Controlling False Discovery Rate and Trials Factors in Searches

Jim Linnemann

MSU

Milagro meeting

University of Maryland

March 28, 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 Astrophysics
  - B. Efron Stanford, J. Rice Berkeley , Y. Benjamani Tel Aviv Statistics
  - Google

Outline

- What is significant enough to report?
  - Multiple Comparison Problem (trials)
- A Multiple Comparison Solution:
  False Discovery Rate (FDR) BH 1995
  - Search for non-background events
  - Need only the background probability distribution
  - Control *fraction* of false positives reported
    - Automatically select how hard to cut, based on that
- FDR Plausibility and Properties
- FDR Example
- References
- Probably no time for:

  GRB comments

  Extensions and details

Significance

- Define "wrong" as reporting false positive:
  - Apparent signal caused by background

- Set α a level of potential wrongness
  - $2\sigma = .05$ $3\sigma = .003$ etc.
    - Probability of going wrong on one test
    - Or, error rate per test
    - Statisticians say: "z value" instead of z σ's

What if you do m tests?

- Search m places
- *Must be able to define "interesting"*
  - *e.g. "not background"*
- Examples from HEP and Astrophysics
  - Look at m histograms, or bins, for a bump
  - Look for events in m decay channels
  - Test standard model with m measurements (not just R_b or g-2)
  - Look at m phase space regions for a Sleuth search (Knuteson)

  - Fit data to m models: What's a bad fit?
  - Reject bad tracks from m candidates from a fitting routine
  - Look for sources among m image pixels
  - Look for "bursts" of signals during m time periods
  - Which of m fit coefficients are nonzero?
    - Which (variables, correlations) are worth including in the model?
    - Which of m systematic effect tests are significant?

      Rather than testing each independently

Must do something about m!

- m is "trials factor" only NE Jour Med demands!
- Don't want to just report m times as many signals
  - P(at least one wrong) = $1 - (1- \alpha)^m \sim m\alpha$
- Use α/m as significance test "Bonferroni correction"
  - This is the main method of control
- *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)
      - Whole set of tests considered together
  - Control by Bonferroni, Bonferroni-Holm, or Random Field Method
    See backup slides for more
- False Discovery Rate (FDR)
  - Fraction of errors in signal candidates
    - Proportion of false positives *among* rejected tests
    - "False Discovery Fraction" might have been clearer?

Decision, based on test statistic:

	Null Retained (can't reject)	Reject Null = Accept Alternative	Total
Null (H_o) True **background**	U	V false positive Type I Error $\alpha = \varepsilon_b$ B false discovery	m_o
Alternative True **signal**	T inefficiency Type II Error $\beta = 1 - \varepsilon_s$	S true positive true detection	m_1
	m-R	R reported signal **= S+B** rejections	m

FDR = $V/R = B/(S+B)$ *if* $R > 0$

0 *if* $R=0$

Goals of FDR

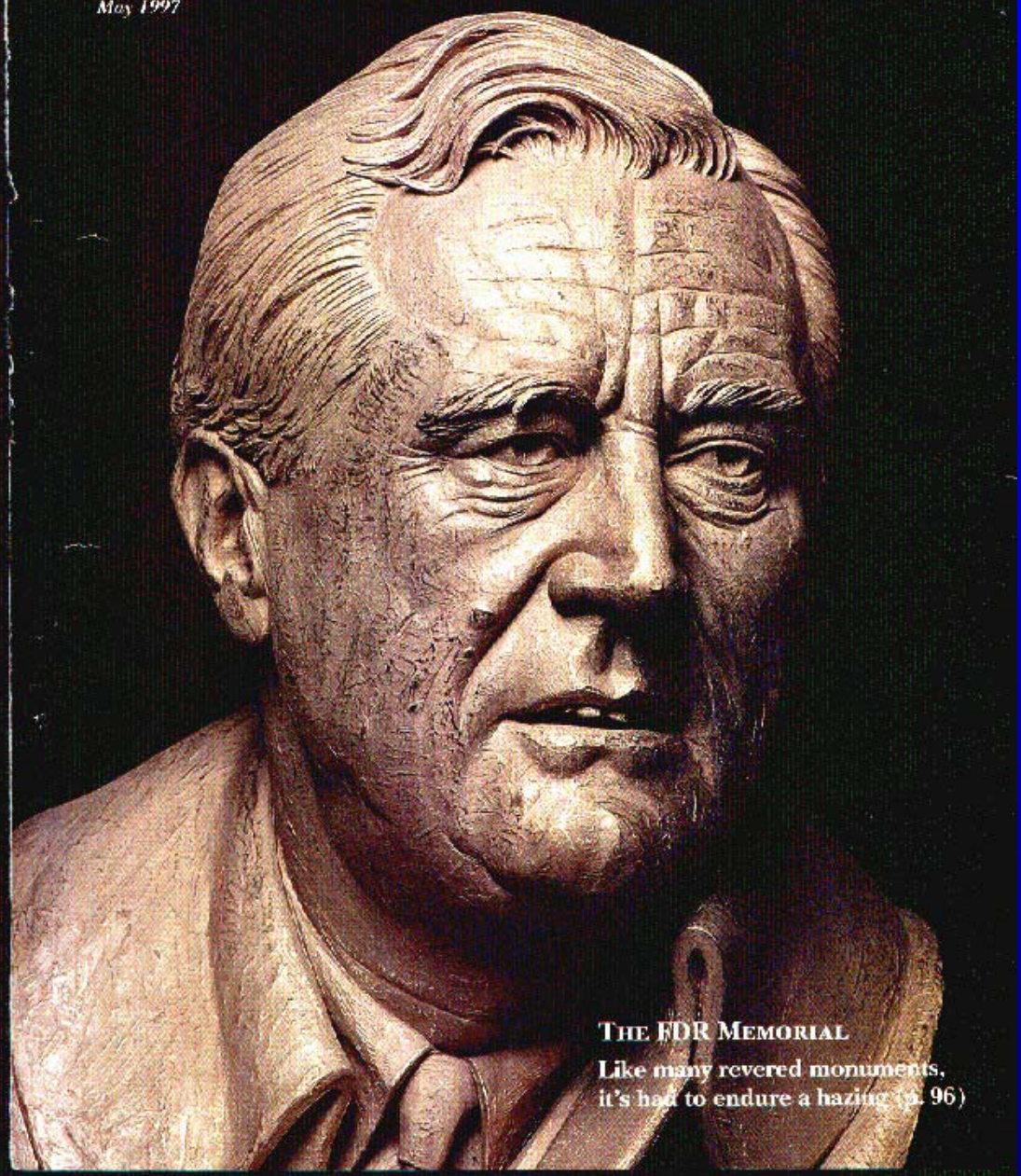
- Tighter than α (single-test)
- Looser than α/m (Bonferroni trials factor)
- Improve sensitivity ("power"; signal efficiency)
- Still control something useful:
  - fraction of false results that you report

    $b/(s+b)$ after your cut = 1 - purity
    - rather than $1-\alpha$ = rejection(b); or efficiency(s)
    - for 1 cut, you only get to pick 1 variable, anyway

- Last, but not least, a catchy TLA

Smithsonian

May 1997



THE FDR MEMORIAL

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

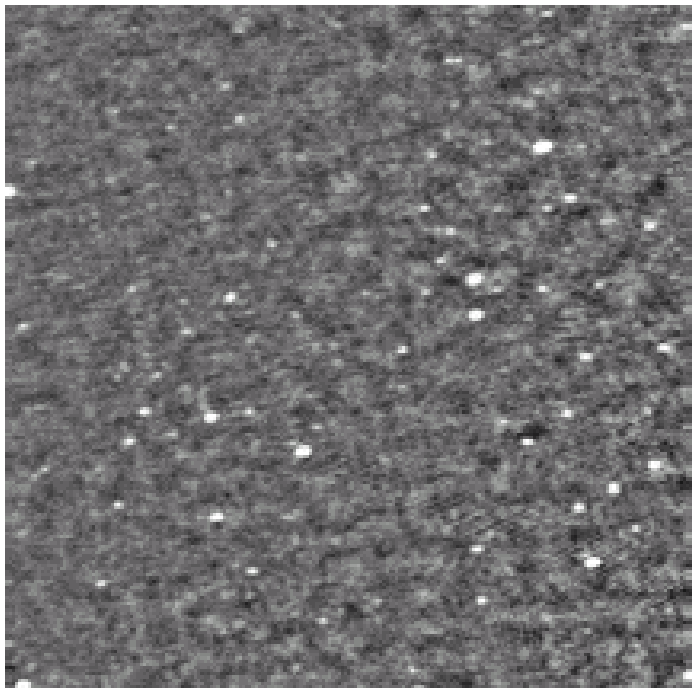
Where did this come from?
Others who have lots of tests!

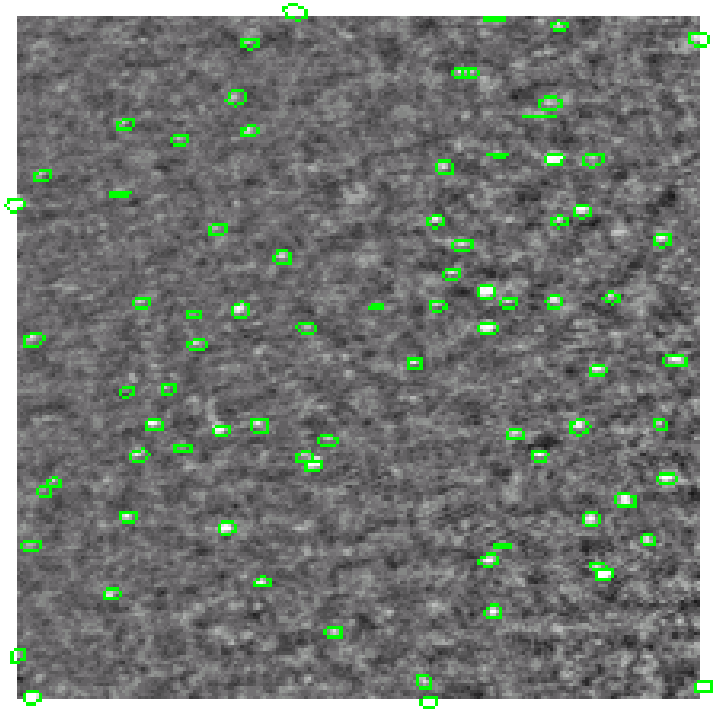
- Screening of chemicals, drugs
- Genetic mapping
- Functional MRI (voxels on during speech processing)
- Data mining (cookies by milk? direct mail)
- Radio telescope images (at last some astronomy!)

- Common factors:
  - *One false positive does not invalidate overall conclusion*
  - 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:

$$\mathrm{Exp}\left(\frac{\text{expense wasted chasing “red herrings”}}{\text{cost of all follow-up studies}}\right) \leq q$$

GRB alerts from Milagro?

What is a p-value?

(*Needed for what's next)*

Observed significance of a measurement

Familiar example: $P(\geq \chi^2 \mid \nu)$ (should be flat)

- Here, probability that event produced by background ("null hypothesis")
  - 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

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

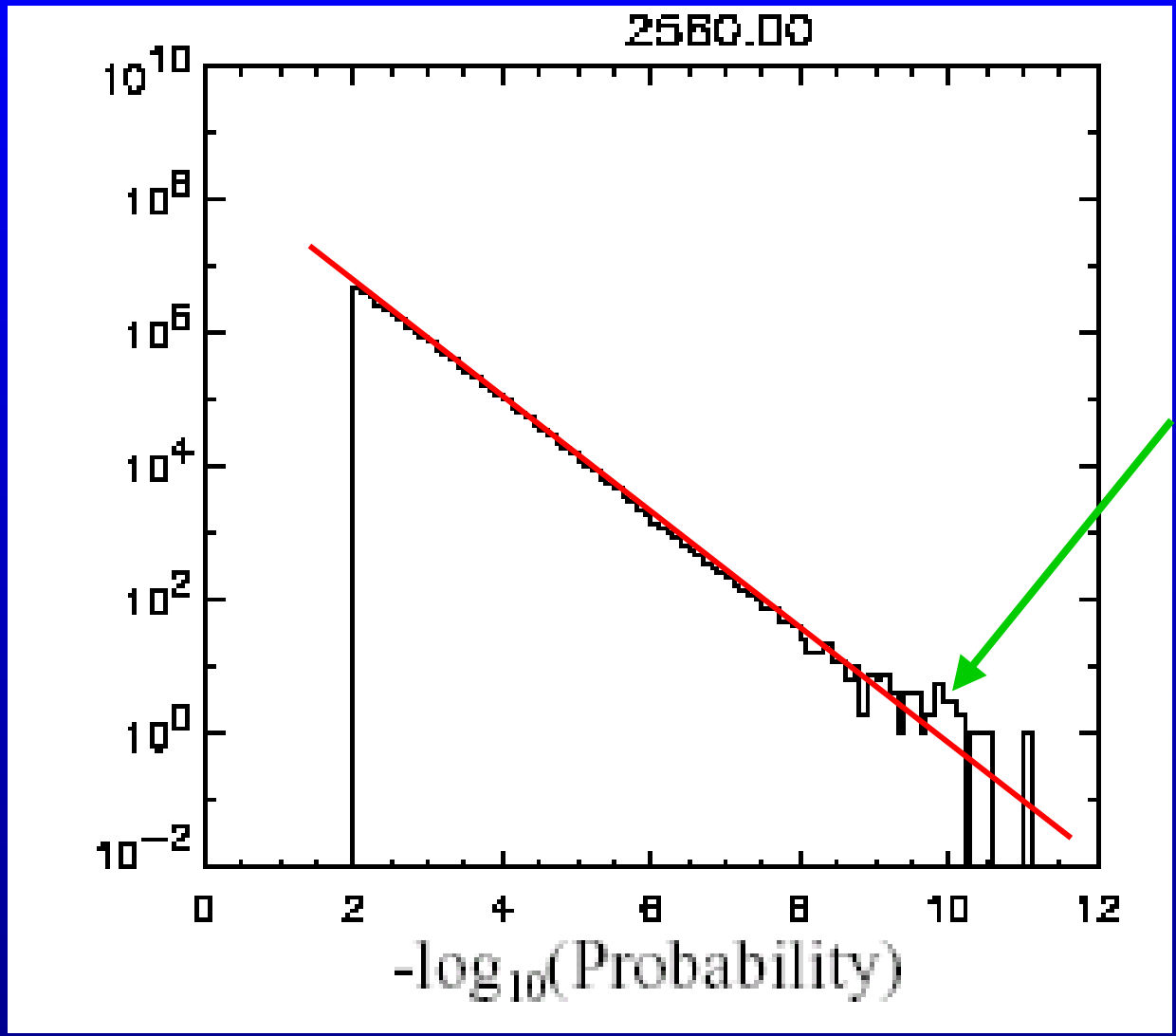
N(p in [a, b]) = (b-a)

So expect $N(p \leq p_{(r)})/m = r/m$ for r-smallest p-value

Flat also means linear in log-log: if y = *ln* p

ln[dn/dy] vs. y is a straight line, with a predicted slope

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)

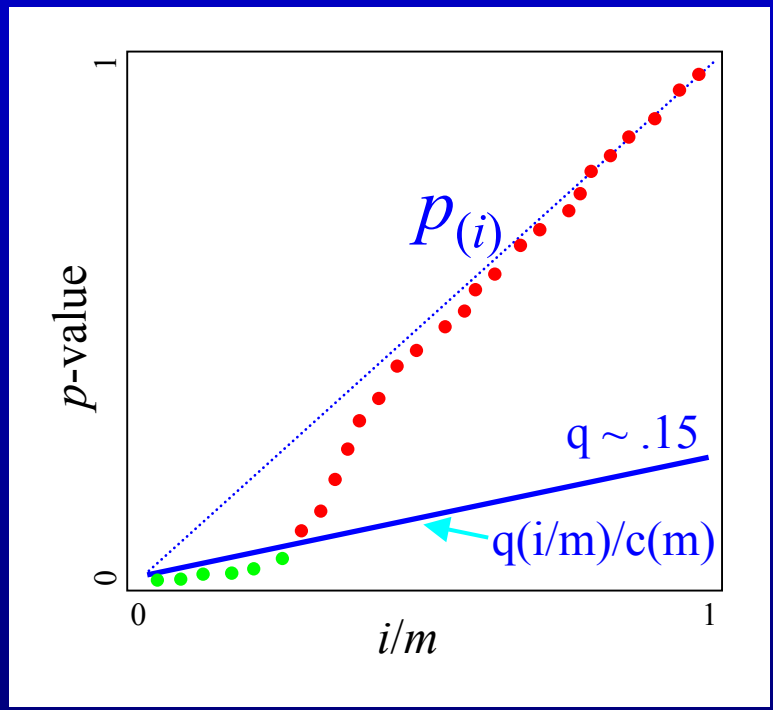
  α is not specified: the method selects it
- Sort the p-values, $p_{(1)} \leq p_{(2)} \leq \ \ldots \leq p_{(m)}$
- Let r be largest i such that

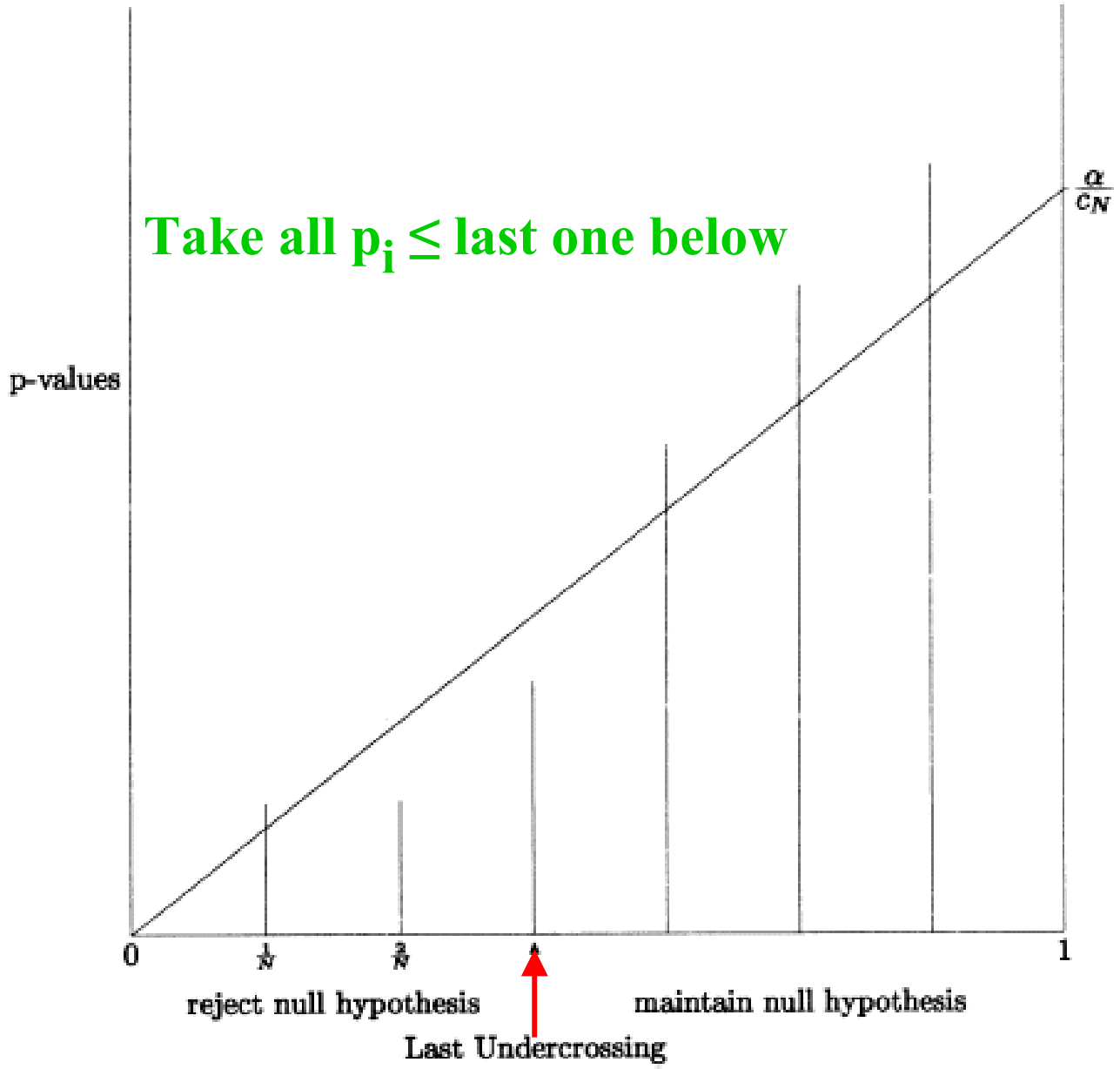
$$p_{(i)} \leq \ q(i/m)/c(m)$$

For now, take c(m)=1

- Reject all null hypotheses corresponding to

  $p_{(1)}, \ \cdots \ , p_{(r)}$.
  - i.e. Accept as signal
- *Proof this works is not obvious!*




Take all $p_i \leq$ last one below
p-values
$\frac{\alpha}{c_N}$
0
$\frac{1}{N}$
$\frac{2}{N}$
1
reject null hypothesis
maintain null hypothesis
Last Undercrossing

Plausibility argument

for easily separable signal of *Miller et al.*

- $p_{(r)} \leq q\ r/m$ (definition of cutoff)
- $\langle p_{(r)} \rangle = q \langle R \rangle / m$ ($\langle r \rangle = \langle R \rangle$: def of # rejects)

- Now assume background uniform
  - AND Say all signal p values << p(background) ≈ 0
- $\langle p_{(r)} \rangle = \langle R_{background} \rangle / m$

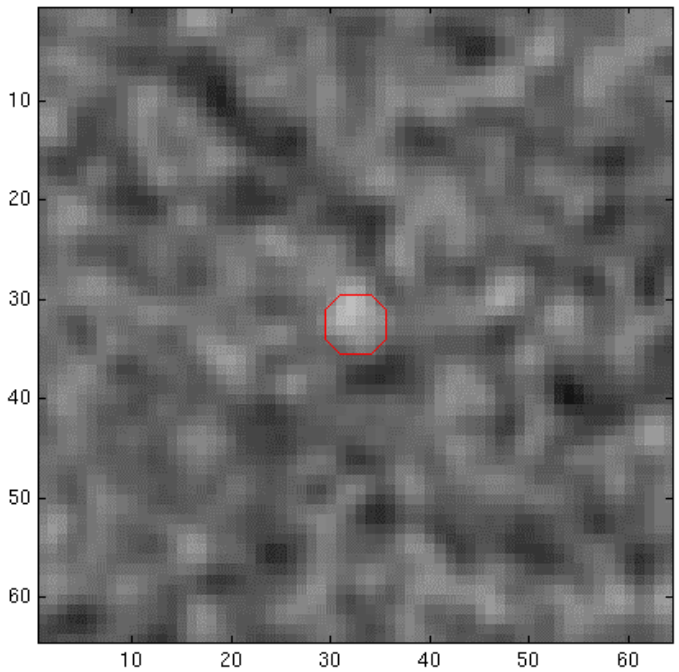
- Solving, $q = \langle R_{background} \rangle / \langle R \rangle$

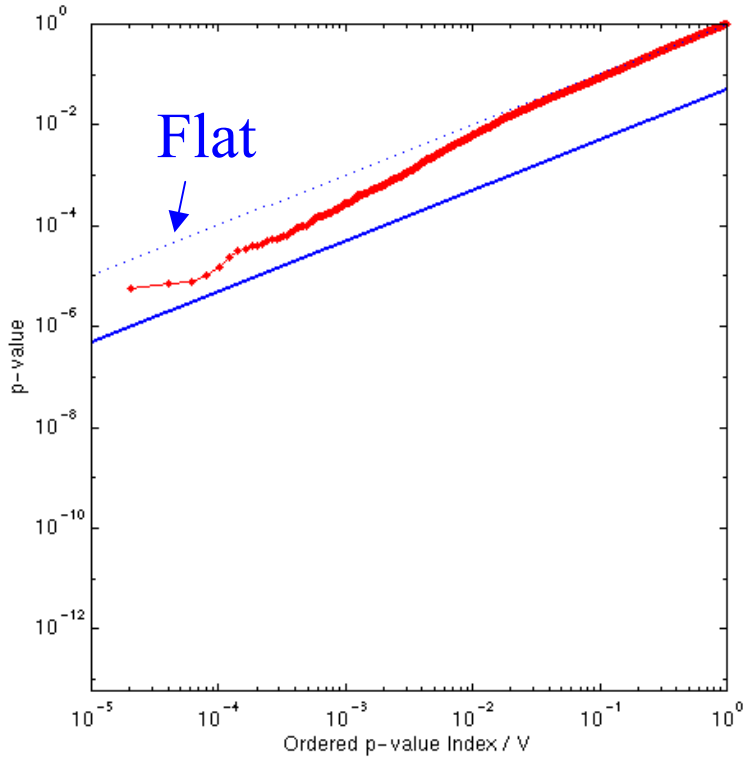
Full proof makes no assumptions on signal

Other than it's distinguishable (p's nearer 0)

Benjamini & Hochberg: Varying Signal Extent (MC)

$p =$ $z =$ (none pass)

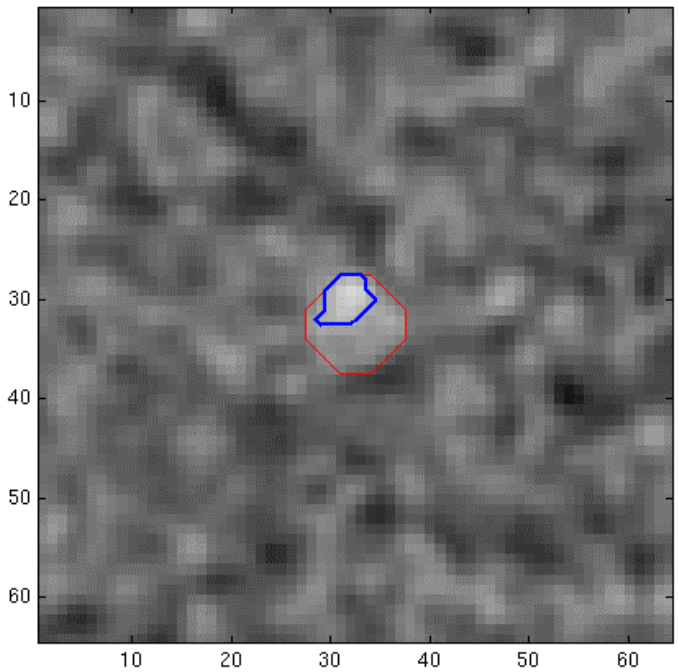


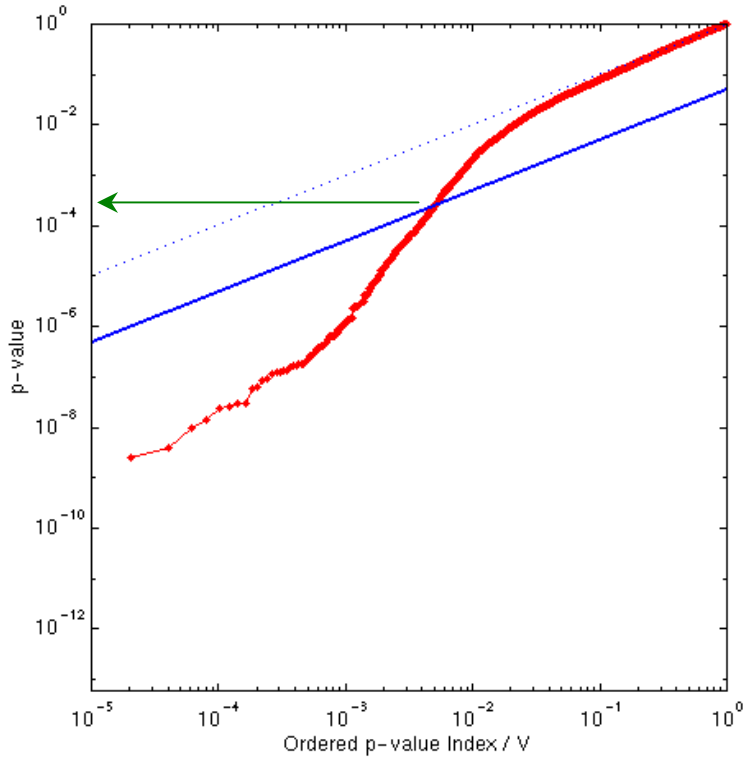


Signal Intensity 3.0 Signal Extent 3.0 Noise Smoothness 3.0

Benjamini & Hochberg: Varying Signal Extent

$p = 0.000252$ $z = 3.48$ (3.5 σ cut chosen by FDR)

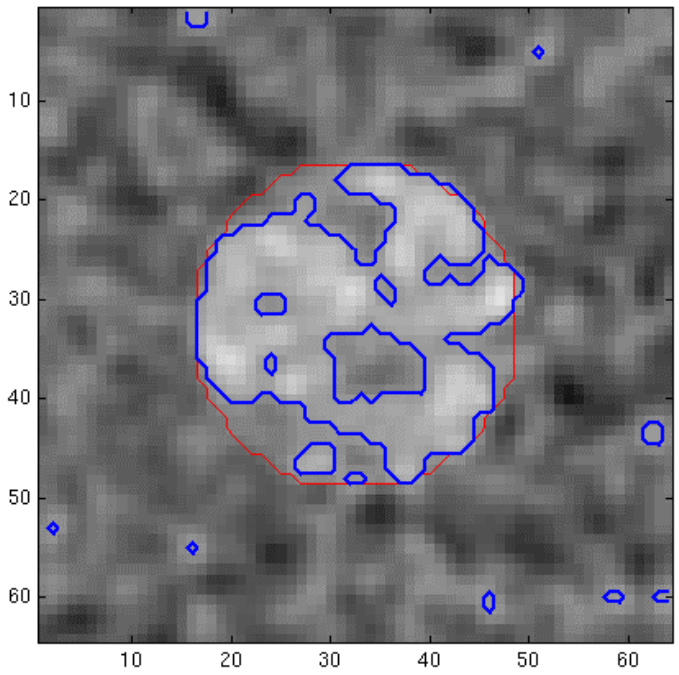


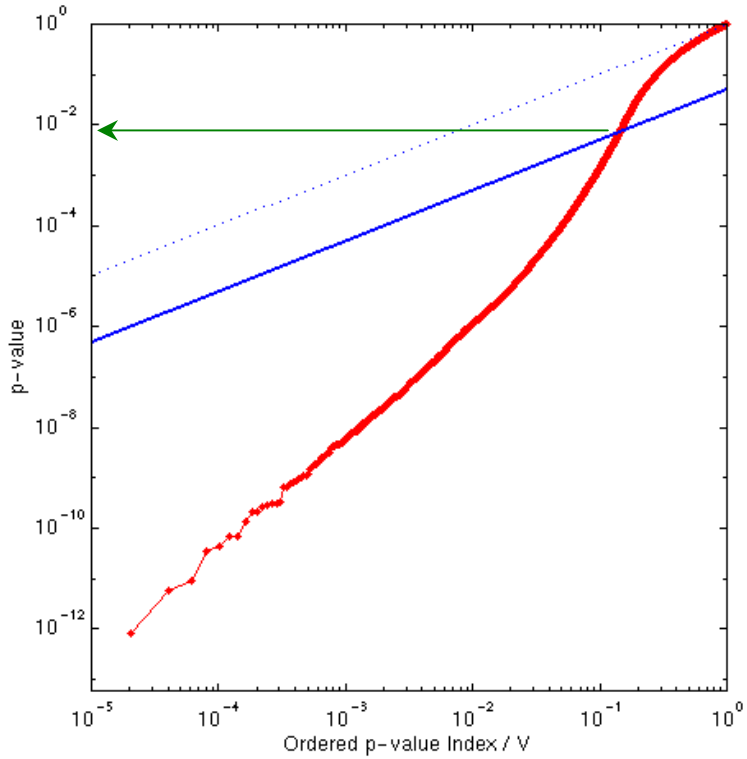


Signal Intensity 3.0 Signal Extent 5.0 Noise Smoothness 3.0

Benjamini & Hochberg: Varying Signal Extent

$p = 0.007157$ $z = 2.45$ (2.5 σ: stronger signal)





Signal Intensity 3.0 Signal Extent 16.5 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
  - Can still use c(m) = 1

Benjamini & Hochberg c(m) factor

- $c(\mathrm{m}) = 1$
  - 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.
- $c(\mathrm{m}) = \Sigma_{i=1,\ldots \mathrm{m}}\ 1/i \ \approx \ \log(\mathrm{m})+0.5772$
  - Arbitrary covariance structure
    - But this is more conservative—tighter cuts

FDR as Hypothesis Test
Quasi distribution-free

- Assumes specific null (flat p-values)
  - in this, like most null hypothesis testing
  - but works for any specific null distribution, not just Gaussian; χ^2
  - distribution-free for alternative hypothesis
    - Distribution-free estimate, control of s/b! A nice surprise
  - Fundamentally Frequentist:
    - Goodness of Fit test to well-specified null hypothesis
    - No crisp alternative to null needed: anti-Bayesian in feeling

Strength: search for ill-specified "something new"

if different enough to give small p-values

- No one claims it's optimal
  - With a specific alternative, could do sharper test
    - Better s/b for same α or vice versa

Comments on FDR

- To use method, you must *not so new!*
  - know trials factor
  - Be able to calculate small p values correctly
  - Have p values of all m tests in hand (retrospective)
    - Or, to use online, a good-enough sample of same mix of s+b
- Lowest p value $p_{(1)}$ always gets tested with q/m (i=1)
    - If no signal , q FDR $\rightarrow$ Bonferroni in $\alpha/m = q/m$
      - FWER = q for FDR α for Bonferroni when no real signal
- Uses distribution of p's
  - Even if $p_{(1)}$ fails
  - FDR sees other $p_{(i)}$ distorting the pure-null shape
  - FRD raises the threshold and accepts $p_{(1)} \ldots p_{(r)}$

Minding your p's and q's

a Frequentist Method with Bayesian Flavor

- p = α = Prob(reject null | null is true) per test; or all m
- q = Prob(null is true | reject null)
  - Intuition: q is "Bayesian posterior p-value"
    - Calculable, given prior signal fraction, signal distribution
- Or: prob any wrong vs. fraction of list wrong

- For any multiple test, can quote both
  - q = <FDR> p = α which FDR selects
  - Or pick α; run FDR backwards: find q giving that α
  - Similar to quoting both efficiency and rejection

FDR: Conclusions

- False Discovery Rate: a new false positive metric
  - Control fraction of false positives in multiple measurements
  - Selects significance cut based on data

- Benjamini & Hochberg FDR Method
  - Straightforward application to imaging, fMRI, gene searches

  - Interesting technique searching for "new" signals
    - Most natural when expect some signal
    - But correct control of false positives even if no signal exists
    - Can report FDR along with significance, no matter how cuts set
      - <b> (significance) , and FDR estimate of <s/(s+b)>

  - Just one way of controlling FDR
    - New methods under development e.g. C. Genovese or J. Storey

Further Developments

- The statistical literature is under active development:
  - understand in terms of mixtures (signal + background)
    - and Bayesian models of these
  - get better sensitivity by correction for mixture
    - more important for larger signal strength fractions
  - Can estimating FDR in an existing data set,
    - or FDR with given cuts
  - calculate confidence bands on FDR

FDR Talks on Web

Users:

- T. Nichol **U Mich** www.sph.umich.edu/~nichols/FDR/ENAR2002.ppt
  Emphasis on Benjamini's viewpoint; Functional MRI
- S. Scheid, MPI http://cmb.molgen.mpg.de/compdiag/docs/storeypp4.pdf
  Emphasis on Storey's viewpoint

Statiticians:

- C. Genovese CMU

http://www.stat.ufl.edu/symposium/2002/icc/web_records/genovese_ufltalk.pdf

- Y. Benjamini Tel Aviv www.math.tau.ac.il/~ybenja/Temple.ppt

Random Field Theory (another approach to smoothed data)

- W. Penny, UCLondon,
  http://www.fil.ion.ucl.ac.uk/~wpenny/talks/infer-japan.ppt

- Matthew Brett, Cambridge
  http://www.mrc-cbu.cam.ac.uk/Imaging/randomfields.html

Some other web pages

- http://medir.ohsu.edu/~geneview/education/Multiple test corrections.pdf
  Brief summary of the main methods

- www.unt.edu/benchmarks/archives/2002/april02/rss.htm
  Gentle introduction to FDR

  www.sph.umich.edu/~nichols/FDR/
  FDR resources and references—imaging

  http://www.math.tau.ac.il/~roee/index.htm
  FDR resource page by discoverer

Some FDR Papers on Web

Astrophysics

arxiv.org/abs/astro-ph/0107034 Miller et. al. **ApJ 122: 3492-3505 Dec 2001**

FDR explained very clearly; heuristic proof for well-separated signal

arxiv.org/abs/astro-ph/0110570 Hopkins et. Al. ApJ 123: 1086-1094 Dec 2002

2d pixel images; compare FDR to other methods

taos.asiaa.sinica.edu.tw/document/chyng_taos_paper.pdf FDR comet search (by occultations)

will set tiny FDR limit 10^{-12} ~ 1/year

Statistics

http://www.math.tau.ac.il/~ybenja/depApr27.pdf Benjamini et al: (invented FDR)

clarifies c(m) for different dependences of data

Benjamani, Hochberg: *JRoyalStatSoc-B* (1995) 57:289-300 paper not on the web

defined FDR, and Bonferroni-Holm procedure

http://www-stat.stanford.edu/~donoho/Reports/2000/AUSCFDR.pdf Benjamani et al

study small signal fraction (sparsity), relate to minimax loss

http://www.stat.cmu.edu/www/cmu-stats/tr/tr762/tr762.pdf Genovese, Wasserman

conf limits for FDR; study for large m; another view of FDR as data-estimated method on mixtures

http://stat-www.berkeley.edu/~storey/ Storey

view in terms of mixtures, Bayes; sharpen with data; some intuition for proof

http://www-stat.stanford.edu/~tibs/research.html Efron, Storey, Tibshirani

show Empirical Bayes equivalent to BH FDR

Some details

- <FDR> = q m_o/m (q × fraction of background)
  - Not just q
- Subtlety in definitions:

  Storey's pFDR = P(Null true|reject null); FDR = pFDR × P(R > 0)

- More plausibility: can view BH differently:

  Use of departure of observed p's from flat:

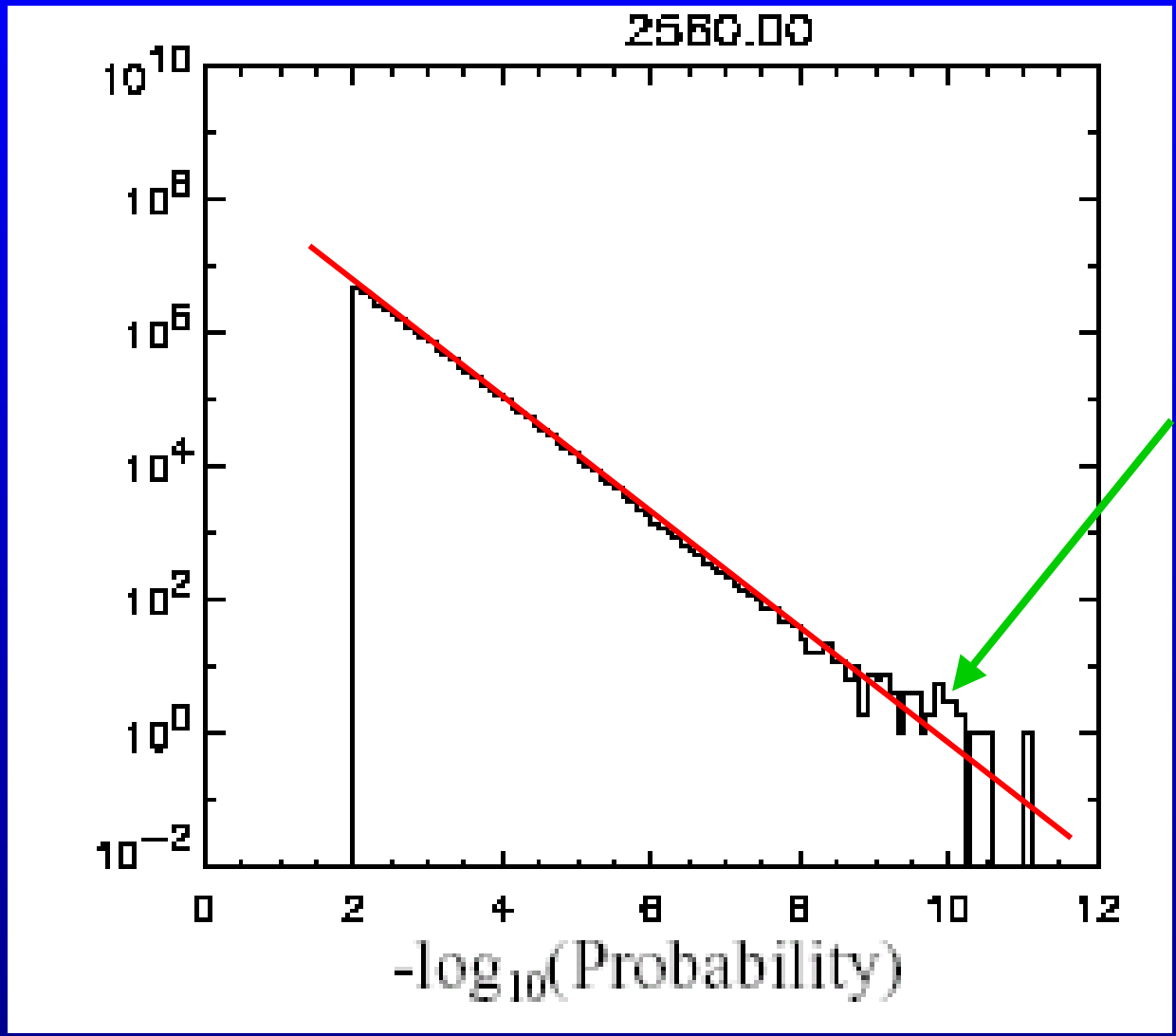
  Implicitly estimates from data m_o/m in a mixture of b(=null) + s

- Improvements (especially for large signals):
  - estimate m_o more directly
  - estimate other parameters of mixture
  - optimum (min MSE) tuning parameters
    - For estimating where to put cut

GRB Paper Comments

- It's not 10^{12} trials: instead chose $\alpha/m = 10^{-12}$
  - Chosen by what criterion? "below ½ of data"
  - What efficiency considerations included?
  - maybe 10^9 with q=.001?
- Do we understand our p distribution?
  - Should predict effect of loosening cuts!
- Looks like limits independent of data?

From GRB paper, fig 1



Signal, statistics, or systematics?

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

Extensions and Details

- FDR Variants
- FDR and c(m): when is c(m)=1?
- Extensions to Bonferroni
  - Bonferroni-Holm
  - Random Field Theory
- More FDR motivational examples
  - And relation to testing theory

Recurring Notation

$m, M_0, N_{1\|0}$	# of tests, true nulls, false discoveries
a	Mixture weight on $a$lternative
$H^m = (H_1, \ldots, H_m)$	Unobserved true classifications
$P^m = (P_1, \ldots, P_m)$	Observed p-values
$P^m_{(\cdot)} = (P_{(1)}, \ldots, P_{(m)})$	Sorted p-values (define $P_{(0)} \equiv 0$)
U	CDF of Uniform$\langle 0, 1 \rangle$
F, f	Alternative CDF and density
$G = (1-a)U + aF$	Marginal CDF of P_i (mixture model)
$\hat{G}$	Estimate of G (e.g., empirical CDF of P^m)

FDR = B/(S+B) after cuts

False discoveries
(false positives)

Background =
null hypothesis

"can't reject" null

b

signal

	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

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

Detected signals

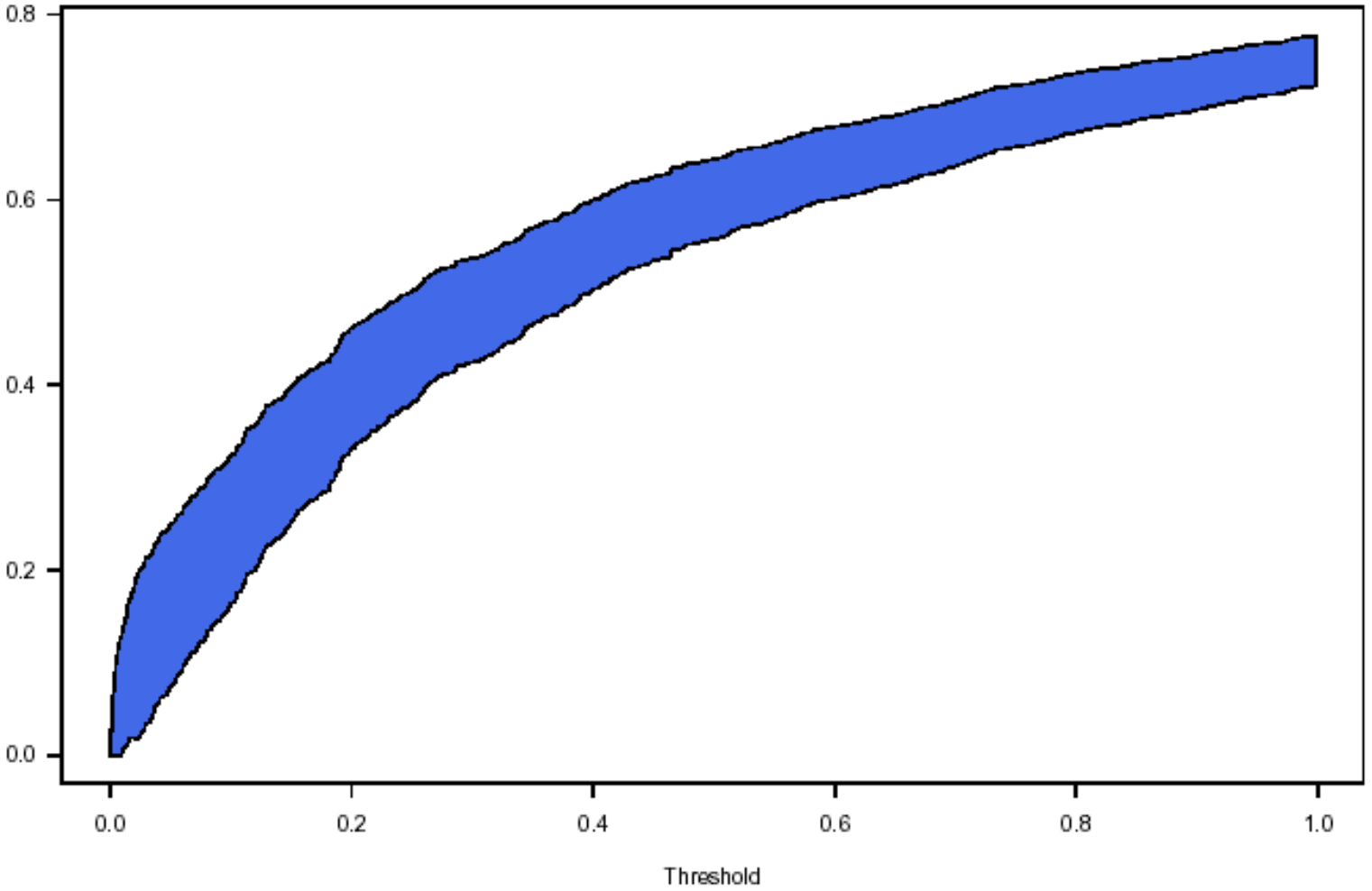
(true positives)

Reported signal
candidates
(rejected nulls)

Reject null = accept alternative

Exact Confidence Thresholds (cont'd)

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



Multiple Testing Procedures

- A multiple testing procedure T is a map $[0,1]^m \to [0,1]$, where the null hypotheses are rejected in all those tests for which $P_i \leq T(P^m)$. Often call T a *threshold*.
- Examples:

Uncorrected testing	$T_{\mathrm{U}}(P^m) = \alpha$
Bonferroni	$T_{\mathrm{B}}(P^m) = \alpha/m$
Fixed threshold at t	$T_t(P^m) = t$
First r	$T_{(r)}(P^m) = P_{(r)}$
Benjamini-Hochberg	$T_{\mathrm{BH}}(P^m) = P_{(R_{\mathrm{BH}})}$ or $\sup\{t : \widehat{G}(t) = t/\alpha\}$
Oracle	$T_{\mathrm{O}}(P^m) = \sup\{t : G(t) = (1-a)t/\alpha\}$
Plug In	$T_{\mathrm{PI}}(P^m) = \sup\{t : \widehat{G}(t) = (1-\widehat{a})t/\alpha\}$
Regression Classifier	$T_{\mathrm{Reg}}(P^m) = \sup\{t : \widehat{\mathsf{P}}\{H_1{=}1 \mid P_1{=}t\} > 1/2\}$

Bayes Oracle: what you could do if you knew

signal fraction and signal distribution

I believe Frequentist would call this Neyman-Pearson test

BH as a Plug-in Procedure

- Let $\hat{G}$ be the empirical cdf of P^m under the mixture model. Ignoring ties, $\hat{G}(P_{(i)}) = i/m$, so BH equivalent to

$$T_{\mathrm{BH}}(P^m) = \arg\max \left\{ t\colon\ \hat{G}(t) = \frac{t}{\alpha} \right\}.$$

- We can think of this as a plug-in procedure for estimating

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

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

FDR and the BH Procedure

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

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

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

  Specifically, the BH procedure guarantees

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

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

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

Storey:

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

  "the rate that false discoveries occur"

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

  "the rate that discoveries are false"

Benjamini (email) argues his definition more appropriate when it's not clear there are any real discoveries to be made

$V = N_{1|0}$

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/`

Yet more details

- FDR controlled at $q \langle m_0/m \rangle$
- more precisely,

  $\langle (V/m_0)/(R/m) \rangle \leq q$

  - For continuous variables, you get $=q$
  - For discrete statistics, only $< q$
- $\langle p_{(i)} \rangle = i/(m+1)$ (not i/m, the naïve value)

- Random remark by Miller et. al.
  - Posterior Bayes Intervals cover (Frequentist) to order 1/n
  - But correspondence breaks down in Hypothesis Testing

Benjamini:

Genovese and Wasserman emphasize the sample quantity V/R

Storey emphasizes $E(V/R \mid R>0)$

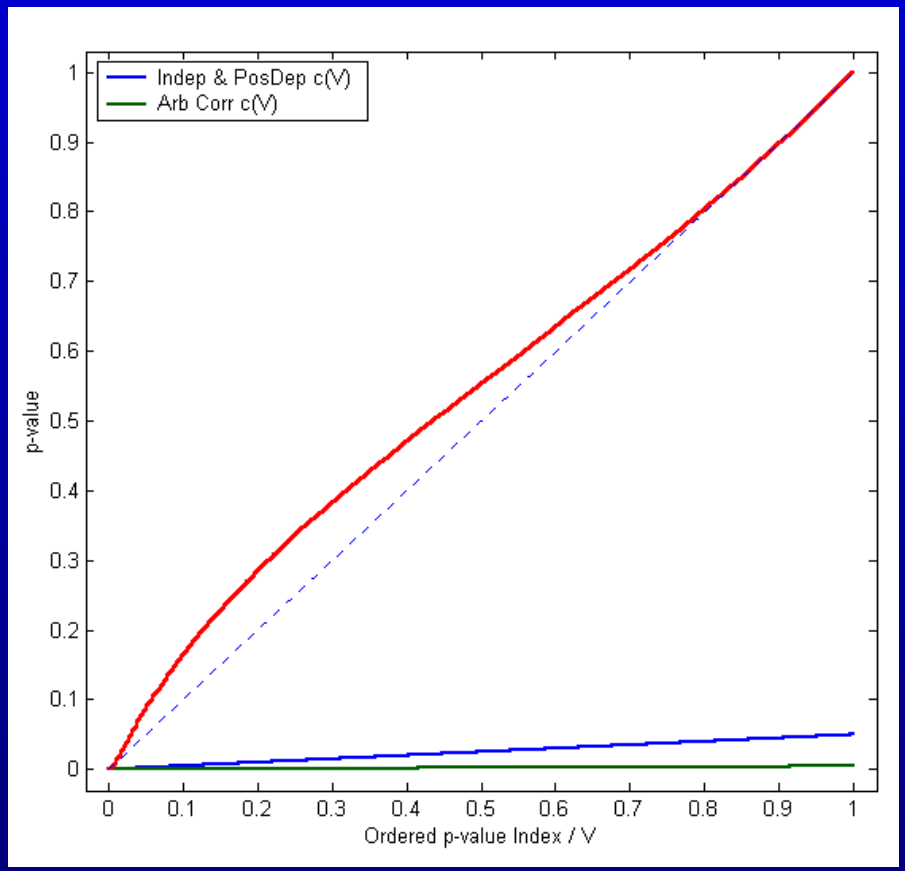
But both keep the term FDR for their versions

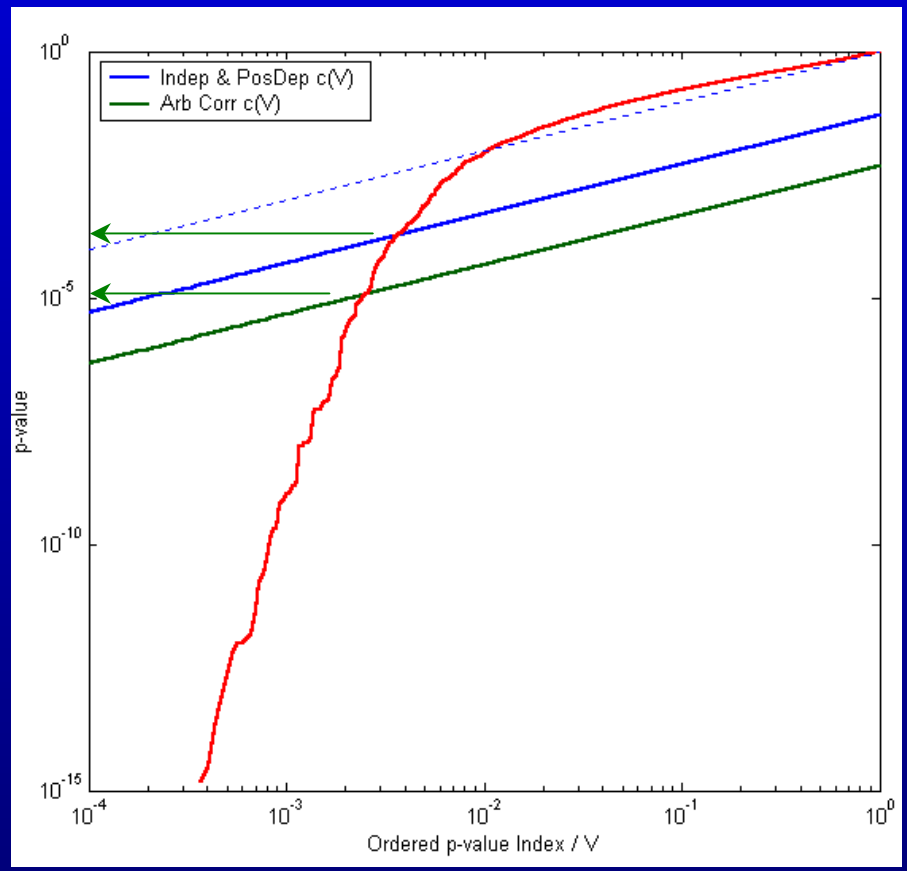
Benjamini & Hochberg c(m) factor

- $c(\mathrm{m}) = 1$
  - 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.
- $c(\mathrm{m}) = \Sigma_{i=1,\ldots \mathrm{m}}\ 1/i \ \approx \ \log(\mathrm{m})+0.5772$
  - Arbitrary covariance structure
    - But this is more conservative—tighter cuts

FDR Example: Plot of FDR Inequality

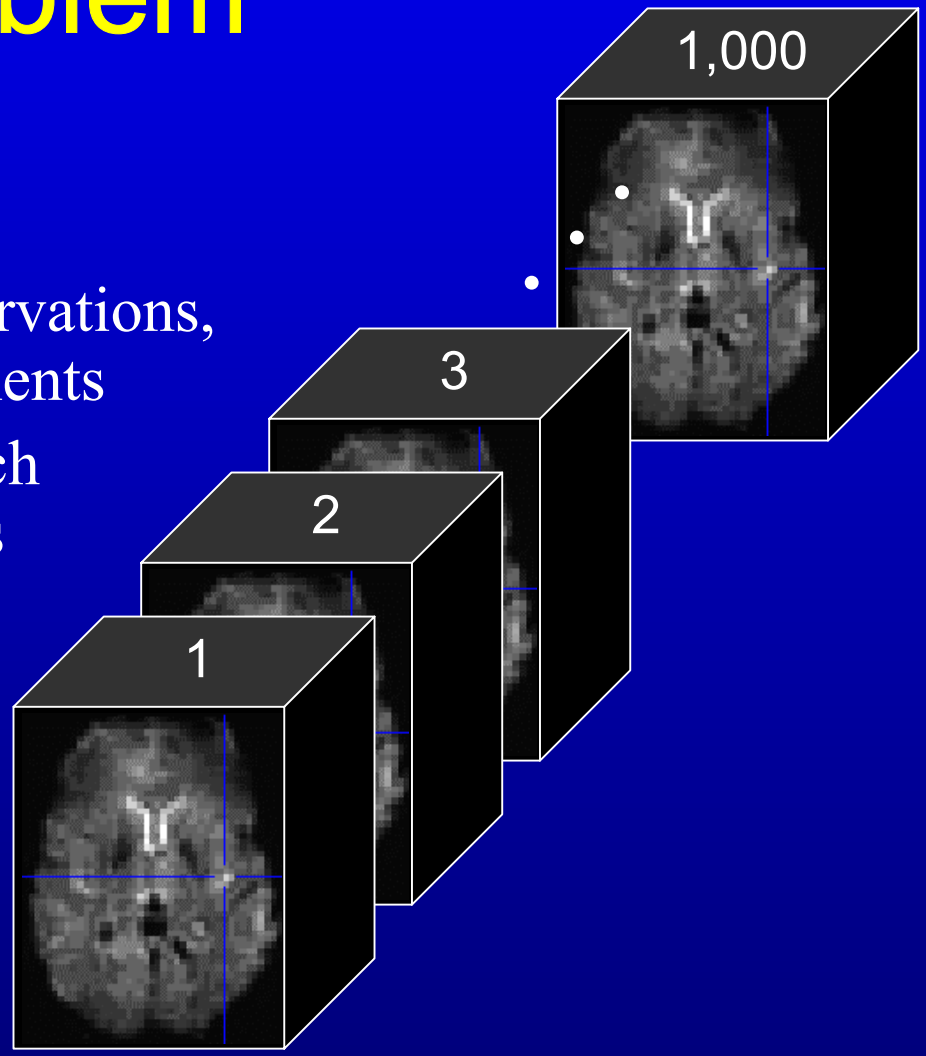
$$p_{(i)} \leq q\,(\,i/m)/c(m)$$



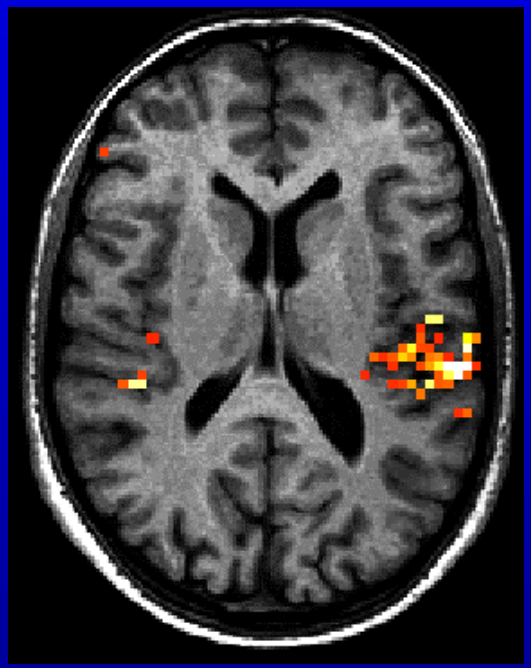


fMRI Multiple Comparisons Problem

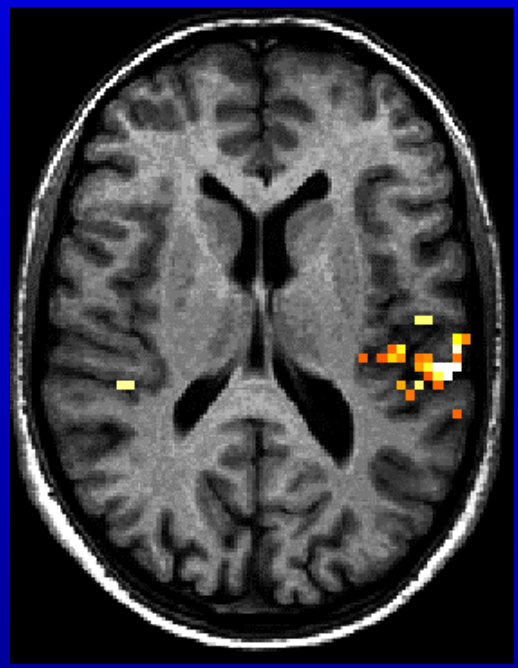
- 4-Dimensional Data
  - 1,000 multivariate observations, each with 100,000 elements
  - 100,000 time series, each with 1,000 observations
- Massively Univariate Approach
  - 100,000 hypothesis tests per image
- Massive MCP!



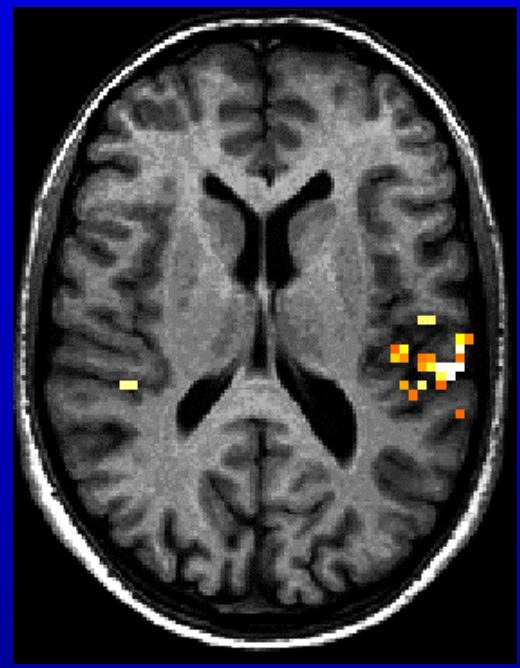
FDR: Example



FDR ≤ 0.05
Indep/PRDS
$t_0 = 3.8119$



FDR ≤ 0.05
Arbitrary Cov.
$t_0 = 5.0747$



FWER ≤ 0.05
Bonferroni
$t_0 = 5.485$

Positive dependency
(conditions for c(m) = 1)

- **P**ositive **R**egression **D**ependency on the **S**ubset of true null hypotheses:
- If the test statistics are $\mathbf{X}=(X_1,X_2,\ldots,X_m)$:
  - For any increasing set D, and H_{0i} true
  - Prob($\mathbf{X}$ in D | X_i=s) is increasing in s
- Important Examples
  - Multivariate Normal with positive correlation
  - Absolute Studentized independent normal
  - (Studentized PRDS distribution, for q<.5)

More about dependency

- If the test statistics are :
  - All Pairwise Comparisons: $x_i - x_j$ $i,j=1,2,\ldots k$

$$FDR \le \frac{m_0}{m} q$$

even though correlations between pairs of comparisons are both + and -

Based on many simulation studies:

Williams, Jones, & Tukey ('94,'99); YB, Hochberg, & Kling ('94+)
Kesselman, Cribbie, &Holland ('99).

And limited theoretical evidence

Yekutieli ('99+)

so the theoretical problem is still open...

Bonferroni-Holm
Sequential Variant of Bonferroni

Small change if m is large

- Like Bonferroni, controls total error to α across all m tests

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

  but stop at the first failure

  loosens cut mildly as more pass

  re-do Bonferroni, remove each rejected p: m → m-1

  identical to α/m if none pass

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

There are other variants: see for example

statwww.epfl.ch/davison/teaching/Microarrays/lec/week10.ppt

W. Penny:

Random Field Method

- For images with heavy correlation among pixels
  - Sampled finer than resolution
    - FWHM > 3 x pixel size (if not, too conservative: could cut harder)
  - Modeled as Gaussian correlation (random field)
- **RFT is nearly same as Bonferroni**

  with m = effective independent pixels (RESELs)

- RFT formula relates m, α, and u (threshold per pixel)

$$\alpha = m\,(4 \ln 2)\,(2\pi)^{-3/2}\,u \exp(-u^2/2) \quad \text{(2-d Gaussian)}$$

Example: $\alpha = .05$; 300 x 300 image; FWHM = 30

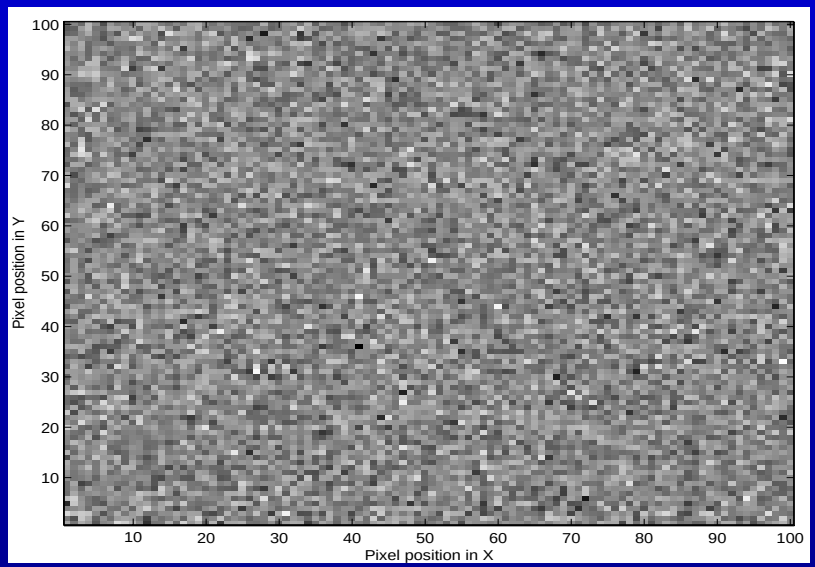
m = 300 x 300 / (30 x 30) = 100

Bonferroni gives u=3.3 RFT gives u = 3.8 (harder cut)

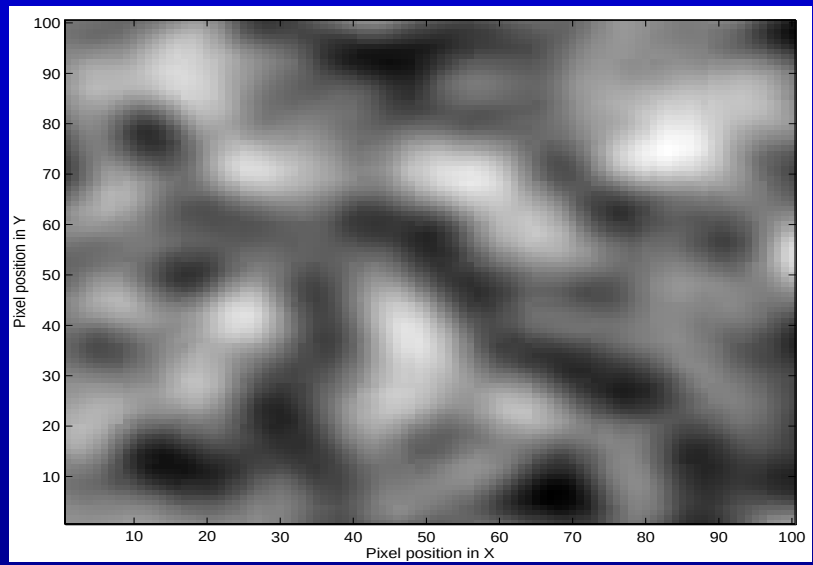
Friston et al. (1991) J. Cer. Bl Fl. M.

Correlated data

Independent Voxels



Spatially Correlated Voxels



Multiple comparisons terminology

- ***Family*** **of hypotheses**
  - $H^k \; k \in \Omega = \{1,\ldots,K\}$
  - $H^\Omega = H^1 \cap H^2 \ldots \cap H^k \cap H^K$
- ***Familywise*** **Type I error**
  - ***weak*** **control –** ***omnibus test***
    - Pr("reject" $H^\Omega \mid H^\Omega$) $\leq \alpha$
    - ***"anything, anywhere" ?***
  - ***strong*** **control –** ***localising test***
    - Pr("reject" $H^W \mid H^W$) $\leq \alpha$
      $\forall \; W: W \subseteq \Omega \;\&\; H^W$
    - ***"anything, & where" ?***

Null: Activation is zero everywhere

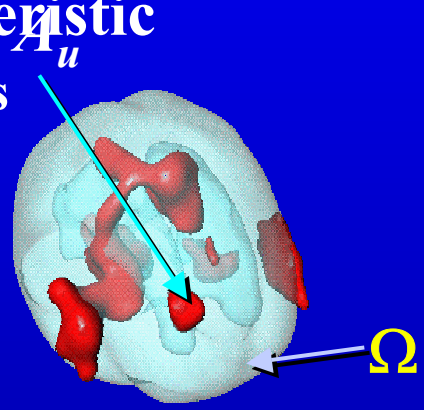
eg. Look at average activation over volume

eg. Look at maxima of statistical field for specific activation sites

Unified Theory: RFT

- **General form for expected Euler characteristic**
  - **χ^2, *F*, & *t* fields • restricted search regions**

$$\alpha = \sum R_d(\Omega)\,\rho_d(u)$$



$R_d(\Omega)$: **RESEL count**

$R_0(\Omega) = \chi(\Omega)$ **Euler characteristic of** Ω
$R_1(\Omega) =$ **resel diameter**
$R_2(\Omega) =$ **resel surface area**
$R_3(\Omega) =$ **resel volume**

$\rho_d(u)$: ***d*-dimensional EC density** –
E.g. Gaussian RF:

$$\rho_0(u) = 1 - \Phi(u)$$
$$\rho_1(u) = (4\ln 2)^{1/2} \exp(-u^2/2) / (2\pi)$$
$$\rho_2(u) = (4\ln 2) \exp(-u^2/2) / (2\pi)^{3/2}$$
$$\rho_3(u) = (4\ln 2)^{3/2} (u^2 - 1) \exp(-u^2/2) / (2\pi)^2$$
$$\rho_4(u) = (4\ln 2)^2 (u^3 - 3u) \exp(-u^2/2) / (2\pi)^{5/2}$$

Worsley et al. (1996), HBM

Benjamini:

Motivating Examples

- High throughput screening
  - Of Chemical compounds
  - Of gene expression
- Data Mining
  - Mining of Association Rules
  - Model Selection

High throughput screening of Chemical Compounds

- Purpose: at early stages of drug development, screen a large number of potential chemical compounds, in order to find any interaction with a given class of compounds (a "hit")
- The classes may be substructures of libraries of compounds involving up to 10^5 members.
- Each potential compound interaction with class member is tested once and only once

Microarrays and Multiplicity

- Neglecting multiplicity issues, i.e. working at the individual 0.05 level, would identify, on the average, 6359*0.05=318 differentially expressed genes, even if really no such gene exists.
- Addressing multiplicity with Bonferroni at 0.05 identifies 8

Table 1: First 12 Largest T-Statistics [1,2]

T-Statistic	P-Value (df=14)
-20.6	$7.0*10^{-12}$
-12.5	$5.6*10^{-9}$
-11.9	$1.1*10^{-8}$
-11.7	$1.3*10^{-8}$
-11.4	$1.8*10^{-8}$
-11.3	$1.9*10^{-8}$
-7.8	$1.8*10^{-6}$
-7.4	$3.6*10^{-6}$
5.0	$1.8*10^{-4}$
-4.5	$4.6*10^{-4}$
-4.5	$4.9*10^{-4}$
-4.4	$6.5*10^{-4}$

1. The t-statistics were ranked according to their absolute values.
2. Bonferroni adjusted p-value is $1.6*10^{-4}$.

Mining of association rules in Basket Analysis

- A basket bought at the food store consists of:

  (Apples, Bread,Coke,Milk,Tissues)

Data on all baskets is available (through cash registers)

The goal: Discover association rules of the form

Bread&Milk => Coke&Tissue

Also called linkage analysis or item analysis

Model Selection

Paralyzed veterans of America

Mailing list of 3.5 M potential donors

200K made their last donation 1-2 years ago

Is there something better than mailing all 200K?

- If all mailed, net donation is $10,500
- FDR-like modeling raised to $14,700

What's in common?

- Size of the problem: large to huge

(m small n large ;m=n large; m large n small)

- Question 1: Is there a real effect at a specific gene/site/location/association rule?
- Question 2: If there is an effect, of what size?
- Discoveries are further studied; negative results are usually ignored
- Results should be communicated compactly to a wide audience
- A threshold is being used for question 1.

Model Selection in large problems

- known approaches to model selection
  - Penalize error rate for using k parameters
  - AIC and Cp

$$SSR\ (k) + \sigma^2 k \cdot 2$$

  - .05 in testing "forward selection" or "backward elimination

$$SSR\ (k) + \sigma^2 k \cdot Z^2_{\frac{.05}{2}}$$

  - The Universal Threshold of Donoho and Johnstone

$$SSR\ (k) + \sigma^2 k \cdot 2 \log\ m$$

Model Selection and FDR - Practical Theory

The theory is being developed for the minimizer of the following penalized Sum of Squared Residuals:

$$SSR(k) + \sigma^2 \sum_{i=1}^{k} z^2_{\frac{iq}{m}}$$

$$\approx SSR(k) + \sigma^2 k \cdot z^2_{\frac{kq}{m}} \approx SSR(k) + \sigma^2 k \cdot 2\log(m2/kq)$$

2 AIC

$2\log(m)$

The Linear Step-Up is Essentially “backwards elimination” (and close to “forward selection”) with the above penalty function :

1. Linear StepUp Procedure

- If the test statistics are :
  - Independent YB&Yekutieli (‘01)

$$FDR \le \frac{m_0}{m} q$$

  - independent and continuous YB&Hochberg (‘95)

$$FDR = \frac{m_0}{m} q$$

  - Positive dependent YB&Yekutieli (‘01)

$$FDR \le \frac{m_0}{m} q$$

  - General YB&Yekutieli (‘01)

$$FDR \le \frac{m_0}{m} q\,(1 + 1/2 + 1/3 + \ldots + 1/m)$$
$$\approx \frac{m_0}{m} q \log(m)$$

Adaptive procedures that control FDR

- Recall the m_0/m factor of conservativeness
- Hence: if m_0 is known using linear step-up procedure with $qi/m(m/m_0) = qi/m_0$ controls the FDR at level q exactly.
- The adaptive procedure BY & Hochberg (‘00):
  - Estimate m_0 from the uniform q-q plot of the p-values
- This is FDR controlling under independence (via simulations)

Testimation - some theory

- In the independent problem
- Consider #(parameters) -> infinity
  - If prop(non-zero coefficients) -> 0,
  - Or If size of sorted coefficients decays fast, (while the others need not be exactly 0).
  - THEN thresholding by FDR testing of the coefficients is adaptively minimax over bodies of sparse signals
  - Where performance measured by any loss $0 < p \leq 2$: #(errors), sum|error|, sum(error)^2, relative to best "oracle" performance.

Abramovich, YB, Donoho, & Johnstone ('00+)