P9120 - Statistical Learning and Data Mining

Lecture 12 - Multiple Testing

Min Qian

Department of Biostatistics, Columbia University

Nov. 21, 2024

Outline

Multiple Testing: Gene Expression Example

$n_1 = 44$, $n_2 = 14$, number of genes $M = 12,625$.

TABLE 18.4. *Subset of the 12,625 genes from microarray study of radiation sensitivity. There are a total of 44 samples in the normal group and 14 in the radiation sensitive group; we only show three samples from each group.*

	Normal				Radiation Sensitive			
Gene 1	7.85	29.74	29.50	...	17.20	-50.75	-18.89	...
Gene 2	15.44	2.70	19.37	...	6.57	-7.41	79.18	...
Gene 3	-1.79	15.52	-3.13	...	-8.32	12.64	4.75	...
Gene 4	-11.74	22.35	-36.11	...	-52.17	7.24	-2.32	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Gene 12,625	-14.09	32.77	57.78	...	-32.84	24.09	-101.44	...

Goal: Find genes which express differently between the radiation sensitive group and the normal group of patients.

Traditional Approach

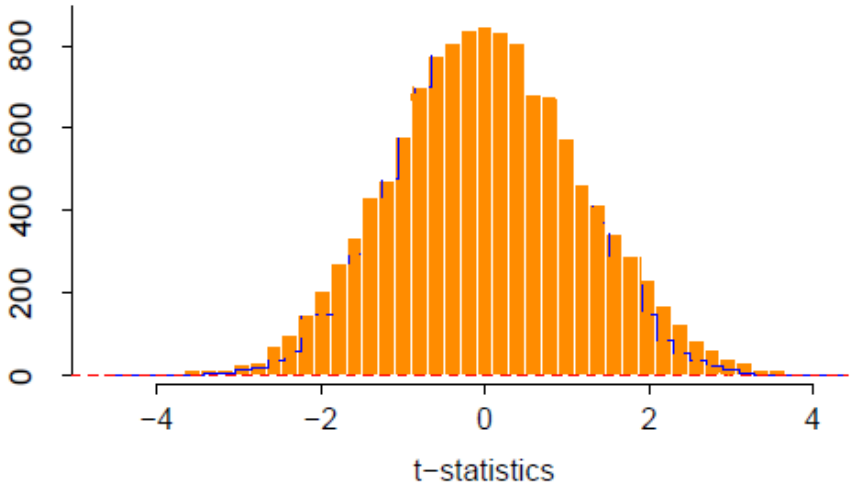
H_{0j} : Gene j doesn't express differently between two groups

vs. H_{1j} : Gene j expresses differently between two groups

- Two sample t-test:

$$t_j = \frac{\bar{x}_{2j} - \bar{x}_{1j}}{se_j}$$

- Reject H_{0j} if $|t_j| \geq 2$ (correspond to a significant level of $\alpha = 5\%$)



There are 1189 genes with $|t_j| \geq 2$.

Any problem?

- The two-sample t-test may not be valid.

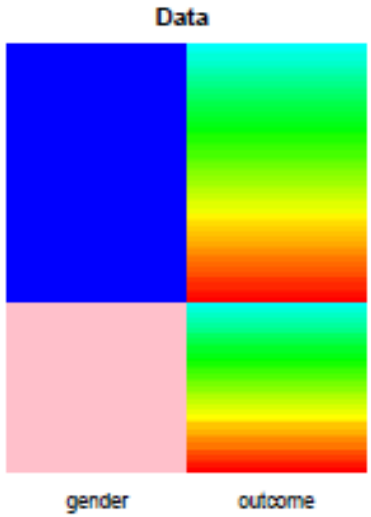
- Many large t values may occur by chance. If the genes were independent, the expected number of falsely significant genes is $M\alpha = 12625 \times 0.05 = 631.3$ if all null hypotheses are true.

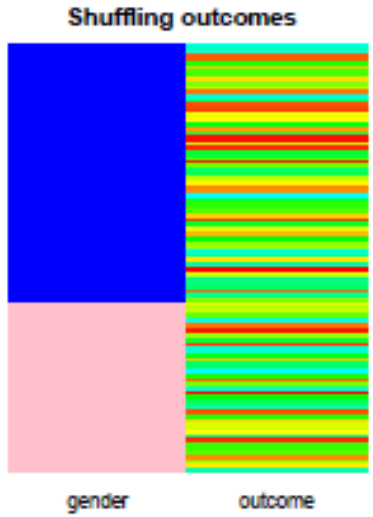
Permutation Test

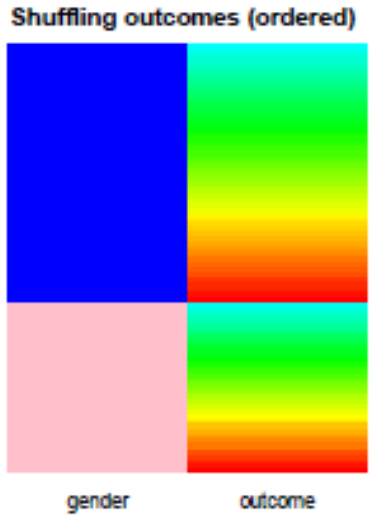
- Use random shuffles of the data to get the correct sampling distribution of a test statistic under the null hypothesis.
- The ranking of the real test statistic among the shuffled test statistics gives a p-value
- Permutation test is used when the distribution of the test statistic is unknown or hard to compute.

Permutation Example

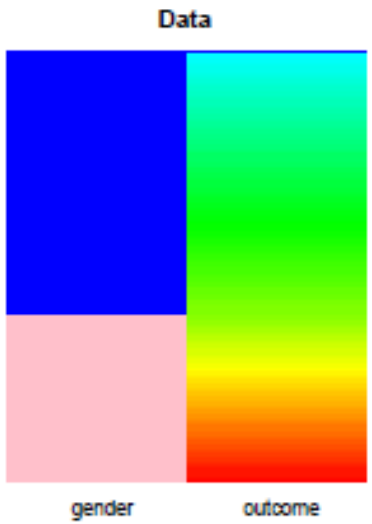
- Null is true

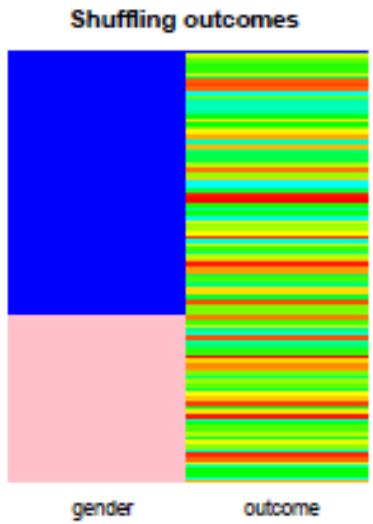


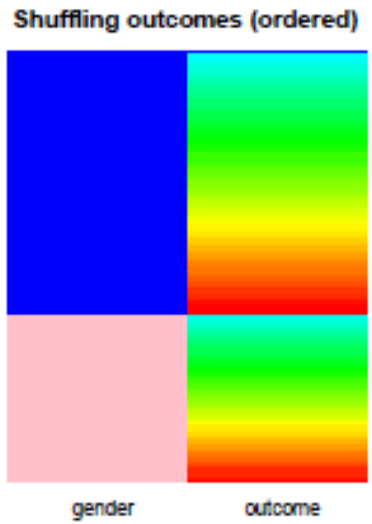




- Null is false







Permutation Test

1. Compute the t-statistic, t_j, from the original data set.
2. Compute all K permutations/combinations of the sample labels and calculate the t-statistic t_j^k for each permutation.
3. Calculate p-value by comparing t_j to t_j^k's:

$$p_j = \frac{1}{K} \sum_{k=1}^{K} I_{|t_j^k| > |t_j|}.$$

4. If the number of total permutations is too large, take a random sample of possible permutations, say $K = 1000$.
5. To exploit the fact that the genes are similar (e.g. measured on the same scale), we can pool the results for all genes in computing the p-values

$$p_j = \frac{1}{MK} \sum_{j'=1}^{M} \sum_{k=1}^{K} I_{|t_{j'}^k| > |t_j|}.$$

Count Errors

With One hypothesis test:

Truth	Decision: Do Not Reject H_0	Decision: Rejct H_0
H_0 True	Correct Decision $1 - \alpha$	Incorrect Decision Type I Error α
H_0 False	Incorrect Decision Type II Error β	Correct Decision $1 - \beta$

With M hypothesis tests:

	Called Not Significant	Called Significant	Total
H_0 True	U	V	M_0
H_0 False	T	S	M_1
Total	$M - R$	R	M

- V =# of type I errors, T =# of type II errors.
- We would like to control V (overall type I error) in some way.

Approaches To Control Type I Errors

- Control Per Comparison Error Rate (PCER)
  - e.g. "uncorrected testing" – reject H_{0j} if $p_j \leq \alpha$.
  - May result in many type I errors.
- Control Family-wise Error Rate (FWER).
  - Guarantees FWER $\triangleq P(V \geq 1) \leq \alpha$.
  - Concept proposed by Tukey (1953) and Ryan (1959).
  - e.g. "Bonferroni correction" – reject H_{0j} if $p_j \leq \alpha/M$.
  - today: Holm procedure.
- Control False Discovery Rate (FDR):
  - Guarantees FDR $\triangleq E(V/R) \leq \alpha$.
  - First defined by Benjamini & Hochberg (BH, 1995, 2000)

FWER

Many procedures have been developed to control FWER (the probability of at least one type I error): $P(V \geq 1)$

Two general types of FWER corrections:

- Single step: equivalent adjustments made to each p-value
- Sequential: adaptive adjustment made to each p-bvalue

Single Step Approach: Bonferroni

Reject any hypothesis with p-value$\leq \alpha/M$.

Among M null hypotheses, let $H_0^{(1)}, \ldots, H_0^{(M_0)}$ be M_0 true null hypotheses.

$$\begin{aligned}\text{FWER} = P(V \geq 1) &\leq \sum_{j=1}^{M_0} P(\text{reject } H_0^{(j)}) \\ &= \sum_{j=1}^{M_0} P\Big(p - value^{(j)} \leq \alpha/M\Big) = \frac{M_0\alpha}{M} \leq \alpha.\end{aligned}$$

- FWER $\approx \alpha$ if the M tests are independent and $M_0 = M$.
- In general, it is highly conservative (lower power). In our example, the threshold is $0.05/12625 = 3.9 \times 10^{-6}$. None of the genes had a p-value this small.

Sequential Approach: Holm Procedure

1. Order the p-values such that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(M)}$
2. Reject $H_{0,(j)}$ if $p_{(k)} \leq \alpha/(M-k+1)$ for all $k = 1, \ldots, j$.

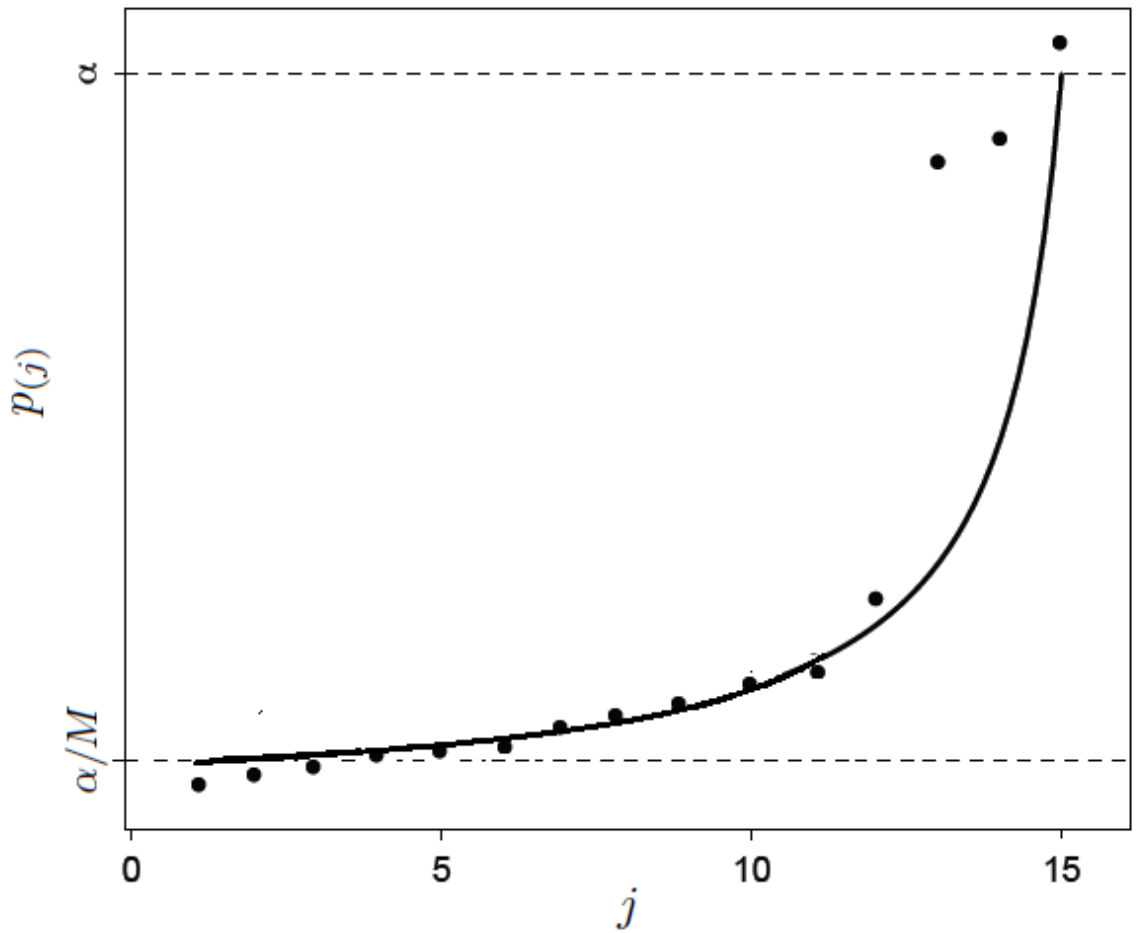
Proof.

- J: set of indices corresponding to true null hypotheses.
- $j_0 = \arg\min_{j \in J} p_{(j)}$: ranking of the first true null.
- $j_0 \leq M - M_0 + 1$, where M_0 is the # of true null.
- Commit a false significance if and only if

$$p_{(1)} \leq \frac{\alpha}{M}, p_{(2)} \leq \frac{\alpha}{M-1}, \ldots, p_{(j_0)} \leq \frac{\alpha}{M-j_0+1}$$

- $\text{FWER} \leq \text{P}\left(\min_{j \in J} p_{(j)} \leq \frac{\alpha}{M_0}\right) \leq \sum_{j \in J} P\left(p_{(j)} \leq \frac{\alpha}{M_0}\right) = \alpha.$

Bonferroni vs. Holm



- Bonferroni: reject $H_{0,(j)}$ if $p_{(j)} \leq \alpha/M$.
- Holm: Reject $H_{0,(j)}$ if $p_{(k)} \leq \alpha/(M-k+1)$ for all $k = 1, \ldots, j$.

Practical Problem with FWER

- While guarantee of FWER-control is appealing, the resulting thresholds often suffer from low power. In practice, this tends to wipe out evidence of the most interesting effects.

- In many cases (particularly in genomics) we can live with a certain number of false positives. FDR control offers a way to increase power while maintaining some principled bound on error.

- With FDR, we say "4 false discoveries out of 10 rejected null hypotheses" is a more serious error than "20 false discoveries out of 100 rejected null hypotheses."

Benjamini and Hochberg FDR

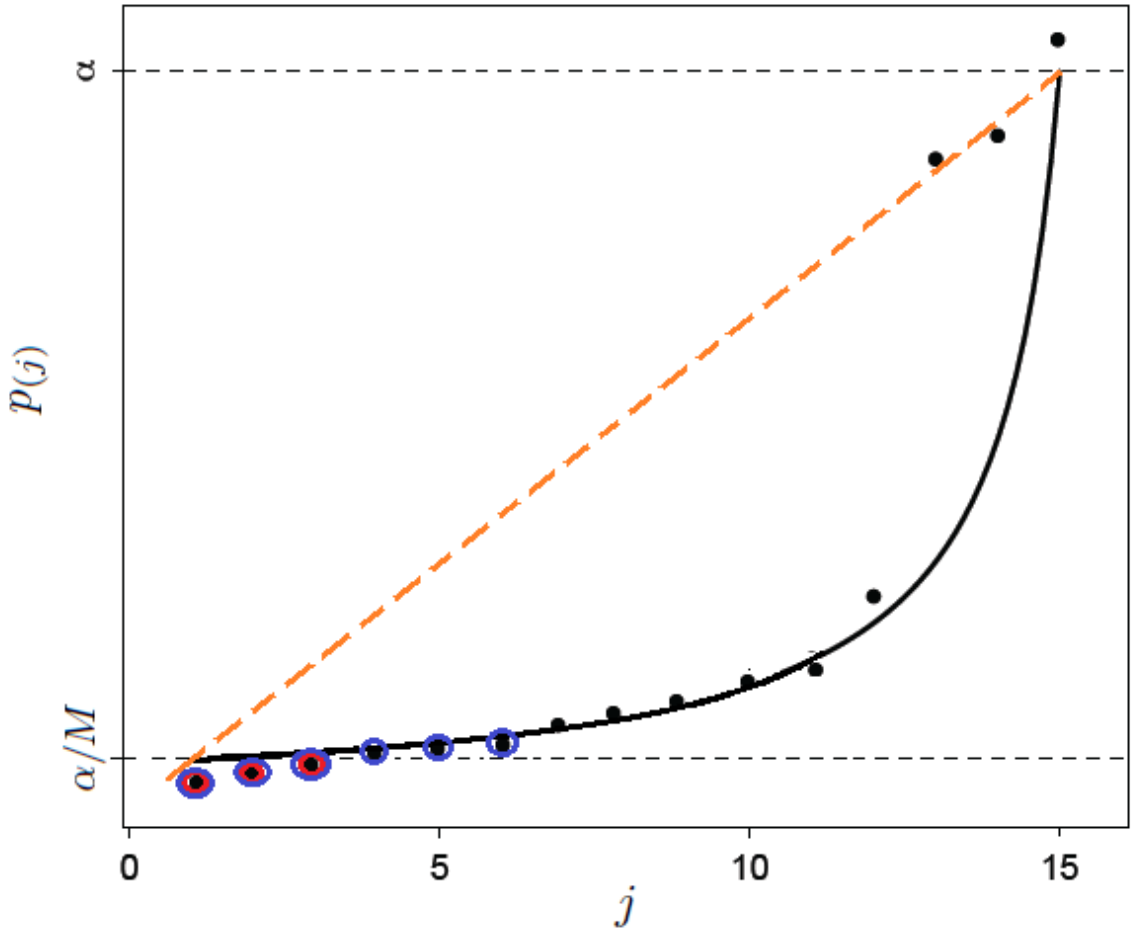
Assume all test statistics are independent. To control FDR at level α,

1. Order the p-values such that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(M)}$
2. Define $j^* = \max\{j : p_{(j)} \leq \frac{j\alpha}{M}\}$.
3. Reject all $H_{0,(j)}$ with $j \leq j^*$.

Intuition:

- BH procedure finds maximal j^* s.t. $\frac{Mp_{(j^*)}}{j^*} \leq \alpha$, and rejects any hypothesis with p-value$\leq p_{(j^*)}$.
- When using $p_{(j^*)}$ as the cut-off value for the raw p-values, the expected number of false positives can be estimated by $M_0 p_{(j^*)}$, since p-value under the null is uniformly distributed.
- Estimate of FDR: $M_0 p_{(j^*)}/j^*$.
- Since M_0 is unknown, M is used as a conservative estimate of M_0.

Toy Example Continued



Gene Expression Example Continued

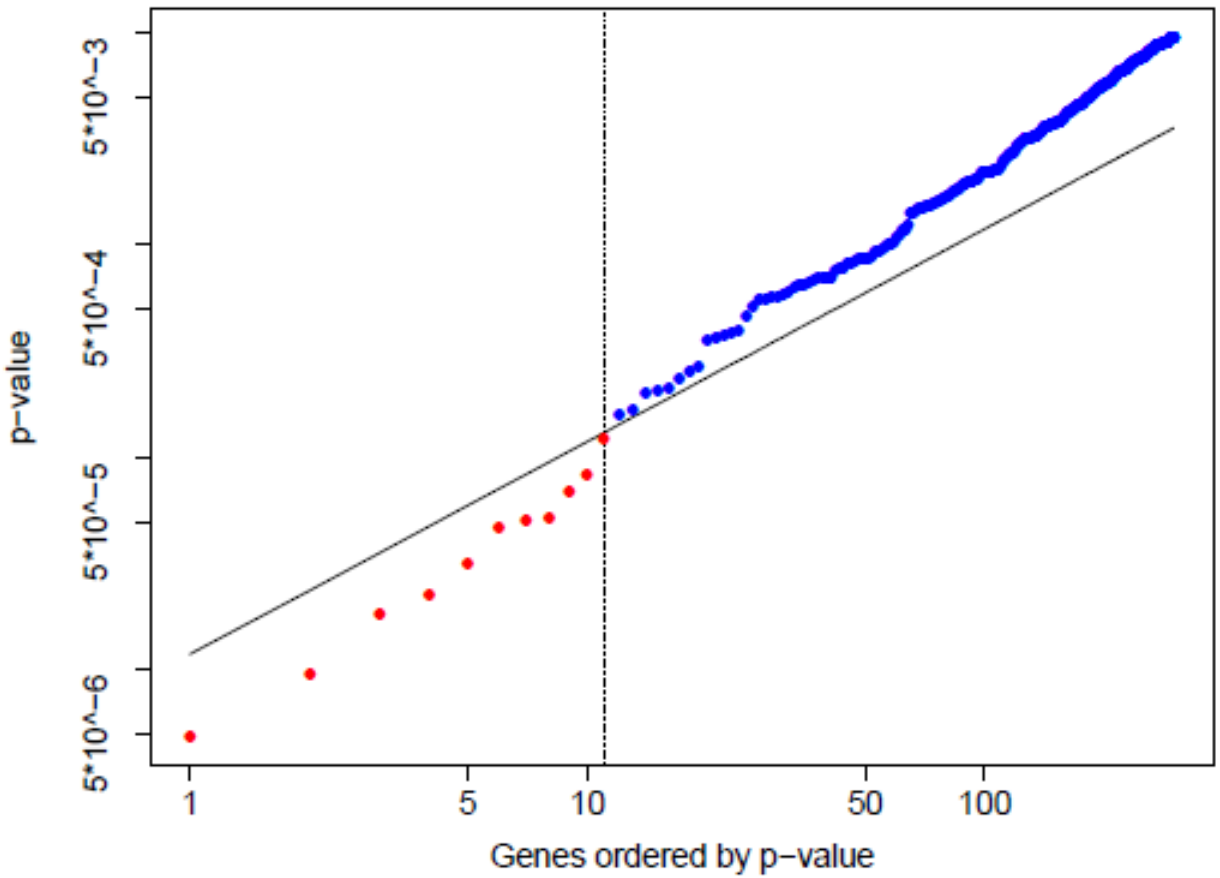


FIGURE 18.19. *Microarray example continued. Shown is a plot of the ordered p-values $p_{(j)}$ and the line $0.15 \cdot (j/12,625)$, for the Benjamini–Hochberg method. The largest j for which the p-value $p_{(j)}$ falls below the line, gives the BH threshold. Here this occurs at $j = 11$, indicated by the vertical line. Thus the BH method calls significant the 11 genes (in red) with smallest p-values.*

FWER vs. FDR

V : number of false discoveries
R : number of discoveries

$$\text{Note that} \qquad \text{FDR} = E\left[\frac{V}{R}\right] \leq P(V \geq 1)$$

$$\text{FDR} = E\left[\frac{V}{R}\right] = E\left[\frac{V}{R}\middle| R > 0\right] P(R > 0)$$

by setting $V/R = 0$ whenever $R = 0$.

- If all null hypotheses are true, then FDR is equivalent to FWER.

  Control of FDR implies control of FWER in a weak sense.

- When $M_0 < M$, any procedure that controls the FWER also controls the FDR. A procedure that controls the FDR but not the FWER can only be less stringent and thus more powerful.

Summary

- Permutation test is based on random shuffles of data to get the correct sampling distribution of a test statistic under the null hypothesis. It is useful when the distribution of the test statistic is unknown or hard to compute.

- When there are multiple comparisons, we need to address the multiple testing problem. The standard approach is to control FWER, e.g. Bonferroni, Holm, etc.

- When the number of tests M is large, FWER control may be too conservative. FDR is a useful alternative.

- The BH method is fast and robust, but it may over-control FDR. The main difficulty is finding a good estimator of FDR or M_0. There are improved adaptive methods, say BKY (2004).

Reminders

- HW #4 is due at 12pm noon on Monday December 9th.
- Quiz for lecture 12 is due at 9pm on Monday November 25th.
- In class group paper presentation (December 5th):
  - CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning
    Team members: Sixuan Chen, Aiying Huang, Yixiao Sun, Zihan Wu, Allison Xia, Jingyi Xu, Tongxi Yu
  - Deep learning for the prediction of early on-treatment response in metastatic colorectal cancer from serial medical imaging
    Team members: Mingzhi Chen, Chenshuo Pan, Zixuan Qiu, Xiaoting Tang, Mia Yu, Shubo Zhang