

Multiple Testing and Thresholding

NITP, 2010

Thanks for the slides Tom Nichols!

Overview

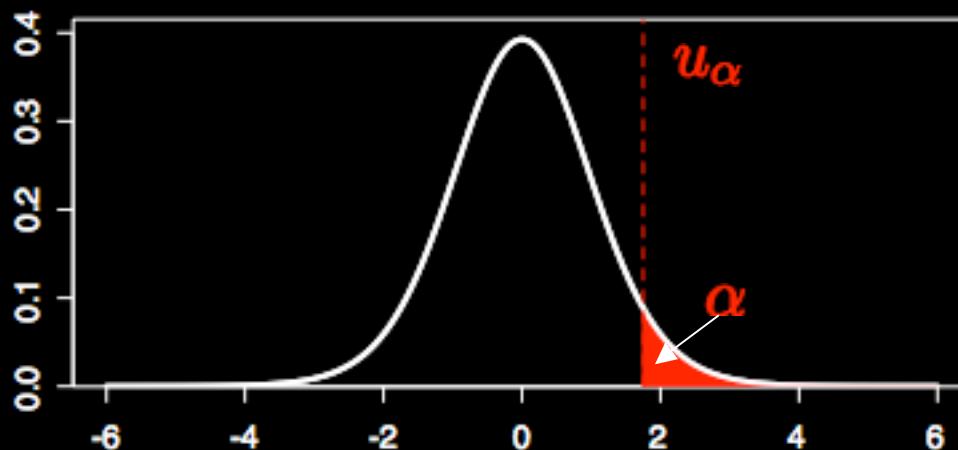
- Multiple Testing Problem
 - Which of my 100,000 voxels are “active”?
- Two methods for controlling false positives
 - Familywise Error Rate
 - Controlling the chance of any false positives
 - Bonferroni, Random Field and Nonparametric Methods
 - False Discovery Rate
 - Controlling the fraction of false positives

Overview

- Multiple Testing Problem
 - Which of my 100,000 voxels are “active”?
- Two methods for controlling false positives
 - Familywise Error Rate
 - Controlling the chance of any false positives
 - Bonferroni, Random Field and Nonparametric Methods
 - False Discovery Rate
 - Controlling the fraction of false positives

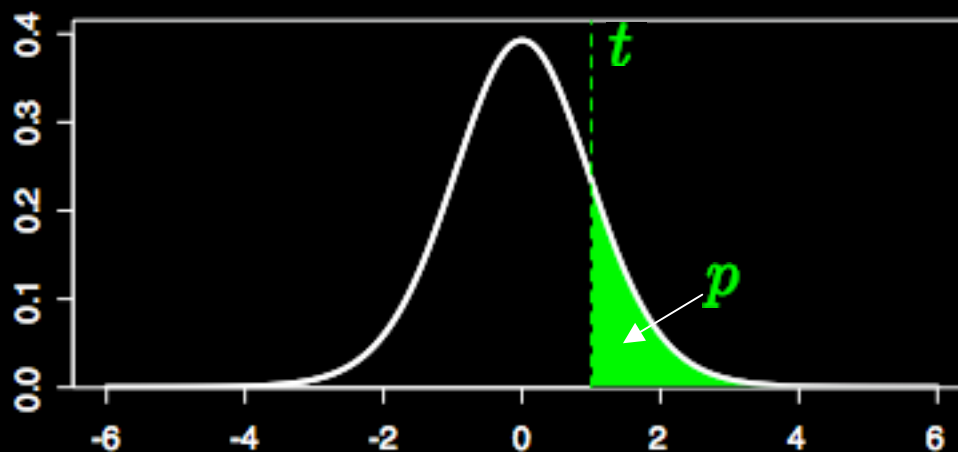
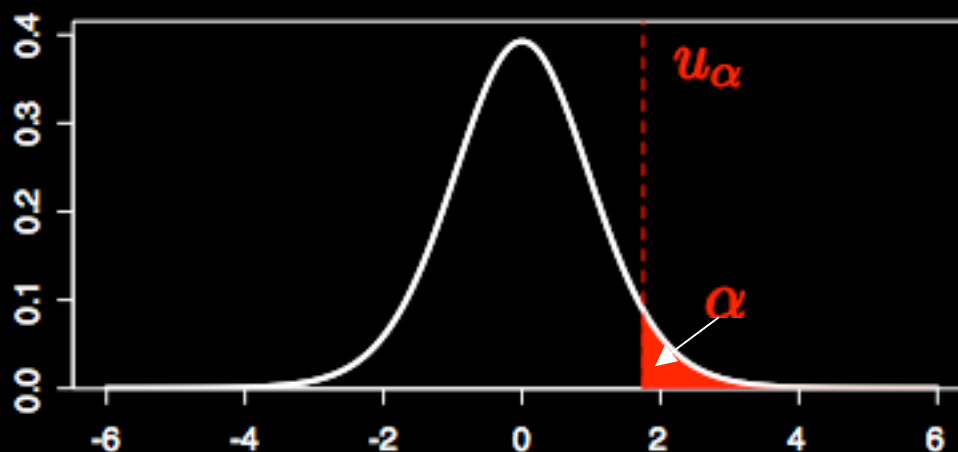
Hypothesis Testing Review

- Establish H_0 : no activation in voxel i
- Establish significance level α
 - Derive threshold u_α



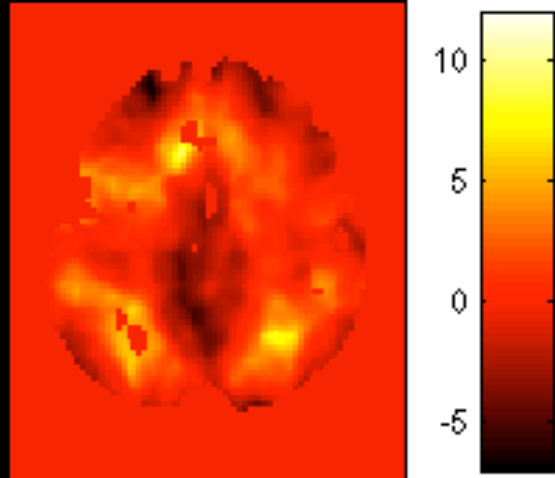
Hypothesis Testing Review

- Establish H_0 : no activation in voxel i
- Establish significance level α
 - Derive threshold u_α
- Calculate test statistic t
- P-value
 - $P(T > t | H_0)$
- Decision: Reject or accept H_0



Hypothesis Testing in fMRI

- Mass Univariate Modeling
 - Fit a separate model for each voxel
 - Look at images of statistics

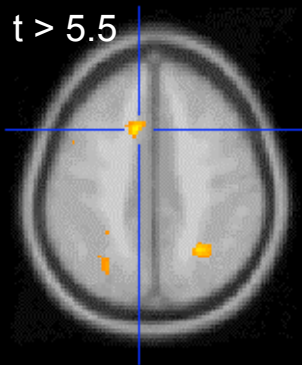


- Apply Threshold...

Assessing Statistic Images

- What threshold will show us signal?

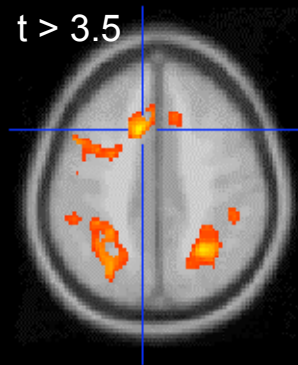
High Threshold



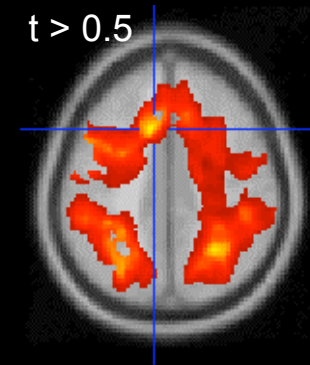
Good Specificity

Poor Power
(risk of false negatives)

Med. Threshold



Low Threshold

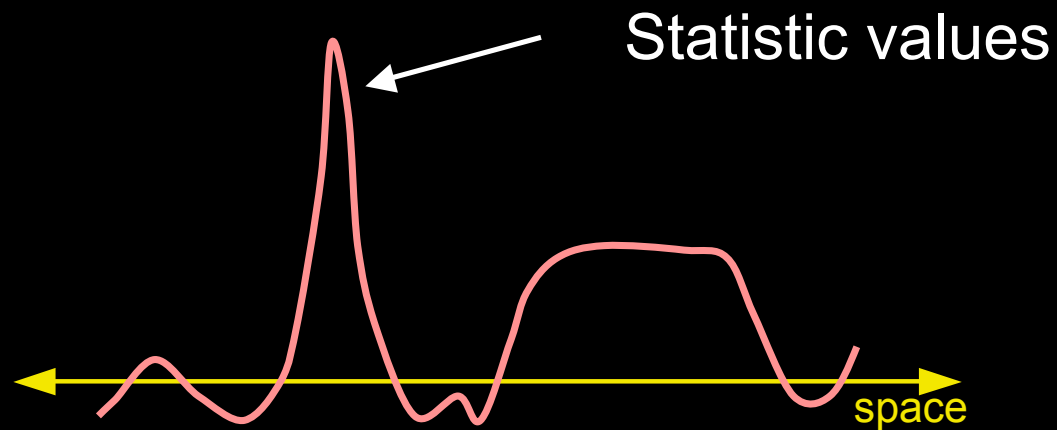


Poor Specificity
(risk of false positives)

Good Power

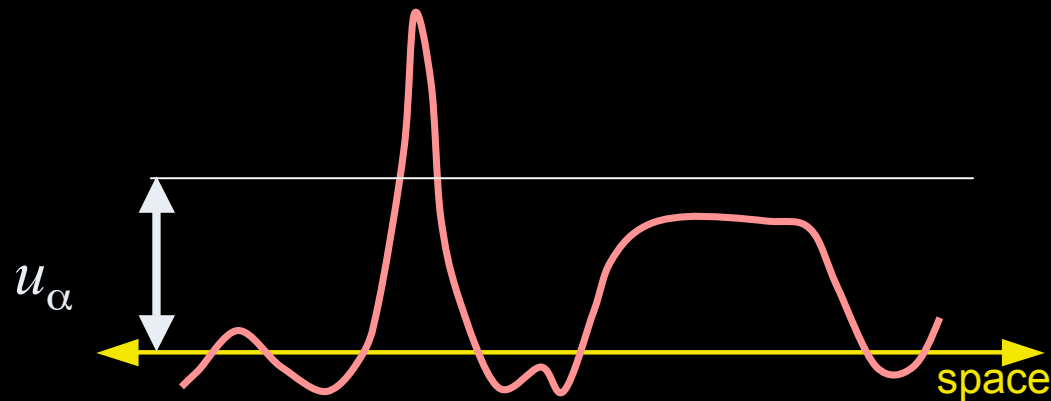
Voxel-level Inference

- Retain voxels above α -level threshold u_α
- Gives best spatial specificity
 - The null hyp. at a single voxel can be rejected



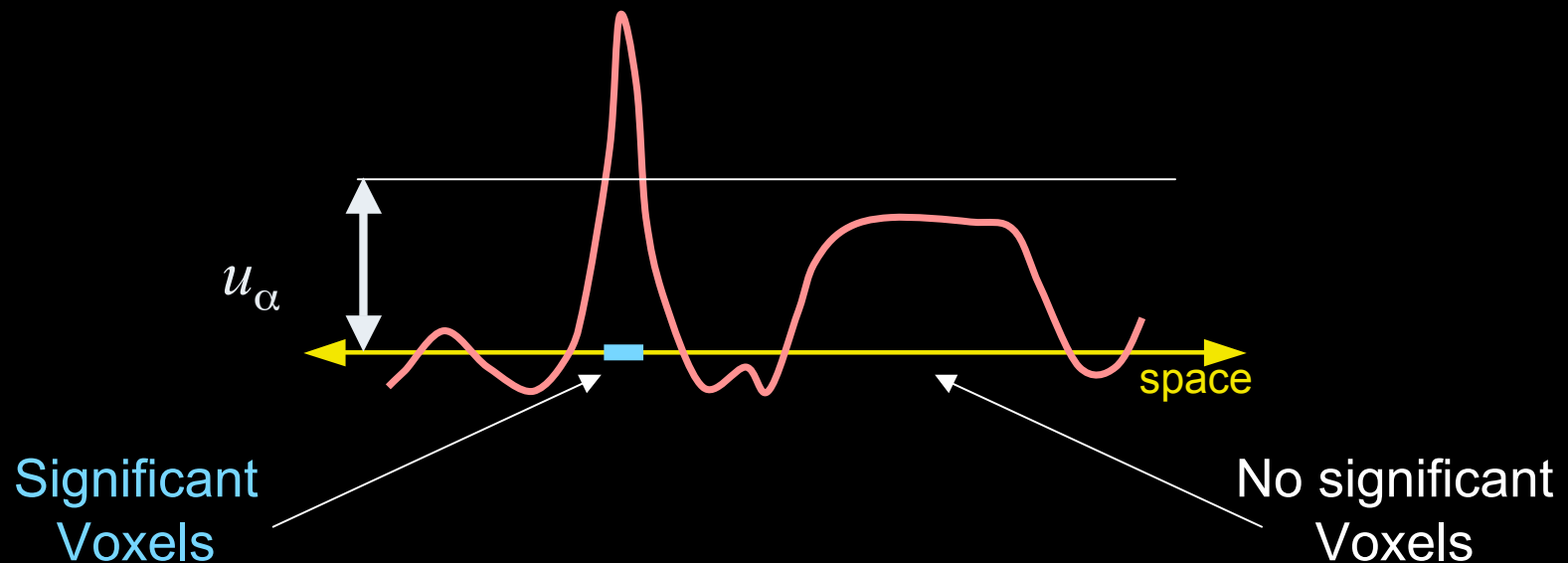
Voxel-level Inference

- Retain voxels above α -level threshold u_α
- Gives best spatial specificity
 - The null hyp. at a single voxel can be rejected



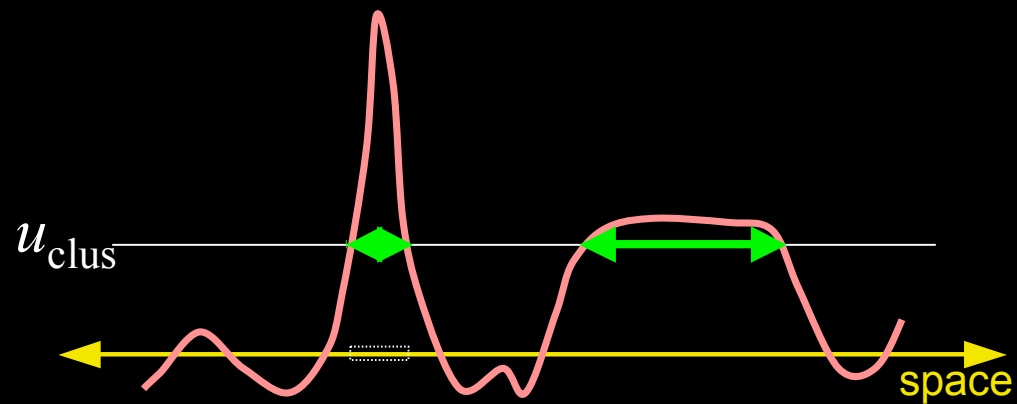
Voxel-level Inference

- Retain voxels above α -level threshold u_α
- Gives best spatial specificity
 - The null hyp. at a single voxel can be rejected



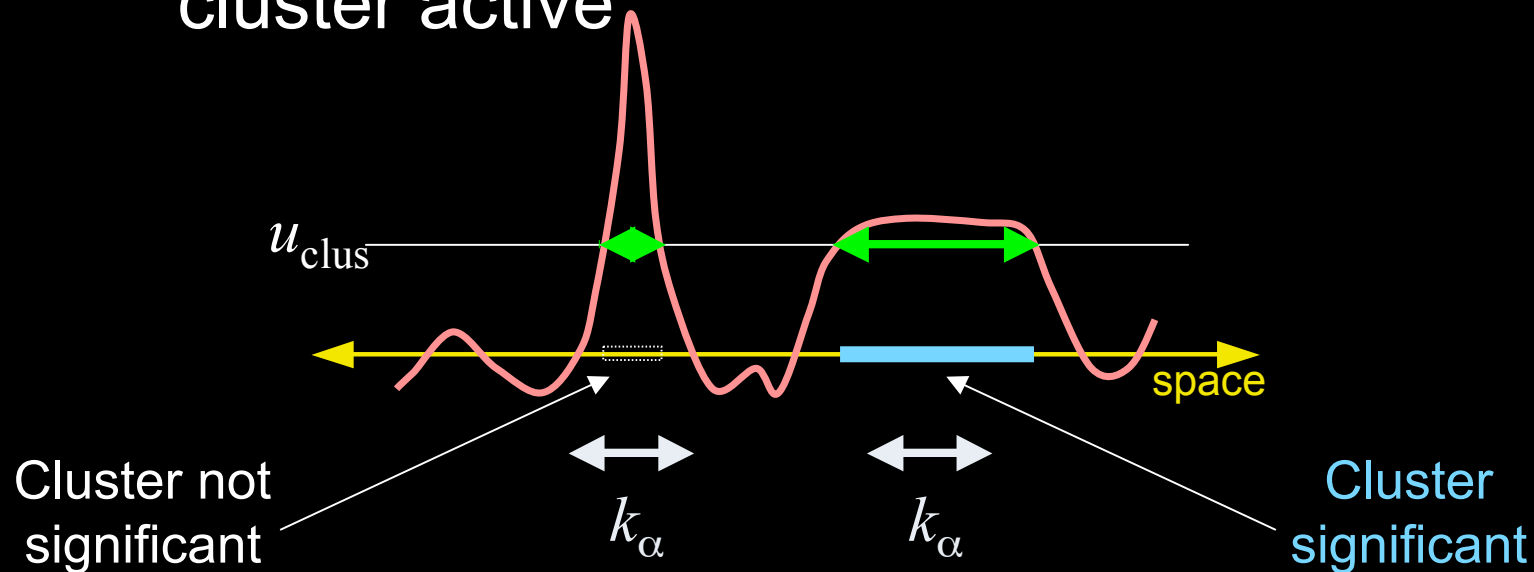
Cluster-level Inference

- Two step-process
 - Define clusters by arbitrary threshold u_{clus}



Cluster-level Inference

- Typically better sensitivity
- Worse spatial specificity
 - The null hyp. of entire cluster is rejected
 - Only means that *one or more* of voxels in cluster active



Voxel-wise Inference & Multiple Testing Problem (MTP)

- Standard Hypothesis Test
 - Controls Type I error of each test, at say 5%
 - But what if I have 100,000 voxels?
 - 5,000 false positives on average!
- Must control false positive rate
 - What false positive rate?
 - Chance of 1 or more Type I errors?
 - Proportion of Type I errors?

Overview

- Multiple Testing Problem
 - Which of my 100,000 voxels are “active”?
- Two methods for controlling false positives
 - Familywise Error Rate
 - Controlling the chance of any false positives
 - Bonferroni, Random Field and Nonparametric Methods
 - False Discovery Rate
 - Controlling the fraction of false positives

FWER MTP Solutions

- Bonferroni
- Maximum Distribution Methods
 - Random Field Theory
 - Permutation

Bonferroni

- Based on the Bonferroni inequality

- $P(E_1 \text{ or } E_2 \text{ or } \dots E_n) \leq \sum_{i=1}^n P(E_i)$

- If $P(Y_i \text{ passes}|H_0) \leq \alpha/n$ then

- $P(\text{some } Y_i \text{ passes}|H_0) \leq \sum P(Y_i \text{ passes}|H_0) \leq \alpha$

- For 100,000 voxels

- $\alpha = 0.05/100,000 = 0.0000005$

Bonferroni

- Can be too conservative
- Bonferroni assumes all tests are independent
- fMRI data tend to be spatially correlated
 - # of independent tests $<$ # voxels

Bonferroni

- Where does spatial correlation come from?
 - How images are constructed from the scanner
 - Physiologic signal
 - Preprocessing steps (realignment, smoothing, etc.)

Why not use a spatial model

- If we can model temporal correlation, why not spatial?
- Need an explicit spatial model
- No routine spatial modeling methods exist
 - High-dimensional mixture modeling problem
 - Activations don't look like Gaussian blobs
 - Need realistic shapes, sparse representation
 - Some work by Hartvig *et al.*, Penny *et al.*

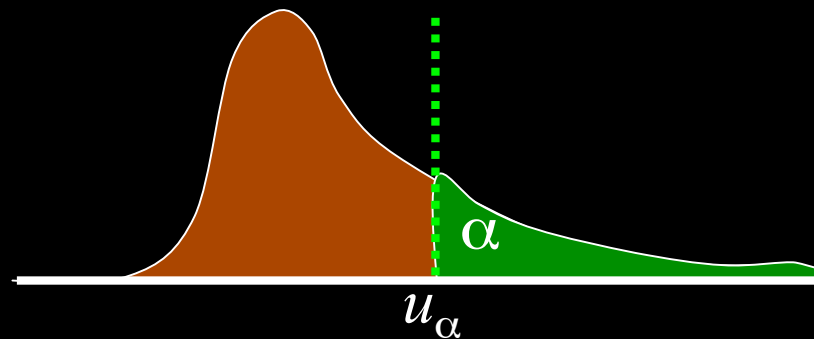
FWER MTP Solutions: Controlling FWER w/ Max

- FWER & distribution of maximum

$$\begin{aligned}\text{FWER} &= P(\text{FWE}) \\ &= P(\text{One or more voxels} \geq u \mid H_0) \\ &= P(\text{Max voxel} \geq u \mid H_0)\end{aligned}$$

- $100(1-\alpha)\%$ ile of max distⁿ controls FWER

$$\text{FWER} = P(\text{Max voxel} \geq u_\alpha \mid H_0) \leq \alpha$$



FWER MTP Solutions: Random Field Theory

- Euler Characteristic χ

- Topological Measure

- #blobs - #holes

- At high thresholds,
just counts blobs

- FWER = $P(\text{Max voxel} \geq u \mid H_0)$

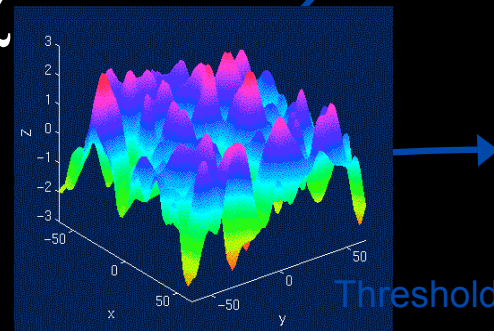
No holes

= $P(\text{One or more blobs} \mid H_0)$

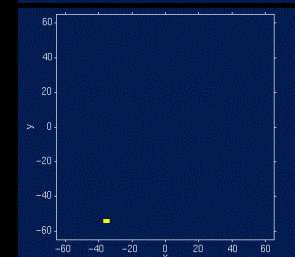
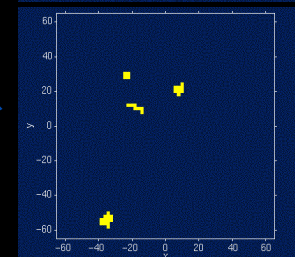
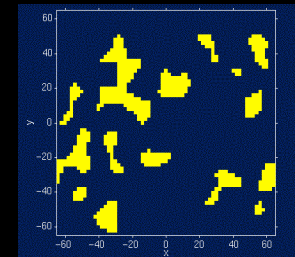
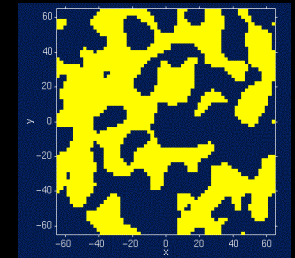
*Never more
than 1 blob*

$\approx P(\chi_u \geq 1 \mid H_0)$

$\approx E(\chi_u \mid H_0)$



Threshold



Suprathreshold Sets

Distribution details

- Math is hairy!
 - Nichols and Hayasaka 2003
 - Cao and Worsley 2001
- What you need to know
 - Depends on “smoothness” of your image
 - Must quantify smoothness and it is important to report when using RFT

General idea

- $E(\chi_u) \approx \text{Mathy stuff} * \text{Volume/Smoothness}$
- We know what the volume is
- What is smoothness?

Smoothness

- How smooth are the data?
 - Measured by $\text{FWHM}=[\text{FWHM}_x, \text{FWHM}_y, \text{FWHM}_z]$
 - Starting with white noise smooth with a gaussian
 - How large does the variance of that gaussian need to be such that the smoothness matches your data?

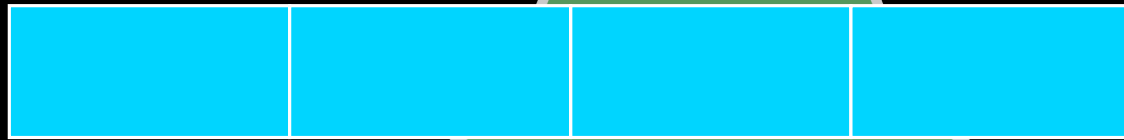
RESEL

- RESolution Element
 - $\text{RESEL} = \text{FWHM}_x \times \text{FWHM}_y \times \text{FWHM}_z$
- RESEL count
 - If your voxels were the size of a RESEL, how many are required to fill your volume?
 - 10 voxels, 2.5 voxel FWHM smoothness
⇒ 4 RESELS

voxels



FWHM=
2.5 voxels



RESEL count=4

Note about RESELS

- Not the number of independent tests
 - Not the magic bullet for a better Bonferroni
- Re-expression of volume in terms of smoothness
- We need it, since it is necessary to calculate our p-values

Revisit distribution

- $E(\chi_u) \approx \text{Mathy stuff} * \text{Volume/Smoothness}$
- Smoothness is defined in RESELS
- $E(\chi_u)$ is our p-value
 - How does a p-value change as volume increases?
 - How does a p-value change as smoothness increases?

RFT adapts

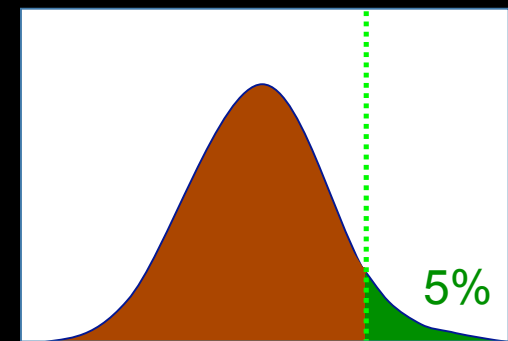
- For larger volumes it is more strict
 - Multiple comparison problem is worse
- For smoother data it is less strict
 - Multiple comparison problem is less severe

Shortcomings of RFT

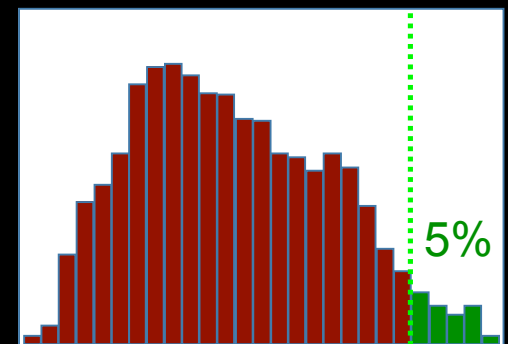
- Requires estimating a lot of parameters
- Random field must be sufficiently smooth
 - If you don't spatially smooth the data enough, RFT doesn't work well
- We'll see comparisons in a bit!

Nonparametric Permutation Test

- Parametric methods
 - Assume distribution of statistic under null hypothesis
- Nonparametric methods
 - Use *data* to find distribution of statistic under null hypothesis
 - Any statistic!



Parametric Null Distribution



Nonparametric Null Distribution

Permutation Test Toy Example

- Data from voxel in visual stim. experiment
A: Active, flashing checkerboard B: Baseline, fixation
6 blocks, ABABAB Just consider block averages...

A	B	A	B	A	B
103.00	90.48	99.93	87.83	99.76	96.06

- Null hypothesis H_0
 - No experimental effect, A & B labels arbitrary
- Statistic
 - Mean difference

Permutation Test Toy Example

- Under H_0
 - Consider all equivalent relabelings

AAABBB

ABABAB

BAAABB

BABBAA

AABABB

ABABBA

BAABAB

BBAAAB

AABBAB

ABBAAB

BAABBA

BBAABA

AABBBA

ABBABA

BABAAB

BBABAA

ABAABB

ABBBA

BABABA

BBBAAA

Permutation Test Toy Example

- Under H_0
 - Consider all equivalent relabelings
 - Compute all possible statistic values

AAABBB	4.82	ABABAB	9.45	BAAABB	-1.48	BABBAA	-6.86
AABABB	-3.25	ABABBA	6.97	BAABAB	1.10	BBAAAB	3.15
AABBAB	-0.67	ABBAAB	1.38	BAABBA	-1.38	BBAABA	0.67
AABBBA	-3.15	ABBABA	-1.10	BABAAB	-6.97	BBABAA	3.25
ABAABB	6.86	ABBBA	1.48	BABABA	-9.45	BBBAAA	-4.82

Permutation Test Toy Example

- Under H_0
 - Consider all equivalent relabelings
 - Compute all possible statistic values
 - Find 95%ile of permutation distribution

AAABBB	4.82	ABABAB	9.45	BAAABB	-1.48	BABBAA	-6.86
AABABB	-3.25	ABABBA	6.97	BAABAB	1.10	BBAAAB	3.15
AABBAB	-0.67	ABBAAB	1.38	BAABBA	-1.38	BBAABA	0.67
AABBBA	-3.15	ABBABA	-1.10	BABAAB	-6.97	BBABAA	3.25
ABAABB	6.86	ABBBA	1.48	BABABA	-9.45	BBBAAA	-4.82

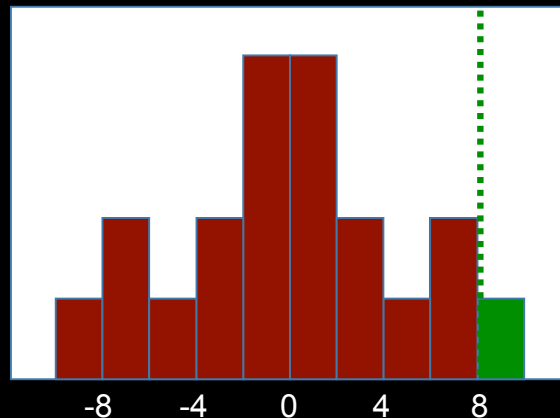
Permutation Test Toy Example

- Under H_0
 - Consider all equivalent relabelings
 - Compute all possible statistic values
 - Find 95%ile of permutation distribution

AAABBB	4.82	ABABAB	9.45	BAAABB	-1.48	BABBAA	-6.86
AABABB	-3.25	ABABBA	6.97	BAABAB	1.10	BBAAAB	3.15
AABBAB	-0.67	ABBAAB	1.38	BAABBA	-1.38	BBAABA	0.67
AABBBA	-3.15	ABBABA	-1.10	BABAAB	-6.97	BBABAA	3.25
ABAABB	6.86	ABBBA	1.48	BABABA	-9.45	BBBAAA	-4.82

Permutation Test Toy Example

- Under H_0
 - Consider all equivalent relabelings
 - Compute all possible statistic values
 - Find 95%ile of permutation distribution

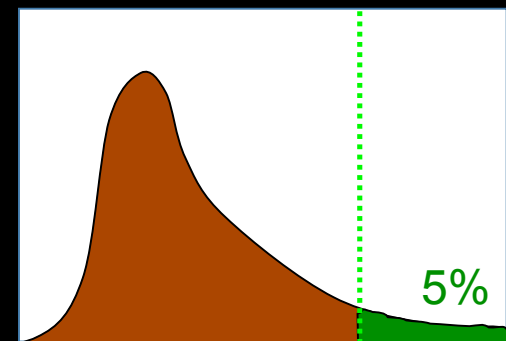


Small Sample Sizes

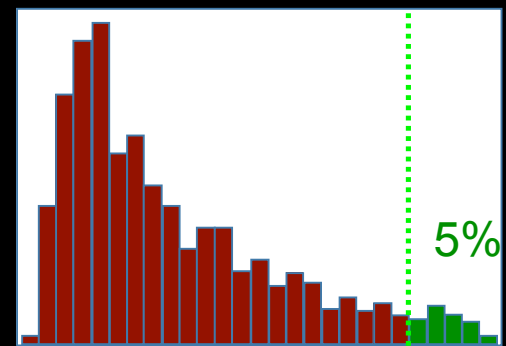
- Permutation test doesn't work well with small sample sizes
 - Possible p-values for previous example:
 - 0.05, 0.1, 0.15, 0.2, etc
 - Tends to be conservative for small sample sizes

Controlling FWER: Permutation Test

- Parametric methods
 - Assume distribution of *max* statistic under null hypothesis
- Nonparametric methods
 - Use *data* to find distribution of *max* statistic under null hypothesis
 - Again, any max statistic!



Parametric Null Max Distribution



Nonparametric Null Max Distribution

Permutation Test & Exchangeability

- Exchangeability is fundamental
 - Def: Distribution of the data unperturbed by permutation
 - Under H_0 , exchangeability justifies permuting data
 - Allows us to build permutation distribution

Permutation Test & Exchangeability

- Subjects are exchangeable
 - Under H_0 , each subject's A/B labels can be flipped
- fMRI scans are not exchangeable under H_0
 - If no signal, can we permute over time?
 - No, permuting disrupts order, temporal autocorrelation

Permutation Test & Exchangeability

- Two sample t test
 - Compare subjects in group 1 to subjects in group 2
 - Randomly assign group labels in permutations
- One sample t test
 - Randomly flip sign of values for some subjects

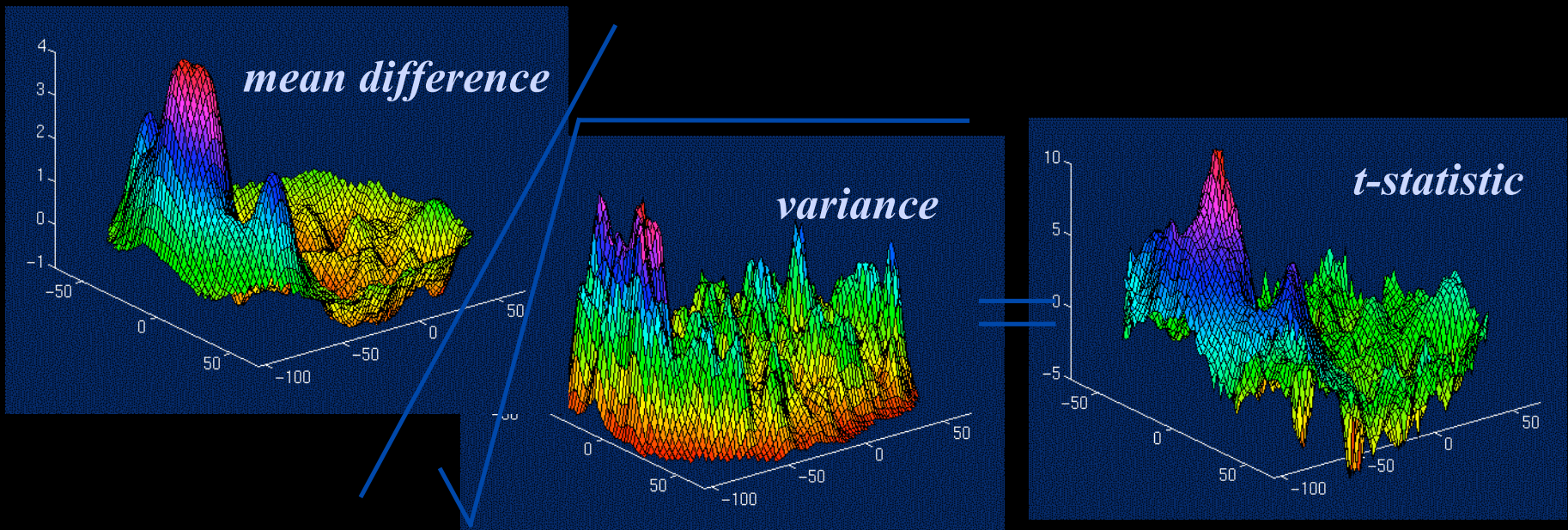
Permutation Test

Other Statistics

- Collect max distribution
 - To find threshold that controls FWER
- Consider smoothed variance t statistic
 - To regularize low-df variance estimate

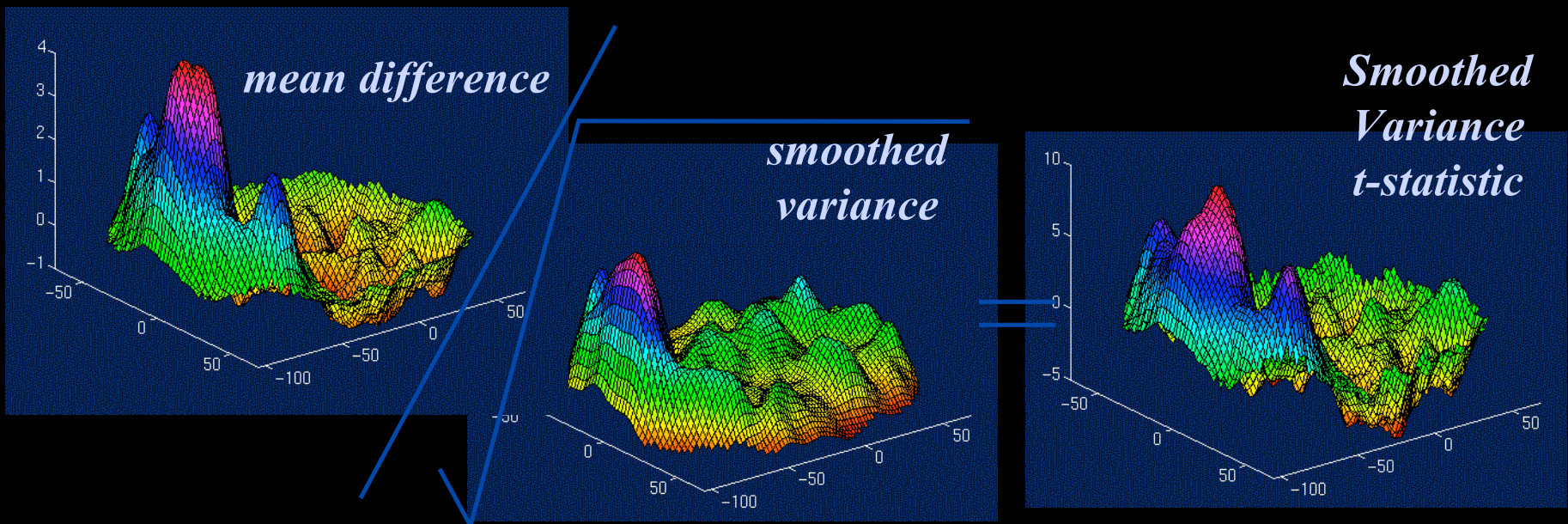
Permutation Test Smoothed Variance t

- Collect max distribution
 - To find threshold that controls FWER
- Consider smoothed variance t statistic



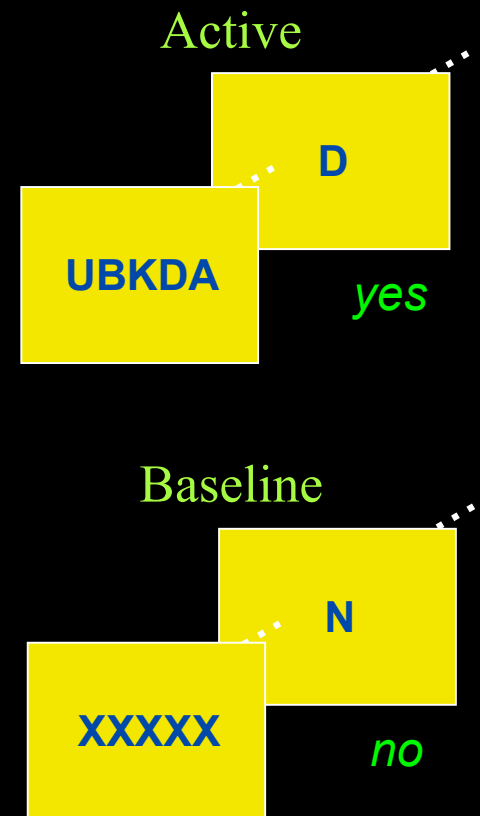
Permutation Test Smoothed Variance t

- Collect max distribution
 - To find threshold that controls FWER
- Consider smoothed variance t statistic



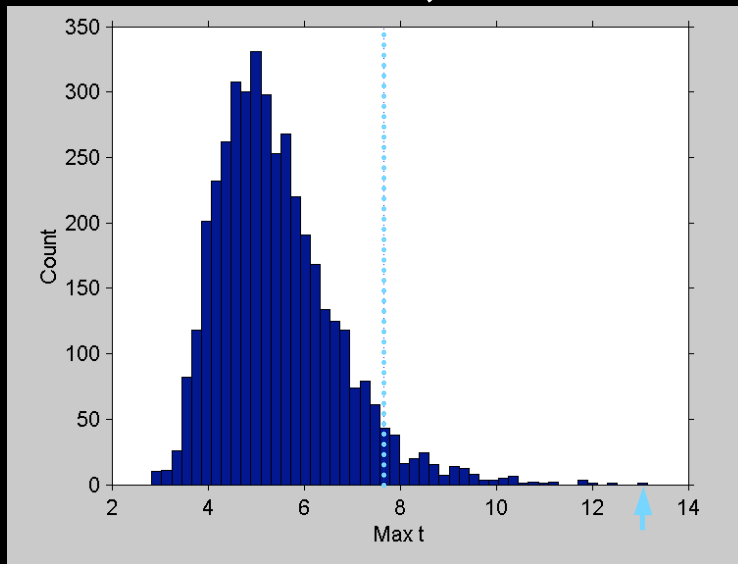
Permutation Test Example

- fMRI Study of Working Memory
 - 12 subjects, block design Marshuetz et al (2000)
 - Item Recognition
 - **Active**: View five letters, 2s pause, view probe letter, **respond**
 - **Baseline**: View XXXXX, 2s pause, view Y or N, **respond**
- Second Level RFX
 - Difference image, A-B constructed for each subject
 - One sample *t* test

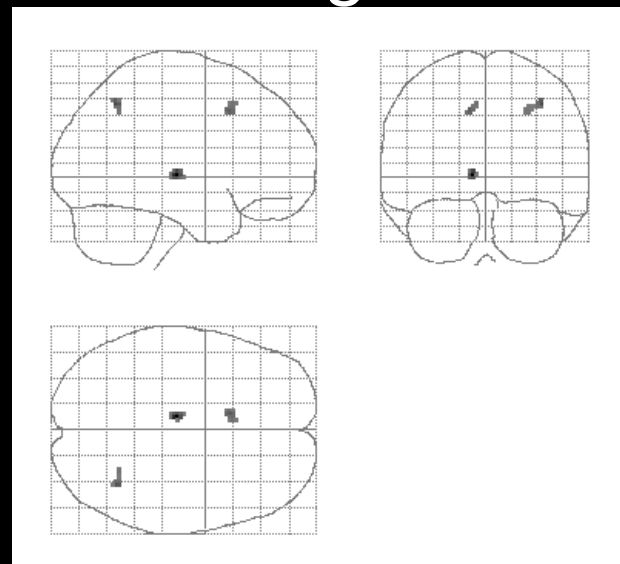


Permutation Test Example

- Permute!
 - $2^{12} = 4,096$ ways to flip 12 A/B labels
 - For each, note maximum of t image



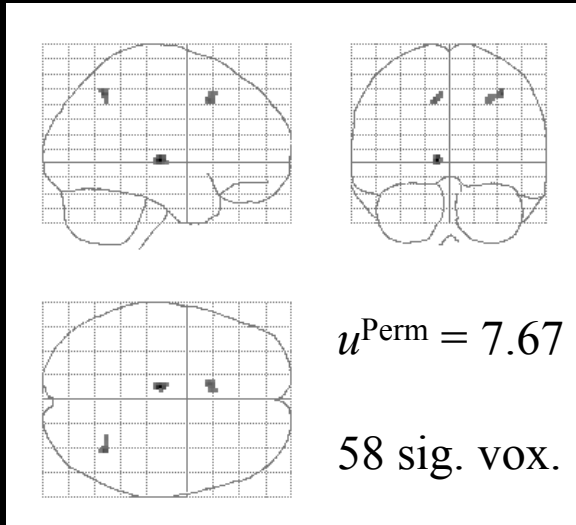
Permutation Distribution
Maximum t



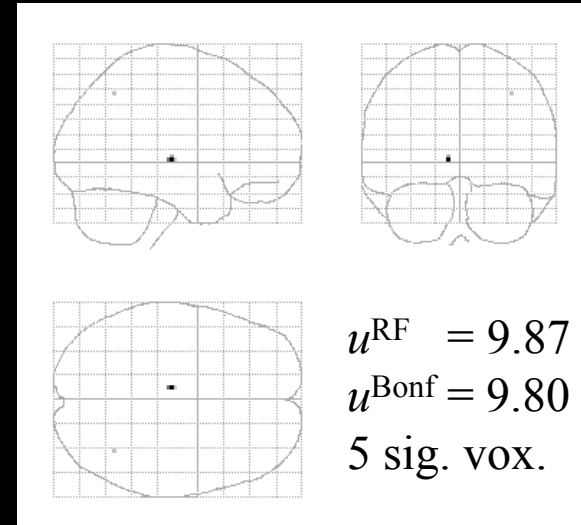
Maximum Intensity Projection
Thresholded t

Permutation Test Example

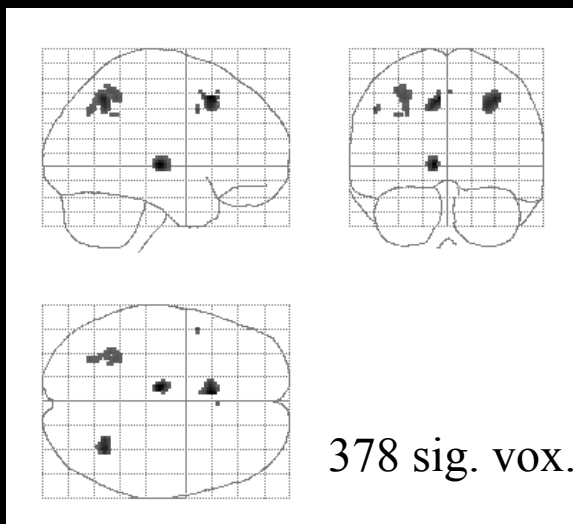
- Compare with Bonferroni
 - $\alpha = 0.05/110,776$
- Compare with parametric RFT
 - 110,776 $2 \times 2 \times 2$ mm voxels
 - $5.1 \times 5.8 \times 6.9$ mm FWHM smoothness
 - 462.9 RESELS
- Compare with smoothed variance T based permutation test



t_{11} Statistic, Nonparametric Threshold



t_{11} Statistic, RF & Bonf. Threshold



Smoothed Variance t Statistic,
Nonparametric Threshold

- RFT threshold is conservative (not smooth enough, d.f. too small)
- Permutation test is more efficient than Bonferroni since it accounts for smoothness
- Smooth variance is more efficient for small d.f.

Does this Generalize?

RFT vs Bonf. vs Perm.

		<i>t</i> Threshold (0.05 Corrected)		
	df	RF	Bonf	Perm
Verbal Fluency	4	4701.32	42.59	10.14
Location Switching	9	11.17	9.07	5.83
Task Switching	9	10.79	10.35	5.10
Faces: Main Effect	11	10.43	9.07	7.92
Faces: Interaction	11	10.70	9.07	8.26
Item Recognition	11	9.87	9.80	7.67
Visual Motion	11	11.07	8.92	8.40
Emotional Pictures	12	8.48	8.41	7.15
Pain: Warning	22	5.93	6.05	4.99
Pain: Anticipation	22	5.87	6.05	5.05

Does this Generalize?

RFT vs Bonf. vs Perm.

	df	No. Significant Voxels (0.05 Corrected)			
		<i>t</i>			SmVar <i>t</i>
		RF	Bonf	Perm	Perm
Verbal Fluency	4	0	0	0	0
Location Switching	9	0	0	158	354
Task Switching	9	4	6	2241	3447
Faces: Main Effect	11	127	371	917	4088
Faces: Interaction	11	0	0	0	0
Item Recognition	11	5	5	58	378
Visual Motion	11	626	1260	1480	4064
Emotional Pictures	12	0	0	0	7
Pain: Warning	22	127	116	221	347
Pain: Anticipation	22	74	55	182	402

Overview

- Multiple Testing Problem
 - Which of my 100,000 voxels are “active”?
- Two methods for controlling false positives
 - Familywise Error Rate
 - Controlling the chance of any false positives
 - Bonferroni, Random Field and Nonparametric Methods
 - False Discovery Rate
 - Controlling the fraction of false positives

False Discovery Rate

- For any threshold, all voxels can be cross-classified:

	Accept Null “Negative”	Reject Null “Positive”	
Null True (no effect)	V_{0N}	V_{0P}	m_0
Null False (true effect)	V_{1N}	V_{1P}	m_1
	V_N	V_P	V

- False Discovery Proportion

$$\text{FDP} = V_{0P}/V_P \quad (\text{FDP}=0 \text{ if } V_P=0)$$

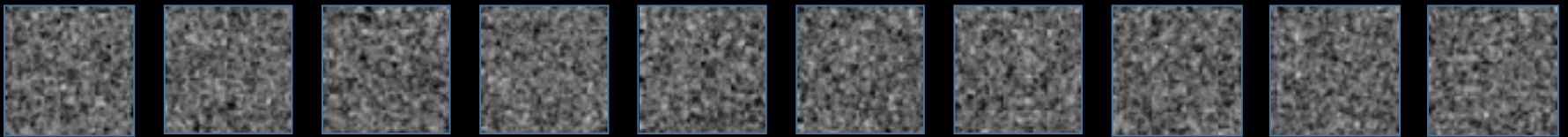
- But only can observe V_P , don't know V_{0P}

– We control the *expected* FDP

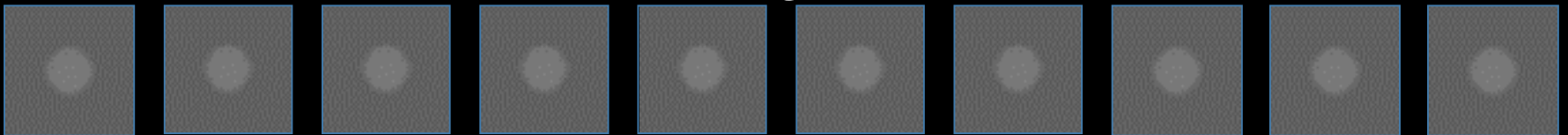
$$\text{FDR} = E(\text{FDP})$$

False Discovery Rate Illustration:

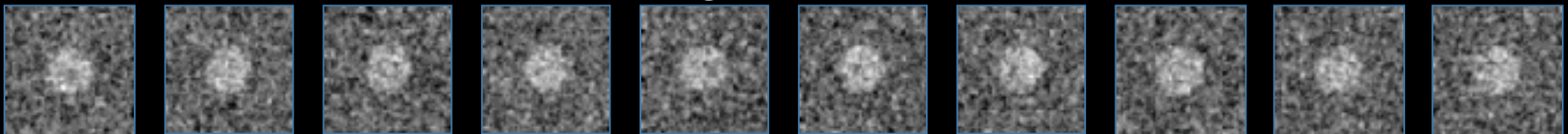
Noise



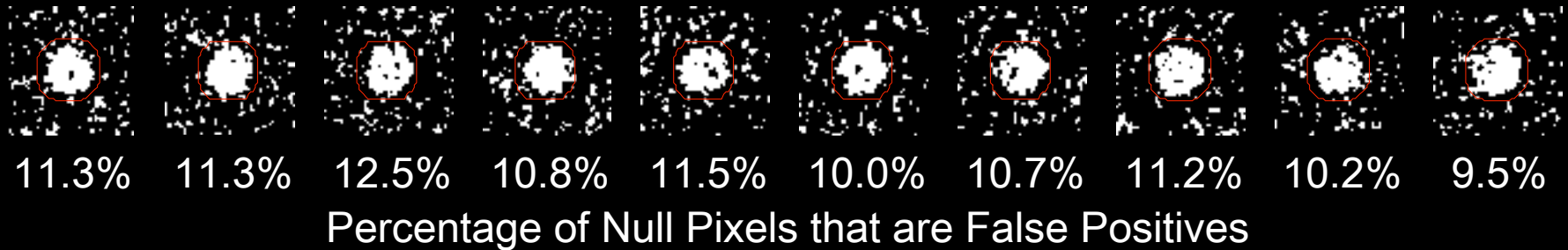
Signal



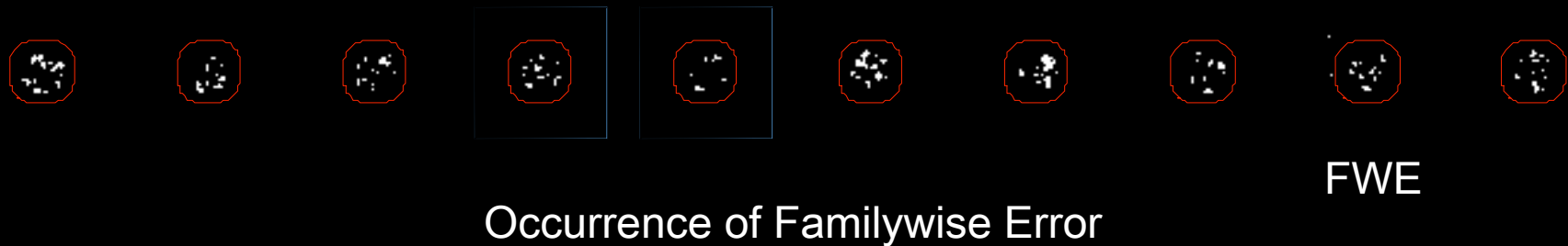
Signal+Noise



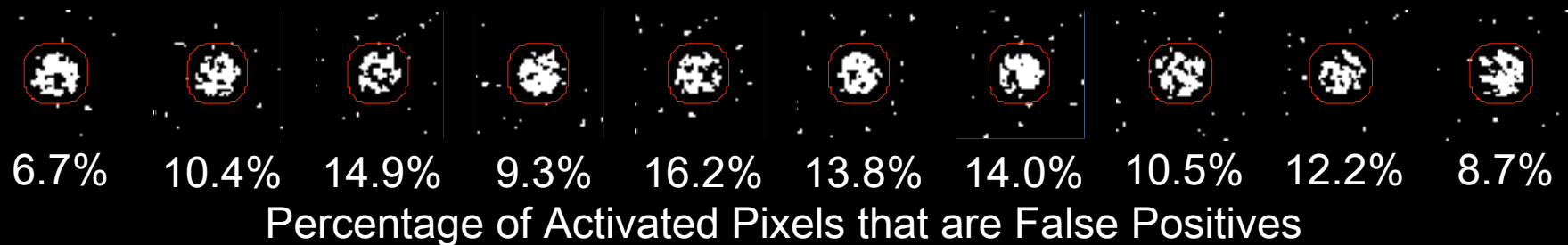
Control of Per Comparison Rate at 10%



Control of Familywise Error Rate at 10%



Control of False Discovery Rate at 10%

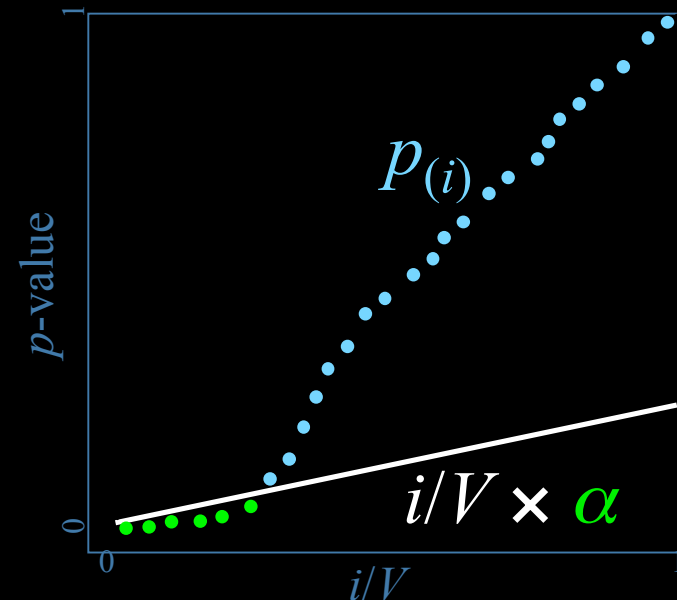


Benjamini & Hochberg Procedure

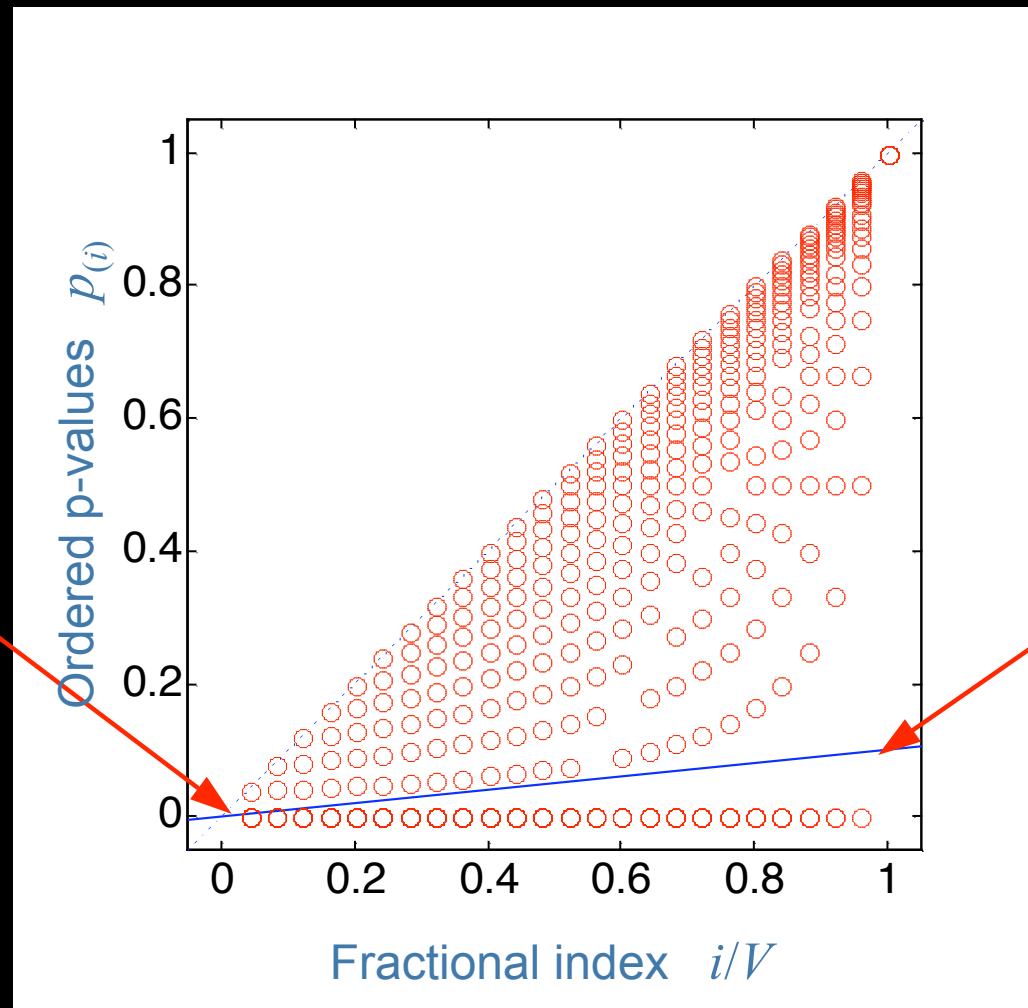
- Select desired limit α on FDR
- Order p-values, $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(V)}$
- Let r be largest i such that

$$p_{(i)} \leq i/V \times \alpha$$

- Reject all hypotheses corresponding to $p_{(1)}, \dots, p_{(r)}$.



Adaptiveness of Benjamini & Hochberg FDR



When no
signal:
P-value
threshold
 α/v

When all
signal:
P-value
threshold
 α

...FDR adapts to the amount of signal in the data

Benjamini & Hochberg: Key Properties

- FDR is controlled

$$E(\text{FDP}) \leq \alpha m_0/v$$

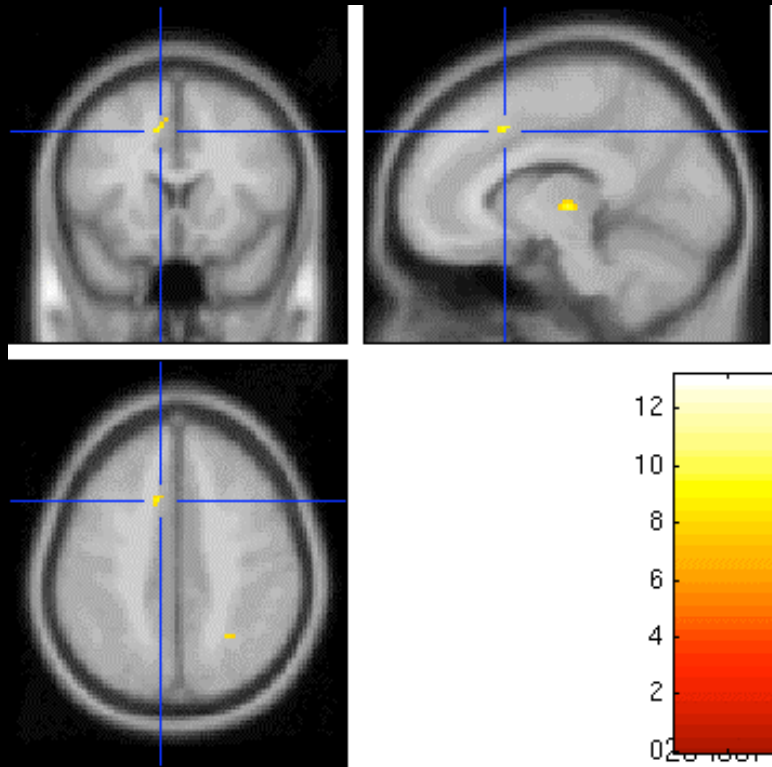
- Conservative, if large fraction of nulls false

- Adaptive

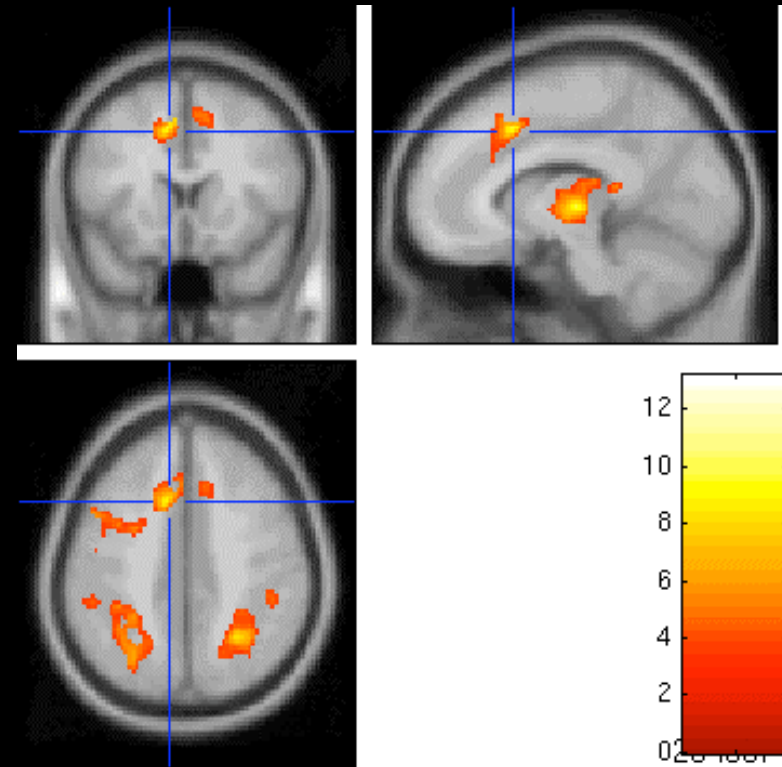
- Threshold depends on amount of signal

- More signal, More small p-values,
More $p_{(i)}$ less than $i/v \times \alpha/c(v)$

FDR Example



FWER Perm. Thresh. = 7.67
58 voxels



FDR Threshold = 3.83
3,073 voxels

Conclusions for voxelwise tests

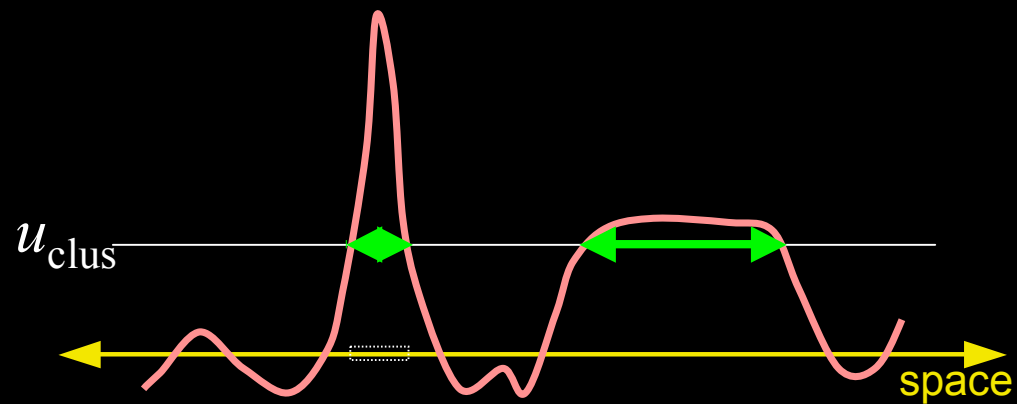
- Multiple Testing Problem
 - Choose a MTP metric (FDR, FWE)
 - Use a powerful method that controls the metric
- Nonparametric Inference
 - More power for small group FWE inferences
- References
 - GRF: Worsley, et al., J.Cerb.Blood.Flow.Met., 1992: 900-918
 - Permutation: Nichols & Holmes, HBM, 2001: 1-20
 - FDR: Nichols & Hayasaka, Stat. Meth. Med. Res., 2003:419-416

Cluster-based inference

- We use RFT all the time, so it can't be as bad as the RFT results we just saw
- Use cluster size as the test statistic for RFT
- Permutation tests use cluster size or cluster mass

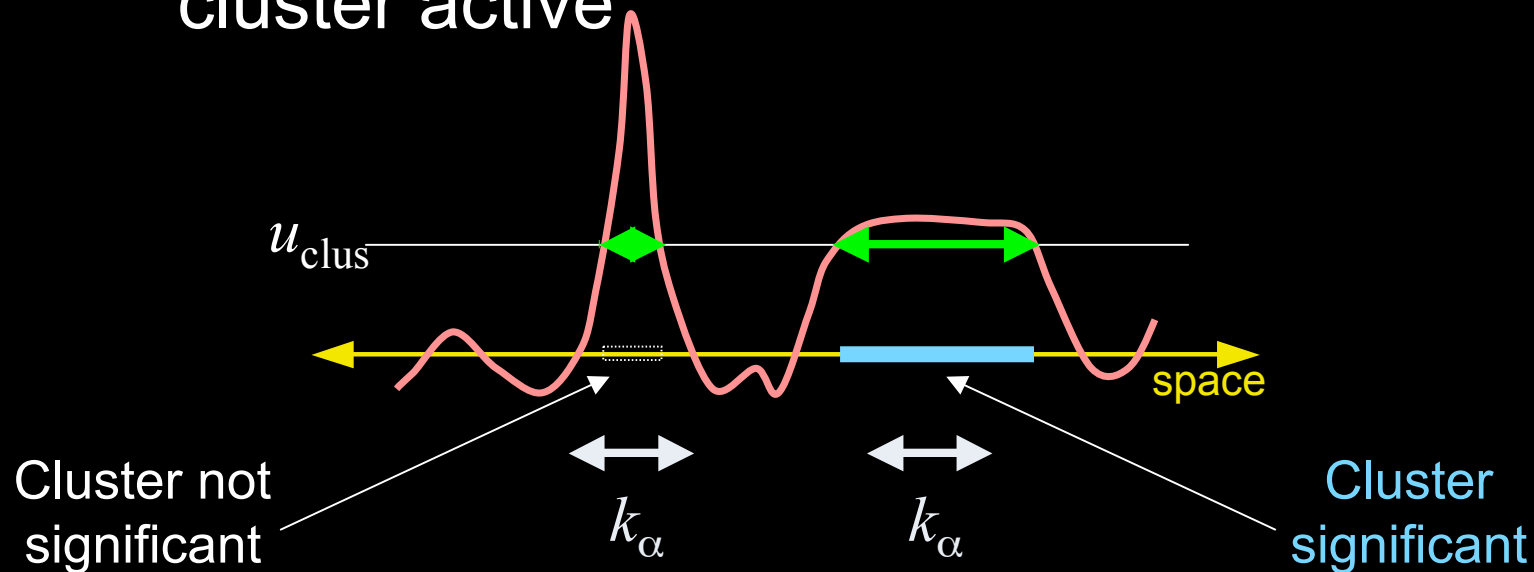
Cluster-level Inference

- Two step-process
 - Define clusters by arbitrary threshold u_{clus}



Cluster-level Inference

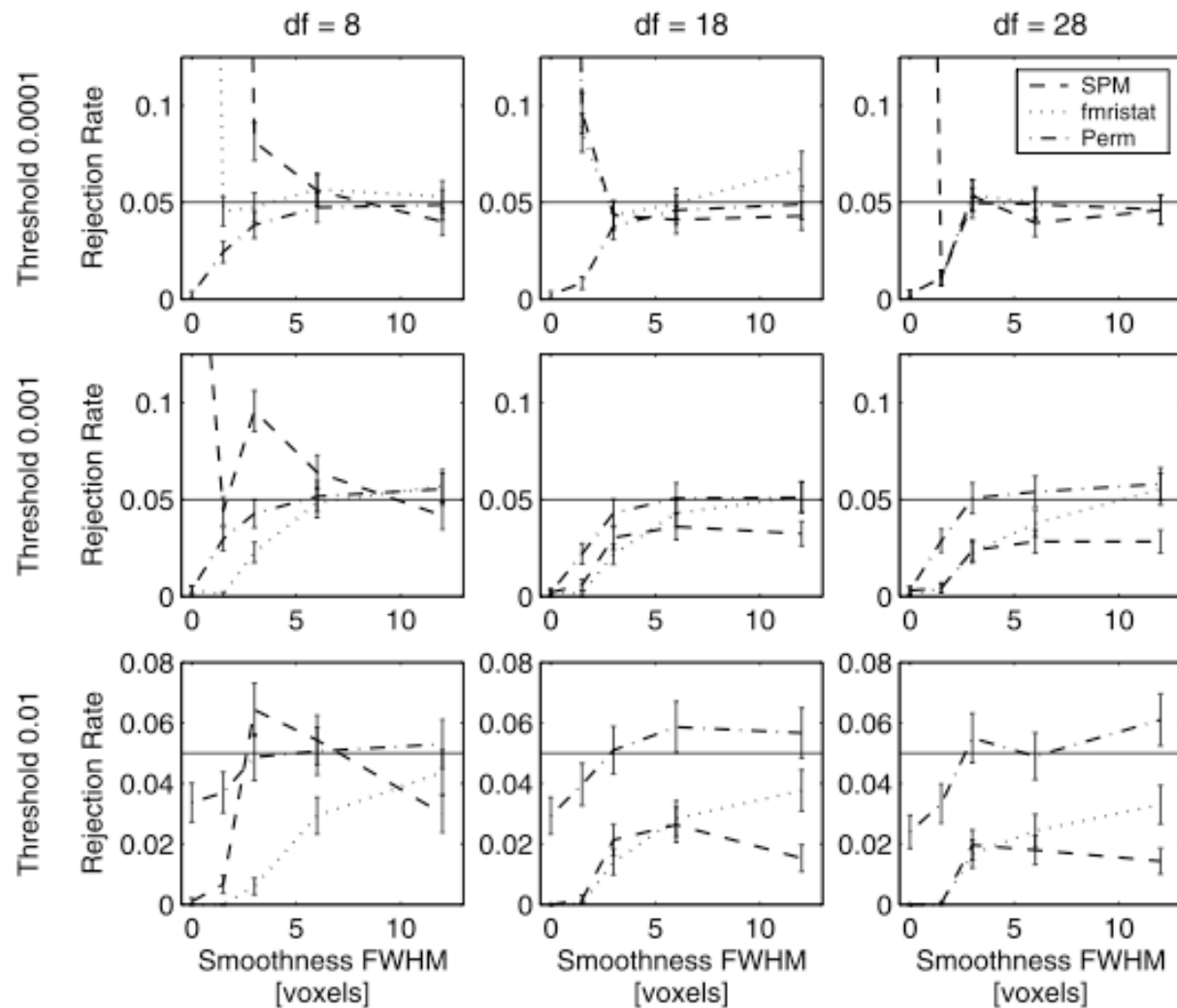
- Typically better sensitivity
- Worse spatial specificity
 - The null hyp. of entire cluster is rejected
 - Only means that *one or more* of voxels in cluster active



Extent vs Mass

- Cluster extent
 - How many voxels are in cluster
 - Sensitive to spatially extended signals
- Cluster mass
 - Combines signal extent and intensity
 - Can be done with FSL's randomise and SnPM
 - Generally works better, but RFT-based distribution is difficult

RF vs Perm: cluster mass



Conclusions

- Cluster extent RF test
 - Generally conservative (especially for low smoothness)
 - Only close to 0.05 for high threshold (0.0001) and smooth data
 - In some cases extremely anticonservative
 - Results seem to worsen with larger sample sizes (not sure why)

Conclusions

- Cluster extent permutation test
 - In general works well for smooth data with sufficient DF
 - Generally conservative due to discreteness of the test

What to do?

- Start with fast RFT-based approaches
- If you think you have something use longer permutation-based thresholding
- Also check out new threshold free cluster enhancement (TFCE) option in FSL
 - No need to choose 2 thresholds!