Inference for fMRI

Jeanette Mumford
University of Wisconsin - Madison

Overview

- Step 1
  - What is type I error?
- Step 2
  - Understand different *statistics* we can use
  - Understand the types of error rates
- Step 3
  - Parametric vs Nonparametric

Type I error

- Assuming the null is true, the probability that we reject the null

Type I error

- Assuming the null is true, the probability that we reject the null

- 5% of the time, we'll have a false positive

Interpretation

	Declared active	Declared inactive	Total
Non-active			
Active			
Total			

Interpretation

What we know (test results)

	Declared active	Declared inactive	Total
Non-active			
Active			
Total			

Interpretation

What we don't know
(truth)

	Declared active	Declared inactive	Total
Non-active			
Active			
Total			

Interpretation

- 1100 total voxels
- 100 voxels have signal (null is false)

- 1000 voxels have no signal (null)

	Declared active	Declared inactive	Total
Non-active			1000
Active			100
Total			1100

Interpretation

- 1100 total voxels
- 100 voxels have signal (null is false)
  - 80% power -> 80 voxels detected
- 1000 voxels have no signal (null)

	Declared active	Declared inactive	Total
Non-active			1000
Active	80	20	100
Total			1100

Interpretation

- 1100 total voxels
- 100 voxels have signal (null is false)
  - 80% power -> 80 voxels detected
- 1000 voxels have no signal (null)

	Declared active	Declared inactive	Total
Non-active			1000
Active	80 (Power)	20 (Type II err.)	100
Total			1100

Interpretation

- 1100 total voxels
- 100 voxels have signal (null is false)
  - 80% power -> 80 voxels detected
- 1000 voxels have no signal (null)
  - 5% type I error -> 50 false positives

	Declared active	Declared inactive	Total
Non-active	50	950	1000
Active	80 (Power)	20 (Type II err.)	100
Total			1100

Interpretation

- 1100 total voxels
- 100 voxels have signal (null is false)
  - 80% power -> 80 voxels detected
- 1000 voxels have no signal (null)
  - 5% type I error -> 50 false positives

	Declared active	Declared inactive	Total
Non-active	50 (Type I err.)	950 (Correct)	1000
Active	80 (Power)	20 (Type II err.)	100
Total			1100

Interpretation

- 1100 total voxels
- 100 voxels have signal (null is false)
  - 80% power -> 80 voxels detected
- 1000 voxels have no signal (null)
  - 5% type I error -> 50 false positives

	Declared active	Declared inactive	Total
Non-active	50	950	1000
Active	80	20	100
Total	130	970	1100

Interpretation

- 1100 total voxels
- 100 voxels have signal (null is false)
  - 80% power -> 80 voxels detected
- 1000 voxels have no signal (null)
  - 5% type I error -> 50 false positives

focus is on controlling this number

	Declared active	Declared inactive	Total
Non-active	$\boxed{50}$	950	1000
Active	80	20	100
Total	130	970	1100

Implication of type I error

- If you run enough tests, you'll find something that is significant
  - This doesn't mean it is truly significant
  - If you run 20 tests with a 5% threshold on type I errors, you *expect* to have at least 1 significant test
    - This would be a false positive

Whole brain: Thresholding the map

- What type of error rate does the researcher want to control?
- What statistic does the researcher want to use?

Whole brain: Thresholding the map

- What type of error rate does the researcher want to control?
  - Per comparison error
    - Use $p < 0.05$ for each statistic

- What statistic does the researcher want to use?

Whole brain: Thresholding the map

- What type of error rate does the researcher want to control?
  - Per comparison error
    - Use $p < 0.05$ for each statistic
  - Family wise error
    - Probability of *any* false positives

- What statistic does the researcher want to use?

Whole brain: Thresholding the map

- What type of error rate does the researcher want to control?
  - Per comparison error
    - Use $p < 0.05$ for each statistic
  - Family wise error
    - Probability of *any* false positives
  - False discovery rate
    - Proportion of voxels found to be active that are false activations
- What statistic does the researcher want to use?

Whole brain: Thresholding the map

- What type of error rate does the researcher want to control?
  - Per comparison error
    - Use $p < 0.05$ for each statistic
  - Family wise error
    - Probability of *any* false positives
  - False discovery rate
    - Proportion of voxels found to be active that are false activations
- What statistic does the researcher want to use?
  - Voxelwise

Whole brain: Thresholding the map

- What type of error rate does the researcher want to control?
  - Per comparison error
    - Use $p < 0.05$ for each statistic
  - Family wise error
    - Probability of *any* false positives
  - False discovery rate
    - Proportion of voxels found to be active that are false activations
- What statistic does the researcher want to use?
  - Voxelwise
  - Clusterwise

Whole brain: Thresholding the map

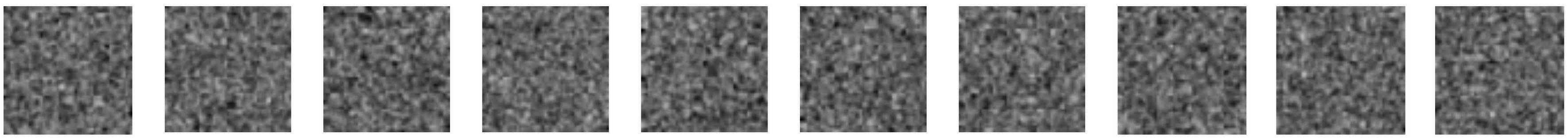
- What type of error rate does the researcher want to control?
  - Per comparison error
    - Use $p < 0.05$ for each statistic
  - Family wise error
    - Probability of *any* false positives
  - False discovery rate
    - Proportion of voxels found to be active that are false activations
- What statistic does the researcher want to use?
  - Voxelwise
  - Clusterwise
  - Peakwise

Whole brain: Thresholding the map

- What type of error rate does the researcher want to control?
  - Per comparison error
    - Use $p < 0.05$ for each statistic
  - Family wise error
    - Probability of *any* false positives
  - False discovery rate
    - Proportion of voxels found to be active that are false activations
- What statistic does the researcher want to use?
  - Voxelwise
  - Clusterwise
  - Peakwise

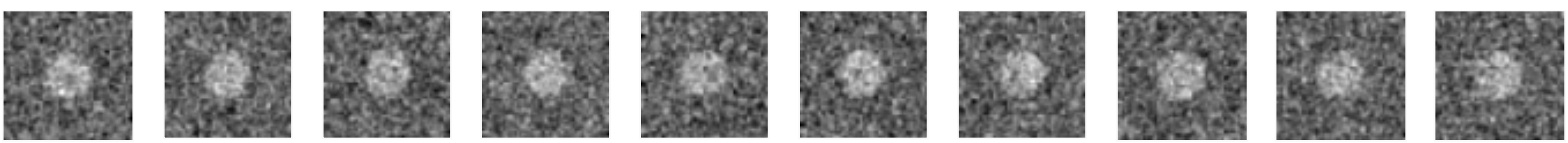
Error Rate Illustration:

Noise

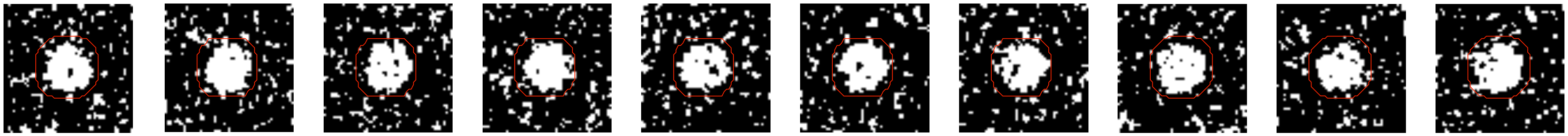


Signal

Signal+Noise



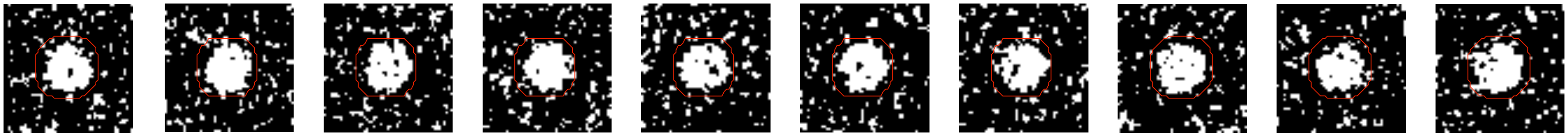
Control of Per Comparison Rate at 10%



11.3% 11.3% 12.5% 10.8% 11.5% 10.0% 10.7% 11.2% 10.2% 9.5%

Percentage of Null Pixels that are False Positives

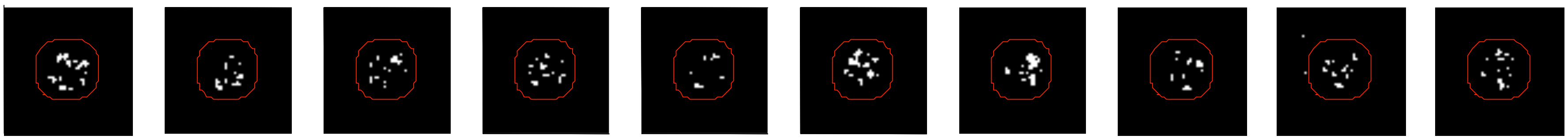
Control of Per Comparison Rate at 10%



11.3% 11.3% 12.5% 10.8% 11.5% 10.0% 10.7% 11.2% 10.2% 9.5%

Percentage of Null Pixels that are False Positives

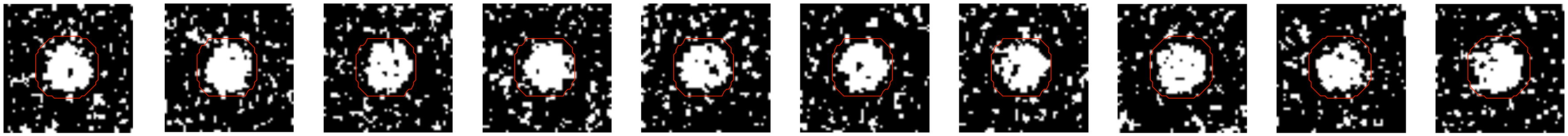
Control of Familywise Error Rate at 10%



FWE

Occurrence of Familywise Error

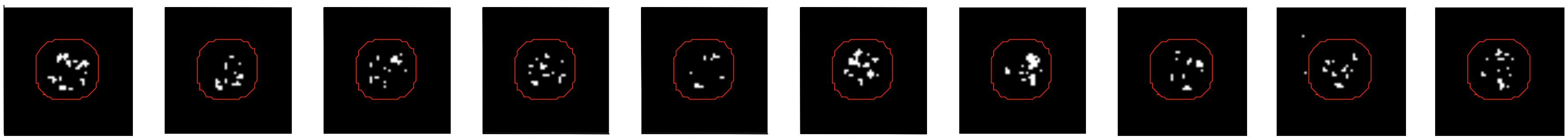
Control of Per Comparison Rate at 10%



11.3% 11.3% 12.5% 10.8% 11.5% 10.0% 10.7% 11.2% 10.2% 9.5%

Percentage of Null Pixels that are False Positives

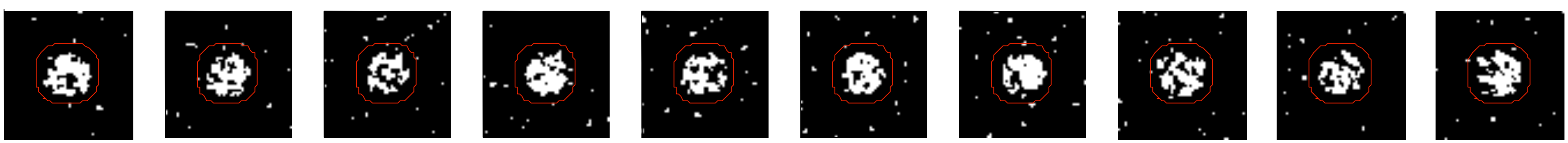
Control of Familywise Error Rate at 10%



FWE

Occurrence of Familywise Error

Control of False Discovery Rate at 10%



6.7% 10.4% 14.9% 9.3% 16.2% 13.8% 14.0% 10.5% 12.2% 8.7%

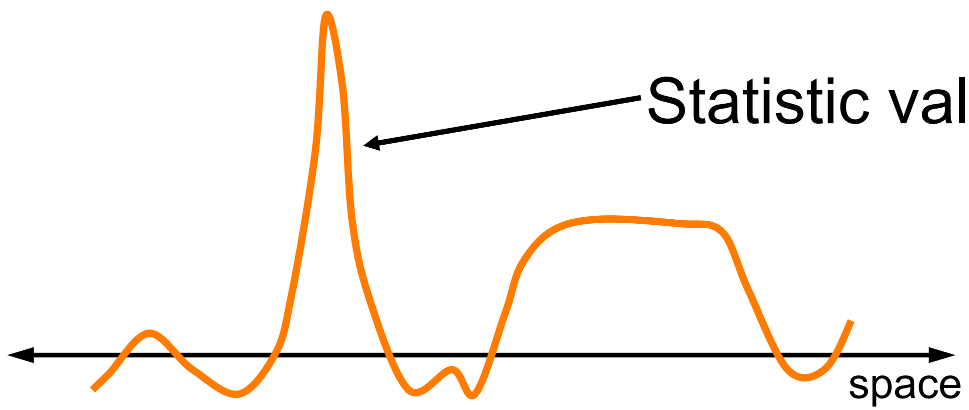
Percentage of Activated Pixels that are False Positives

Levels of inference

- Voxel level
- Cluster level
- Peak level

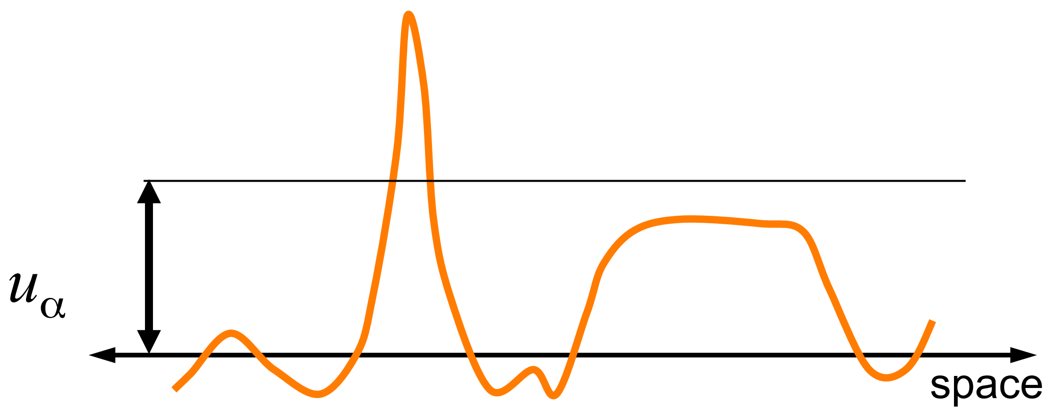
Voxel-level Inference

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



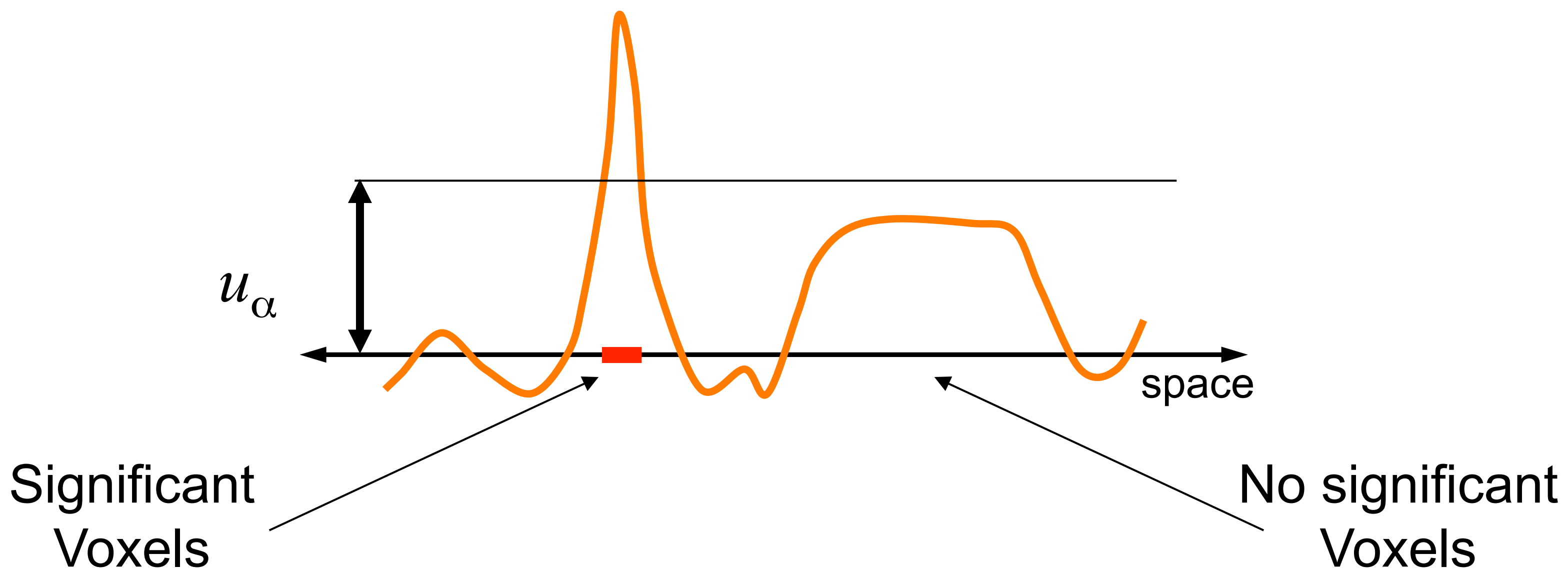
Voxel-level Inference

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



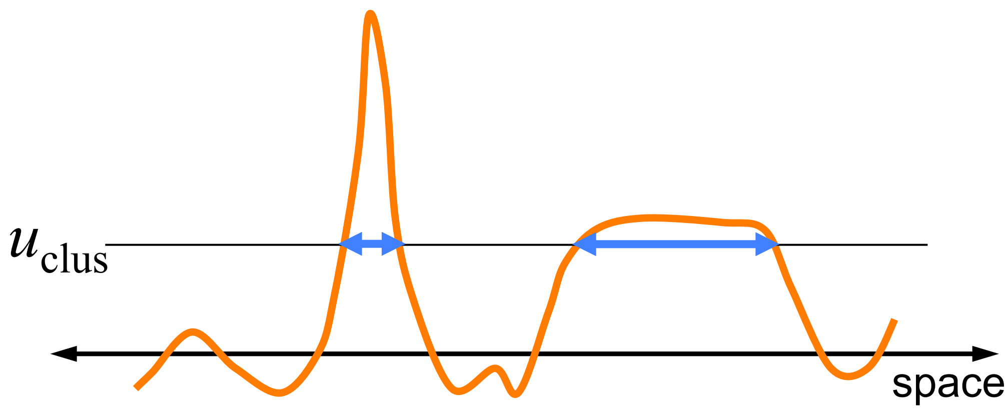
Voxel-level Inference

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



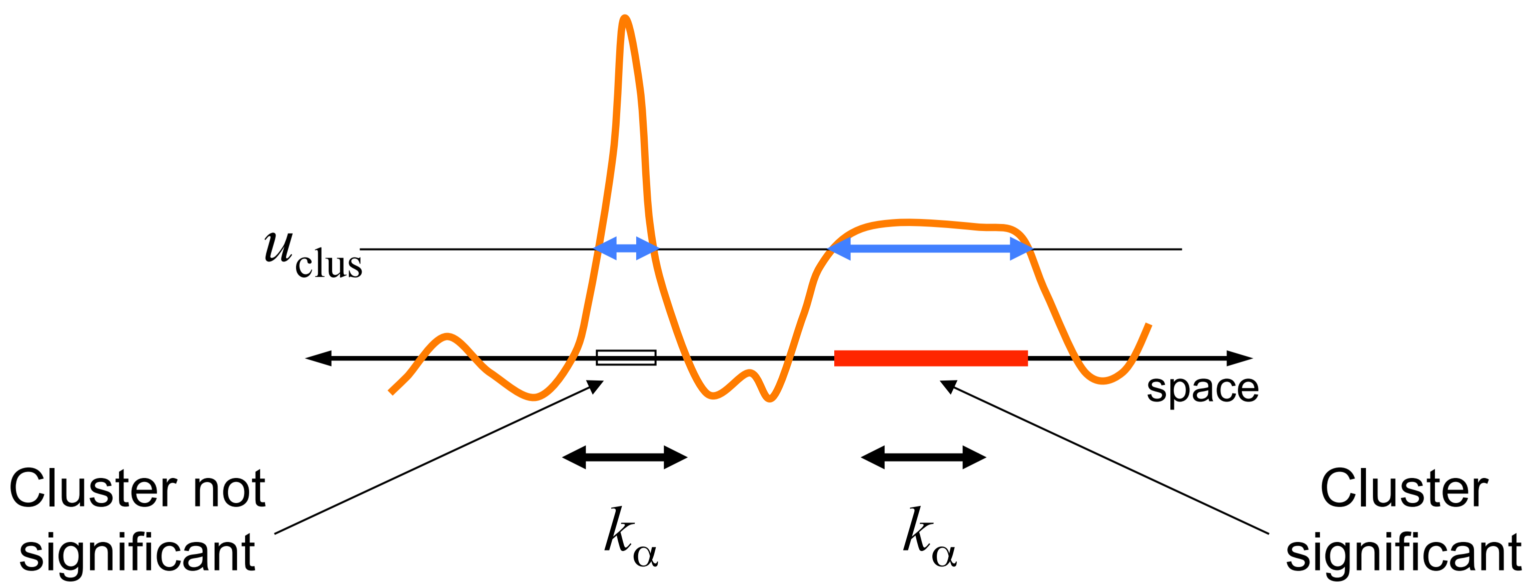
Cluster-level Inference

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



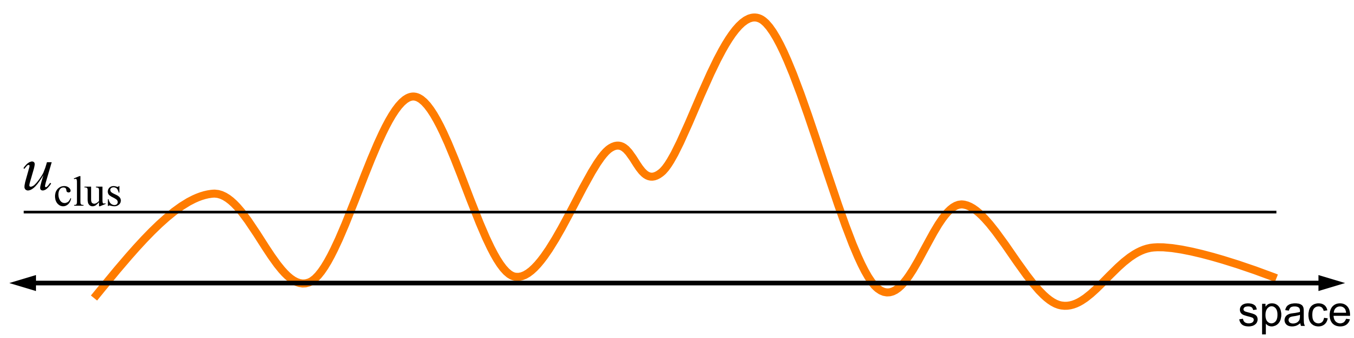
Cluster-level Inference

- Two step-process
  - Define clusters by arbitrary threshold u_{clus}
  - Retain clusters larger than α-level threshold k_α



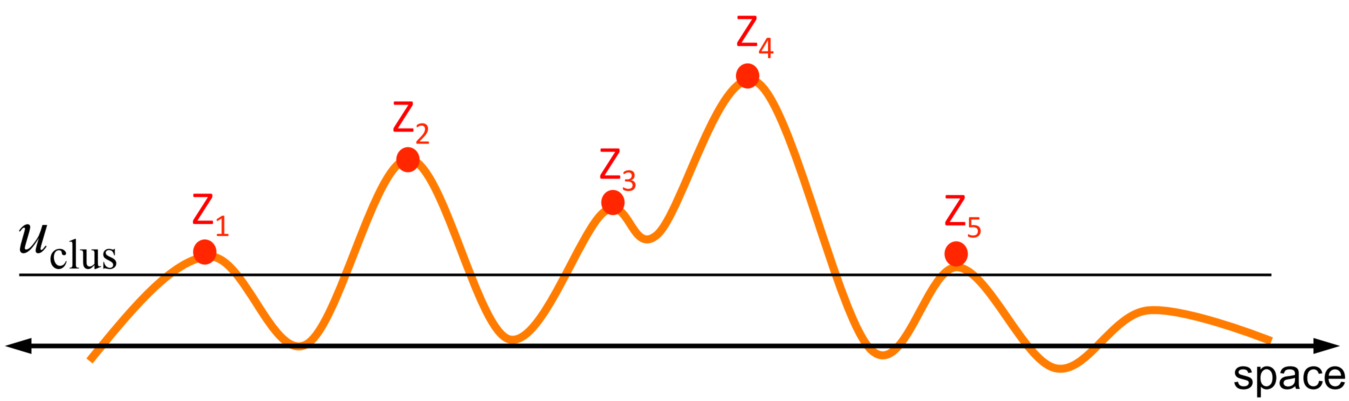
Peak level inference

- Again start with a cluster forming threshold



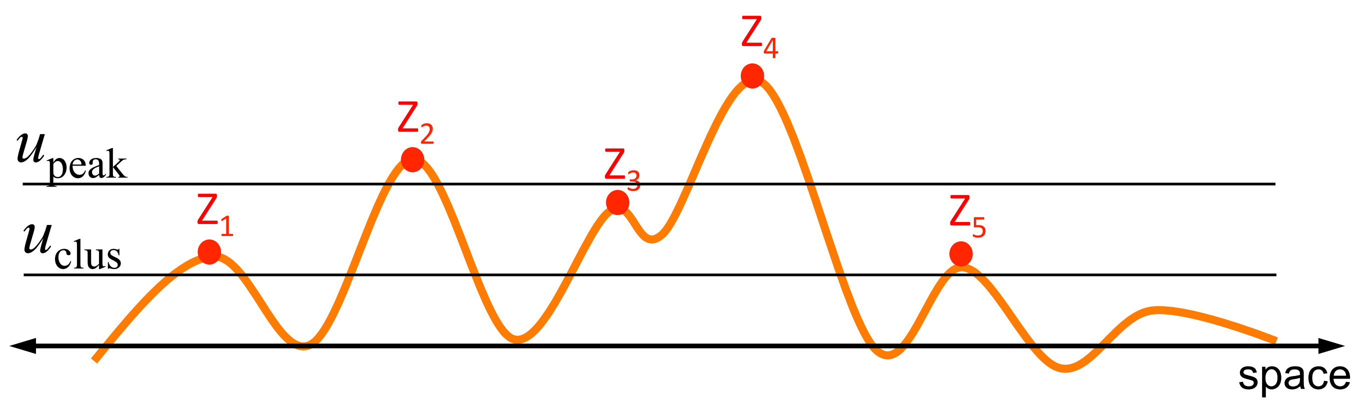
Peak level inference

- Again start with a cluster forming threshold
- Instead of cluster size, focus on peak height



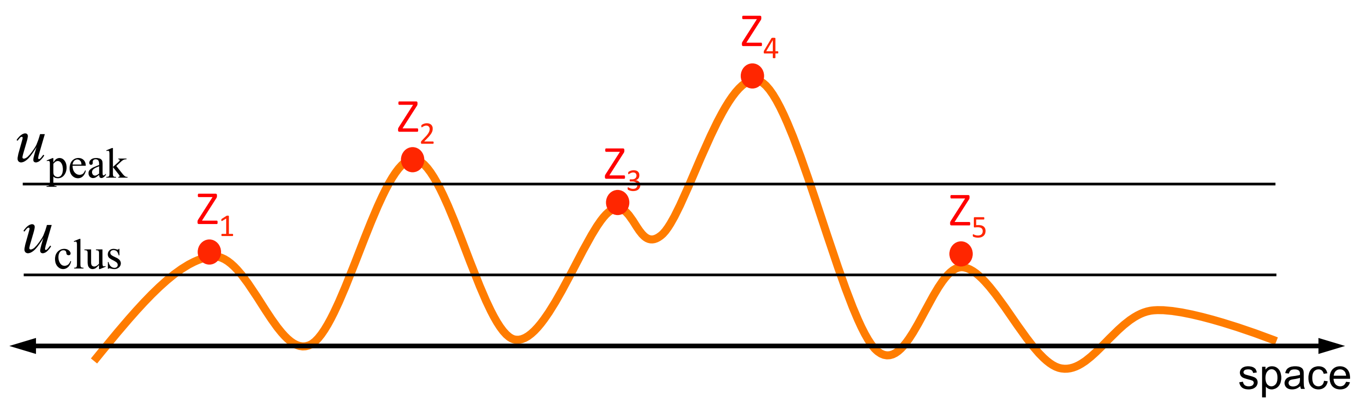
Peak level inference

- Again start with a cluster forming threshold
- Instead of cluster size, focus on peak height



Peak level inference

- Again start with a cluster forming threshold
- Instead of cluster size, focus on peak height
- Similarly to cluster level inference, significance applies to a set of voxels
  - The peak and its neighbors

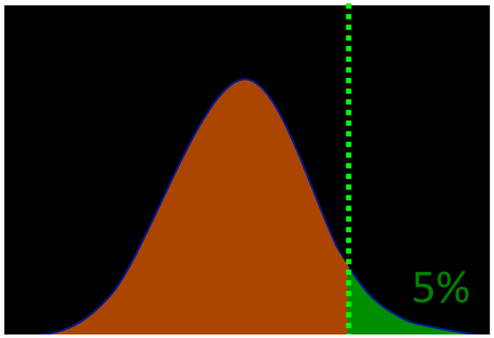


What gets used?

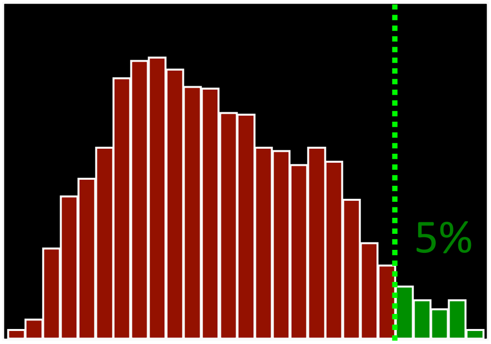
- Most common
  - Cluster-based statistics thresholded to control FWER with random field theory
- Also common
  - Peak-wise inference controlling FWER with random field theory

Permutation test

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



Parametric Null Distribution



Nonparametric Null Distribution

Very quick overview of parametric methods

- Random field theory assumptions
  - Spatial smoothness of signal is constant over the brain
  - Spatial autocorrelation follows a specific distribution (squared exponential)
  - Cluster forming threshold needs to be high

Very quick overview of parametric methods

- Cluster forming threshold must be “high enough”
  - Main issue in Eklund paper (Cluster failure)
  - 3.1 is high enough in FSL

- Faster than nonparametric

- Usually run by default if you’re using Feat

Nonparametric approach

- randomise is the FSL tool
- Takes longer to run
- Fewer assumptions
  - Exchangeability
  - There still is one and it can still be violated!
  - Is not a magic fix for outliers

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_o
  - No experimental effect, A & B labels arbitrary
- Statistic
  - Mean difference

Permutation Test
Toy Example

- Under H_o
  - Consider all equivalent relabelings

AAABBB	ABABAB	BAAABB	BABBAA
AABABB	ABABBA	BAABAB	BBAAAB
AABBAB	ABBAAB	BAABBA	BBAABA
AABBBA	ABBABA	BABAAB	BBABAA
ABAABB	ABBBAA	BABABA	BBBAAA

Permutation Test
Toy Example

- Under H_o
  - 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	ABBBAA	1.48	BABABA	-9.45	BBBAAA	-4.82

Permutation Test
Toy Example

- Under H_o
  - 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	ABBBAA	1.48	BABABA	-9.45	BBBAAA	-4.82

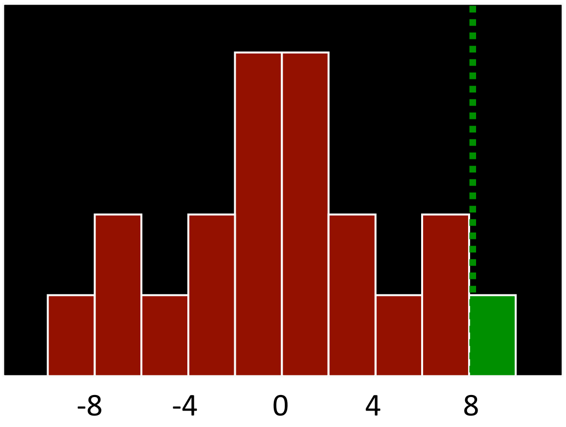
Permutation Test
Toy Example

- Under H_o
  - 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	ABBBAA	1.48	BABABA	-9.45	BBBAAA	-4.82

Permutation Test
Toy Example

- Under H_o
  - 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

What should you use?

- Typically Cluster-based approaches
  - Voxelwise approaches are very conservative
  - Peakwise is harder to interpret and not used as much

- Parametric with a cluster forming threshold of Z=3.1 or p=0.001 is fine
- Nonparametric with any cluster forming threhold (or threshold free cluster enhancement in randomise) is also fine!

Thank you

- Questions?