

Controlling False Discovery Rate and Trials factors in searches

Jim Linnemann

MSU

Milagro

March 11, 2003

Thanks to:

- Slides from web:
 - T. Nichol **UMich**; C. Genovese CMU
 - Y. Benjamini Tel Aviv, S. Scheid, MPI
- Email advice and pointers to literature
 - C. Miller CMU Astro
 - B. Efron Stanford, J. Rice Berkeley Stat
 - Google

Outline

- What is significant enough to report?
 - Multiple Comparison Problem (trials)
- A Multiple Comparison Solution:
False Discovery Rate (FDR)
- FDR Properties
- FDR Example

Significance

- Define “wrong” as reporting false positive:
 - Apparent signal caused by background
- Set α a level of potential wrongness
 - $2 \sigma = .05$ $3 \sigma = .01$ etc.
 - Probability of going wrong on **one test**
 - Or, error rate per test

What if you do m tests?

- m is “trials factor” **only NE Jour Med demands!**
- Don’t want to just report m times as many signals!
 - $P(\text{at least one wrong}) = 1 - (1 - \alpha)^m$
- Use α / m as significance test “Bonferroni”
- *Keeps to α the probability of reporting 1 or more wrong on whole ensemble of m tests*
- **Good:** control publishing rubbish
- **Bad:** lower sensitivity (must have more obvious signal)
 - For some purposes, have we given up too much?

Bonferroni Who?

- *"Good Heavens! For more than forty years I have been speaking prose without knowing it."*

-Monsieur Jourdan in

"Le Bourgeoise Gentilhomme" by Moliere

I believe that translates to Jordan Goodman?

“Multiple Comparisons”

- Must Control False Positives
 - How to measure multiple false positives?
- Chance of *any* false positives in whole set
 - Jargon: Familywise Error Rate (FWER)
 - Control by Bonferroni, Bonferroni-Holm, & “Random Field Method” = ???
- False Discovery Rate (FDR)
 - Fraction of errors in signal candidates
 - Proportion of false positives *among* rejected tests

Background =
null hypothesis

False discoveries
(false positives)

b
s

	H_0 Retained	H_0 Rejected	Total
H_0 True	$N_{0 0}$	$N_{1 0}$	M_0
H_0 False	$N_{0 1}$	$N_{1 1}$	M_1
Total	$m - R$	R	m

inefficiency

$$\text{FDR} = \begin{cases} \frac{N_{1|0}}{R} & \text{if } R > 0, \\ 0, & \text{if } R = 0. \end{cases}$$

Detected signals
(true positives)

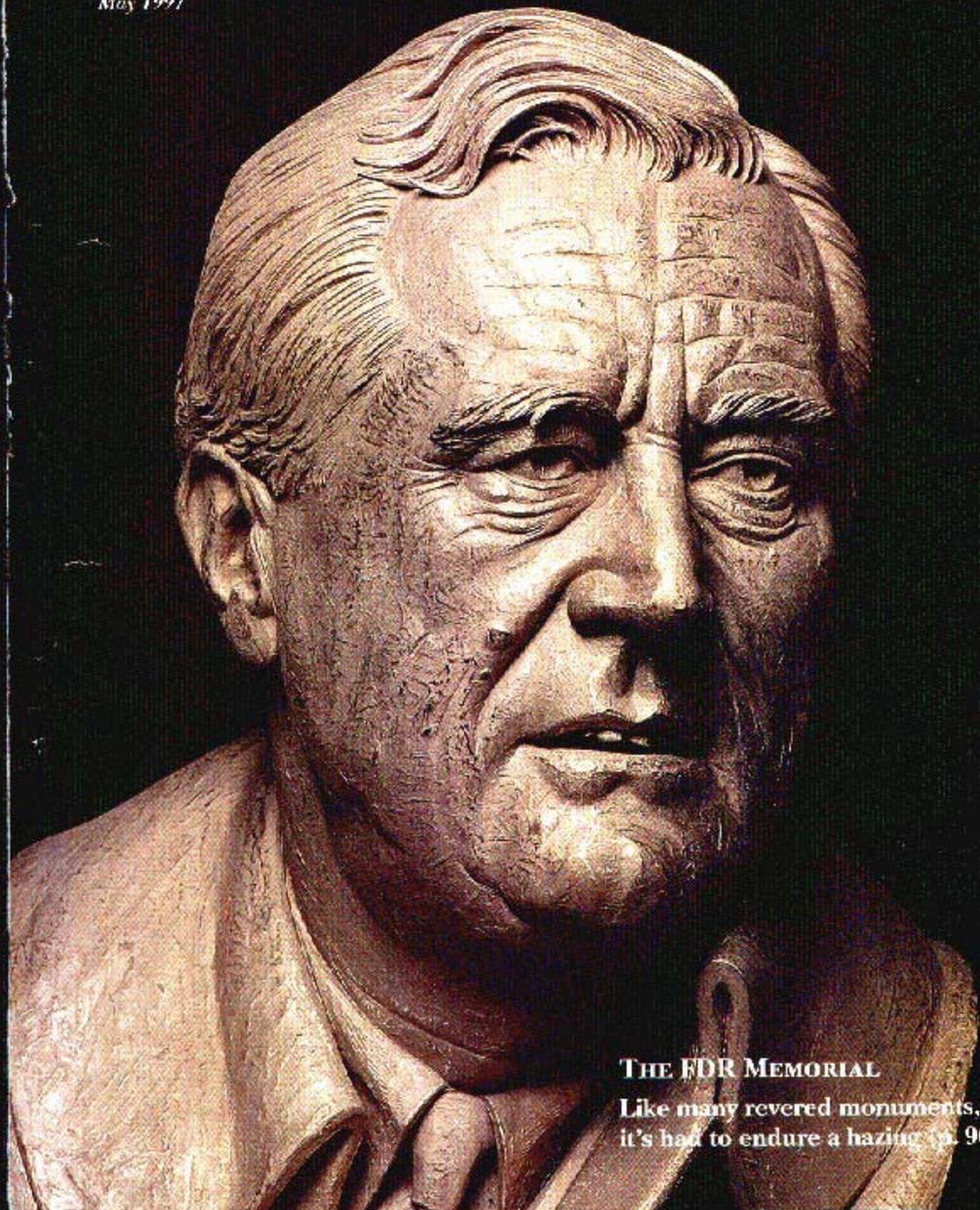
Reported signal
candidates
(rejected nulls)

Goals of FDR

- Tighter than α (single-test)
- Looser than α/m (trials factor/Bonferroni)
- Improve sensitivity (“power”)
- Still control something useful:
 - **fraction** of false results that you report
- **Catchy TLA**

Smithsonian

May 1997



THE FDR MEMORIAL

Like many revered monuments,
it's had to endure a hazing (p. 96)

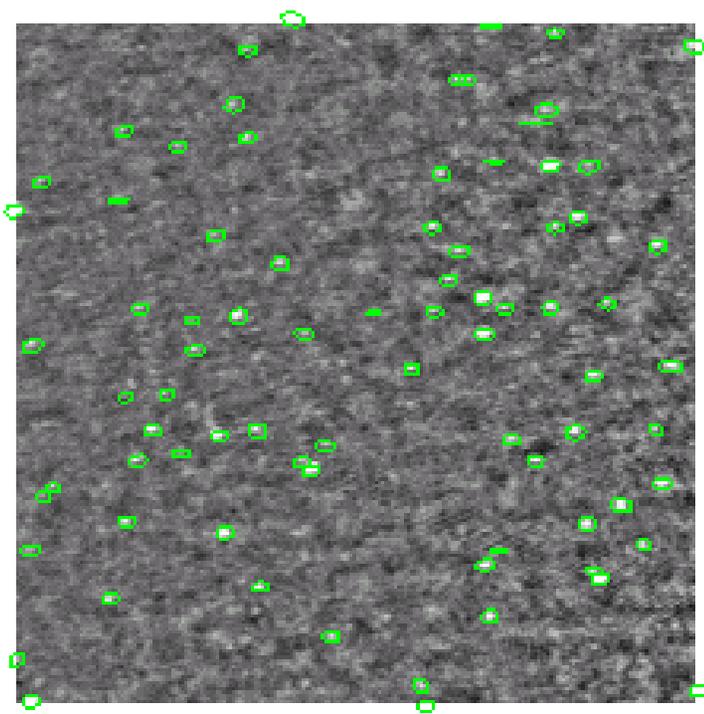
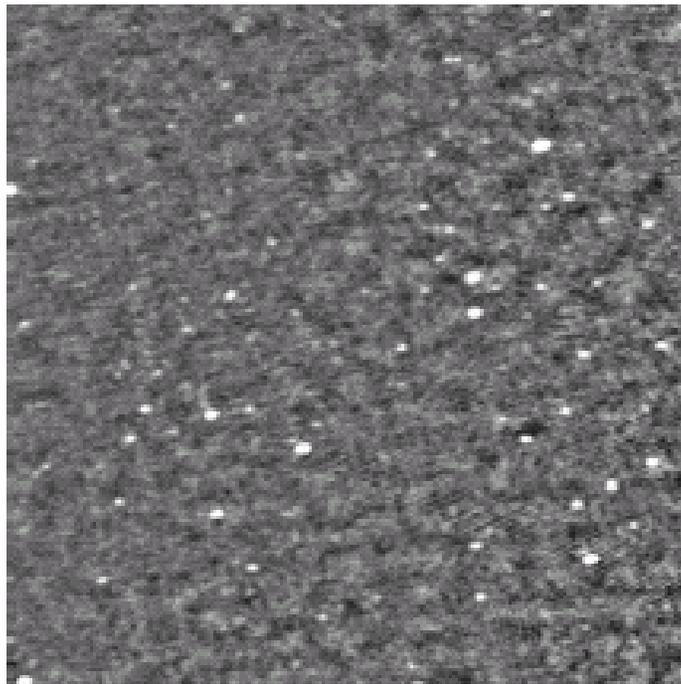
Where did this come from?

Others who have lots of tests!

- Screening of chemicals, drugs
- Genetic mapping
- Functional MRI (voxels on during speech processing)
- Data mining (cookies by milk? direct mail)
- Radio telescope images (at last some astronomy!)
- Common factors:
 - Usually expect some real effects
 - Can follow up by other means
 - trigger next phase with mostly real stuff

Motivating Example #2: Source Detection

- Interferometric radio telescope observations processed into digital image of the sky in radio frequencies.
- Signal at each pixel is a mixture of source and background signals.



FDR in High Throughput Screening

An interpretation of FDR:

$$\text{Exp}\left(\frac{\text{expenses wasted chasing "red herrings"}}{\text{expenses made on follow-up studies}}\right) \leq q$$

Our GRB alerts?

What is a p-value?

(Needed for what's next!)

- Crudely, probability that event produced by background (“null hypothesis”)
 - *significance* of result, measured in probability
 - Same as “sigmas”—different units, that’s all

P value properties: If all events are background

Distribution of p values = dn/dp should be flat
and have a linearly rising cumulative distribution

$$N(x) = \int_0^x dp (dn/dp) = x$$

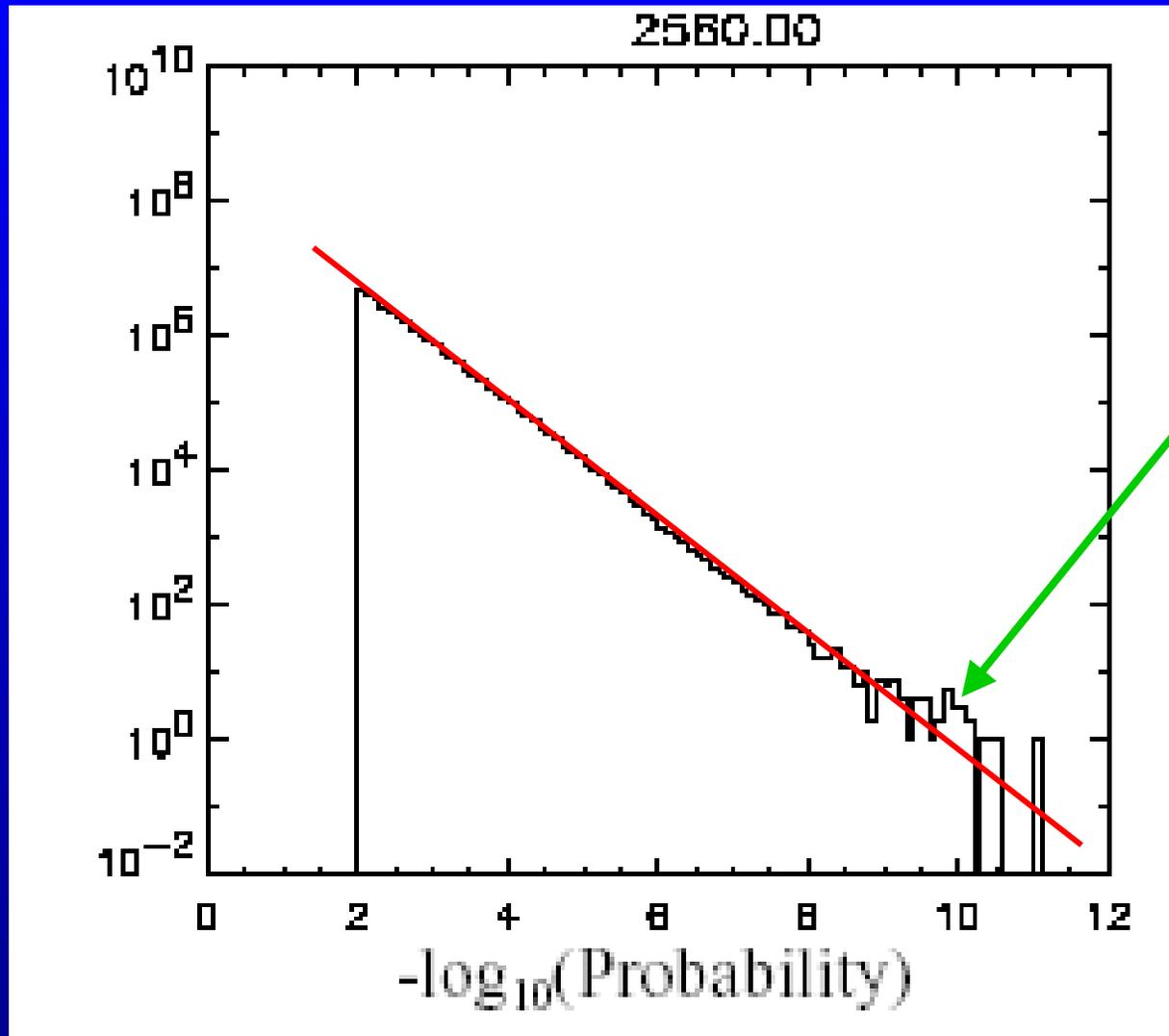
$$N(p \text{ in } [a, b]) = (b-a)$$

$$\text{So expect } N(p < p_k) = k/m$$

Flat also means linear in log-log: if $y = \ln p$
 $\ln[dn/dy]$ vs. y is a straight line

See figure 1 in GRB paper

From GRB
paper, fig 1



Signal,
statistics, or
systematics?

“Best” of 9 plots

Note: A histogram is a binned sorting of the p-values

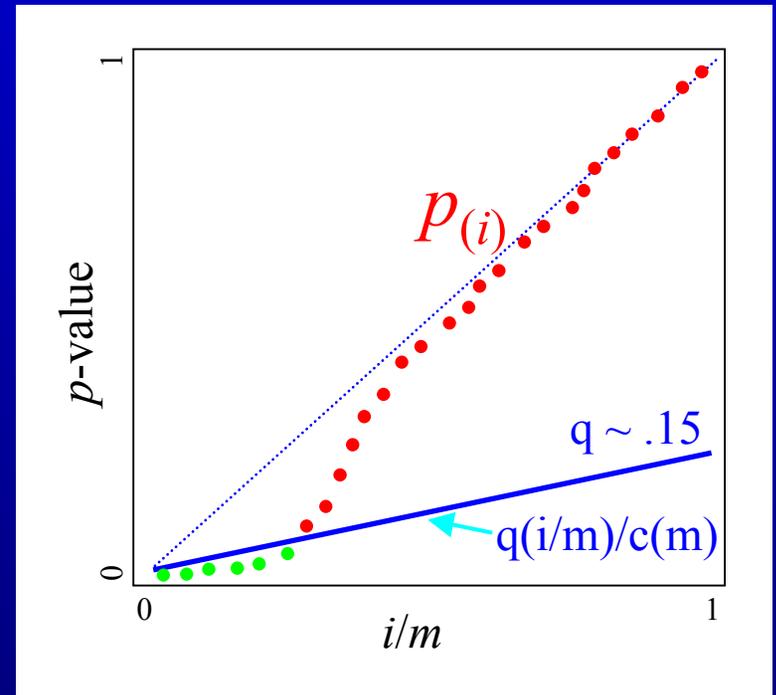
Benjamini & Hochberg

JRSS-B (1995) 57:289-300

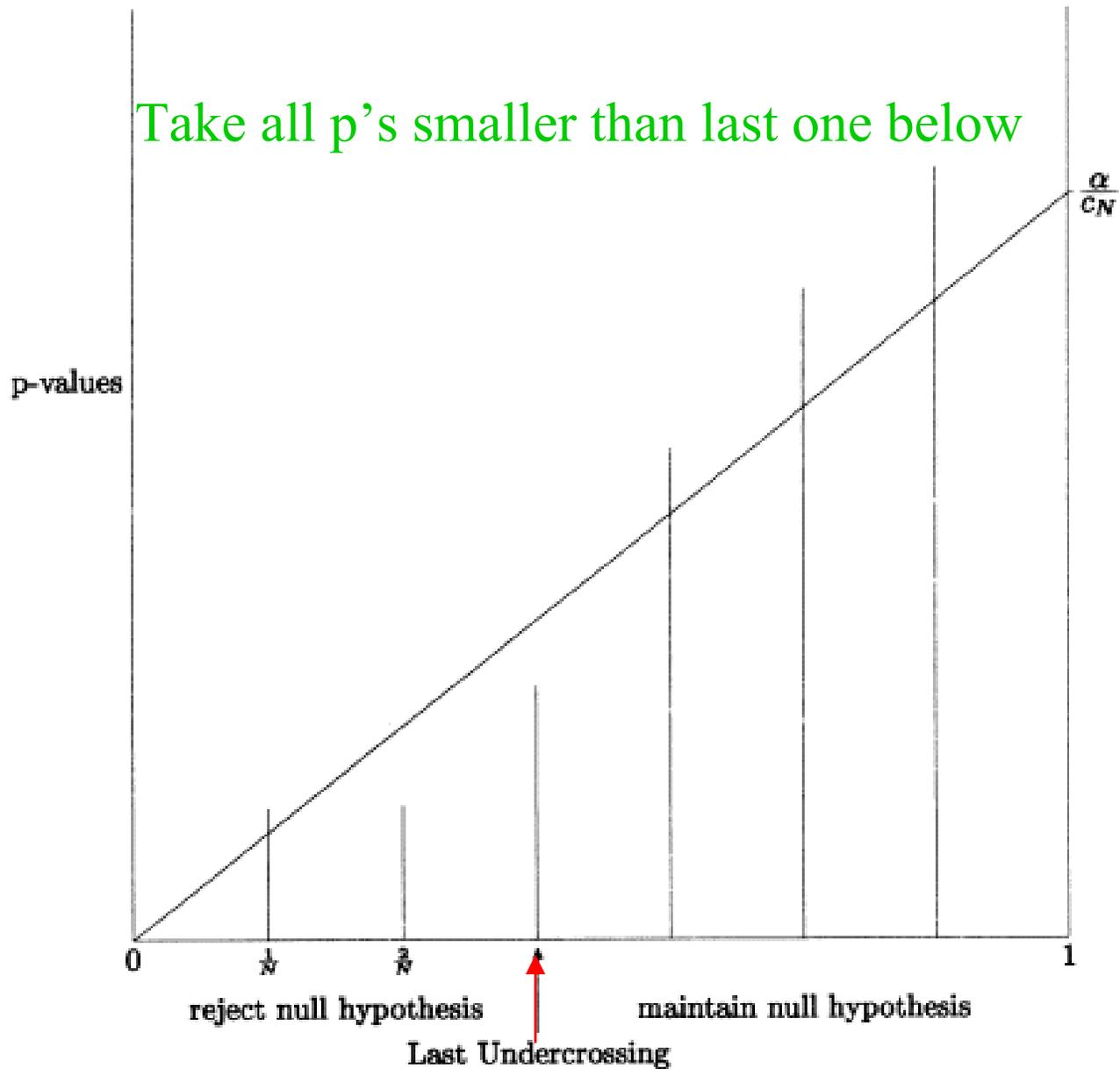
- Select desired limit q on Expectation(FDR)
 α is not specified!!
- Sort the p-values, $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$
- Let r be largest i such that

$$p_{(i)} \leq q(i/m)/c(m)$$

- Reject all hypotheses corresponding to $p_{(1)}, \dots, p_{(r)}$.
- *Proof this works is not obvious!*



Take all p's smaller than last one below



Comments on FDR

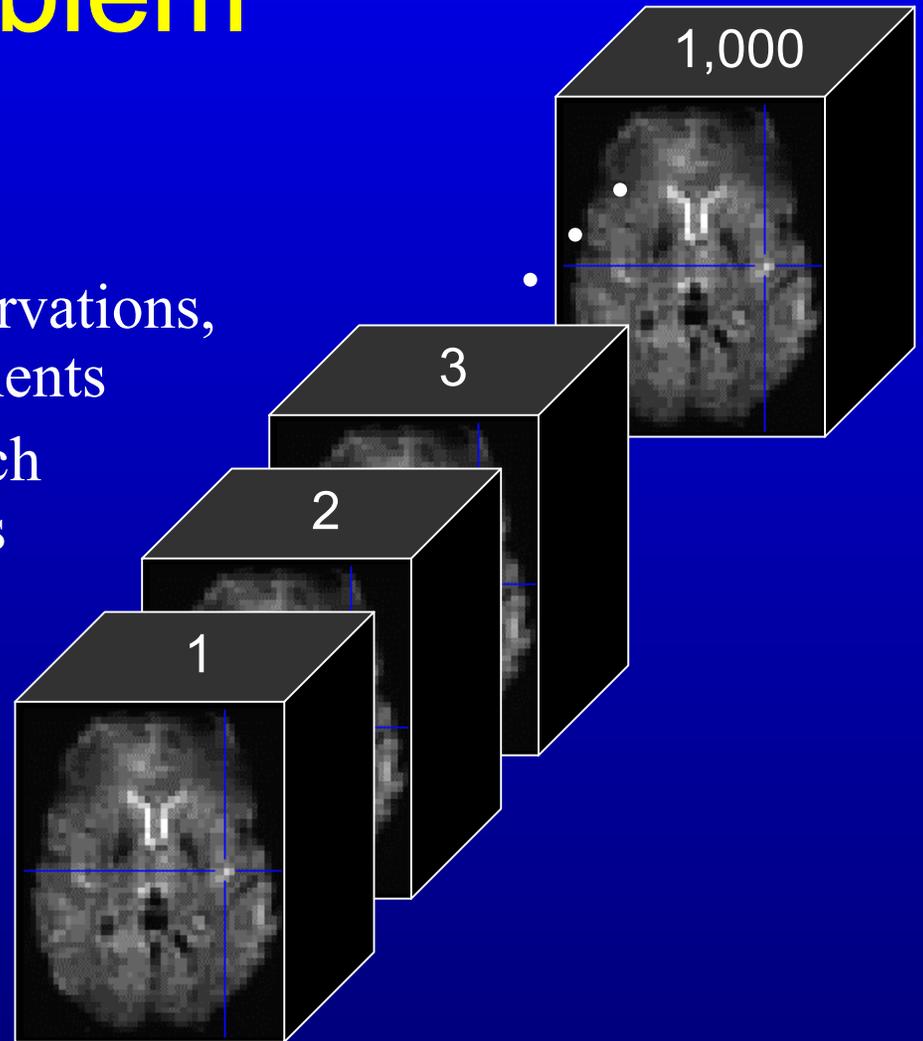
- To use method, you must *not really new!*
 - know trials factor
 - Be able to calculate small p values correctly
- Lowest p value $p_{(1)}$ always gets tested with q/m
- Even if $p_{(1)}$ fails, FDR allows other $p_{(i)}$ distorting the pure-null shape to raise the threshold and accept the $p_{(1)} \dots p_{(j)}$: you depend on **distribution**
- Suspect as $q \rightarrow 0$, FDR \rightarrow Bonferroni in q/m
- You can always quote both α/m and $q = \langle \text{FDR} \rangle$
 - Pick α ; run backwards: find q giving that α

Benjamini & Hochberg Procedure

- $c(m) = 1$
 - Positive Regression Dependency on Subsets
 - Technical condition, special cases include
 - Independent data
 - Multivariate Normal with all positive correlations
 - Result by Benjamini & Yekutieli, *Annals of Statistics*, in press.
- $c(m) = \sum_{i=1, \dots, m} 1/i \approx \log(m) + 0.5772$
 - Arbitrary covariance structure
 - But this is more conservative—tighter cuts

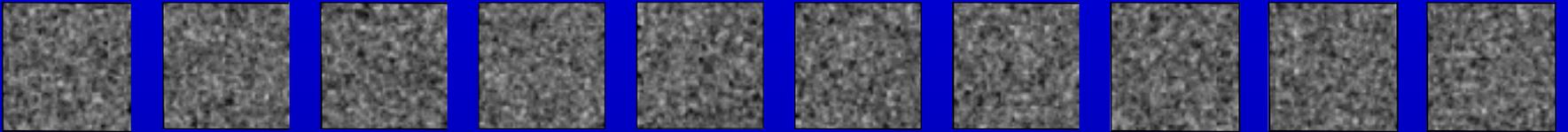
fMRI Multiple Comparisons Problem

- 4-Dimensional Data
 - 1,000 multivariate observations, each with 100,000 elements
 - 100,000 time series, each with 1,000 observations
- Massively Univariate Approach
 - 100,000 hypothesis tests
- Massive MCP!

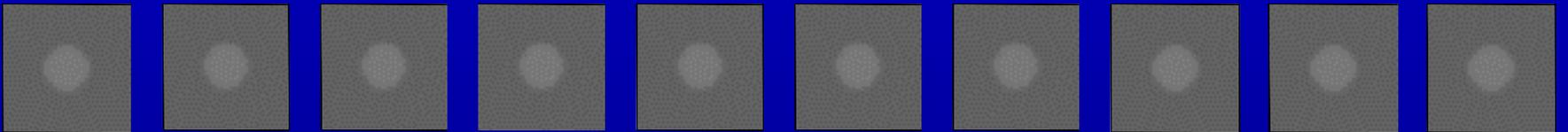


False Discovery Rate Illustration:

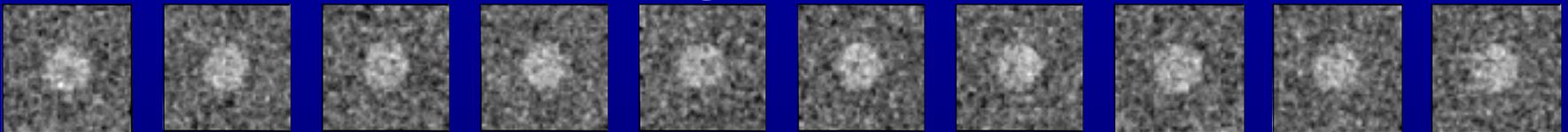
Noise



Signal



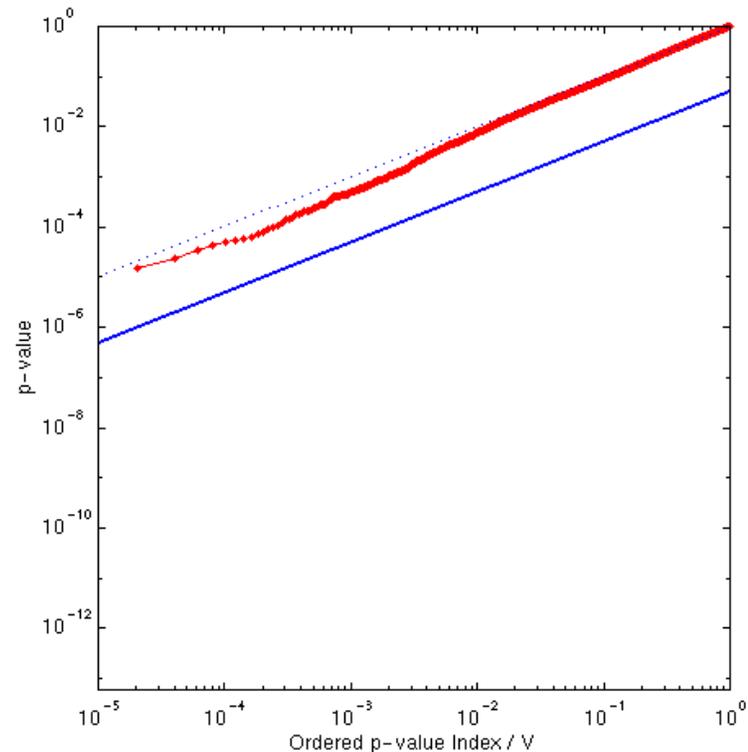
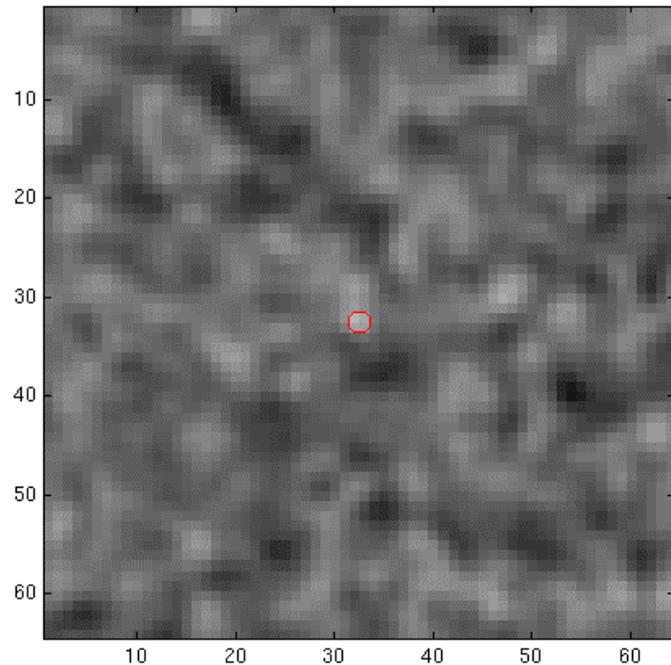
Signal+Noise



Benjamini & Hochberg: Varying Signal Extent

$p =$

$z =$

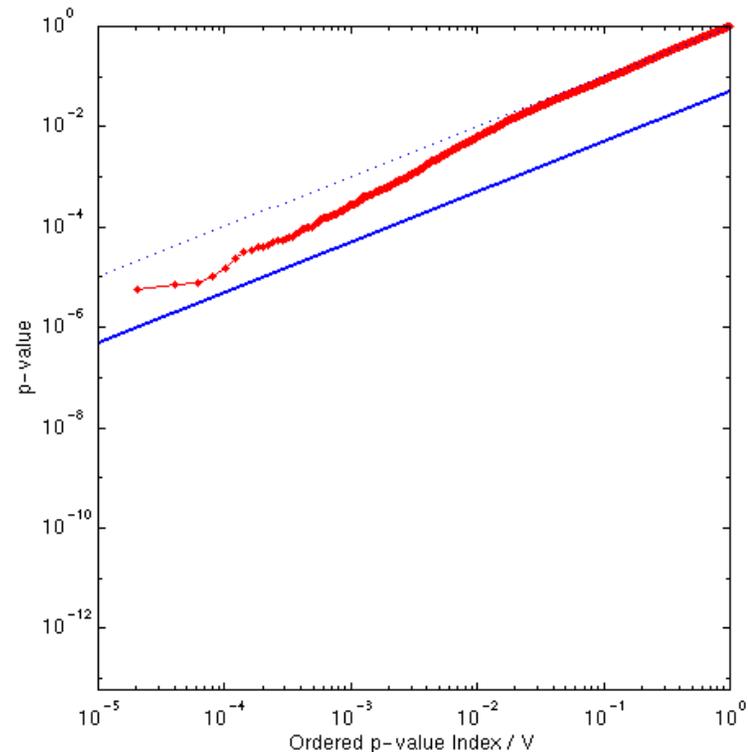
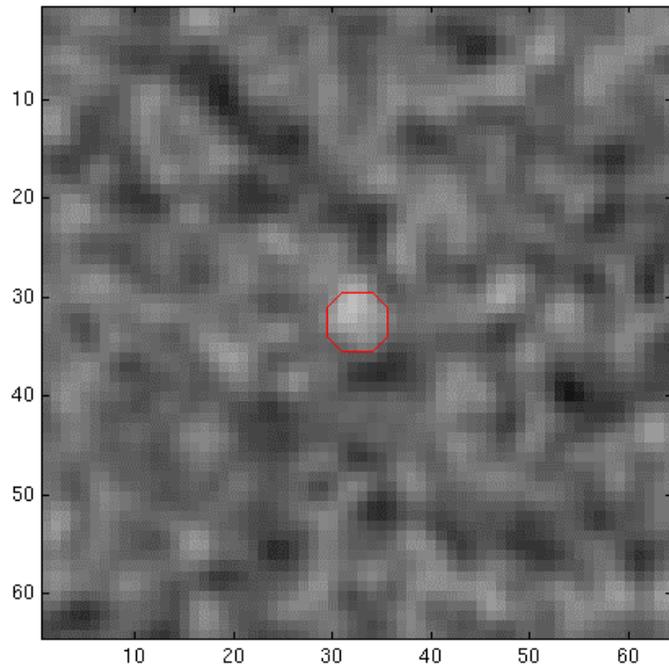


Signal Intensity 3.0 Signal Extent 1.0 Noise Smoothness 3.0

Benjamini & Hochberg: Varying Signal Extent

$p =$

$z =$

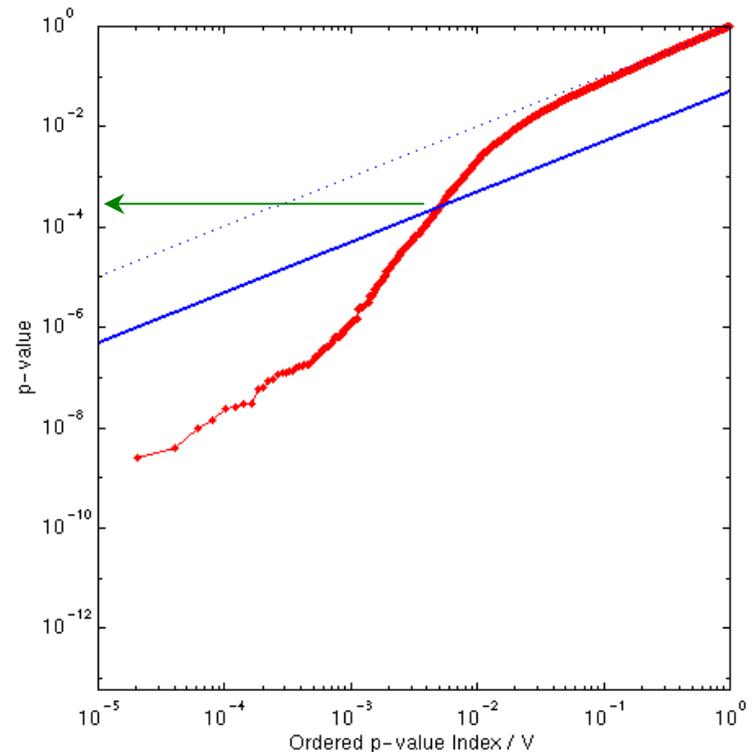
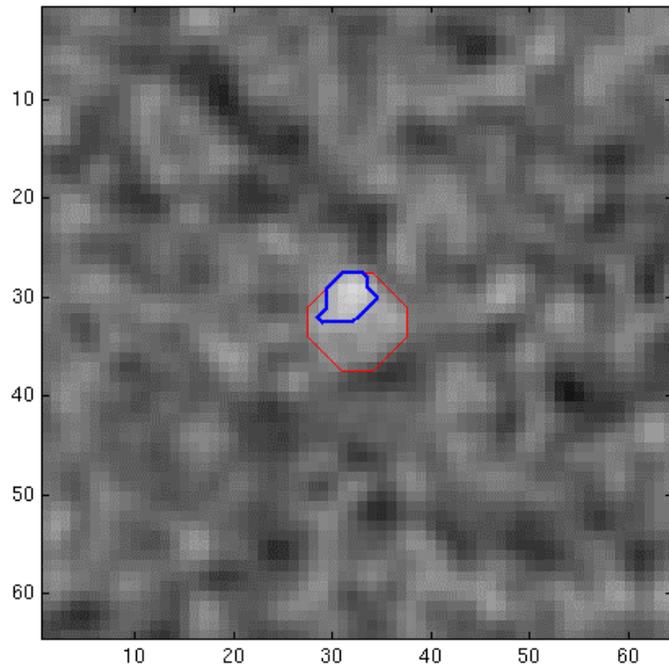


Signal Intensity 3.0 Signal Extent 3.0 Noise Smoothness 3.0

Benjamini & Hochberg: Varying Signal Extent

$$p = 0.000252$$

$$z = 3.48 \text{ (} 3.5 \sigma \text{)}$$

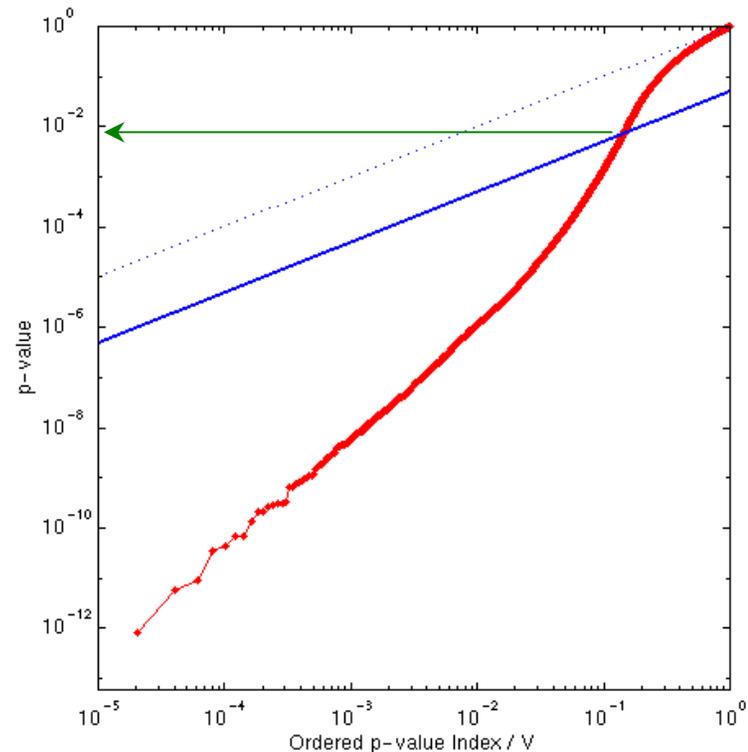
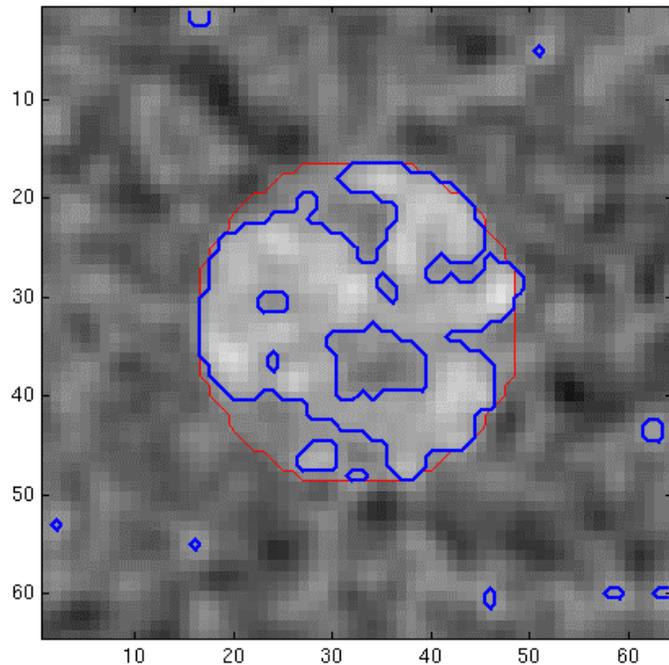


Signal Intensity 3.0 Signal Extent 5.0 Noise Smoothness 3.0

Benjamini & Hochberg: Varying Signal Extent

$$p = 0.007157$$

$$z = 2.45 \text{ (} 2.5 \sigma \text{: stronger signal)}$$

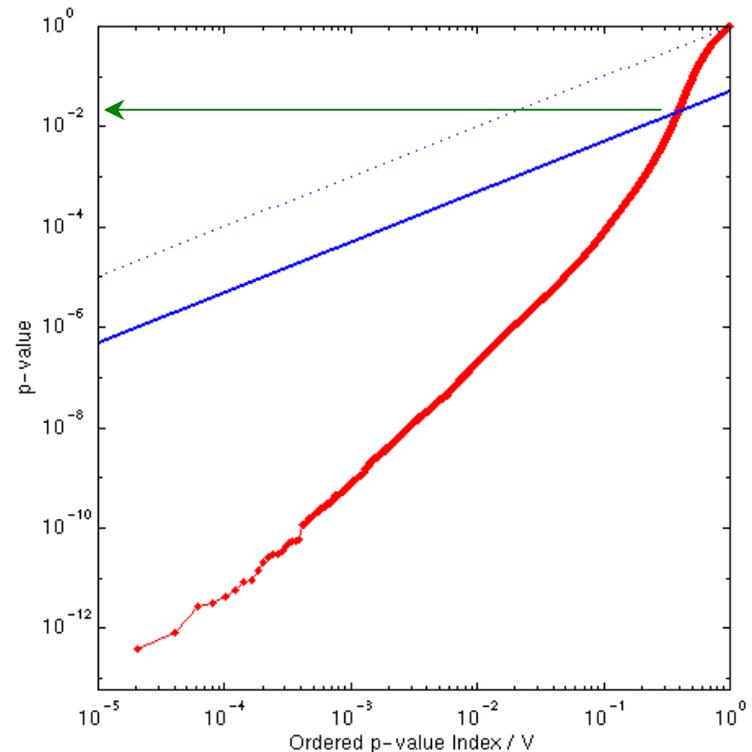
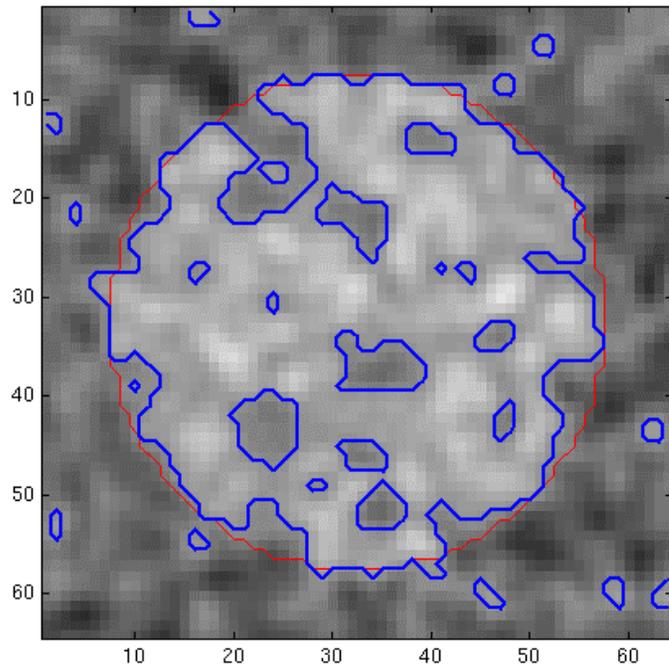


Signal Intensity 3.0 Signal Extent 16.5 Noise Smoothness 3.0

Benjamini & Hochberg: Varying Signal Extent

$$p = 0.019274$$

$$z = 2.07 \text{ (2.1 } \sigma \text{: stronger signal)}$$



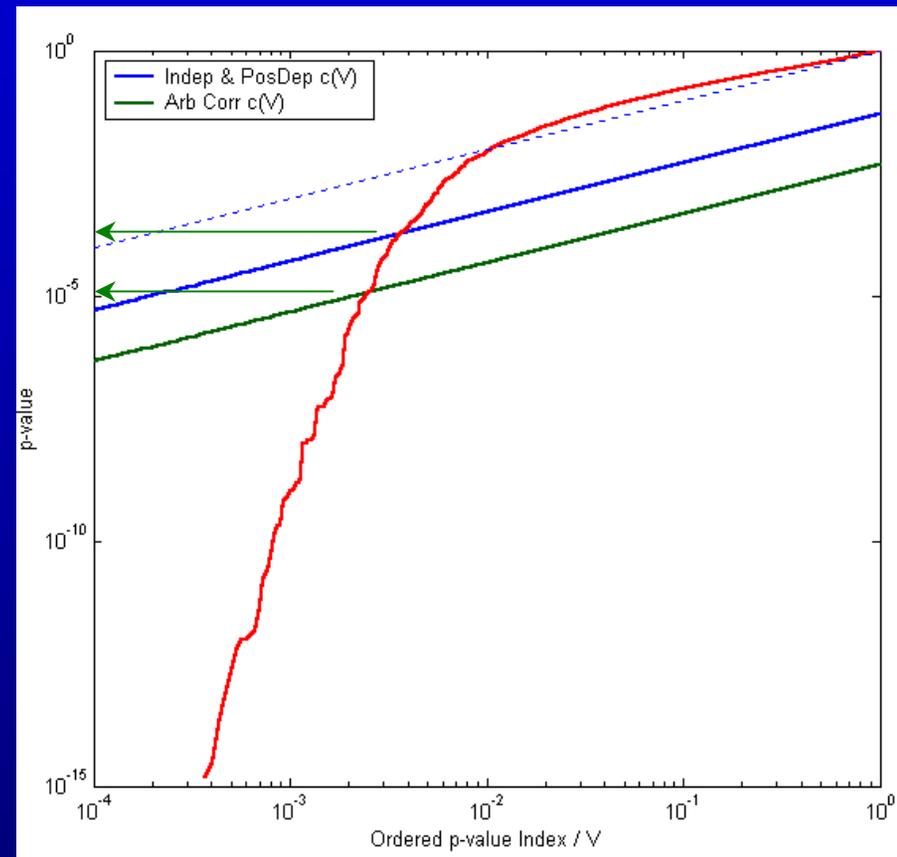
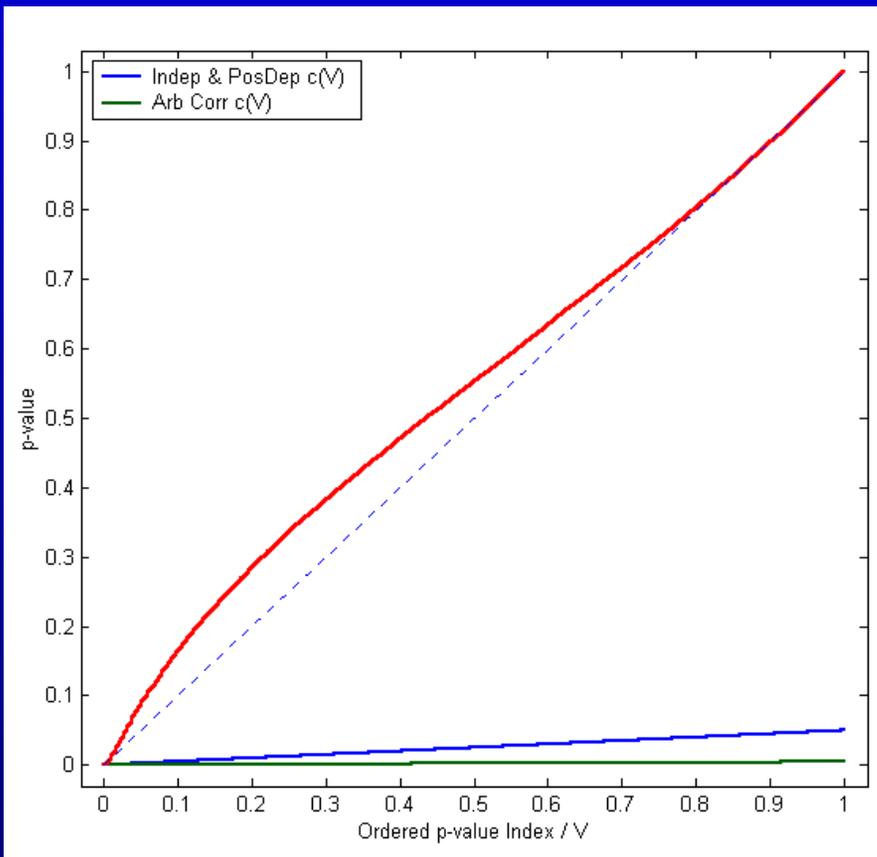
Signal Intensity 3.0 Signal Extent 25.0 Noise Smoothness 3.0

Benjamini & Hochberg: Properties

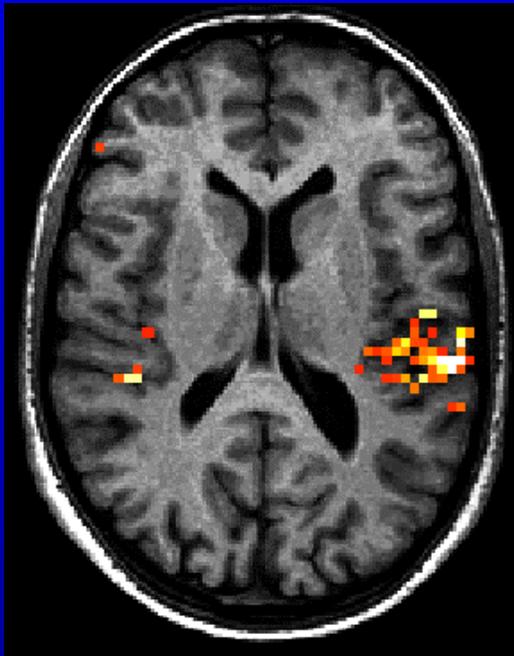
- Adaptive
 - Larger the signal, the lower the threshold
 - Larger the signal, the more false positives
 - False positives constant as fraction of rejected tests
 - Not a problem with imaging's sparse signals
- Smoothness OK
 - Smoothing introduces positive correlations

FDR Example: Plot of FDR Inequality

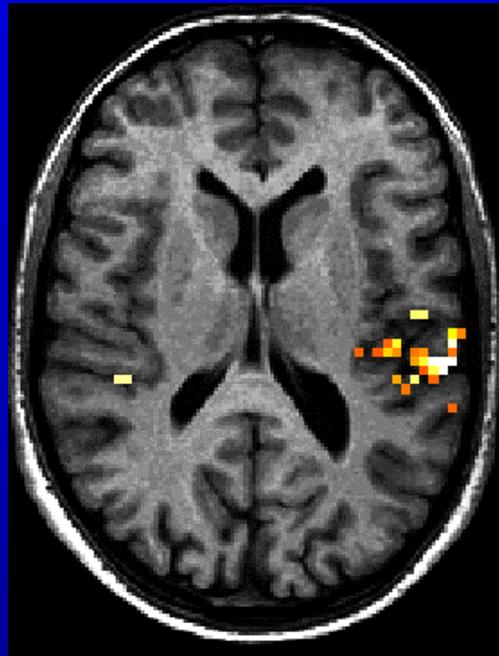
$$p_{(i)} \leq q (i/m)/c(m)$$



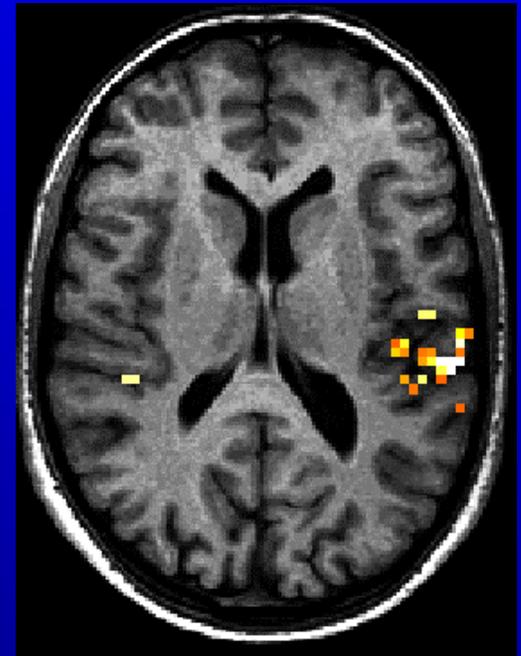
FDR: Example



FDR ≤ 0.05
Indep/PRDS
 $t_0 = 3.8119$



FDR ≤ 0.05
Arbitrary Cov.
 $t_0 = 5.0747$



FWER ≤ 0.05
Bonferroni
 $t_0 = 5.485$

FDR: Conclusions

- False Discovery Rate
 - A new false positive metric
- Benjamini & Hochberg FDR Method
 - Straightforward solution to fNI MCP
 - Just one way of controlling FDR
 - New methods under development
e.g. C. Genovese or J. Storey
- Limitations: best for independent data
 - Arbitrary dependence means less sensitive test

Sequential Variant of Bonferroni

Bonferroni-Holm

- Like Bonferroni, control total error α across all tests

Threshold at $\alpha/(m+1-i)$ starting at $p_{(1)}$

but stop at the first failure

loosens cut mildly as more pass;

identical to α/m if none pass

$$\alpha/(m+1-i) \approx (\alpha/m)\{1+(i-1)/m\} \ll \alpha(i/m) = \text{FDR}(\alpha)$$

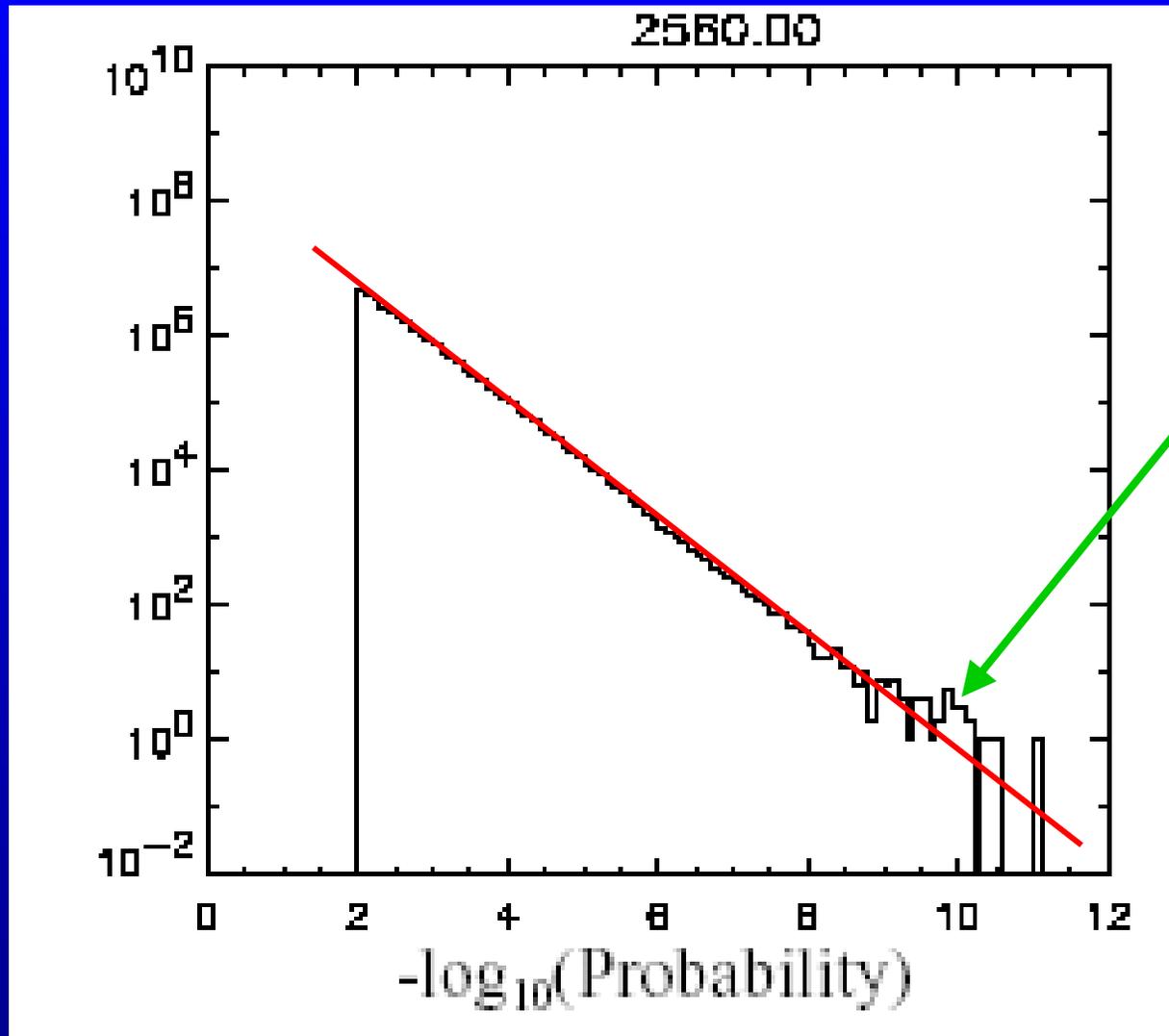
References

- ApJ 122: 3492-3505 Dec 2001 (I have pdf)
- ApJ 123: 1086-1094 Dec 2002 (I have pdf)
- The statistical literature is under active development:
 - understand in terms of mixtures (signal + background) and Bayes
 - get better sensitivity by correction for mixture
 - estimating FDR in an existing data set, or FDR with given cuts
 - calculate confidence bands on FDR
- The statistics papers are harder to read; can provide...

GRB Paper Comments

- It's **not 10^{12} trials**; rather chose $\alpha/m = 10^{-12}$
 - maybe 10^9 with $q=.001$?
 - Chosen by what criterion?
 - What efficiency considerations included?
- Do we understand our p distribution?
 - Should **predict effect of loosening cuts!**
- Looks like limits independent of data?

From GRB
paper, fig 1



Note: A histogram is a binned sorting of the p-values

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

“the rate that false discoveries occur”

- Storey: $pFDR = \mathbb{E} \left[\frac{V}{R} \mid R > 0 \right]$

“the rate that discoveries are false”

Articles

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

Storey, J.D. (2001b): **A Direct Approach to False Discovery Rates**, submitted

Storey, J.D., Tibshirani, R. (2001): **Estimating False Discovery Rates Under Dependence, with Applications to DNA Microarrays**, submitted

<http://www-stat.stanford.edu/~jstorey/>

FDR and the BH Procedure

- Define the *realized* False Discovery Rate (FDR) by

$$\text{FDR} = \begin{cases} \frac{N_{1|0}}{R} & \text{if } R > 0, \\ 0, & \text{if } R = 0. \end{cases}$$

- Benjamini & Hochberg (1995) define a sequential p-value procedure that controls *expected* FDR.

Specifically, the BH procedure guarantees

$$E(\text{FDR}) \leq \frac{M_0}{m} \alpha \leq \alpha$$

for a pre-specified $0 < \alpha < 1$.

(The first inequality is an equality in the continuous case.)

Genovese and Wasserman emphasize the
sample quantity $N_{1|0}/R$

Storey emphasizes $E(N_{1|0}/R \mid R > 0)$

But both keep the term FDR for their versions

Exact Confidence Thresholds (cont'd)

\mathcal{U} yields a confidence envelope for FDR(t) sample paths.

