A Tutorial on False Discovery Control

Christopher R. Genovese

Department of Statistics

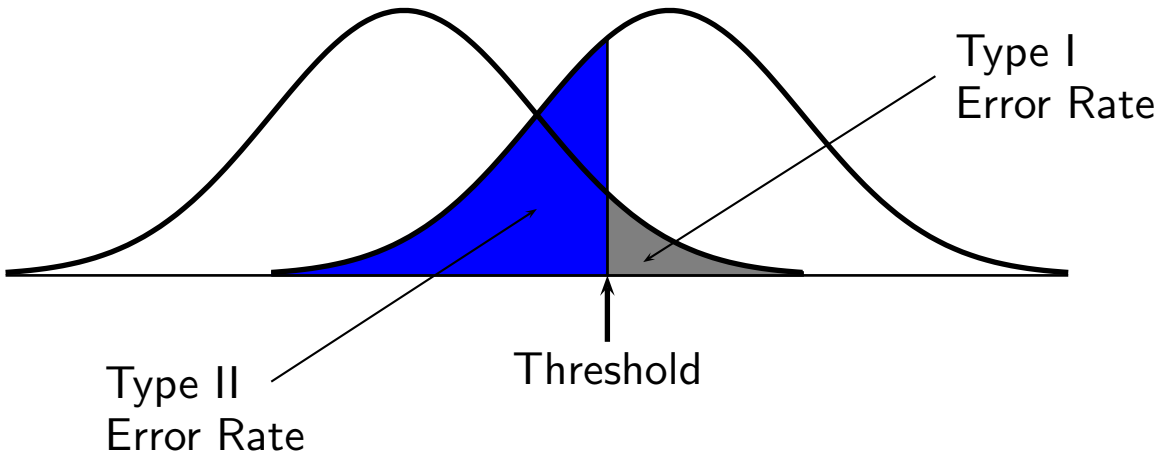
Carnegie Mellon University

http://www.stat.cmu.edu/~genovese/

This work partially supported by NSF Grants SES 9866147 and NIH Grant 1 R01 NS047493-01.

One Test, One Threshold

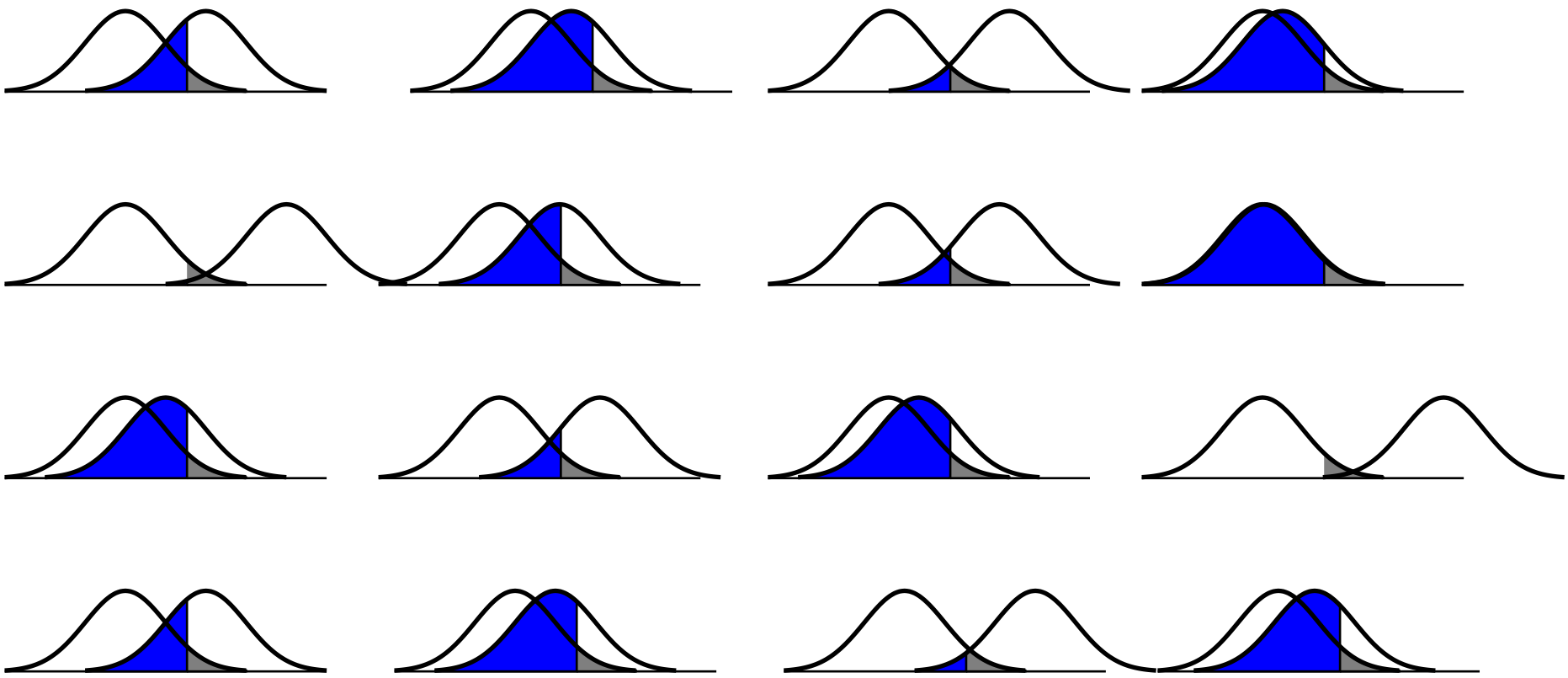
With a single hypothesis test, we choose a rejection threshold to control the Type I error rate,



while achieving a desirable Type II error rate for relevant alternatives.

Many Tests, One Threshold

With multiple tests, the problem is more complicated

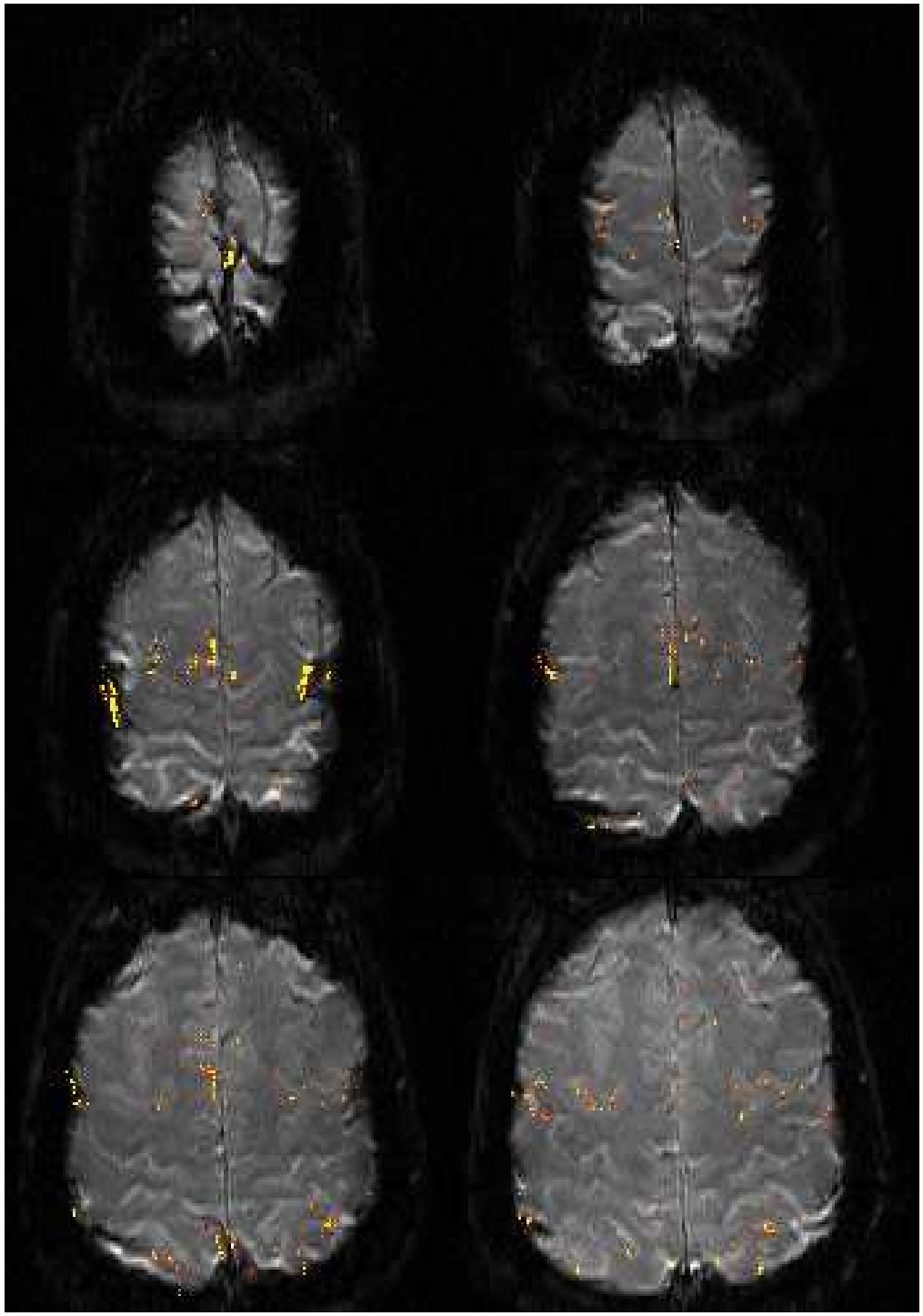


Each test has possible Type I and Type II errors, and there are many possible ways to combine them. The probability of a Type I error grows with the number of tests.

Many, Many Tests

It has become quite common in applications to perform *many thousands, even millions*, of simultaneous hypothesis tests.

Power is critical in these applications because the most interesting effects are usually at the edge of detection.



Plan

1. The Multiple Testing Problem
   - Error Criteria and Power
   - False Discovery Control and the BH Method
2. Why BH Works
   - A Useful Model
   - A Stochastic Process Perspective
   - Performance Characteristics
3. Variations on BH
   - Improving Power
   - Dependence
   - Alternative Formulations

Plan

1. The Multiple Testing Problem
   - Error Criteria and Power
   - False Discovery Control and the BH Method
2. Why BH Works
   - A Useful Model
   - A Stochastic Process Perspective
   - Performance Characteristics
3. Variations on BH
   - Improving Power
   - Dependence
   - Alternative Formulations

The Multiple Testing Problem

- Perform m simultaneous hypothesis tests with a common procedure.
- For any given procedure, classify the results as follows:

	H_0 Retained	H_0 Rejected	Total
H_0 True	TN	FD	T_0
H_0 False	FN	TD	T_1
Total	N	D	m

Mnemonics: T/F = True/False, D/N = Discovery/Nondiscovery

All quantities except m, D, and N are *unobserved*.

- The problem is to choose a procedure that balances the competing demands of sensitivity and specificity.

How to Choose a Threshold?

- Control Per-Comparison Type I Error (PCER)
  - a.k.a. "uncorrected testing," many type I errors
  - Gives $\mathbb{P}\{FD_i > 0\} \leq \alpha$ marginally for all $1 \leq i \leq m$

- Control Familywise Type I Error (FWER)
  - e.g.: Bonferroni: use per-comparison significance level α/m
  - Guarantees $\mathbb{P}\{FD > 0\} \leq \alpha$

- Control False Discovery Rate (FDR)
  - first defined by Benjamini & Hochberg (BH, 1995, 2000)
  - Guarantees FDR $\equiv \mathbb{E}\left(\frac{FD}{D}\right) \leq \alpha$

- ...

A Practical Problem

- 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.

- FDR control offers a way to increase power while maintaining some principled bound on error.

  It is based on the assessment that

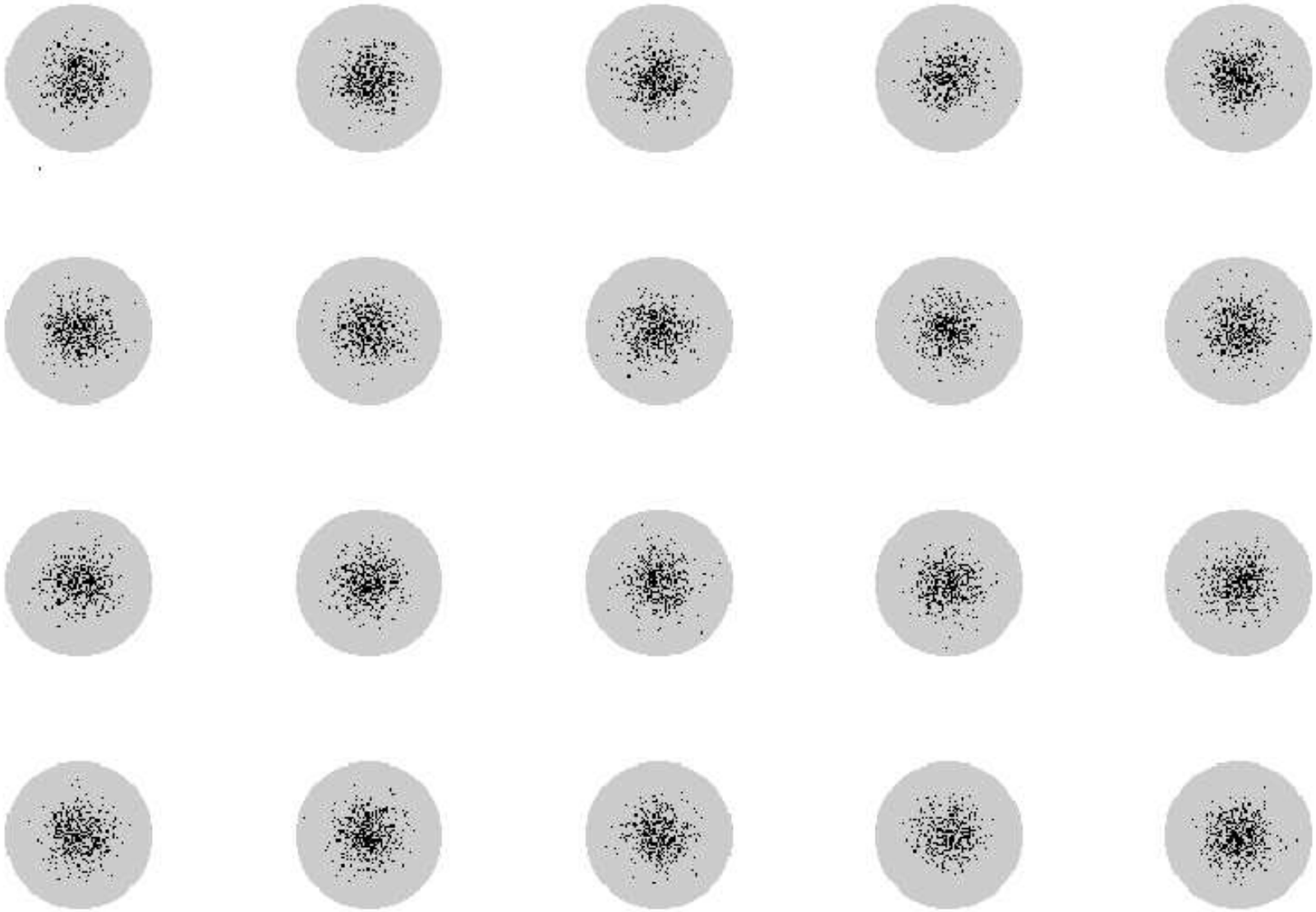
  4 false discoveries out of 10 rejected null hypotheses

  is a more serious error than

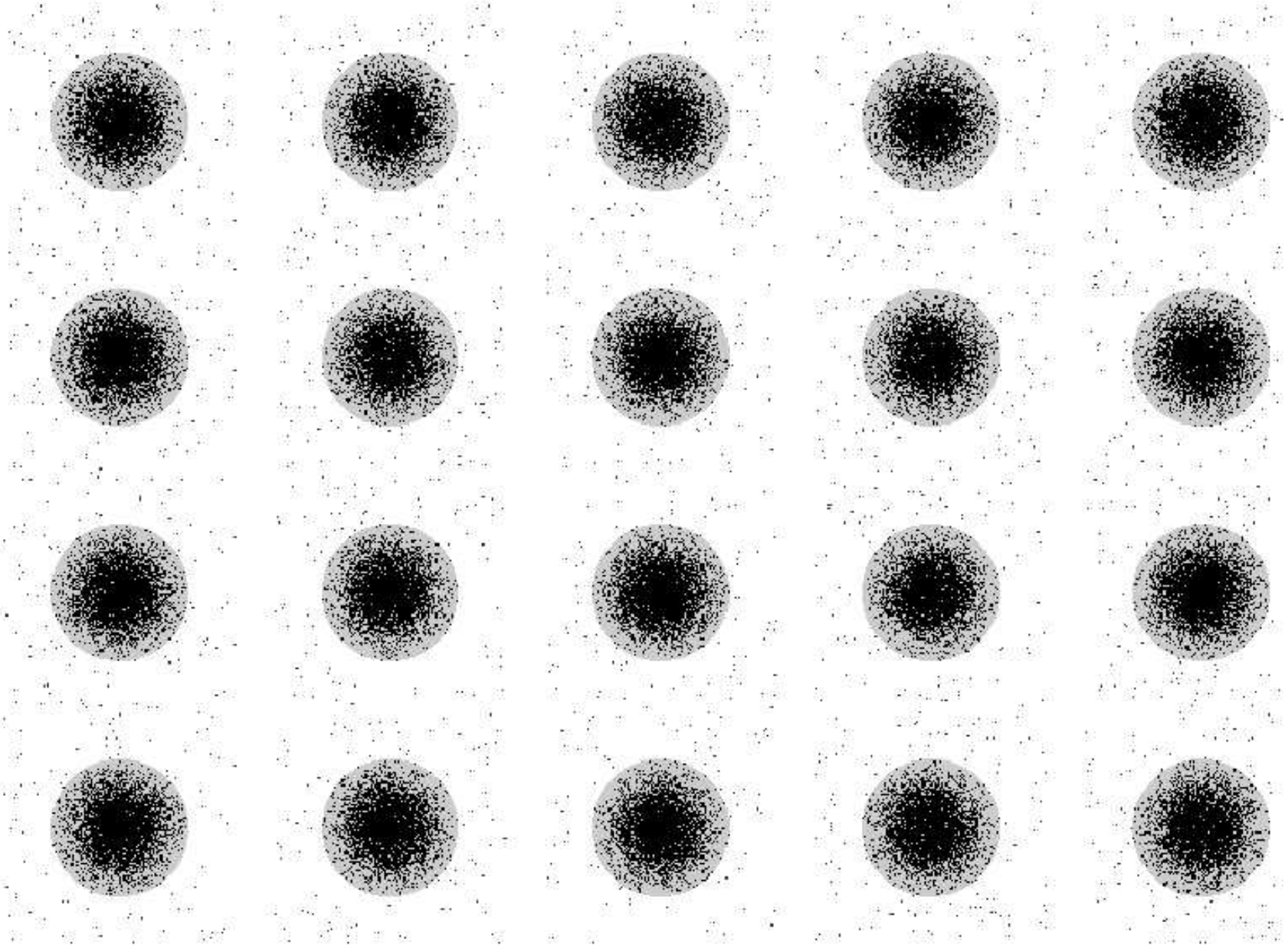
  20 false discoveries out of 100 rejected null hypotheses.

- A simple illustration ...

FWER Control

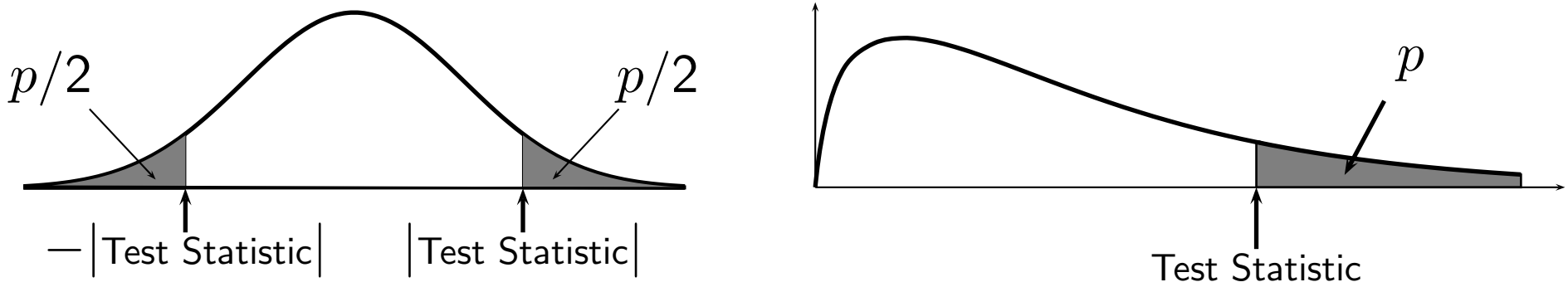


FDR Control



Recurring Notation

- Define p-values $P^m = (P_1, \ldots, P_m)$ for the m tests.



- Let $P_{(0)} \equiv 0$ and order the p-values

$$P_{(0)} = 0 < P_{(1)} < \cdots < P_{(m)}.$$

- Define hypothesis indicators $H^m = (H_1, \ldots, H_m)$, where $H_i = 0$ when the ith null hypothesis is true and $H_i = 1$ when the ith alternative is true.
- A multiple testing threshold T is a map $[0,1]^m \to [0,1]$, where we reject each null hypothesis with $P_i \leq T(P^m)$.

The False Discovery Rate

- Define the False Discovery Proportion (FDP) to be the (unobserved) *proportion of false discoveries among total rejections*.

  As a function of threshold t (and implicitly P^m and H^m), write this as

$$\text{FDP}(t) = \frac{\sum_i 1\Big\{P_i \leq t\Big\}\,(1 - H_i)}{\sum_i 1\Big\{P_i \leq t\Big\} + 1\Big\{\text{all } P_i > t\Big\}} = \frac{\#\text{False Discoveries}}{\#\text{Discoveries}}$$

- The False Discovery Rate (FDR) for a multiple testing threshold T is defined as the expected FDP using that procedure:

$$\text{FDR} = \mathbb{E}\Big(\text{FDP}(T)\Big)\,.$$

Aside: The False *Non*-Discovery Rate

- We can define a dual quantity to the FDR, the False Nondiscovery Rate (FNR).
- Begin with the False Nondiscovery Proprotion (FNP): the proportion of missed discoveries among those tests for which the null is retained.

$$\text{FNP}(t) = \frac{\sum_i 1\{P_i > t\} H_i}{\sum_i 1\{P_i > t\} + 1\{\text{all } P_i \leq t\}} = \frac{\#\text{False Nondiscoveries}}{\#\text{Nondiscoveries}}$$

- Then, the False Nondiscovery Rate (FNR) is given by

$$\text{FNR} = \mathbb{E}\left(\text{FNP}(T)\right).$$

The Benjamini-Hochberg Procedure

- Benjamini and Hochberg (1995) introduced the FDR and show that a procedure of Eklund, and independently Simes (1986), controls it.

  The procedure – which I'll call the BH procedure – is simple to compute but at first appears somewhat mysterious.

- The BH threshold is defined for pre-specified $0 < \alpha < 1$ as

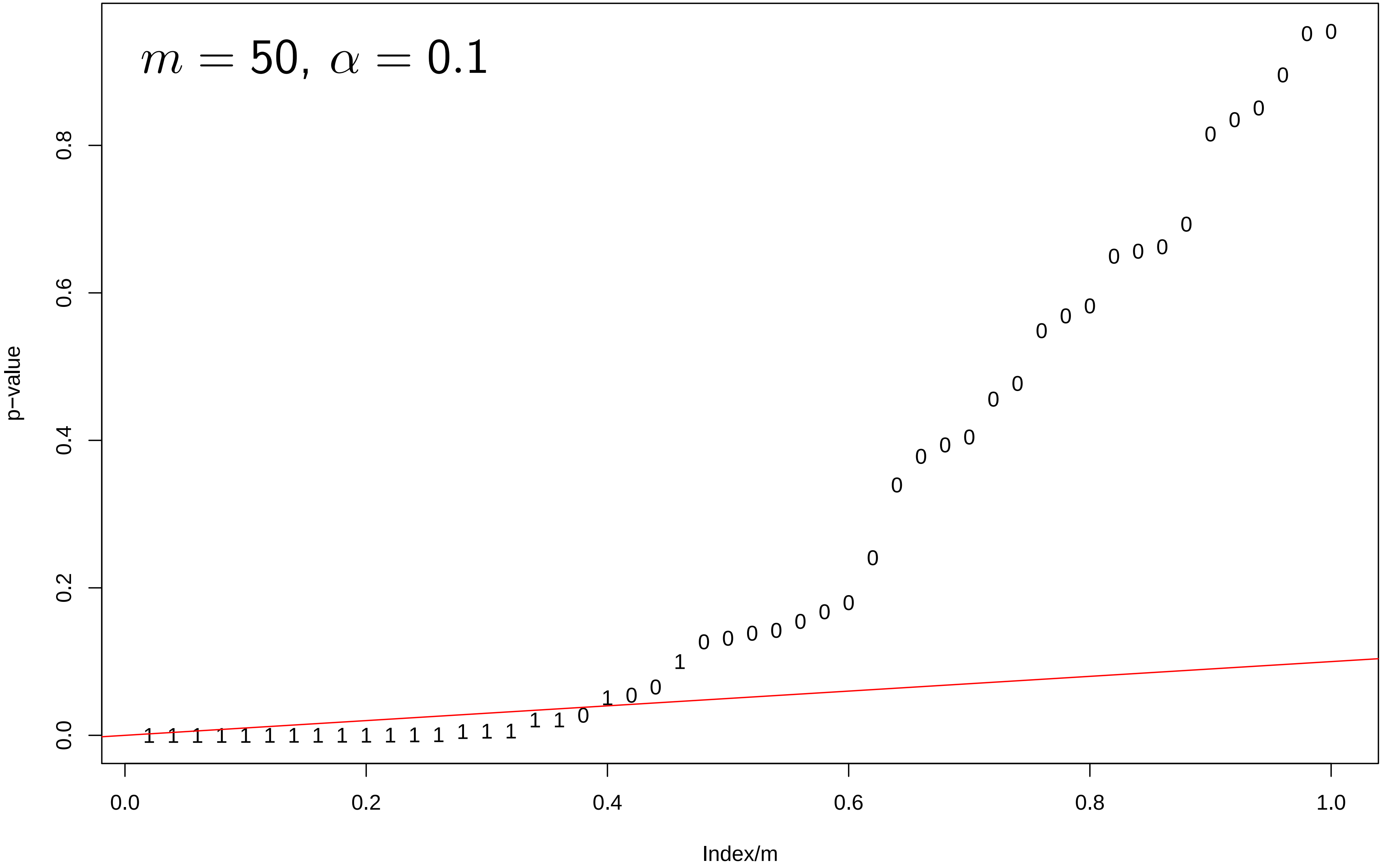
$$T_{\mathrm{BH}} = \mathsf{max}\left\{ P_{(i)}\colon \;\; P_{(i)} \leq \alpha \frac{i}{m},\; 0 \leq i \leq m \right\}.$$

- BH (1995) proved (for independent tests) that using this procedure guarantees – *for any alternative distributions* – that

$$\mathsf{FDR} \equiv \mathbb{E}\left(\mathsf{FDP}(T_{\mathrm{BH}})\right) \leq \frac{T_0}{m}\,\alpha$$

  where the inequality is an equality with continuous p-value distributions.

The Benjamini-Hochberg Procedure (cont'd)



Plan

1. The Multiple Testing Problem
   - Error Criteria and Power
   - False Discovery Control and the BH Method
2. Why BH Works
   - A Useful Model
   - A Stochastic Process Perspective
   - Performance Characteristics
3. Variations on BH
   - Improving Power
   - Dependence
   - Alternative Formulations

A Useful Mixture Model

- The following model is helpful for understanding and analyzing BH and its variants:

$$
\begin{aligned}
H_1, \ldots, H_m &\text{ iid Bernoulli}\langle a \rangle \\
\Xi_1, \ldots, \Xi_m &\text{ iid } \mathcal{L}_{\mathcal{F}} \\
P_i \mid H_i = 0, \Xi_i = \xi_i &\sim \text{Uniform}\langle 0, 1 \rangle \\
P_i \mid H_i = 1, \Xi_i = \xi_i &\sim \xi_i.
\end{aligned}
$$

  where $\mathcal{L}_{\mathcal{F}}$ denotes a probability distribution on a class $\mathcal{F}$ of distributions on $[0, 1]$.

- Typical examples for the class $\mathcal{F}$:
  - Parametric family: $\mathcal{F}_\Theta = \{F_\theta\colon \ \theta \in \Theta\}$
  - Concave, continuous distributions

$$\mathcal{F}_C = \{F\colon \ F \text{ concave, continuous cdf with } F \geq U\}.$$

A Useful Mixture Model (cont'd)

- Under this model, the m p-values $P^m = (P_1, \ldots, P_m)$ are *marginally* IID from

$$G = (1 - a)U + aF,$$

where: 1. $0 \leq a \leq 1$ is the frequency of alternatives,

2. U is the Uniform$\langle 0, 1 \rangle$ cdf, and

3. $F = \int \xi \, d\mathcal{L}_{\mathcal{F}}(\xi)$ is a distribution on [0,1].

- The marginal alternative distribution F comes up again and again, but its use does not preclude having different alternatives for different tests.
- Although the model posits IID Bernoulli$\langle a \rangle$ H_is, all the theory carries through with fixed H_is as well.

BH Revisited

Let's use this model to understand FDR and BH.

At any fixed threshold t, we have

$$\begin{aligned}
\mathsf{FDR}(t) &= \mathbb{E}\,\frac{\sum_i 1\{P_i \le t\}(1-H_i)}{\sum_i 1\{P_i \le t\} + 1\{\text{all } P_i > t\}} \\
&\approx \frac{\mathbb{E}\,\frac{1}{m}\sum_i 1\{P_i \le t\}(1-H_i)}{\mathbb{E}\,\frac{1}{m}\sum_i 1\{P_i \le t\} + \frac{1}{m}\mathbb{P}\{\text{all } P_i > t\}} \\
&= \frac{(1-a)t}{G(t) + \frac{1}{m}(1-G(t))^m} \approx \frac{(1-a)t}{G(t)}.
\end{aligned}$$

BH Revisited (cont'd)

Now, let

$$\hat{G}_m(t) = \frac{1}{m} \sum_i 1\Big\{P_i \le t\Big\}$$

be the empirical cdf of P^m.

In the continuous case, we can ignore ties, so $\hat{G}_m(P_{(i)}) = \frac{i}{m}$.

BH is thus equivalent to the following:

$$\begin{aligned} T_{\mathrm{BH}}(P^m) &= \sup\left\{t\colon \ t \le \alpha \hat{G}_m(t)\right\} \\ &= \sup\left\{t\colon \ \hat{G}_m(t) = \frac{t}{\alpha}\right\} \\ &= \sup\left\{t\colon \ \frac{t}{\hat{G}_m(t)} = \alpha\right\}. \end{aligned}$$

BH Revisited (cont'd)

One can think of this in two ways.

First, the BH procedure equates estimated FDR to the target α.

This estimator,

$$\widehat{\mathsf{FDR}}(t) = \frac{t}{\widehat{G}_m(t)},$$

uses $\widehat{G}_m$ in place of G and $\widehat{a} \equiv 0$ in place of a.

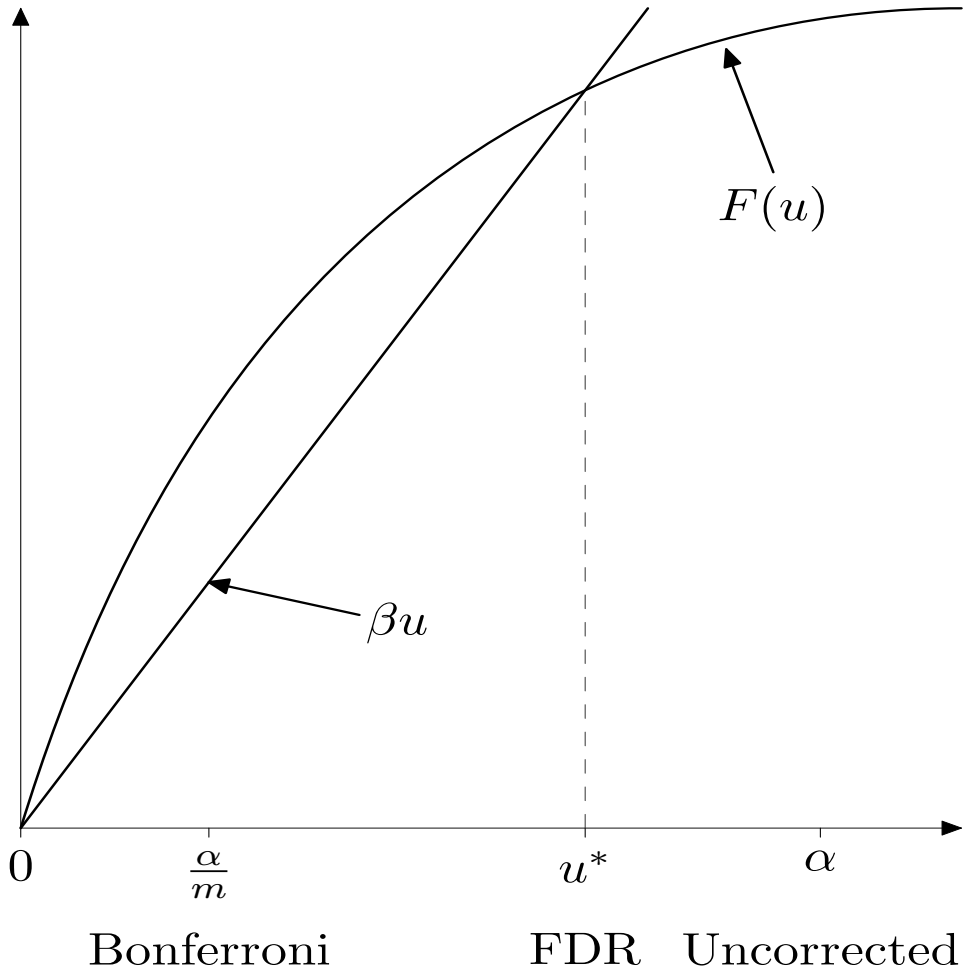
Second, the BH threshold is a plug-in estimator of

$$\begin{aligned} u^*(a, G) &= \mathsf{max}\left\{t\colon \;\; G(t) = \frac{t}{\alpha}\right\} \\ &= \mathsf{max}\left\{t\colon \;\; F(t) = \beta t\right\}, \end{aligned}$$

where $\beta = (1 - \alpha + \alpha a)/\alpha a$.

Asymptotic Behavior of BH Procedure

This yields the following picture:



BH Performance

- BH generally gives more power than FWER control and fewer Type I errors than uncorrected testing.

- BH performs best in very sparse cases ($T_0 \approx m$).

  For example, under the mixture model and in the continuous case,

  $$\mathbb{E}(\text{FDP}(T_{\text{BH}})) = (1 - a)\alpha.$$

  The BH procedure thus *overcontrols* FDR and thus will not in general minimize FNR.

- Power can be improved in non-sparse cases by more complicated adaptive procedures.

BH Performance (cont'd)

- When all m null hypotheses are true, BH is equivalent to FWER control.

- The BH FDR bound holds for certain classes of dependent tests, as we will see.

  In practice, it is quite hard to "break".

- $D \cdot \alpha$ need not bound the number of false discoveries.

  This is a common misconception for end users.

Operating Characteristics of the BH Method

- Define the misclassification risk of a procedure T by

$$R_M(T) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E} \left| 1\Big\{ P_i \leq T(P^m) \Big\} - H_i \right|.$$

  This is the average fraction of errors of both types.

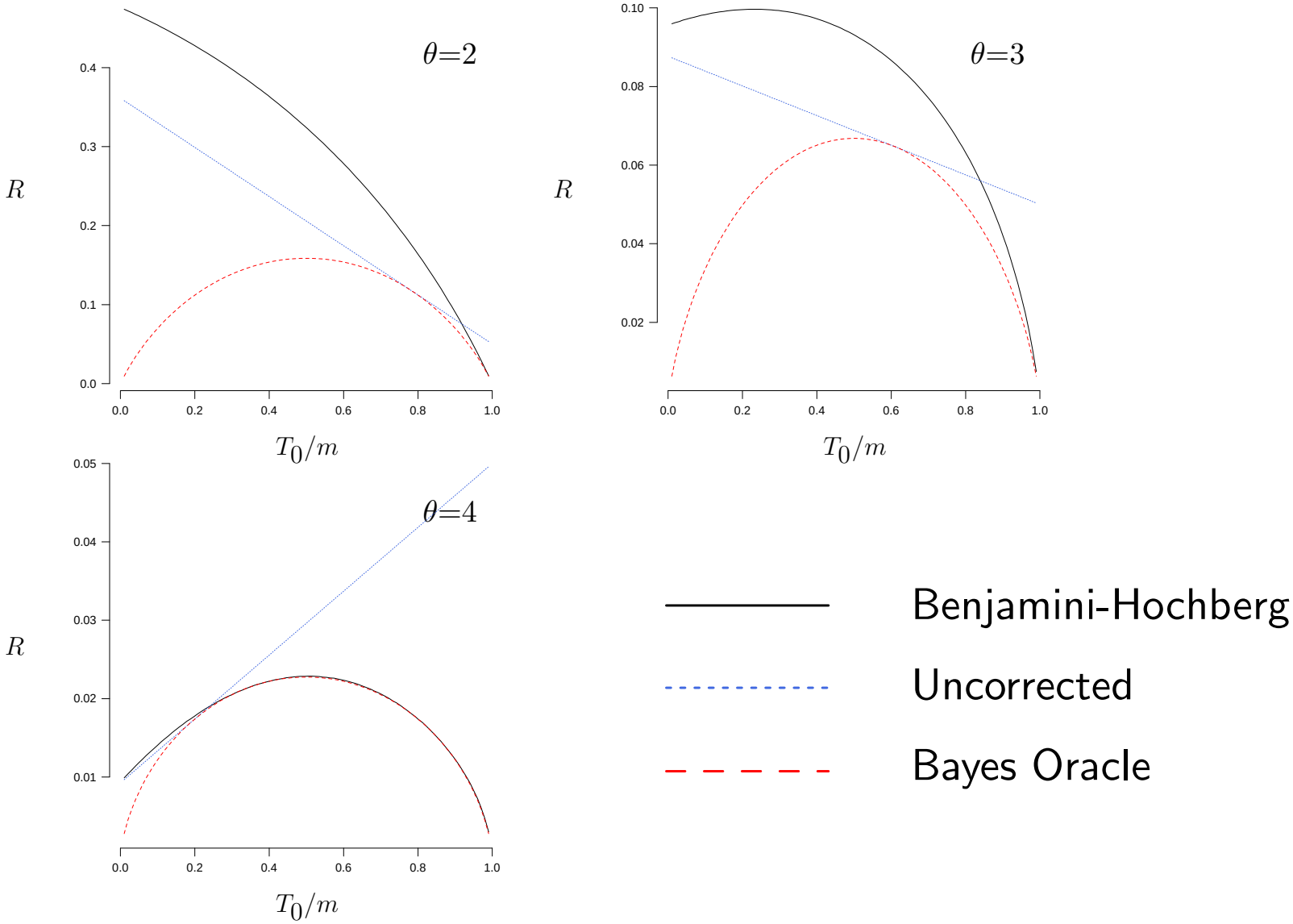
- Then $R_M(T_{\mathrm{BH}}) \sim R(a, F)$ as $m \to \infty$, where

$$R(a, F) = (1-a)u^* + a(1 - F(u^*)) = (1-a)u^* + a(1 - \beta u^*).$$

- Compare this to Uncorrected and Bonferroni and the Bayes' oracle rule $T_{\mathrm{BO}}(P^m) = b$ where b solves $f(b) = (1-a)/a$.

$$\begin{aligned} R_M(T_{\mathrm{U}}) &= (1-a)\,\alpha \;\; + \; a\,(1 - F(\alpha)) \\ R_M(T_{\mathrm{B}}) &= (1-a)\,\frac{\alpha}{m} \; + \; a\left(1 - F\left(\frac{\alpha}{m}\right)\right) \\ R_M(T_{\mathrm{BO}}) &= (1-a)\,b \;\; + \; a\,(1 - F(b))\,. \end{aligned}$$

Normal$\langle\theta, 1\rangle$ Model, $\alpha = 0.05$



FDP and FNP as Stochastic Processes

- Both the FDP(t) and FNP(t) stochastic processes converge to Gaussian processes outside a neighborhood of 0 and 1 respectively.
- For example, define

$$Z_m(t) = \sqrt{m}\,(\mathrm{FDP}(t) - Q(t))\,, \quad \delta \le t \le 1,$$

  where $0 < \delta < 1$ and $Q(t) = (1-a)U/G$.
- Let Z be a mean 0 Gaussian process on $[\delta, 1]$ with covariance kernel

$$K(s,t) = a(1-a)\,\frac{(1-a)stF(s\wedge t) \,+\, aF(s)F(t)(s\wedge t)}{G^2(s)\,G^2(t)}.$$

- Then, $Z_m \rightsquigarrow Z$.

Plan

1. The Multiple Testing Problem
   - Error Criteria and Power
   - False Discovery Control and the BH Method
2. Why BH Works
   - A Useful Model
   - A Stochastic Process Perspective
   - Performance Characteristics
3. Variations on BH
   - Improving Power
   - Dependence
   - Alternative Formulations

Optimal Thresholds

- The equality

$$\mathbb{E}(\mathsf{FDP}(T_{\mathrm{BH}})) = (1-a)\alpha$$

  implies that if we knew a, we could improve power by applying BH at level $\alpha/(1-a)$.

- This suggests using T_{PI}, the plug-in estimator for

$$\begin{aligned} t^*(a, G) &= \mathsf{max}\left\{t\colon \;\; G(t) = \frac{(1-a)t}{\alpha}\right\} \\ &= \mathsf{max}\left\{t\colon \;\; F(t) = (\beta - 1/\alpha)t\right\}, \end{aligned}$$

  where $\beta - 1/\alpha = (1-a)(1-\alpha)/a\alpha$.

- Note that $t^* \geq u^*$.

Optimal Thresholds (cont'd)

- For each $0 \leq t \leq 1$,

$$\mathbb{E}(\mathsf{FDP}(t)) = \frac{(1-a)\,t}{G(t)} + O\left((1-t)^m\right)$$

$$\mathbb{E}(\mathsf{FNP}(t)) = a\,\frac{1-F(t)}{1-G(t)} + O\left((a+(1-a)t)^m\right).$$

- Ignore $O()$ terms and choose t to minimize $\mathbb{E}(\mathsf{FNP}(t))$ subject to $\mathbb{E}(\mathsf{FDP}(t)) \leq \alpha$.

  This yields $t^*(a, G)$ as the optimal threshold.

- Genovese and Wasserman (2002) show that

$$\mathbb{E}(\mathsf{FDP}(t^*(\widehat{a}, \widehat{G}))) \leq \alpha + O(m^{-1/2})$$

  under weak conditions on $\widehat{a}$.

Improving Power

- In practice, the main difficulty here is finding a good estimator of $1 - a$, or alternatively, a good estimator of T_0.

  Part of the challenge is guaranteeing FDR control with the increased variability induced by the estimator.

- Adaptive estimators for improving power in FWER-controlling methods go back to Schweder and Spjotvol (1982) and Hochberg and Benjamini (1990).

Improving Power (cont'd)

- Benjamini and Hochberg (2000) introduced the idea of using the BH procedure to estimate T_0.
  - Use BH at level α. If no rejections, stop.
  - Otherwise, define $\widehat{T}_{0,k} = \dfrac{m+1-k}{1-P_{(k)}}$, for $k = 1, \ldots, m$.
  - Find first $k^* \geq 2$ such that $\widehat{T}_{0,k} > \widehat{T}_{0,k-1}$.
  - Estimate $\widehat{T}_0 = \min(m, \lceil \widehat{T}_{0,k^*} \rceil)$.
  - Use BH at level $\alpha' = \alpha m / \widehat{T}_0$.
- Here, the intermediate estimators $\widehat{T}_{0,k}$ are derived from the number of rejections at fixed threshold $P_{(k)}$, adjusted for the expected $T_0 \cdot P_{(k)}$ false rejections.
- This procedure controls FDR and has good power under independence.

Improving Power (cont'd)

- Storey (2002) gave an alternative adaptive procedure that uses

$$\widehat{1-a} = \frac{1-\widehat{G}(\lambda)}{1-\lambda},$$

for some fixed λ, often $\lambda = 1/2$. The rationale for this estimator is that most of the p-values near 1 should be null, implying $1 - G(\lambda) \approx (1-a)(1-\lambda)$.

- Storey et al. (2003) modified this estimator for theoretical reasons to

$$\widehat{1-a} = \frac{1+\frac{1}{m}-\widehat{G}(\lambda)}{1-\lambda},$$

with the proviso that only nulls with $P_{(i)} \leq \lambda$ can be rejected.

- With this modification, this procedure tends to have higher power than BH2000 *under independence*.

Improving Power (cont'd)

- Genovese and Wasserman (2002) show that this procedure controls FDR asymptotically.

  Storey et al. (2003) show by a nice martingale argument that it controls FDR for a finite number of independent tests.

  They also extended it to a particular form of dependence among the tests.

- Efron et al. (2001) considered a variant with λ set to the median p-value.

  This was motivated primarily toward computing their empirical Bayes local FDR.

Improving Power (cont'd)

- Benjamini, Krieger, and Yekutieli (BKY, 2004) give a comprehensive numerical comparison of adaptive procedures and introduce new procedures, with an elegant proof of FDR control.
- Their two stage method is as follows:
  - Use BH at level β_1. Let r_1 be the number of rejected null hypotheses.
  - If $r_1 = 0$, stop.
  - Otherwise, let $\widehat{T}_0 = m - r_1$.
  - Use BH at level $\alpha' = \beta_2 m / \widehat{T}_0$.
- The initial procedure takes $\beta_1 = \beta_2 = \alpha/(1+\alpha)$, but they also have success with $\beta_1 = \alpha$ and $\beta_2 = \alpha/(1+\alpha)$.
- This method has good power and remains valid under certain kinds of dependence, as we will see.

Dependence

- Benjamini and Yekutieli (2001) show that the original BH method still controls FDR at the nominal level even for dependent tests.

  In particular: under *positive regression dependence on a subset*.

  While this is a somewhat technical condition, a simple case is Gaussian variables with totally positive covariance matrix.

- Under the most general dependence structure, the BH method controls FDR at level

$$\alpha \frac{T_0}{m} \sum_{i=1}^{m} \frac{1}{i}.$$

  Thus, a distribution-free procedure for FDR control is to apply BH at level $\alpha / \sum_{i=1}^{m} \frac{1}{i}$. Unfortunately, this is typically very conservative, sometimes even more so than Bonferroni.

- Practically speaking, BH is quite hard to break even beyond what as been proven.

Dependence (cont'd)

- The challenge of dependence for adaptive procedures is finding an estimator of $1 - a$ or T_0 that performs well under various dependence structures.

  This turns out to be far from easy.

- Storey et al. (2003) show that their procedure controls FDR asymptotically (as $m \to \infty$) under a weak (ergodic) dependence structure.

  BKY (2004) argue, however, that this does not include many cases of practical interest, including equally correlated test statistics.

  Although Storey et al. (2003) see little effect of dependence on the validity of their procedure in simulations, BKY (2004) find that the FDR can be almost double the bound even under positive dependence.

Dependence (cont'd)

- The BH (2000) procedure also loses FDR control with increasing (positive) dependence, though less seriously.

- BKY (2004) show that their two stage procedure continues to control FDR under positive dependence.

  They argue, convingingly, that this is the best option when the degree of dependence is unknown.

- There are also advantages to be explored in using the estimated dependence structure itself to improve performance.

pFDR and Bayesian Connections

- Storey (2001) considers the "positive FDR," defined by

$$\text{pFDR}(t) = \mathbb{E}\left(\text{FDP}(t) \,\middle|\, D(t) > 0\right).$$

  Note that $\text{FDR}(t) = \text{pFDR}(t) \cdot \mathbb{P}\Big\{D(t) > 0\Big\} \leq \text{pFDR}(t)$.

- Storey (2001) makes a nice Bayesian connection.

  Taking a under the mixture model to be the prior probability that a null hypothesis is false, it follows taht

$$\text{pFDR}(t) = \frac{(1-a)t}{G(t)} = \frac{(1-a)\mathbb{P}\Big\{P \leq t \mid H = 0\Big\}}{\mathbb{P}\Big\{P \leq t\Big\}} = \mathbb{P}\Big\{H = 0 \mid P \leq t\Big\}.$$

- Storey (2003) also introduces the *q-value* as the minimum pFDR for which the given statistic is rejected.

  This has a Bayesian interpretation as a "posterior Bayesian p-value".

EBT (Empirical Bayes Testing)

- Efron et al (2001) construct an empirical Bayes measure of "local FDR". They note that

$$\mathbb{P}\Big\{H_i = 0 \mid P^m\Big\} = \frac{(1-a)}{g(P_i)} \equiv q(P_i),$$

  where $g = G'$.

- This suggests a rejection rule $q(p) \leq \alpha$.
- If $f = F'$, then for a, f unknown, $f \geq 0$ implies that

$$a \geq 1 - \min_p g(p) \Longrightarrow \widehat{a} = 1 - \min_p \widehat{g}(p).$$

- Then,

$$\widehat{q}(p) = \frac{1-\widehat{a}}{\widehat{g}(p)} = \frac{\min_s \widehat{g}(s)}{\widehat{g}(p)}$$

EBT (cont'd)

- If we reject when $\mathbb{P}\Big\{H_i = 0 \mid P^m\Big\} \leq \alpha$, how many errors are we making?

- Under weak conditions, can show that

$$q(t) \leq \alpha \text{ implies } Q(t) < \alpha$$

  So EBT is conservative.

- Because g is in general unknown, we need to estimate q, so the performance depends on the behavior of $\widehat{q}$.

EBT (cont'd)

- THEOREM. Let $\widehat{q}(t) = \frac{(1-a)}{\widehat{g}(t)}$. Suppose that

$$m^{\alpha}(\widehat{g}(t) - g(t)) \rightsquigarrow W$$

for some $\alpha > 0$, where W is a mean 0 Gaussian process with covariance kernel $\tau(v, w)$. Then

$$m^{\alpha}\ (\widehat{q}(t) - q(t)) \rightsquigarrow Z$$

where Z is a Gaussian process with mean 0 and covariance kernel

$$K_q(v, w) = \frac{(1-a)^2 \tau(v, w)}{g(v)^4 g(w)^4}.$$

EBT (cont'd)

- Parametric Case: $g \equiv g_\theta = (1-a) + a f_\theta(v)$ Then,

$$\text{rel}(v) = \frac{\widehat{\text{se}}(\widehat{q}(v))}{q(v)} \approx O\left(\frac{1}{\sqrt{m}}\right) \left|\frac{\partial \log g_\theta}{\partial d\theta}\right| = O\left(\frac{1}{\sqrt{m}}\right) |v-\theta| \quad \text{Normal case}$$

- Nonparametric Case

$$\widehat{g}(t) = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{h_m} K\left(\frac{t-P_i}{h_m}\right)$$

$h_m = cm^{-\beta}$ where $\beta > 1/5$ (undersmooth). Then

$$\text{rel}_v = \frac{c}{m^{(1-\beta)/2}\sqrt{g(v)}}.$$

Exceedance Control

- Genovese and Wasserman (2002, 2004) introduce the idea of "exceedance control" where we bould $\mathbb{P}\Big\{\mathsf{FDP} > \gamma\Big\}$ rather than $\mathbb{E}\,\mathsf{FDP}$.

  This is the subject of my later talk.

- van der Laan, Dudoit, and Pollard (2004) introduce FDR and exceedance controlling procedures based on "augmenting" a FWER-controlling test.

- Motivates the term "False Discovery Control" since we're no longer just controlling FDR.

Plan

1. The Multiple Testing Problem
   - Error Criteria and Power
   - False Discovery Control and the BH Method
2. Why BH Works
   - A Useful Model
   - A Stochastic Process Perspective
   - Performance Characteristics
3. Variations on BH
   - Improving Power
   - Dependence
   - Alternative Formulations

Take-Home Points

- False Discovery Control provides a useful alternative to traditional multiple testing methods.
- The BH method is fast and robust, but it overcontrols FDR.
- Good adaptive methods exist that can increase power.

  The BKY (2004) two stage method is recommended when (positive) dependence might be nontrivial.
- Many open problems remain, and alternative formulations – such as exceedance control – offer some advantages.

  Important problems include explicitly accounting for dependence and taking advantage of spatial structure in the alternatives.

Selected References

Abramovich, F., Benjamini, Y., Donoho, D. and Johnstone, I. (2000). Adapting to unknown sparsity by controlling the false discovery rate. Technical report 2000-19. Department of Statistics. Stanford University.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289-300.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational Behavior, Statistics*, **25**, 60–83.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29, 1165-1188.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96, 1151-1160.

Finner, H. and Roters, M. (2002). Multiple hypotheses testing and expected number of type I errors. *Annals of Statistics*, 30, 220–238.

Selected References (cont'd)

Genovese, C. R. and Wasserman, L. (2001). False discovery rates. Technical Report, Department of Statistics, Carnegie Mellon.

Genovese, C. R. and Wasserman, L. (2002). Operating characteristics and extensions of the FDR procedure. *Journal of the Royal Statistical Society B*, **64**, 499–518.

Genovese, C. R. and Wasserman, L. (2003). A stochastic process approach to false discovery control. *Annals of Statistics*, in press.

Harvönek & Chytil (1983). Mechanizing hypotheses formation – a way for computerized exploratory data analysis? *Bulletin of the International Statistical Institution*, **50**, 104–121.

Helperin, M., Lan, G.K.K., and Hamdy, M.I. (1988). Some implications of an alternative definition of the multiple comparison problem. *Biometrika*, **75**, 773–778.

Hochberg, Y. and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statistics in Medicine*, **9**, 811–818.

Hommel, G. and Hoffman, T. (1987) Controlled uncertainty. In P. Bauer, G. Hommel, and E. Sonnemann, (Eds.), *Multiple hypothesis testing* (pp. 154–161). Heidelberg: Springer.

Selected References (cont'd)

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Annals of Statistics*, 30, 239–257.

Schweder, T. and Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, **69**, 493–502.

Simes, J. R. 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73: 75–754.

Shafer, J. (1995). Multiple hypothesis testing. *Annual Reviews in Psychology*, **46**:561–584.

Storey, J. D. (2001). The positive False Discovery Rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, in press.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society B*, **64**, 479–498.

Storey, J.D., Taylor, J. E., and Siegmund, D. (2002). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society B*, in press.

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, **82**, 171–196.