Bonferroni to Benjamini and back to Bayes

Daniel Yekutieli

Statistics and OR
Tel Aviv University

New perspectives in multiple testing - 80 years after Bonferroni

Padua, October 9, 2015

Plan of my talk

1. Benjamini (and Hochberg) vs. Bonferroni
2. Connection to post hoc analysis
3. Back to (e)Bayes

"What has been will be again, what has been done will be done again; there is nothing new under the sun." *Ecclesiastes 1:9 (NIV)*

Multiple hypotheses testing framework

- m tested null hypotheses $H_1 \cdots H_m$
- m_0 null hypotheses $(P_i \sim U[0,1])$,
  $m_1 = m - m_0$ false null hypotheses $(P_i \leq U[0,1])$
- Rejecting a null hypothesis is a discovery, a false discovery is erroneously rejecting a true null hypothesis
- R is the number of discoveries and V is the number of false discoveries

$$FWE := \Pr(V > 0)$$

$$FDR := EQ, \quad Q = \begin{cases} 0 & \text{if } R = 0 \\ V/R & \text{if } R > 0 \end{cases}$$

Level q $(= 0.05)$ BH procedure

1. Sort the p-values $P_{(1)} \leq P_{2)} \leq \cdots \leq P_{(m)}$
2. Compare $P_{(i)}$ with $i \cdot q/m$
3. Let $r = \max\{i : P_{(i)} \leq i \cdot q/m\}$
4. Reject $H_{(1)} \cdots H_{(r)}$

Level α $(= 0.05)$ Bonferroni procedure

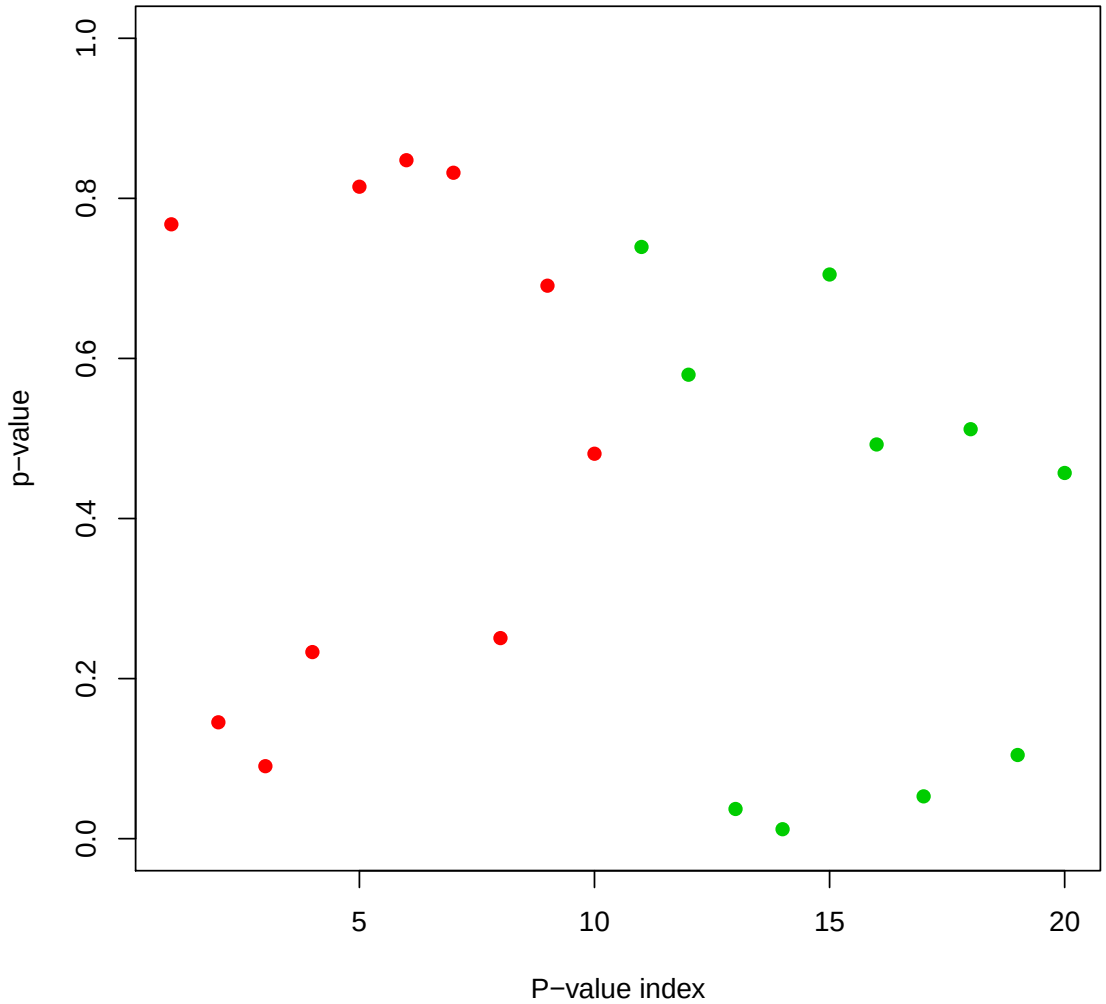
1. Reject H_i if $P_i \leq \alpha/m$

FWE control via Bonferroni inequality:

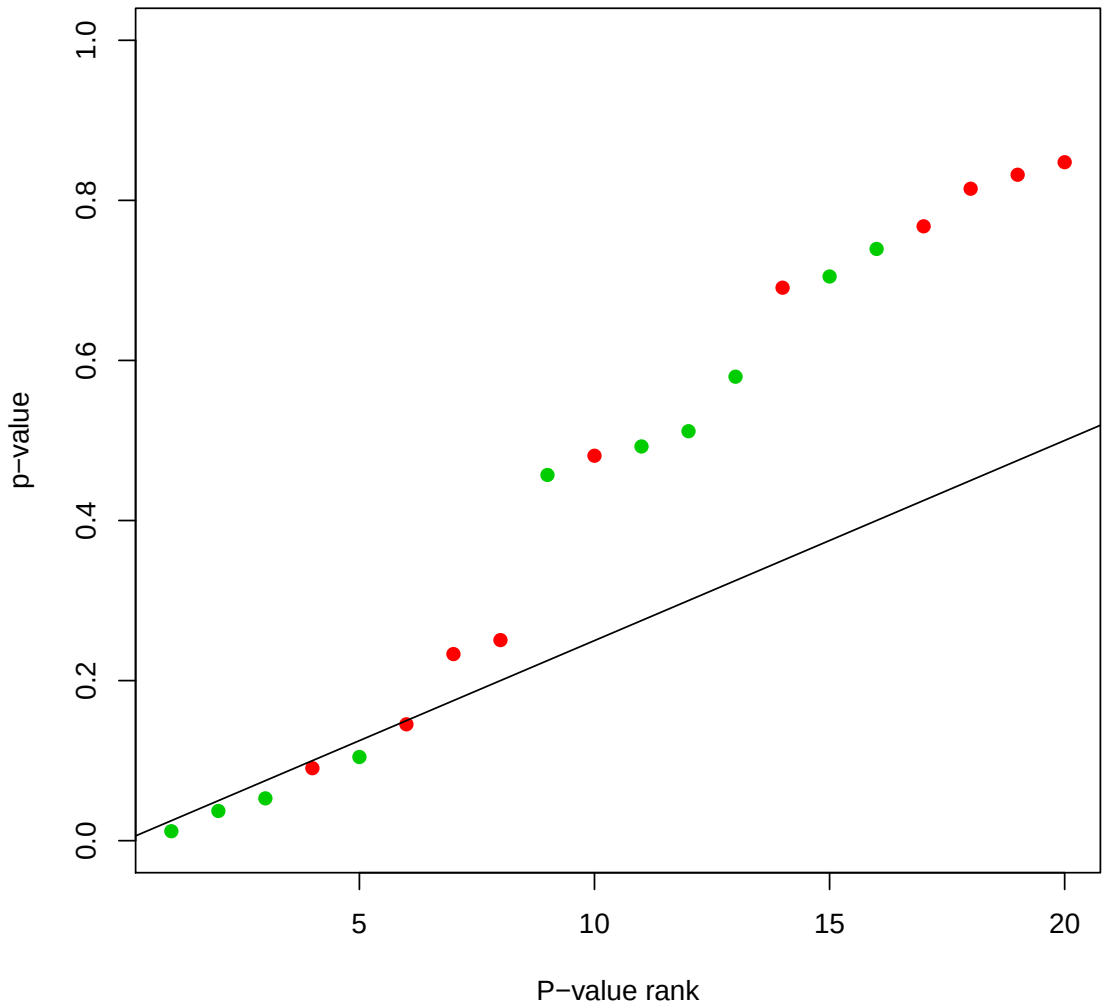
$$\Pr(\cup_{i \in I_0} P_i \leq \alpha/m) \leq \sum_{i \in I_0} \Pr(P_i \leq \alpha/m) = m_0 \cdot \alpha/m \leq \alpha$$

$I_0 \subseteq \{1 \cdots m\}$ is subset of true null hypotheses

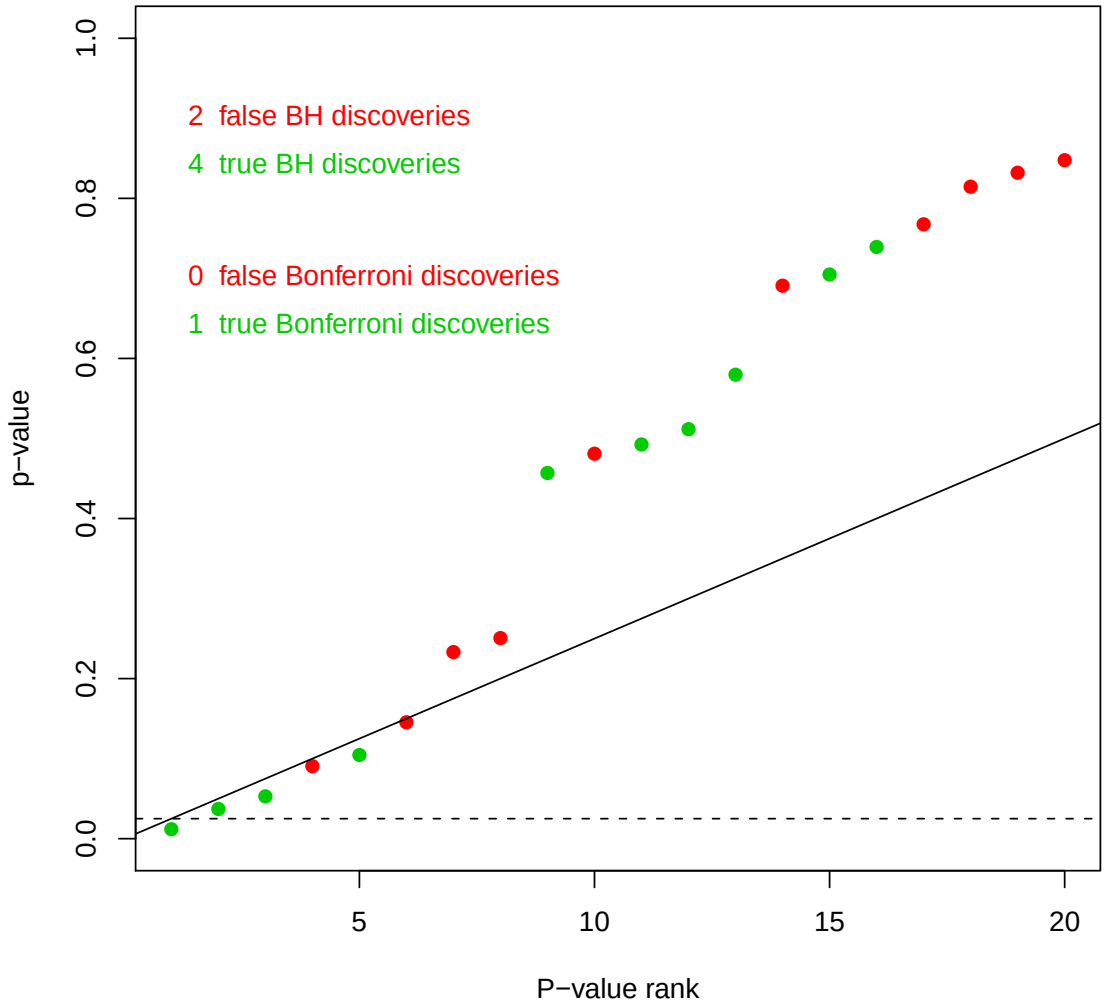
$P_1 \cdots P_{10}$ true null p-values, $P_{11} \cdots P_{20}$ false null p-values



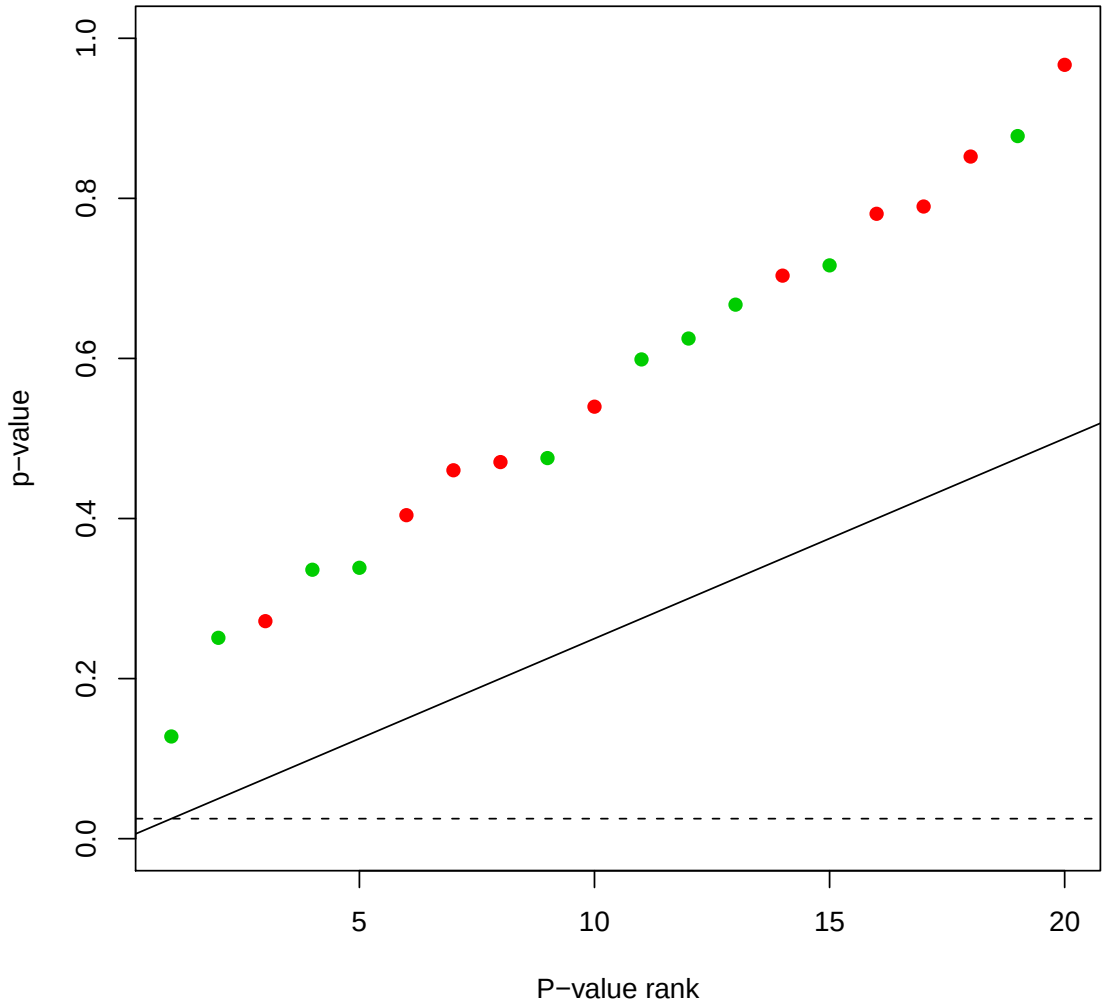
Level $q = 0.5(!?)$ BH procedure



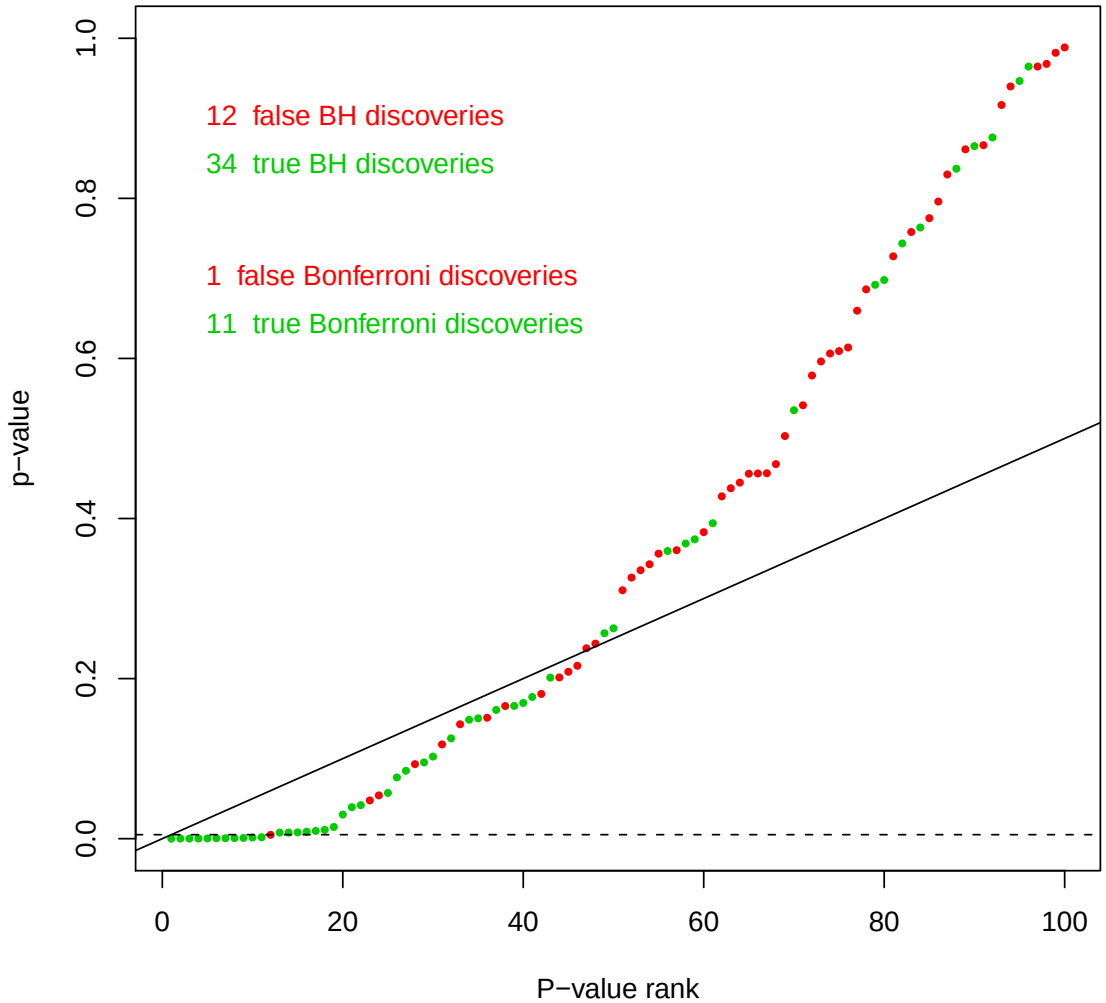
Compare with $\alpha = 0.5$ Bonferroni procedure



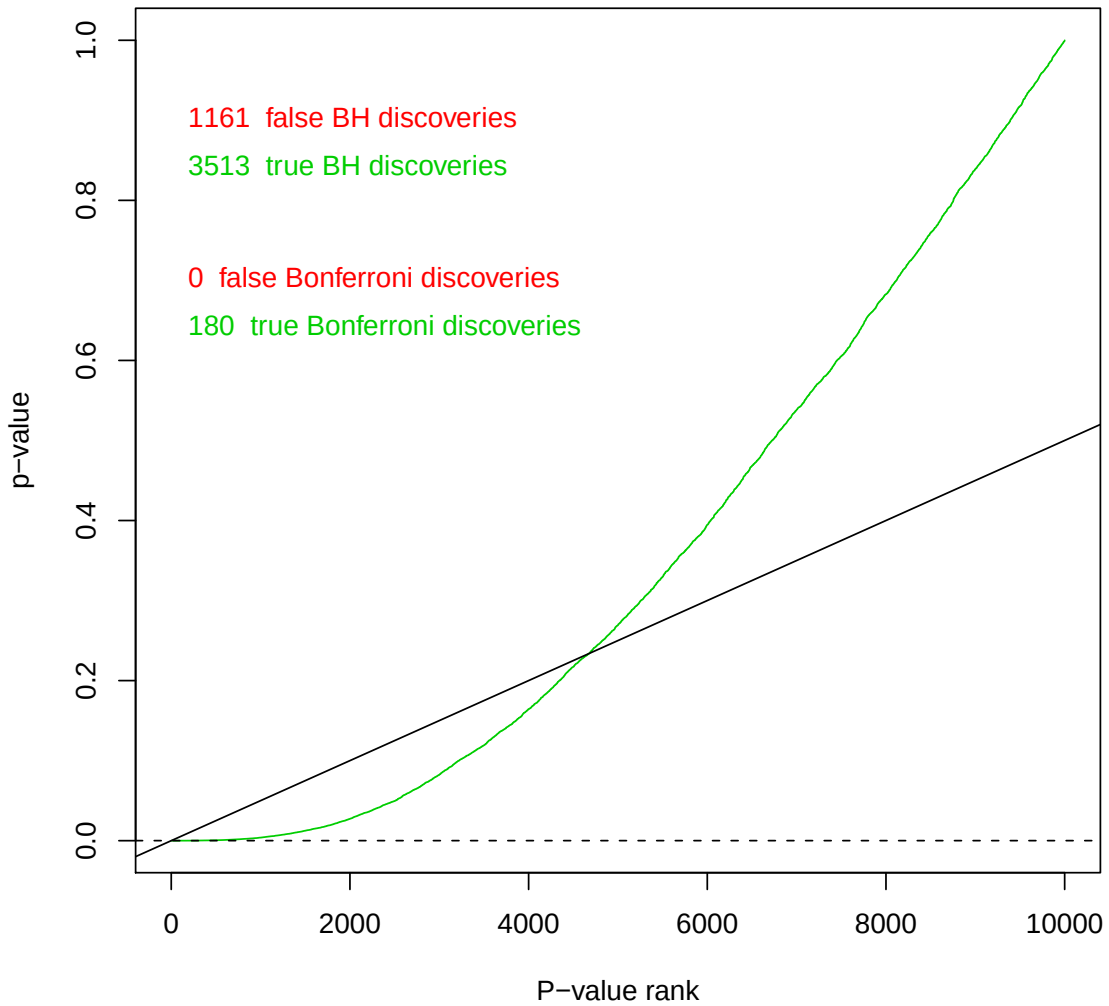
$P_1 \cdots P_{20}$ true null p-values



$P_1 \cdots P_{50}$ true nulls, $P_{51} \cdots P_{100}$ false nulls



Same simulation but now $m = 10000$



Is the BH procedure new?

Simes '86

1. Sort the p-values $P_{(1)} \leq P_{2)} \leq \cdots \leq P_{(m)}$
2. Compare $P_{(i)}$ with $i \cdot q/m$
3. If $\exists i : P_{(i)} \leq i \cdot q/m$ then reject the global null that $H_1 \cdots H_m$ are true null hypotheses.

BH procedure $\rightarrow$ FDR control

BH '95

- Independence $FDR \leq m_0 \cdot q/m$

Benjamini and Yekutieli (2001)

- Independence $FDR = m_0 \cdot q/m$
- Positive dependence $FDR \leq m_0 \cdot q/m$
- Geneal dependence $FDR \leq (1 + 1/2 + \cdots + 1/m) \cdot m_0 \cdot q/m$

Simulations

- Robustness to dependence

Interpretation of level 0.05 FDR control

All noise regime ($m_0 = m$)

- Any discovery is false – $FDR \equiv FWE$
- 0.95 probability of not making any false discovery

Signal and noise regime ($m_0 < m$)

- Many discoveries $\approx$ 0.05 false – $FDR < FWE$
- A randomly selected discovery is true with prob. 0.95

Selective vs. simultaneous inference

Benjamini and Yekutieli '05: two types of problems can arise when providing inferences in studies with multiple parameters . . .

- *Selective inference* – need to provide marginal inferences for parameters that are selected after viewing the data (e.g. microarray analysis) – solution FDR control
- *Simultaneous inference* – need to provide inferences that apply to all the parameters (e.g. subgroup analysis) – solution FWE control

Selective inference a new idea?

- Soric, JASA ‘89
  “ . . . It is mainly the discoveries that are reported and included into science . . . unless the proportion of false discoveries is kept small there is danger that a large part of science is untrue”

- Ioannidis, Plos Medicine ‘05
  “Why Most Published Research Findings Are False”

- Tukey and Scheffe ‘53
  post-hoc analysis

Post-hoc analysis – Scheffe's method

- $\boldsymbol{\mu} = (\mu_1 \cdots \mu_k)$ is a vector of k treatment effects, mean response in i'th treatment group is $\hat{\mu}_i \sim N(\mu_i, \sigma^2/n)$
- After viewing the data (and ANOVA) a contrast, $\boldsymbol{a}\boldsymbol{\mu} = a_1\mu_1 + \cdots a_k\mu_k$ with $a_1 + \cdots + a_k = 0$, is selected

Selective inference problem: how do we use data to select $\boldsymbol{a}\boldsymbol{\mu}$ and then test its significance or construct a confidence interval for it?

Solution: base inference on confidence interval

$$CI_{Scheffe}(\boldsymbol{a}, \alpha) := \boldsymbol{a}\hat{\boldsymbol{\mu}} \pm \frac{\hat{\sigma} \cdot \|\boldsymbol{a}\|}{\sqrt{n}} \cdot \sqrt{(k-1) \cdot F_{1-\alpha,k-1,N-k}}$$

that offer simultaneous coverage for all contrasts

$$\Pr_{\boldsymbol{\mu}}\{\forall \boldsymbol{a} : \ \boldsymbol{a}\boldsymbol{\mu} \in CI_{Scheffe}(\boldsymbol{a}, \alpha)\} \geq 1 - \alpha.$$

Scheffe's method – FWE control

- Family of null hypotheses: $\forall \boldsymbol{a}, H_{\boldsymbol{a}}^0 : \boldsymbol{a}\boldsymbol{\mu} = 0$
- True null hypotheses: $\{\boldsymbol{a} : \boldsymbol{a}\boldsymbol{\mu} = 0\}$
- Test: reject $H_{\boldsymbol{a}}^0 : \boldsymbol{a}\boldsymbol{\mu} = 0$ if $0 \notin CI_{Scheffe}(\boldsymbol{a}, \alpha)$
- Coverage for all contrasts $\rightarrow$ FWE control for family of null hypotheses

Bonferroni method for Post-hoc analysis

"Multiple Comparisons Among Means" O. J. Dunn *JASA* '61

- Instead of considering all possible contrasts, suggests specifying *in advance* m contrasts $\boldsymbol{a}^1\boldsymbol{\mu}\cdots\boldsymbol{a}^m\boldsymbol{\mu}$
- Construct $1-\alpha/m$ confidence interval for the contrast $\boldsymbol{a}^i\boldsymbol{\mu}$ that is selected post-hoc
- Thus, according to the Bonferroni inequality the confidence interval will cover $\boldsymbol{a}^i\boldsymbol{\mu}$ wprob $\geq 1-\alpha$ regardless of joint distribution $\boldsymbol{a}^1\hat{\boldsymbol{\mu}}\cdots\boldsymbol{a}^m\hat{\boldsymbol{\mu}}$
- If m is not too large $1-\alpha/m$ confidence intervals will be smaller than the Scheffe confidence intervals

FDR control – slightly different objective

- Post hoc analysis is concerned with valid inference for a single contrast (possibly the most significant contrast) that is specified according to the data
- Conditional marginal validity property of FDR more appropriate in contemporary applications (microarrays / GWAS / fMRI / nonparametric regression) that are concerned with deriving valid marginal inferences for multiple parameters that are selected after first considering m pre-specified parameters
- And indeed ... Williams, Jones and Tukey '99 suggest using the BH procedure for discovering state-to-state (pairwise) differences in educational achievement.

The two group mixture model

Introduced in Efron et al. '01, see also Efron '10:

- *Random* hypotheses vector $\boldsymbol{H} = (H_1, \cdots, H_m)$
- $H_i \overset{iid}{\sim} Bernouli(1-\pi_0)$
- Data vector $\boldsymbol{Z} = (Z_1, \cdots, Z_m)$
- For $H_i = 0$, $Z_i \sim f_0$ where usually $f_0 = N(0,1)$
- For $H_i = 1$, $Z_i \sim f_1$
- Z_i usually scalar.
- For data summarized by p-values consider $Z_i = \Phi^{-1}(P_i) \quad \rightarrow \quad f_0 = N(0,1)$

Bayesian FDR

- Efron et al '01 (= pFDR in Storey '02-'03):

$$Fdr = Pr(H_i = 0 | R_i = 1)$$

With a little work get $Fdr = E_{\boldsymbol{H},\boldsymbol{Z}}(V/R | R > 0)$

- Relation between Bayesian FDR and the BH FDR

$$Fdr = E_{\boldsymbol{H}}\{E_{\boldsymbol{Z}|\boldsymbol{H}}(V/R \mid R > 0)\} = E_{\boldsymbol{H}}\{FDR \;/\; \Pr_{\boldsymbol{Z}|\boldsymbol{H}}(R > 0)\}$$

Thus $Fdr \approx FDR$

Bayes rule for classification in the two group mixture model

- Classifier (= Test): classify H_i as R_i

$$R_i(\mathbf{Z}, T_i, \delta) = I\{T_i(\mathbf{Z}) \leq \delta\}$$

- Easy to see that Bayes rule for classifying H_i can be expressed

$$R_i \;=\; I\{\, fdr(z_i) \leq \frac{\lambda_2}{\lambda_1 + \lambda_2}\},$$

for the local FDR (Efron '01)

$$fdr(z_i) = \pi_0 \cdot f_0(z_i)/f(z_i) = \Pr(H_i = 0|z_i)$$

for

$$f(z_i) = \pi_0 \cdot f_0(z_i) + \pi_1 \cdot f_1(z_i)$$

$Fdr = q$ classifier

- A $Fdr = q$ classifier is simply $R_i(\boldsymbol{Z}; T_i, \delta(q))$ for which

$$Fdr \;= \Pr(H_i = 0 | R_i = 1) \;=\; q.$$

- In particular, the Bayes classifier can be specified by its Fdr level, rather than by λ_2 and λ_1. i.e. the $Fdr = q$ Bayes classifier is

$$R_i(Z_i;\, fdr,\, \delta(q)) = I\{\, fdr(z_i) \leq \delta(q)\}$$

with $Fdr = q$

The $Fdr = q$ Bayes classifier is optimal

Of all $R_i(\boldsymbol{Z}; T, \delta)$ with $Fdr = q$, the Bayes classifier has

- Maximum power to make discoveries

$$\Pr(R_i = 1)$$

- and minimum type II error

$$Fnr = \Pr(H_i = 1 | R_i = 0)$$

(Storey '07; Sun and Cai '07; Efron '10; Heller and Yekutieli '14)

How do we control Fdr?

With empirical Bayes:

1. Efron's R *locfdr* package applied to $Z_1 \cdots Z_m$ to estimate π_0, f_0, f_1 and use them to compute $Fdr(z)$, the Fdr of the rejection rule $R_i = I(z \leq Z_i)$
2. Storey's R *qvalue* package applied to $P_1 \cdots P_m$ to estimate $qvalue(p)$, the Fdr of the rejection rule $R_i = I(P_i \leq p)$, that equals

$$\Pr(H_i = 0 | P_i \leq p) = \frac{Pr(P_i \leq p | H_i = 0) \cdot Pr(H_i = 0)}{Pr(P_i \leq p)} = \frac{p \cdot \pi_0}{Pr(P_i \leq p)}$$

by

$$\widehat{qvalue}(p) = \frac{p \cdot \hat{\pi}_0}{\#\{i : \ P_i \leq p\}/m}$$

Connection to BH procedure

From a Bayesian perspective, the BH procedure is a $Fdr = q$ classifier $R_i(Z_i; T, \hat{\delta}(q)) = I\{T(Z_i) \leq \hat{\delta}(q)\}$ for which

1. the test statistic is the p-value $T(Z_i) = P_i$
2. the critical value is $\hat{\delta}(\boldsymbol{z}; q) = p_{(r)}$ maximal p-value for which

$$p_{(r)} \leq \frac{q \cdot r}{m} \quad \Leftrightarrow \quad \frac{p_{(r)}}{\#\{i:\ P_i \leq p_{(r)}\}/m} \leq q$$

If we further apply the adaptive BH procedure that includes an estimate of π_0 we get the *qvalue* critical value:

$$\hat{p} \text{ such that } \frac{\hat{p} \cdot \hat{\pi}_0}{\#\{i:\ P_i \leq \hat{p}\}/m} \leq q$$

Generalization of the two group mixture model

Heller and Yekutieli '13:

- Parameter $\boldsymbol{H} = (H_1, \cdots, H_m)$

  For $h \in \mathcal{H}$, $H_i = h$ $wprob$ $\pi(h)$

- Observation $\boldsymbol{Z} = (Z_1, \cdots, Z_m)$

  $ind\ Z_i$: $Z_i | H_i = h \sim f_h$

- Instead of a null parameter value, there is a null set $\mathcal{H}_0 \subset \mathcal{H}$
- The discovery $R_i = 1$ corresponds to declaring that $H_i \notin \mathcal{H}_0$
- Thus the false discovery indicator is $V_i = I(R_i = 1, H_i \in \mathcal{H}_0)$

Level q Bayes classifier

- The Bayesian FDR for any classifier is

$$Fdr = \Pr(H_i \in \mathcal{H}_0 | \; R_i = 1)$$

- The local FDR is

$$fdr(z_i) = \Pr(H_i \in \mathcal{H}_0 | \; z_i)$$

- The level q Bayes classifier is

$$R_i(Z_i;\, fdr,\, \delta(q)) = I\{\, fdr(z_i) \leq \delta(q)\} \quad \text{with} \quad Fdr = q$$

How does the BH procedure look in this case

- $P_i = P(Z_i)$, for $i = 1 \cdots m$, valid p-value

$$\forall h \in \mathcal{H}_0, \quad \Pr(P_i \leq p | H_i = h) \leq p$$

- The BH procedure is $R_i(\boldsymbol{Z}, P(Z_i), \hat{p}(q; \boldsymbol{z}))$ with

$$\hat{p}(q; \boldsymbol{Z}) = max\{p : \frac{p}{\#\{i : P_i \leq p\}/m} \leq q\}$$

BH procedure has much smaller discovery probability than the Bayes classifier

When Z_i is not a scalar or $\mathcal{H}_0$ is a composite null hypothesis the BH procedure has smaller discovery probability than the Bayes classifier

1. All p-value based classifier is suboptimal (Bayes classifier is optimal)
2. BH procedure not calibrated correctly: As $\Pr(H_i \in \mathcal{H}_0) < 1$ and since $\Pr(P_i \leq p) < p$ for composite nulls, the BH procedure over estimates the Fdr making the rejection threshold too small!

Replicability in T2D GWAS – work with Ruth Heller

Genome-wide Association Studies try to identify genetic variants that are associated with a given phenotype in some population of interest. We analyze data from six GWAS testing association with T2D for same 2.5×10^6 SNPs in six different populations.

1. Meta-analysis combine GWAS for increased power to discover SNP associated with outcome in at least one study (for each SNP, test null hypothesis that the SNP is associated with the phenotype in 0 studies)
2. Replicability analysis aims to discover associations between SNP and phenotype that are present in more than one study (test null hypothesis that the SNP is associated with the phenotype in studies ≤ 1)

Replicability? (a) Kraft et al. '09: effects may be small as genetic biases, important to see associations in several GWAS (b) Scientifically Interesting

Analyses of 6 T2D GWAS

- Frequentist FDR analysis (Benjamini, Heller and Yekutieli '09)
  1. Compute p-value for each SNP to test (1) no association (2) no-replication
  2. Apply BH procedure at level 0.05 to each set of 2.5M p-values
  3. Results: 466 associated SNP, replicated associations for 113 SNP in 5 genomic regions
- Bayesian FDR analysis (Heller and Yekutieli '14)
  1. eBayes level 0.05 FDR controlling approach for testing (1) no association (2) no-replication
  2. Results: 803 associated SNP, replicated associations for 219 SNP in 17 genomic regions

Bayes procedure considerably more power than the BH procedure!

Is it real?

Extensive simulation:

- Bayesian FDR procedure has more power than BH procedure for discovering associations, and considerably more (7-15 fold) power for discovering replicated associations!
- Bayesian FDR procedure controls the FDR at nominal level (simulation mean FDP = 0.05) for large studies, slightly under-conservative (simulation mean FDP = 0.07) for smaller studies.
- BH procedure slightly over-conservative (simulation mean FDP = 0.04) for testing no association, highly over-conservative (simulation mean FDP < 0.001) for testing no replication.

– THANKS –