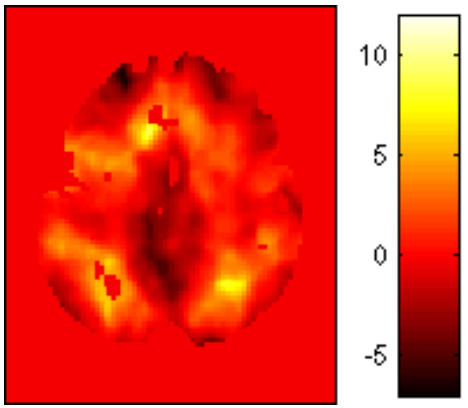
Multiple Comparisons in Neuroimaging

Martin Lindquist
Department of Statistics
Columbia University

Voxel-wise Analysis

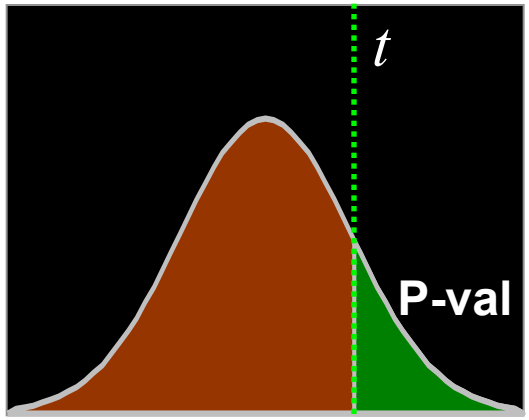
1. Fit a statistical model (e.g., the GLM) to each voxel in the brain.
2. Use the estimated model parameters to test for an effect of interest in that voxel (e.g., H_0: β_1-β_2=0).
3. Summarize the resulting test statistic in a statistical image (e.g., a t-map).

4. Determine which voxels show a statistically significant effect, i.e. threshold the statistical image.

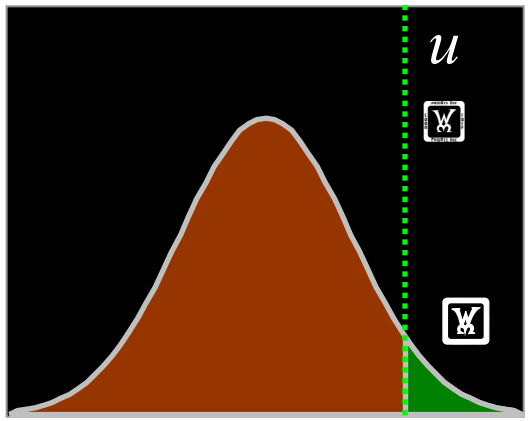


Hypothesis Testing

- Null Hypothesis H_0
  - Statement of no effect (e.g., μ_1-μ_2=0).
- Test statistic T
  - Measures compatibility between the null hypothesis and the data.
- P-value
  - Probability that the test statistic would take a value as or more extreme than that actually observed if H_0 is true, i.e. P(T > t | H_0).
- Significance level
  - Threshold u_μ controls false positive rate at level μ = P(T>u_μ | H_0)



Null Distribution of T



Null Distribution of T

Making Errors

- There are two types of errors one can make when performing hypothesis tests:
  - Type I error
    - H_0 is true, but we mistakenly reject it (False positive).
    - Controlled by the significance level α.
  - Type II error
    - H_0 is false, but we fail to reject it (False negative)

- The probability that a hypothesis test will correctly reject a false null hypothesis is the power of the test.

Multiple Comparisons Problem

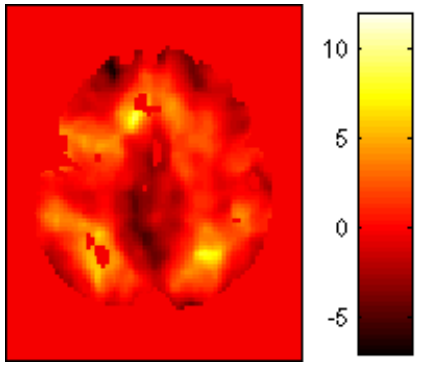
- The question of choosing an appropriate threshold becomes more complicated when dealing with a family of tests.

- If more than one hypothesis test is performed, the risk of making at least one Type I error is greater than the α value for a single test.

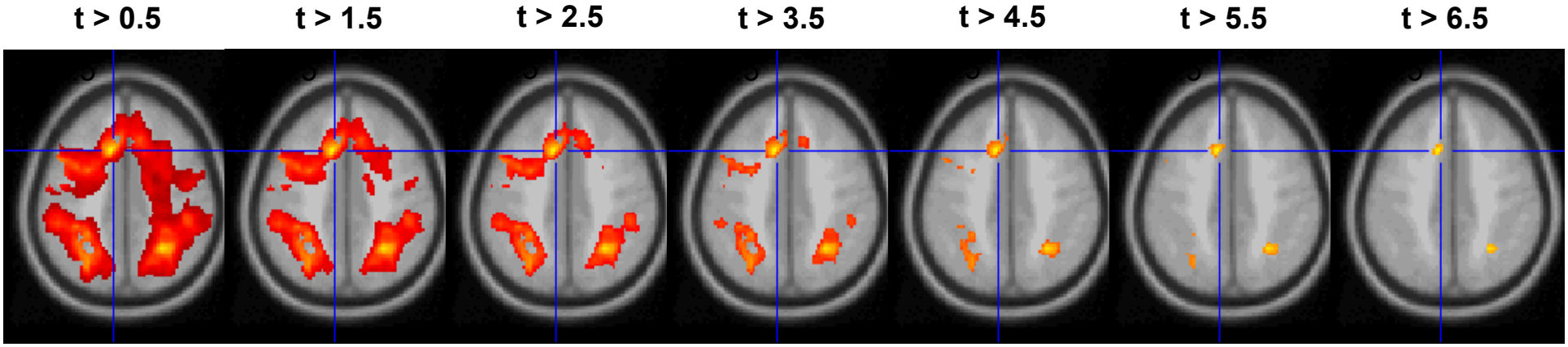
- The more tests one performs, the greater the likelihood of getting at least one false positive.

Multiple Comparisons Problem

- Which of 100,000 voxels are significant?
  - α=0.05 α 5,000 false positive voxels



- Choosing a threshold is a balance between sensitivity (true positive rate) and specificity (true negative rate).



Measures of False Positives

- There exist several ways of quantifying the likelihood of obtaining false positives.

- Family-Wise Error Rate (FWER)
  - Probability of any false positives

- False Discovery Rate (FDR)
  - Proportion of false positives among rejected tests

Family-Wise Error Rate

- The family-wise error rate (FWER) is the probability of making one or more Type I errors in a family of tests, under the null hypothesis.

- FWER controlling methods:
  - Bonferroni correction
  - Random Field Theory
  - Permutation Tests

Problem Formulation

- Let H_{0i} be the hypothesis that there is no activation in voxel i, where i = 1,….m.
- Let T_i be the value of the test statistic at voxel i.
- The family-wise null hypothesis, H_0, states that there is no activation in any of the m voxels.

$$H_0 = \bigcap_i H_{0i}$$

- If we reject a single voxel null hypothesis, H_{0i}, we reject the family-wise null hypothesis.
  - We reject H_0 if one or more T_i ≥ u for a given threshold u.

- Assuming H_0 is true, we want the probability of falsely rejecting H_0 to be controlled by α, i.e.

$$P\left(\bigcup_i \{T_i \geq u\} \mid H_0\right) \leq \alpha$$

- Want to choose u to attain this goal.

Bonferroni Correction

- Choose the threshold u so that

$$P(T_i \geq u \mid H_0) \leq \frac{\alpha}{m}$$

- Hence,

$$FWER = P\left(\bigcup_i \{T_i \geq u\} \mid H_0 \right)$$

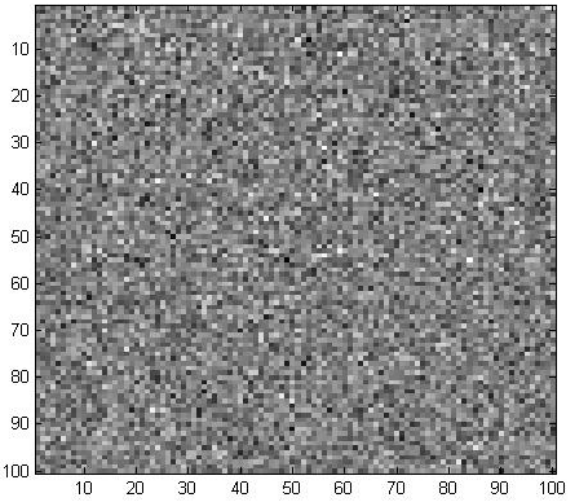
$$\leq \sum_i P(T_i \geq u \mid H_0)$$

Boole's Inequality

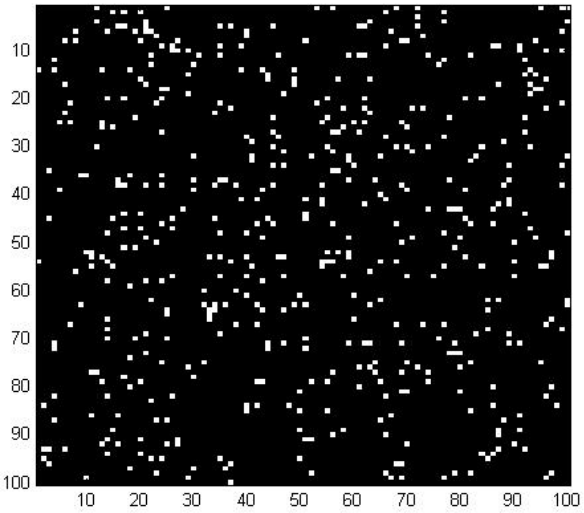
$$\leq \sum_i \frac{\alpha}{m} = \alpha$$

Example

Generate 100×100 voxels from an iid N(0,1) distribution



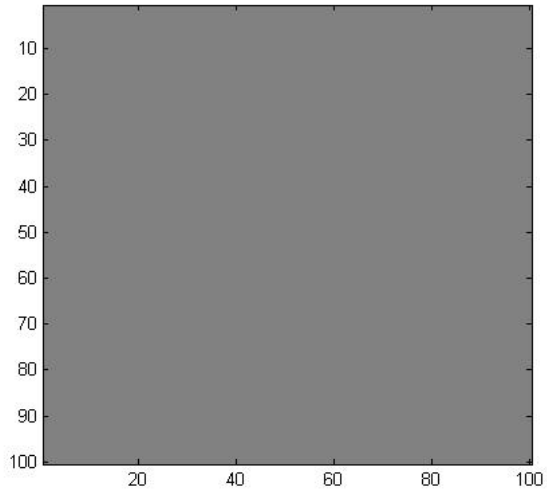
Threshold at u=1.645



Approximately 500 false positives.

To control the FWER at 0.05, the Bonferroni correction is 0.05/10,000.

This corresponds to u=4.42.



No false positives

Only 5 out of every 100 maps generated in this fashion will have one or more values above u.

Bonferroni Correction

- The Bonferroni correction is very conservative, i.e. it results in very strict significance levels.
- It decreases the power of the test (probability of correctly rejecting a false null hypothesis) and greatly increases the chance of false negatives.
- In addition, it is not optimal for correlated data, and most fMRI data has significant spatial correlation.

Spatial Correlation

- We can choose a better threshold by using information about the spatial correlation in the data.

- Random field theory allows one to incorporate the spatial correlation into the calculation of the appropriate threshold.

- It is based on approximating the distribution of the maximum statistic over the whole image.

Maximum Statistic

- Link between FWER and max statistic.

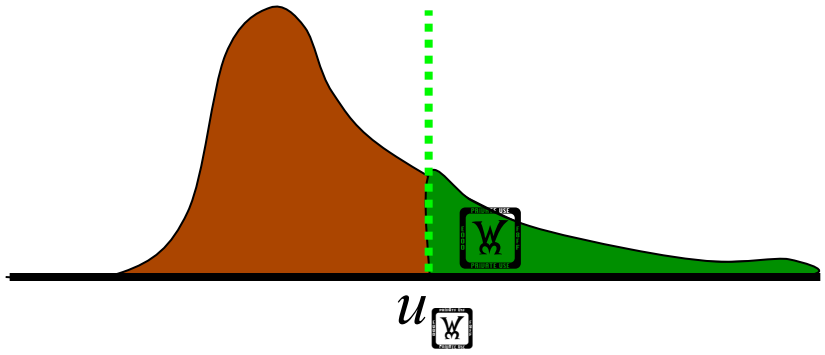
$$\text{FWER} = P(\bigcup_i \{T_i \geq u\} \mid H_o)$$

P(any t-value exceeds u under null)

$$= P(\max_i T_i \geq u \mid H_o)$$

P(max t-value exceeds u under null)

Choose the threshold u such that the max only exceeds it with probability α.



Distribution of $\max_i T_i$

Random Field Theory

- Consider a statistical image to be a lattice representation of a continuous random field.

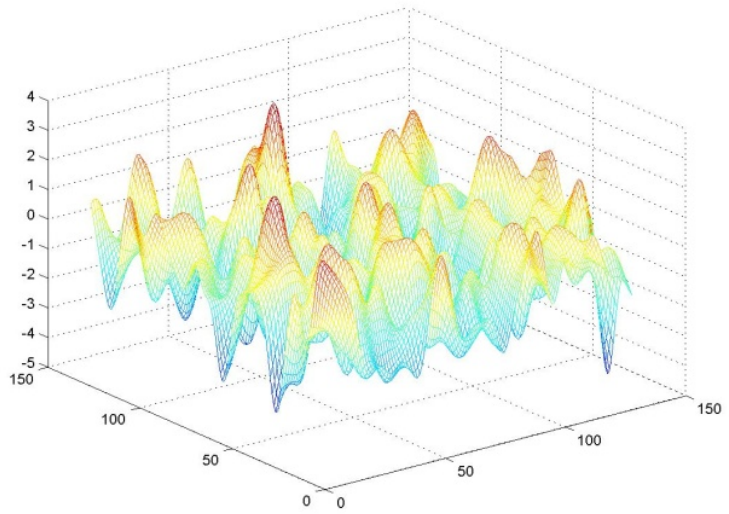
- Random field methods are able to:

  – approximate the upper tail of the maximum distribution, which is the part needed to find appropriate thresholds, and

  – account for the spatial dependence in the data.

Euler Characteristic

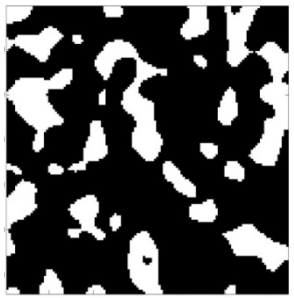
- Euler Characteristic χ_u
  - A property of an image after it has been thresholded.
  - Counts #blobs - #holes
  - At high thresholds, just counts #blobs



Random Field

$\chi_u = 28 - 1 = 27$

$u = 0.5$

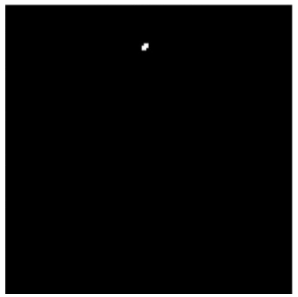


$\chi_u = 2$

$u = 2.75$

$\chi_u = 1$

$u = 3.5$



Controlling the FWER

- Link between FWER and Euler Characteristic.

$$\text{FWER} = P(\max_i T_i \geq u \mid H_o)$$

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

no holes exist

$$\leq P(\chi_u \geq 1 \mid H_o)$$

never more than 1 blob

$$\leq E(\chi_u \mid H_o)$$

- Closed form results exist for $E(\chi_u)$ for Z, t, F and χ^2 continuous random fields.

3D Gaussian Random Fields

For large search regions:

$$E(\chi_u) \approx R(4\log 2)^{3/2}(u^2-1)e^{-u^2/2}(2\pi)^{-2}$$

where

$$R = \frac{V}{FWHM_x FWHM_y FWHM_z}$$

Here V is the volume of the search region and the full width at half maximum (FWHM) represents the smoothness of the image estimated from the data.

R = Resolution Element (Resel)

Controlling the FWER

For large u:

$$FWER \approx R(4\log 2)^{3/2}(u^2-1)e^{-u^2/2}(2\pi)^{-2}$$

where

$$R = \frac{V}{FWHM_x FWHM_y FWHM_z}$$

Properties:

- As u increases, FWER decreases (Note u large).
- As V increases, FWER increases.
- As smoothness increases, FWER decreases.

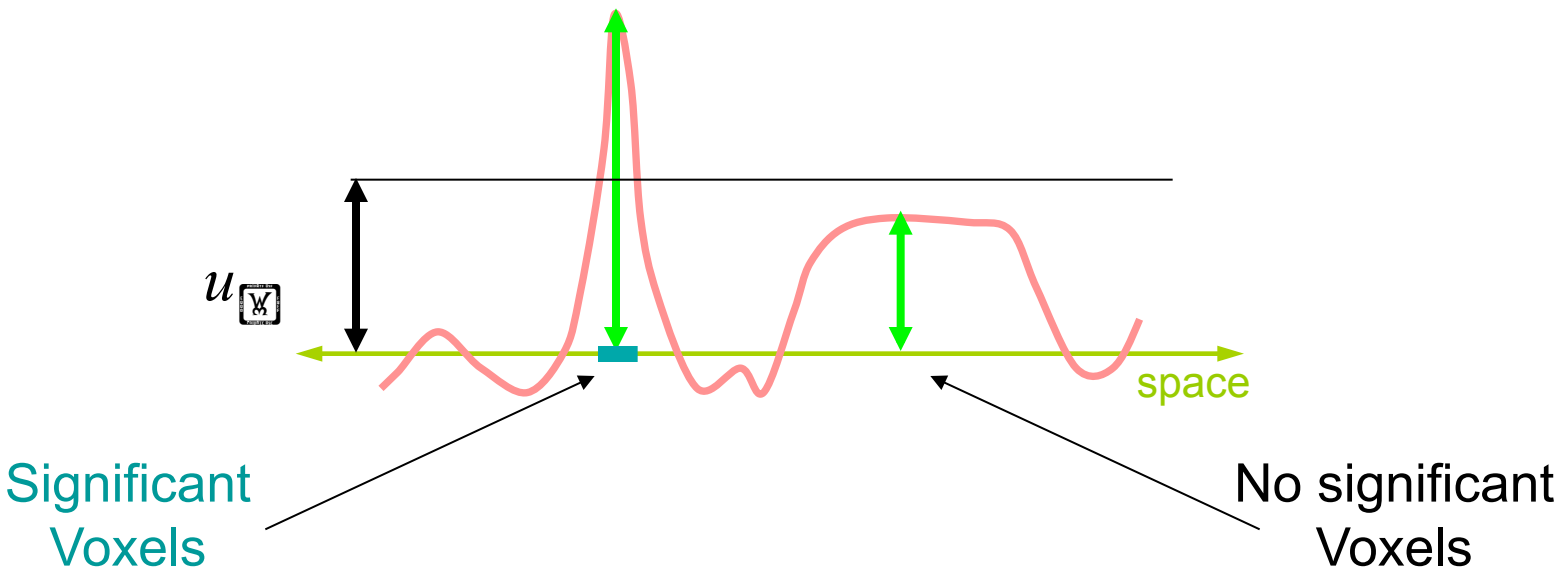
Levels of Inference

- A statistical map has many topological attributes:
  - The height of a peak
  - The number of peaks
  - The volume or extent of an excursion set

- Correction can be performed on several levels, including on the:
  - Voxel-level
  - Cluster-level

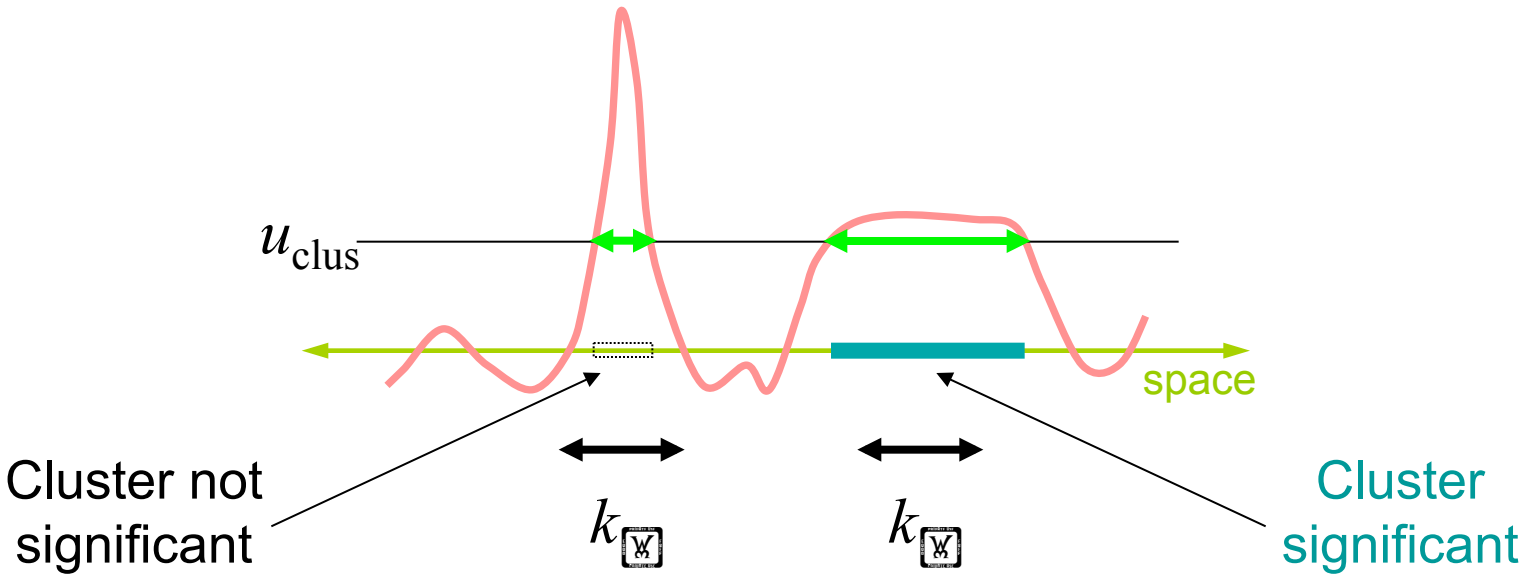
Voxel-level Inference

- Retain voxels above [illegible]-level threshold $u_{[illegible]}$
- Gives best spatial specificity
  - The null hypothesis at a single voxel can be rejected



Cluster-level Inference

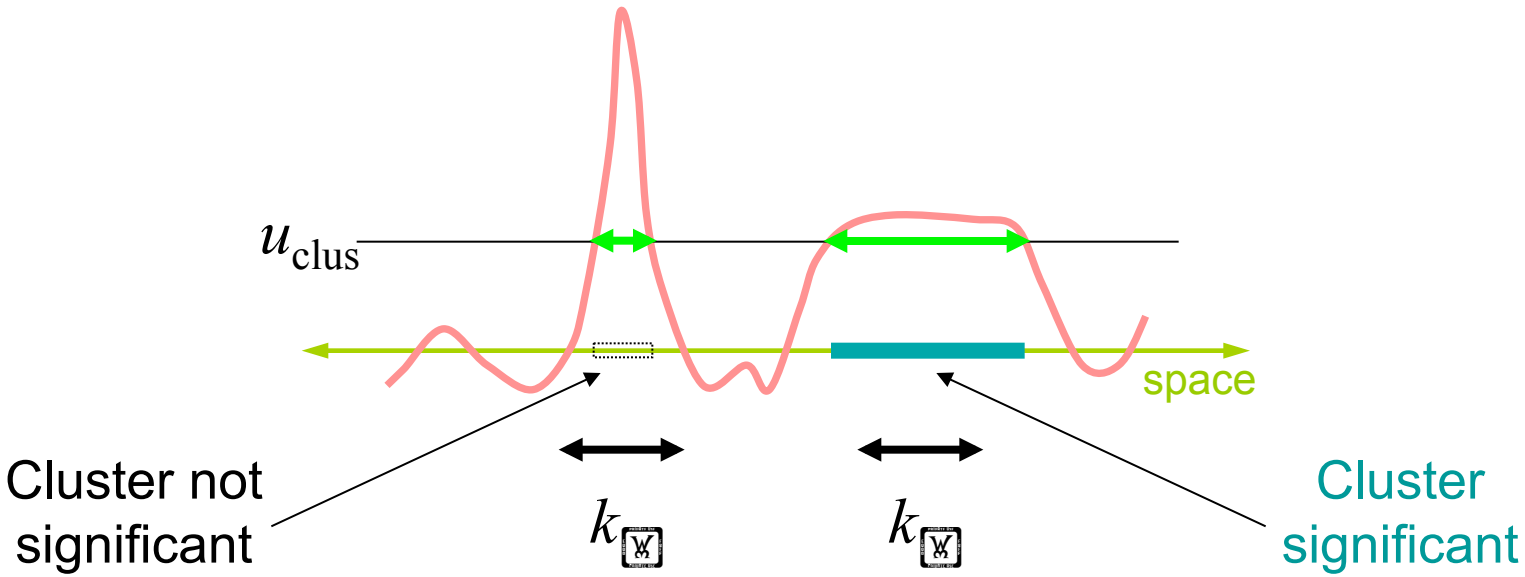
- Two step-process
  - Define clusters by arbitrary threshold u_{clus}
  - Retain clusters larger than [illegible]-level threshold $k_{[illegible]}$



Cluster-level Inference

- Typically better sensitivity (higher true positive rate)

- Worse spatial specificity
  - The null hypothesis of entire cluster is rejected
  - Means that *one or more* voxels in cluster active



Cluster-level Inference

- Can use RFT to perform cluster level inference.
  - Assume clusters behave like a multidimensional Poisson point process.
  - Then, the probability of getting one or more clusters of volume k is given by:

$$p = 1 - \exp\{-E(\chi_u)P(n \geq k)\}$$

  - Here E(χ_u) is the expected Euler Characteristic and n represents the volume of a cluster.

RFT Assumptions

- The entire image is either multivariate Gaussian or derived from multivariate Gaussian images.
- The statistical image must be sufficiently smooth to approximate a continuous random field.
  - FWHM smoothness 3-4×voxel size.
- The amount of smoothness is assumed known.
  - Estimate is biased when images not sufficiently smooth.
- Several layers of approximations.

Issues with FWER

- Methods that control the FWER (Bonferroni, RFT, Permutation Tests) provide a strong control over the number of false positives.

- While this is appealing the resulting thresholds often lead to tests that suffer from low power.

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

False Discovery Rate

- The false discovery rate (FDR) is a recent development in multiple comparison problems due to Benjamini and Hochberg (1995).

  Benjamini, Y. and Hochberg, Y. 1995. Controlling the False Discovery Rate: A practical and Powerful Approach to Multiple Testing. *J.R. Stat. Soc. Ser. B* 57: 289-300.

- While the FWER controls the probability of any false positives, the FDR controls the proportion of false positives among all rejected tests.

Notation

Suppose we perform tests on m voxels.

	Declared Inactive	**Declared Active**	
Truly inactive	U	V	m_0
Truly active	T	S	$m-m_0$
	m-R	R	m

U, V, T and S are unobservable random variables.

R is an observable random variable.

Definitions

- In this notation:

$$FWER = P(V \geq 1)$$

- False discovery rate:

$$FDR = E\left(\frac{V}{R}\right)$$

- The FDR is defined to be 0 if R=0.

Properties

- A procedure controlling the FDR ensures that on average the FDR is no bigger than a pre-specified rate q which lies between 0 and 1.

- However, for any given data set the FDR need not be below the bound.

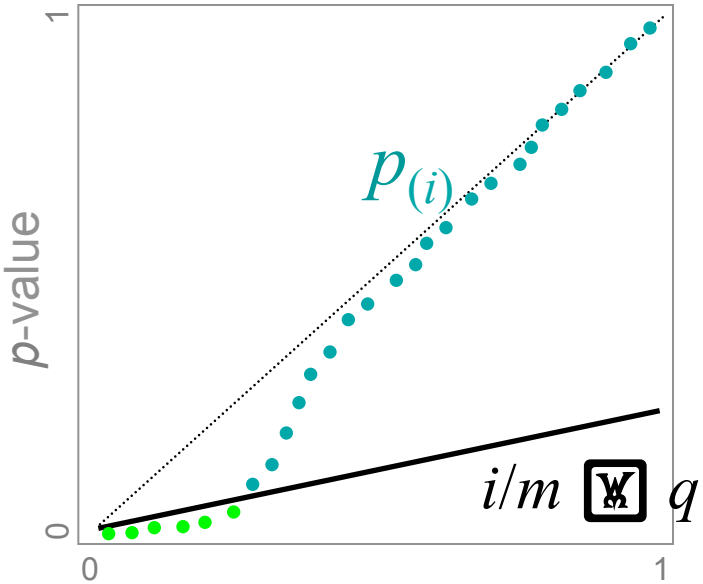
BH Procedure

1. Select desired limit q on FDR (e.g., 0.05)

2. Rank p-values, $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$

3. Let r be largest i such that
$$p_{(i)} \leq i/m \leq q$$

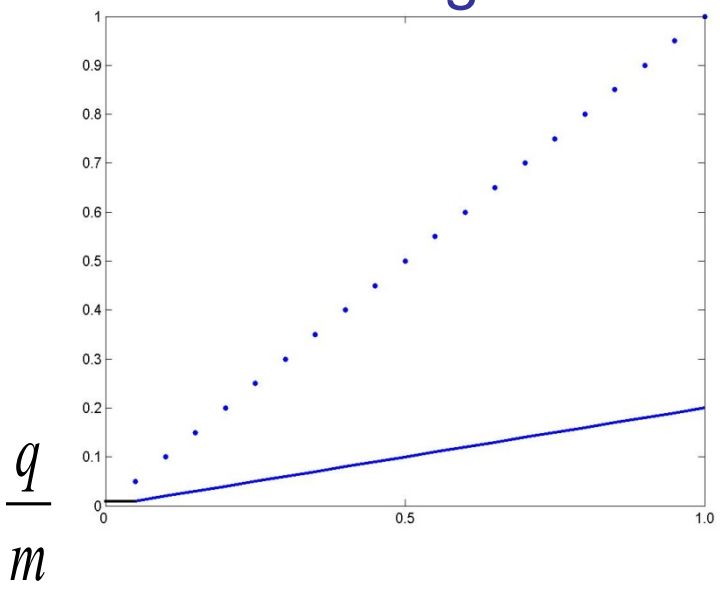
4. Reject all hypotheses corresponding to $p_{(1)}, \ldots, p_{(r)}$.



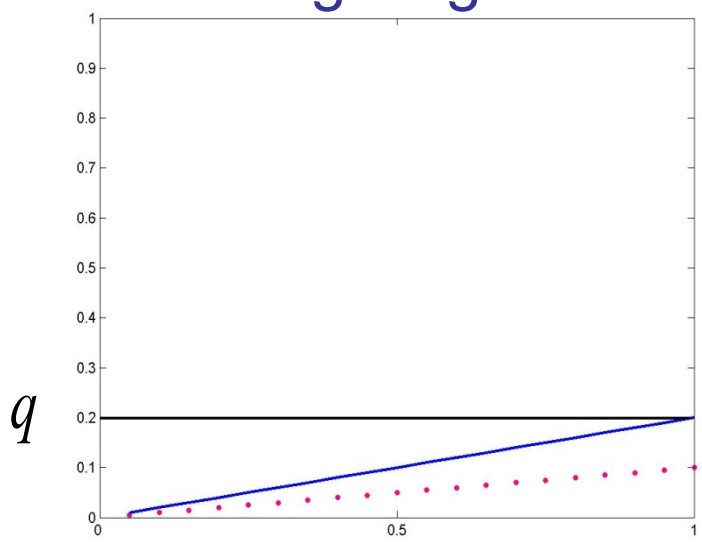
The BH procedure is adaptive in the sense that the larger the signal, the lower the threshold.

As it adapts its threshold to the features of the data, this eliminates an unnecessary excess of errors.

Low signal



High signal



Comments

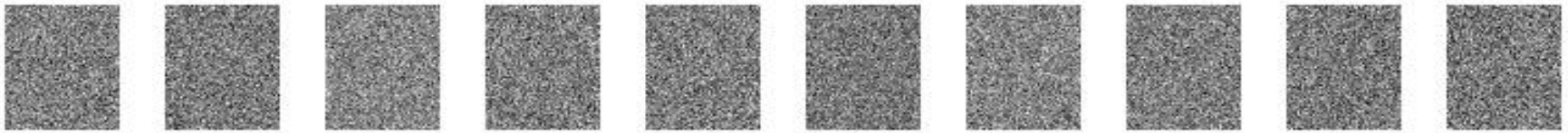
- If all null hypothesis are true, the FDR is equivalent to the FWER.
- Any procedure that controls the FWER also controls the FDR. A procedure that controls the FDR only can be less stringent and lead to a gain in power.
- Since FDR controlling procedures work only on the p-values and not on the actual test statistics, it can be applied to any valid statistical test.

Example

Signal

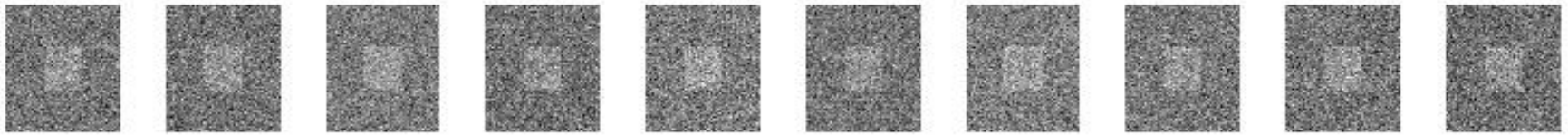
+

Noise

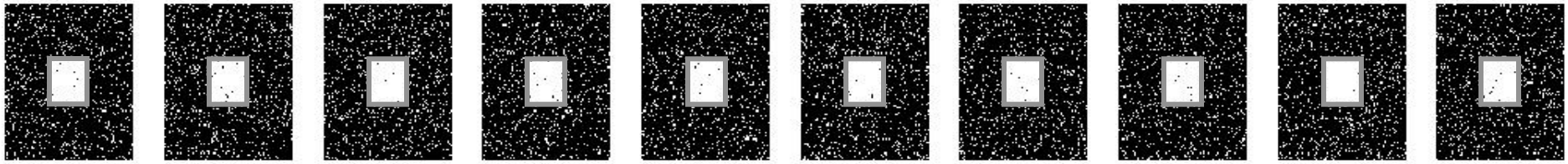


=

Signal+Noise



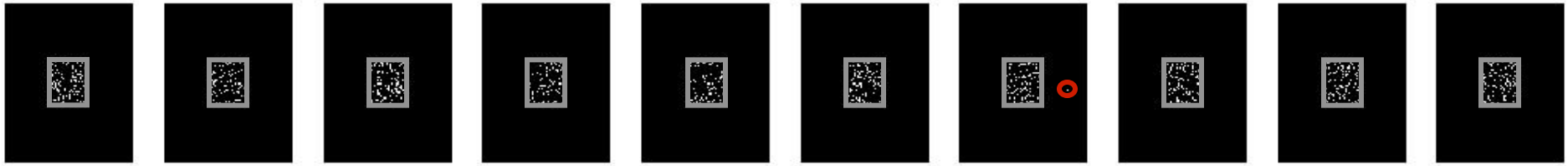
=0.10, No correction



0.0974 0.1008 0.1029 0.0988 0.0968 0.0993 0.0976 0.0956 0.1022 0.0965

Percentage of false positives

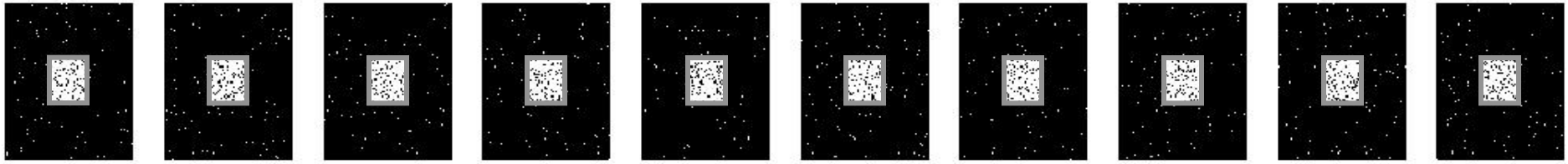
FWER control at 10%



FWER

Occurrence of false positive

FDR control at 10%



0.0871 0.0952 0.0790 0.0908 0.0761 0.1090 0.0851 0.0894 0.1020 0.0992

Percentage of active voxels that are false positives

Uncorrected Thresholds

- Most published PET and fMRI studies use arbitrary uncorrected thresholds (e.g., $p<0.001$).
  - A likely reason is that with the available sample sizes, corrected thresholds are so stringent that power is extremely low.

- Using uncorrected thresholds is problematic when interpreting conclusions from individual studies, as many of the activated regions may be false positives.

- Null findings are hard to disseminate, hence it is difficult to refute false positives established in the literature.

Extent Threshold

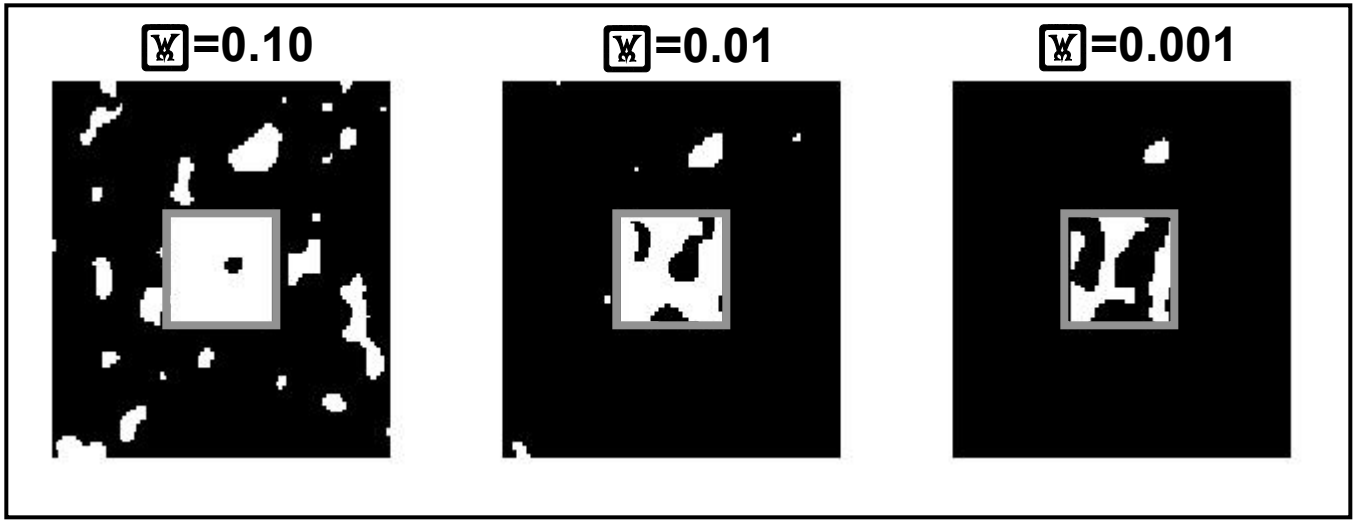
- Sometimes an arbitrary extent threshold is used when reporting results.

- Here a voxel is only deemed truly active if it belongs to a cluster of k contiguous active voxels (e.g., $p<0.001$, 10 contingent voxels).

- Unfortunately, this does not necessarily correct the problem because imaging data are spatially smooth and therefore false positives may appear in clusters.

Example

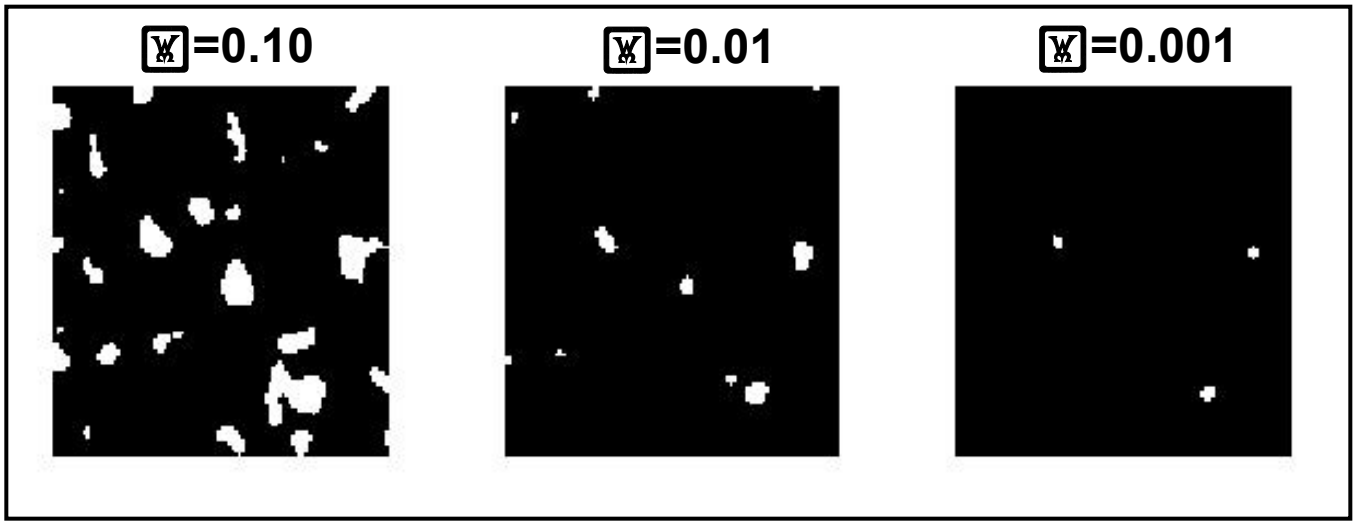
- Activation maps with spatially correlated noise thresholded at three different significance levels. Due to the smoothness, the false-positive activation are contiguous regions of multiple voxels.



Note: All images smoothed with FWHM=12mm

Example

- Similar activation maps using null data.



Note: All images smoothed with FWHM=12mm