4} = \sum_k \mathrm{E}[\mathbf{S}^{(k)}]$$

Aggregation vs. Stratification

- Fixed rejection region: α fixed ($\mathbf{R} = \{\#\text{ p-value} \leq \alpha\}$).

$$\text{FDR} = \sum_k w^{(k)} \text{FDR}^{(k)}.$$

$$w^{(k)} = \text{E}[R^{(k)}] / \sum_j \text{E}[R^{(j)}],$$

  - If $\pi_0^{(k)} = \pi_0$, $1 - \overline{\beta(\alpha)}^{(k)} = 1 - \overline{\beta(\alpha)}$: $\text{FDR}^{(k)} = \text{FDR}$.
  - If $\pi_0^{(k)} < \pi_0$, $1 - \overline{\beta(\alpha)}^{(k)} > 1 - \overline{\beta(\alpha)}$: $\text{FDR}^{(k)} < \text{FDR}$.
  - If $\pi_0^{(k)} > \pi_0$, $1 - \overline{\beta(\alpha)}^{(k)} < 1 - \overline{\beta(\alpha)}$: $\text{FDR}^{(k)} > \text{FDR}$.

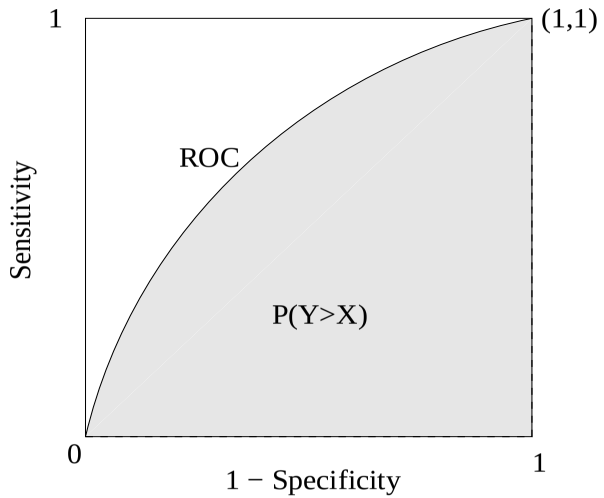
- Fixed FDR: γ fixed ($\text{E}[\mathbf{V}/\mathbf{R}] = \gamma$).

$$\text{E}[\mathbf{R}] \leq \sum_k \text{E}[\mathbf{R}^{(k)}].$$

ROC curves

- The traditional ROC curve is used for diagnostic accuracy
- X is the diagnostic tool measurement for controls (true null) and Y is the diagnostic tool measurement for cases (false null).
- Specificity: probability that a control is classified as normal.

  Sensitivity: probability that a case is classified as diseased.
- ROC plots Sensitivity vs 1 - Specificity.


1
(1,1)
ROC
Sensitivity
P(Y>X)
0
1
1 − Specificity

ROC-like comparison for FDC methods

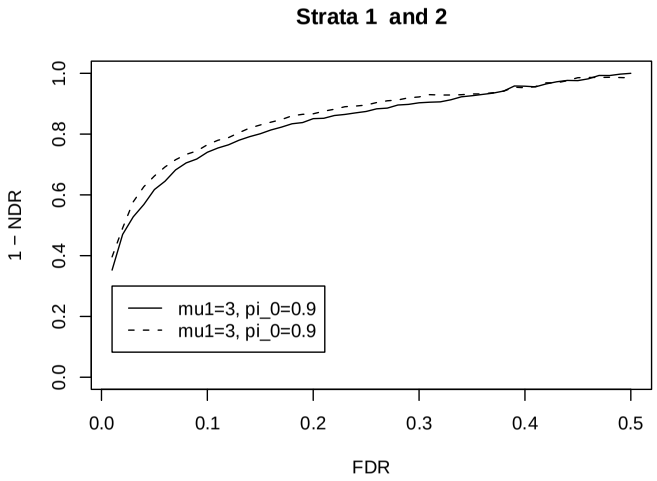
Stratified FDR and "classical" FDR can be compared as FDC procedures defined on the whole (unstratified) set of p-values.

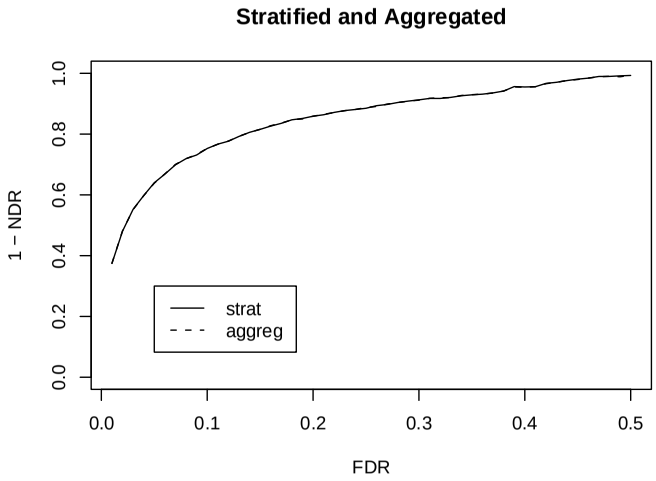
- 1-NDR = Sensitivity; FDR = 1 - Specificity.
- Simulation study of two strata each with different m, π_0 and signals of different strength.
- Aggregated NDR can be obtained in two ways:

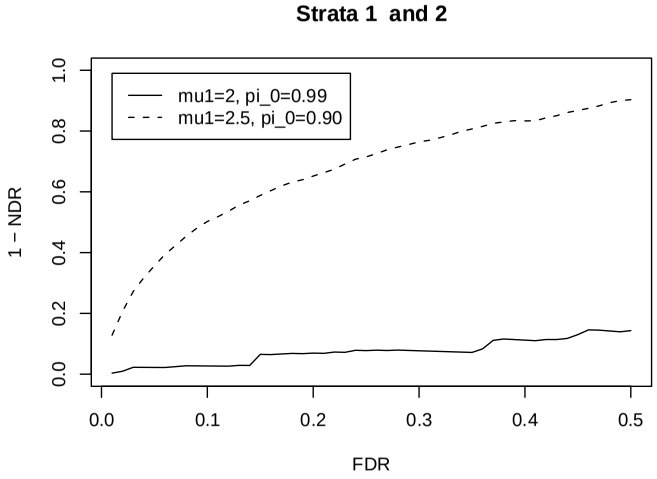
  1) Work with all the p-values (ignore stratification).

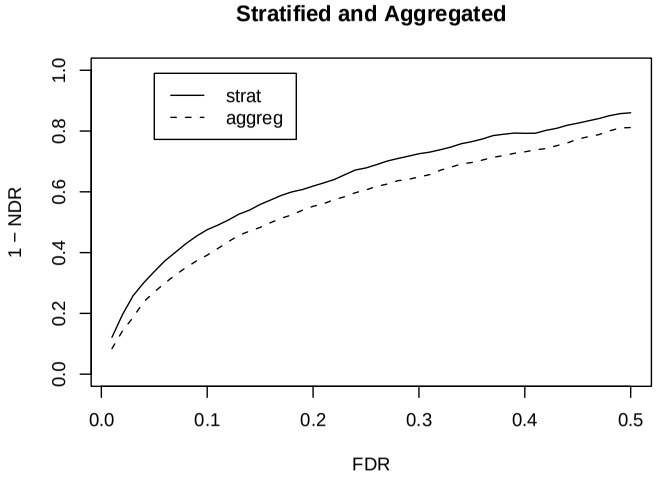
  2) Combine the strata-specific NDR's into a unified measure:

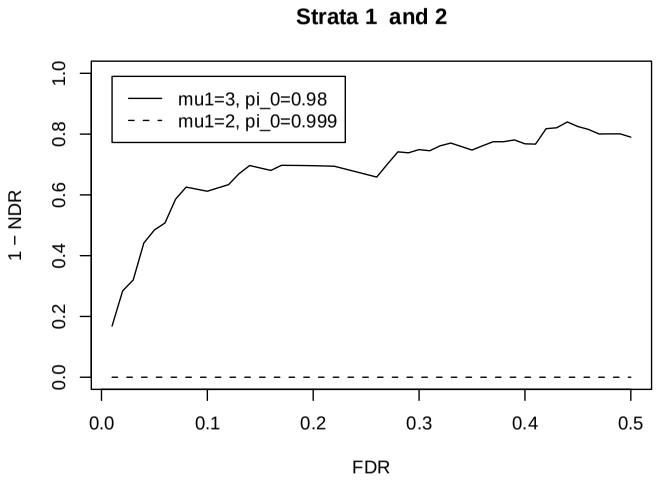
$$NDR_s = \frac{m_1^{(1)} NDR^{(1)} + m_1^{(2)} NDR^{(2)}}{m_1^{(1)} + m_1^{(2)}}$$

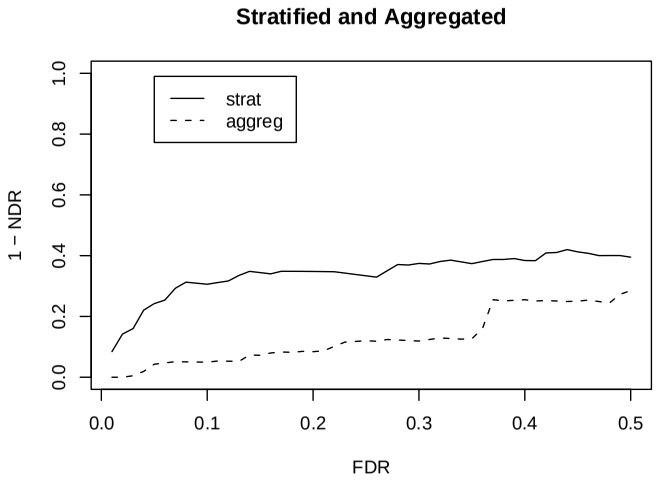

Strata 1 and 2
1 – NDR
FDR
mu1=3, pi_0=0.9
mu1=3, pi_0=0.9
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5

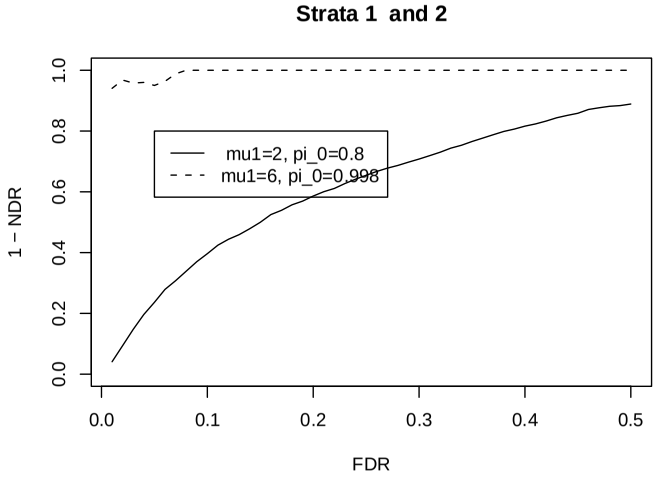

Stratified and Aggregated
1 – NDR
FDR
strat
aggreg
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5

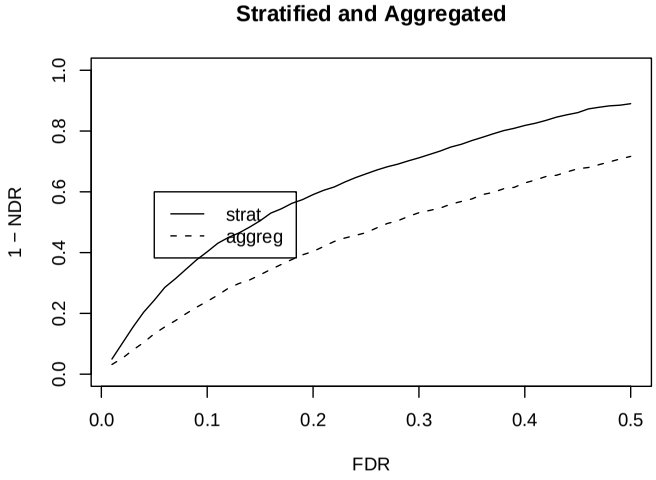

Strata 1 and 2
1 – NDR
FDR
mu1=2, pi_0=0.99
mu1=2.5, pi_0=0.90
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5


Stratified and Aggregated
1 – NDR
FDR
strat
aggreg
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5


Strata 1 and 2
mu1=3, pi_0=0.98
mu1=2, pi_0=0.999
1 – NDR
FDR
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5


Stratified and Aggregated
strat
aggreg
1 – NDR
FDR
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5


Strata 1 and 2
mu1=2, pi_0=0.8
mu1=6, pi_0=0.998
1 – NDR
FDR
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5


Stratified and Aggregated
strat
aggreg
1 – NDR
FDR
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5

References

- Craiu RV, Sun L (2005). Choosing the lesser evil: trade-off between false discovery rate and non-discovery rate. Technical Report #0504, Department of Statistics, University of Toronto.
- Sun L, Bull SB, Craiu VR, Paterson AD (2006). Stratified false discovery control for large scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology*, to appear.