

Multiple Testing in Neuroscience

from Bonferroni to Random Field Theory

Livio Finos

University of Padova

Contents

1	Introduction	2
1.1	Biblio	2
1.2	Thresholding	3
1.3	Motivation	3
2	FamilyWise Error Rate	4
2.1	FamilyWise Error Rate (FWER)	4
2.2	Sidak Correction	5
2.3	Sidak Correction	5
2.4	Bonferroni	5
3	False Discovery Rate	6
3.1	False Discovery Rate (FDR)	6
3.2	BH procedure	6
3.3	A toy example	7
3.4	FDR with Dependent tests	8
3.5	FWER or FDR?	8
3.6	Subsets of Rejections	10
4	Three levels of inference in neuroscience	10
4.1	Levels of inference	10
4.2	Voxel-level Inference	11
4.3	Cluster-level Inference	11
4.4	Set-level Inference	11

5	Voxel-level Inference	12
5.1	Do you Bonferroni... or not?	12
5.2	Smoothed Images - Autocorrelation	12
5.3	RESEL	12
5.4	Max-T distribution	13
5.5	Random Field Theory	13
5.6	How it works	14
6	Cluster-level inference	14
6.1	Spatial extent - motivation	14
6.2	Cluster-level Inference	14
6.3	Voxel/Cluster-level in a glance	16
6.4	Remarks	16
6.5	Limitations	17

1 Introduction

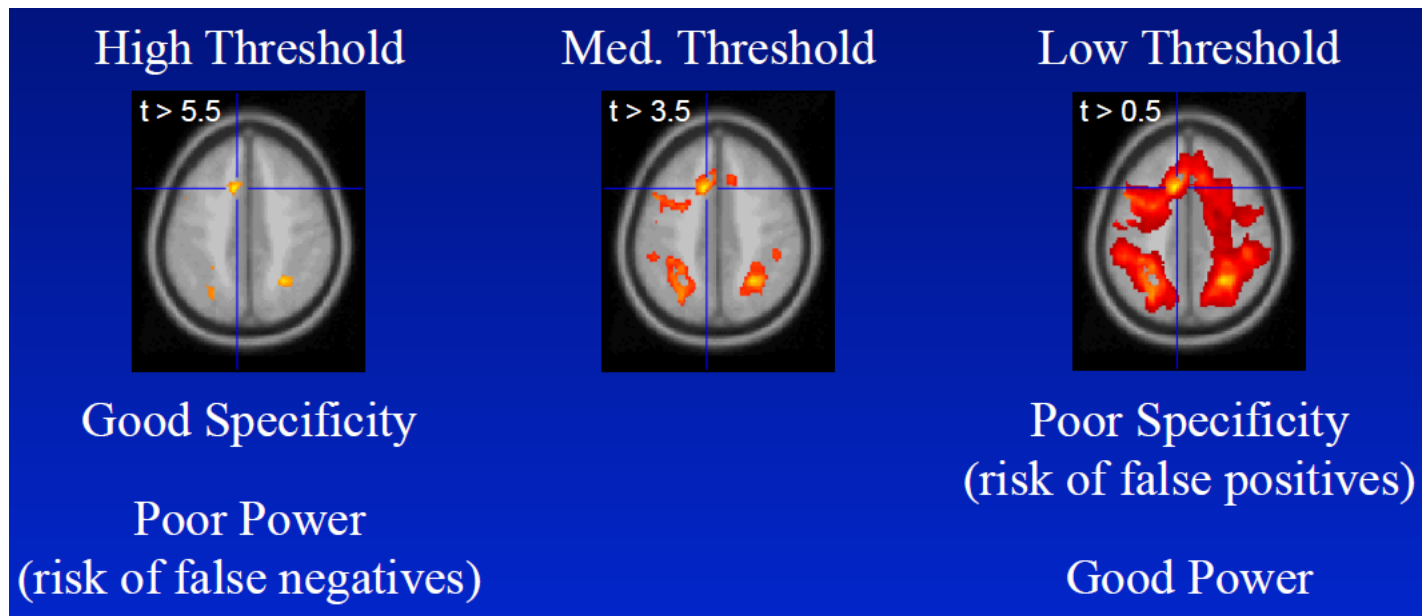
1.1 Biblio

- J Ashburner, K Friston, W Penny (2003) Human Brain Function - 2nd Ed. Academic Press (preprint online: <https://www.fil.ion.ucl.ac.uk/spm/doc/books/hbf2/>)
 - [SPM14] Chapter 14: An introduction to Random Field Theory. Brett M., Penny W. and Keibel S.
 - [SPM15] Chapter 15: Developments in Random Field Theory. K.J. Worsley
- [L] Lazar, Nicole A. (2008) The statistical analysis of functional MRI data. Springer
- [PMN] Russell A. Poldrack, Jeanette A. Mumford, Thomas E. Nichols. (2011) Handbook of functional MRI data analysis. Cambridge
- Friston, Holmes, Polin, Price and Frith (1996). Detecting Activations in PET and fMRI: Levels of Inference and Power. Neuroimage
- Goeman & Solari (2014) Tutorial in biostatistics: multiple hypothesis testing in genomics. Statistics in medicine
- MRC - Cambridge University: <http://imaging.mrc-cbu.cam.ac.uk/imaging/PrinciplesRandomFields>
- wiki http://en.wikipedia.org/wiki/Random_field

The following material is largely borrowed by:

- https://warwick.ac.uk/fac/sci/statistics/staff/academic-research/nichols/presentations/ohbm2012/Nichols_Thresholding.pdf (New and best-practice approaches to thresholding. by T. Nichols)
- https://fsl.fmrib.ox.ac.uk/fslcourse/lectures/feat2_part2.pdf (FSL Course by FSL Group)
- http://www.sbirc.ed.ac.uk/cyril/SPM-course/Talks/2015/10_multiple%20testing.pdf (Cyril Pernet)

1.2 Thresholding



1.3 Motivation

Bennett et al. (2012)

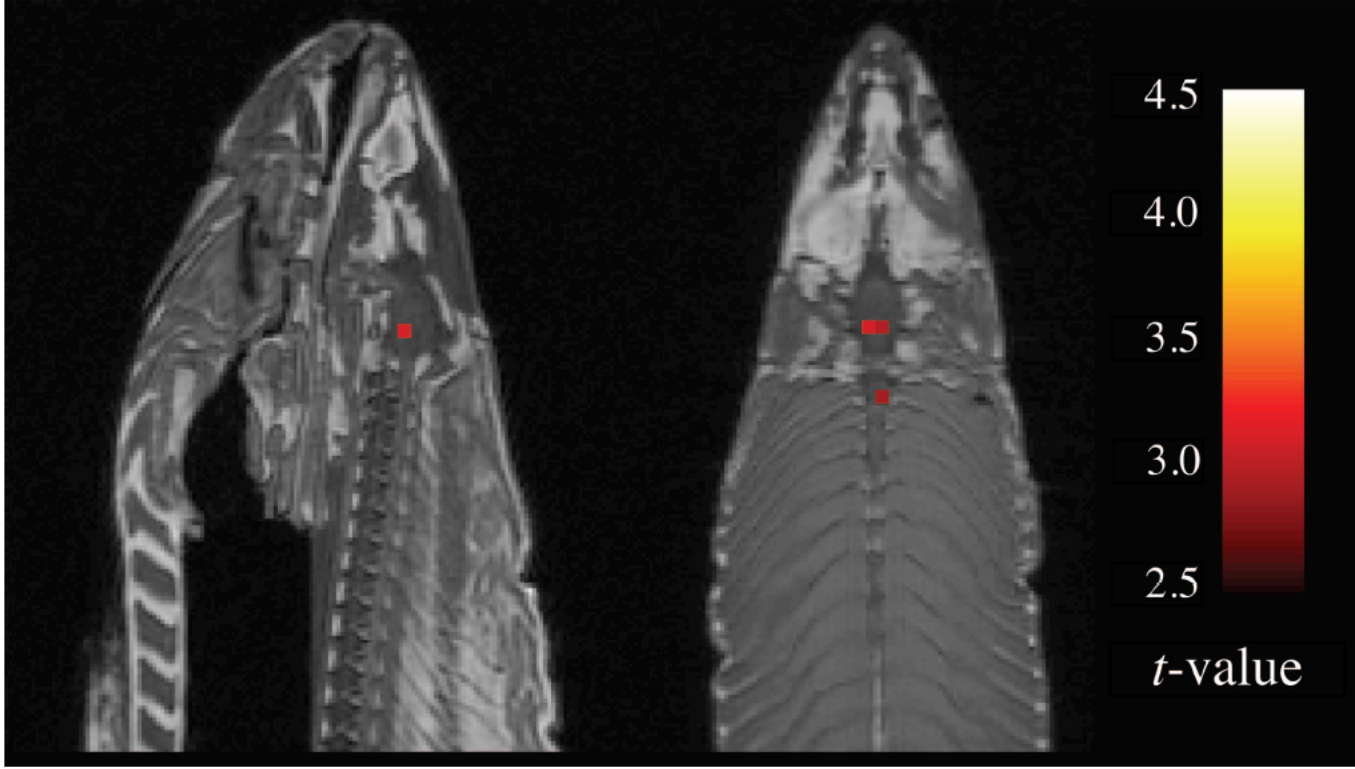


Fig. 1. Sagittal and axial images of significant brain voxels in the task > rest contrast. The parameters for this contrast were $t(131) > 3.15$, $p(\text{uncorrected}) < 0.001$, 3 voxel extent threshold. Two clusters were observed in the salient nervous system. One cluster was observed in the medial brain cavity and another was observed in the upper spinal cord.

We need a method that ensure a given (good) Specificity and as much Power it can.

2 FamilyWise Error Rate

2.1 FamilyWise Error Rate (FWER)

Probability of AT LEAST one false rejection

$$\begin{aligned} \text{FWER} = \alpha &= P(p_i \leq \tilde{\alpha} \text{ for at least one true null hypothesis}) \\ &= P\left(\bigcup_{i \in \{\text{true null hypos}\}} \{p_i \leq \tilde{\alpha}\}\right) \end{aligned}$$

Procedure:

- Fix α (usually $\alpha = .05$ or $.01$)
- Compute $\tilde{\alpha}$
- Derive the threshold u from $\tilde{\alpha}$ (e.g. for z -scores: $u_{\tilde{\alpha}} = \Phi^{-1}(1 - \tilde{\alpha})$)

2.2 Sidak Correction

If one want to control (the probability of) *FWER* at level α , what is the $\tilde{\alpha}$ -level to be used for each test?

When the m tests are **independent** (or with some form positive dependence):

$$\begin{aligned}
 \text{FWER} = \alpha &= P(p_i \leq \tilde{\alpha} \text{ for at least one true null hypo}) = \\
 &= P\left(\bigcup_{i \in \{\text{true null hypos}\}} \{p_i \leq \tilde{\alpha}\}\right) = \\
 &= 1 - P\left(\bigcap_{i \in \{\text{true null hypos}\}} \{p_i > \tilde{\alpha}\}\right) = \\
 &\text{(deMorgan)} \\
 &= 1 - (1 - \tilde{\alpha})^{m_0} \quad (m_0 = \text{numb of true null hypos}) \\
 &\text{(we don't know } m_0, \text{ but we know that } m_0 \leq m) \\
 &\leq 1 - (1 - \tilde{\alpha})^m
 \end{aligned}$$

2.3 Sidak Correction

Hence:

$$\begin{aligned}
 1 - \alpha &= (1 - \tilde{\alpha})^m \\
 (1 - \alpha)^{1/m} &= (1 - \tilde{\alpha}) \\
 \tilde{\alpha} &= 1 - (1 - \alpha)^{1/m}
 \end{aligned}$$

So, we define $\tilde{\alpha} = 1 - (1 - \alpha)^{1/m}$

Declare Active all voxles with statistic $z \geq u_{\tilde{\alpha}}^{-}$ (m = number of hypotheses)

Unfortunately, this solution is valid only when the p-values are INDEPENDENT (or have a positive dependence).

In most cases, tests have a dependence induced by the original variables.

2.4 Bonferroni

FWER: Probability of AT LEAST one false rejection:

Bonferroni: $\tilde{\alpha} = \alpha/m$

Declare Active all voxles with statistic $z \geq u_{\tilde{\alpha}}^{-}$ (m = number of hypotheses)

FWER under control:

$$\begin{aligned}
 \text{FWER} &= P(p_i \leq \alpha/m \text{ for at least one True null hypo}) \\
 &= P\left(\bigcup_{i \in \{\text{true null hypotheses}\}} \{p_i \leq \alpha/m\}\right) \\
 &\leq \sum_{i \in \{\text{true null hypotheses}\}} P(p_i \leq \alpha/m) \\
 &\leq \#\{\text{true null hypotheses}\} \frac{\alpha}{m} \leq \alpha
 \end{aligned}$$

Pros

- Very easy
- Control the FWER under any dependence

Cons

- Conservative (Adj. P-values very high, few rejections)

3 False Discovery Rate

3.1 False Discovery Rate (FDR)

	# Not Rejected	# Rejected	Total
# H_0	A_0	R_0	m_0
# H_1	A_1	R_1	m_1
	A	R	m

To control the **False Discovery Rate (FDR)** means defining a procedure s.t.

$$\text{mean}\left(\frac{\#False\ Rej.s}{\#Rej.s}\right) = \text{mean}\left(\frac{R_0}{R}\right) \leq \alpha$$

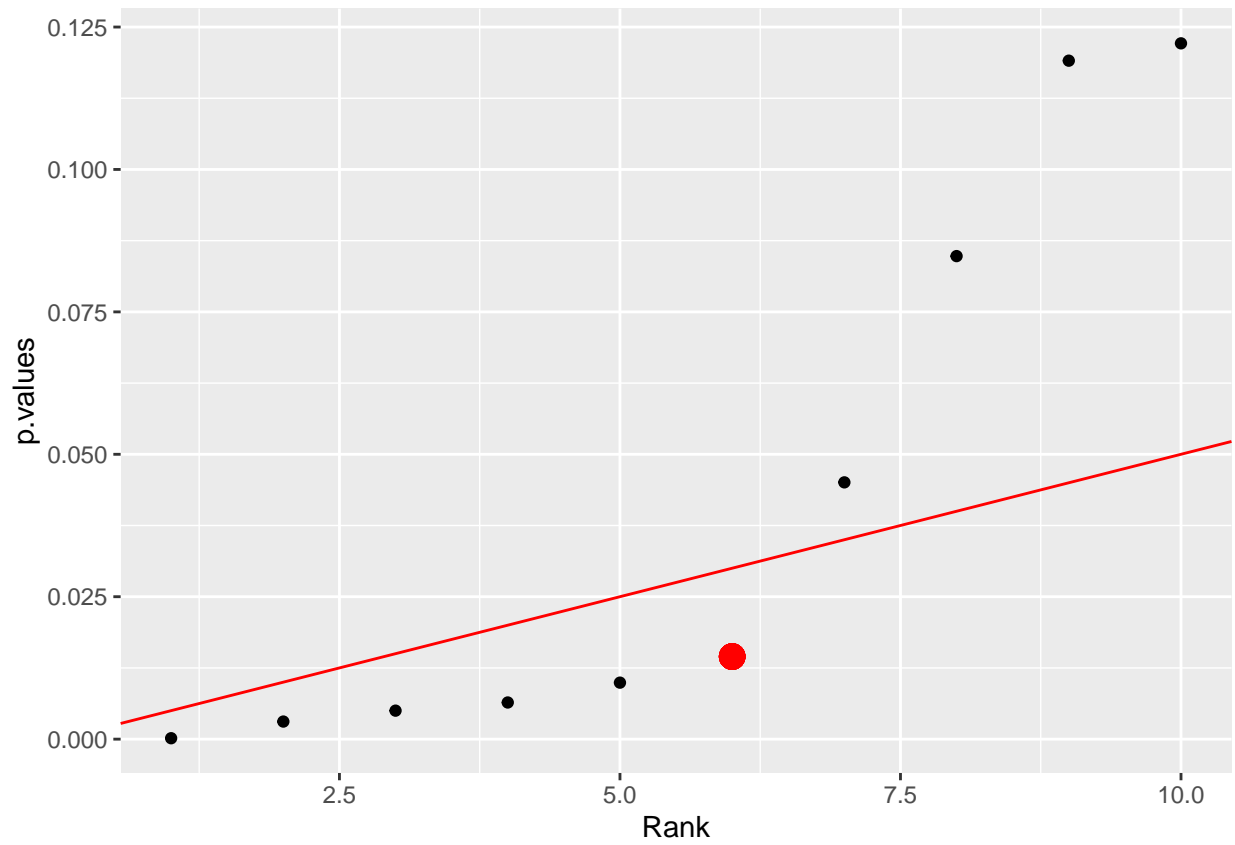
usually $\alpha = .05$

Remark: $FWER = P\left(\frac{R_0}{R} > 0\right) \leq \alpha$

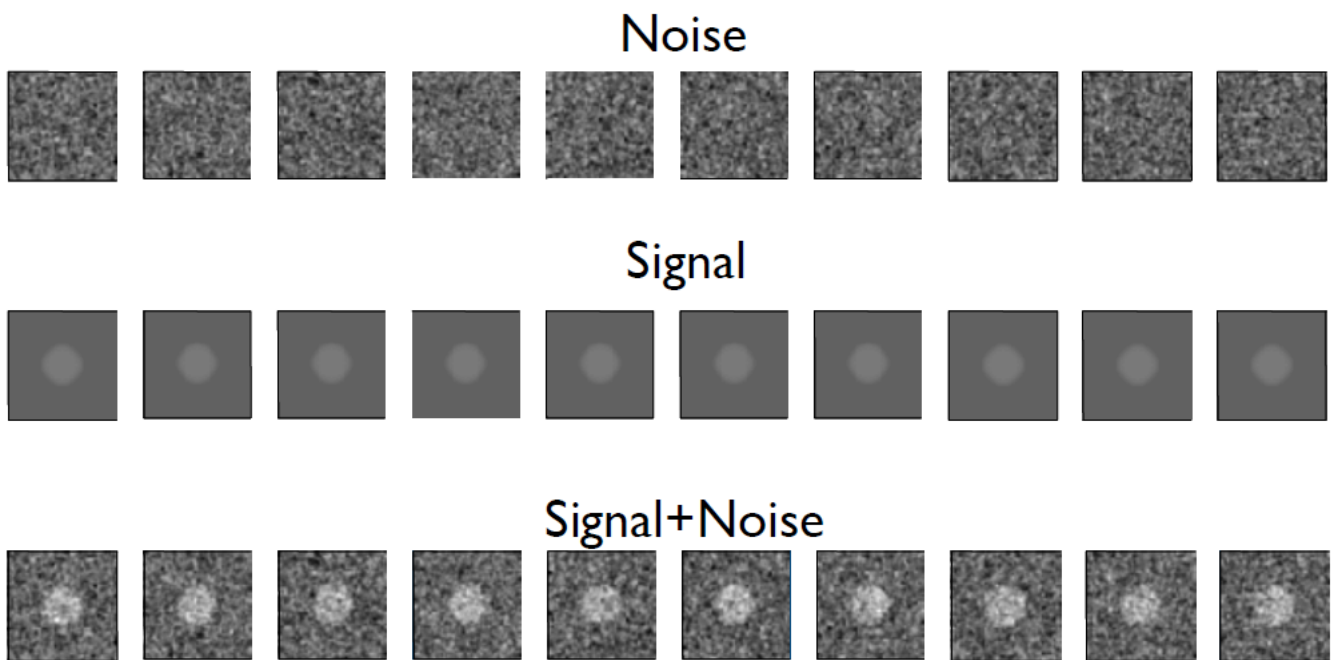
Benjamini and Hochberg (1995). Journal of the Royal Statistical Society, Series B (Methodological) 57 (1): 289–300.

3.2 BH procedure

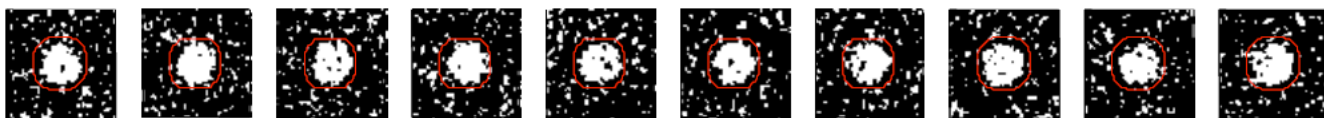
- Find the largest sorted p-value such that $p_{(k)} \leq \frac{k}{m}\alpha$ (m = number of hypotheses)
- Define $\tilde{\alpha} = p_{(k)}$
- Declare Active all voxles with statistic $z \geq u_{\tilde{\alpha}}$



3.3 A toy example



uncorrected voxelwise control of FP rate at 10%



percentage of all null pixels that are False Positives

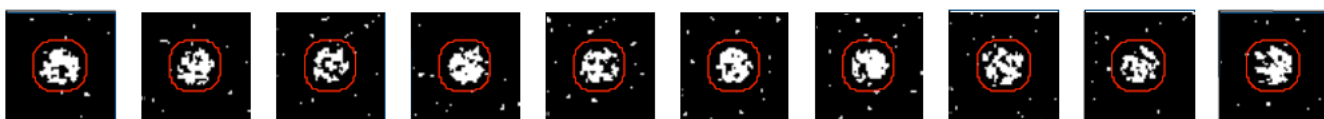
control of FamilyWise Error rate at 10%



occurrence of FamilyWise Error

FWE

control of False Discovery Rate at 10%



percentage of activated (reported) pixels that are False Positives

3.4 FDR with Dependent tests

BH is valid under assumption of independence between the p-value and **Positive Regression Dependency** on each subset of true null hypos
(eg normal with positive correlation)

Usually valid in fMRI data

For ANY dependence: **BY**

Benjamini Y, Yekutieli D. (2001) The control of the false discovery rate in multiple testing under dependency. Annals of statistics 29 (4): 1165-1188

But usually very conservative (sometime more than Bonferroni)

3.5 FWER or FDR?

Assumptions implied by FDR

The assumptions are exchangeable:

True Rejections can compensate False Rejections

I don't think that the FDR is adequate in fMRI data.

Problems

- Cheating
- Subsets

Cheating I can add not interesting hypotheses with significant with p-values to compensate false rejections.

Subsets

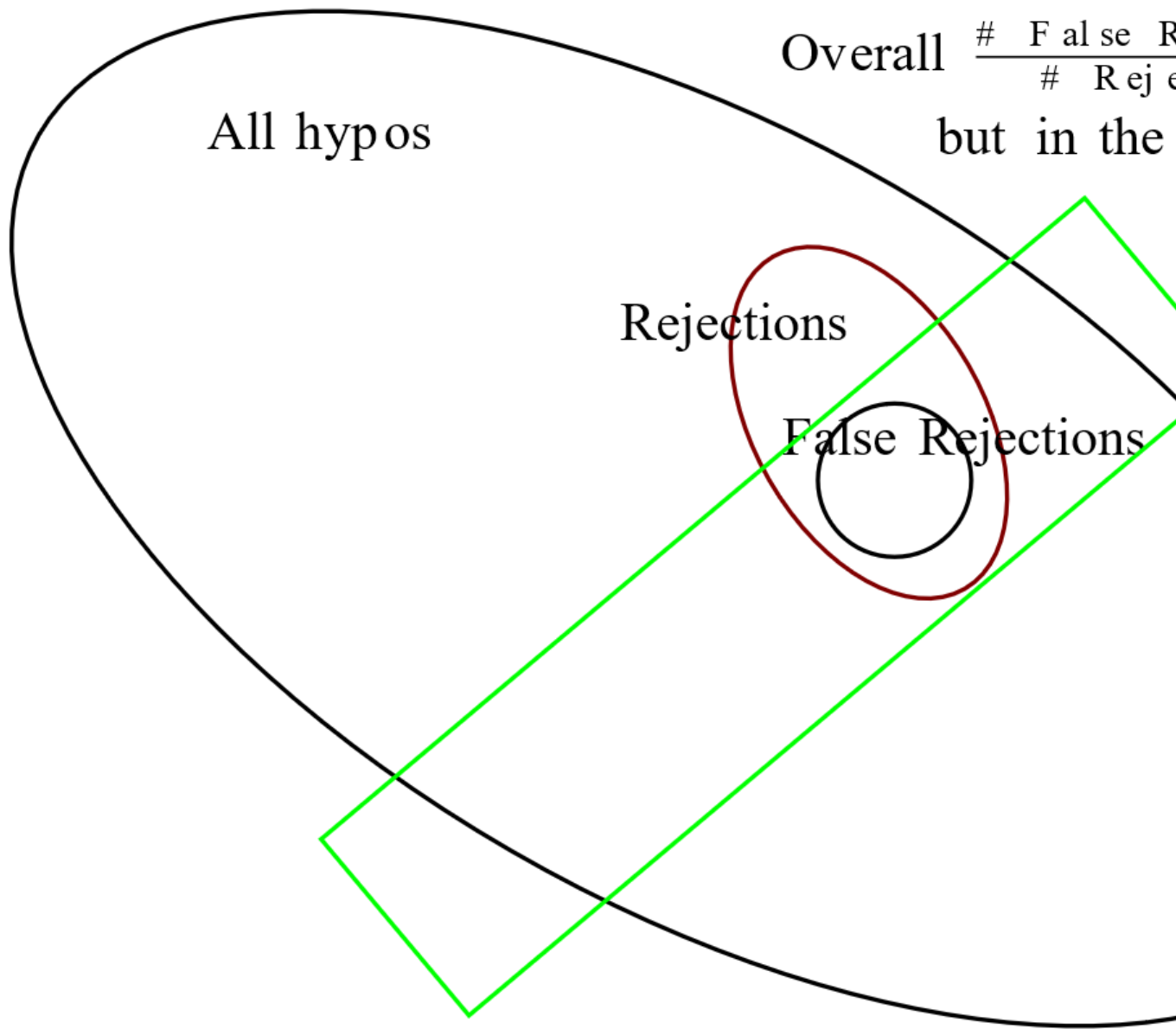
FDR control does NOT imply control of FDR in all subsets

eg: I correct all the tests, while discussing only those that I know how to better explain.

- FDR control on all subsets = FWER control
- FWER control on all subsets = FWER control

Finner H, Roters M. (2001) On the false discovery rate and expected type I errors. biometrical Journal; 43 (8): 985-1005

3.6 Subsets of Rejections

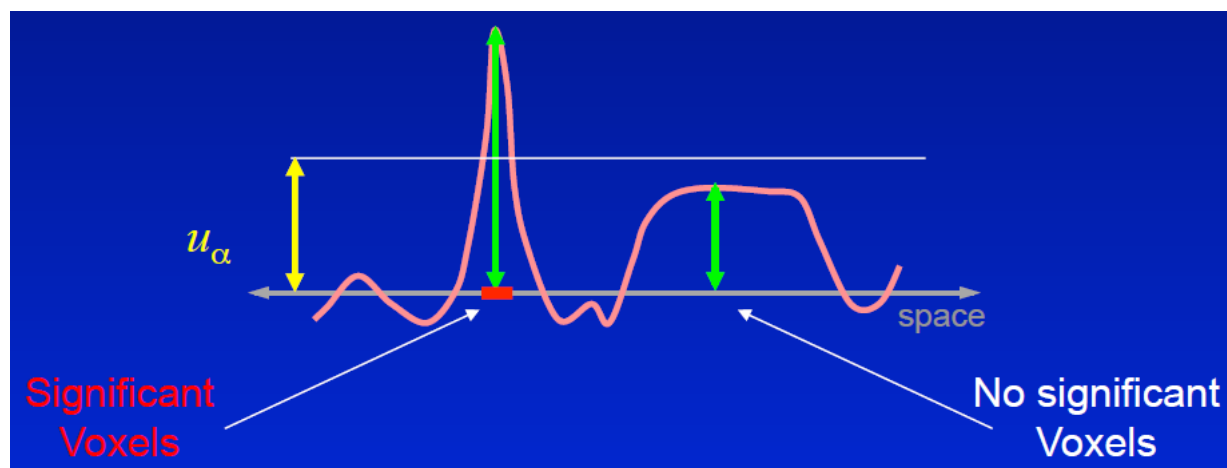


4 Three levels of inference in neuroscience

4.1 Levels of inference

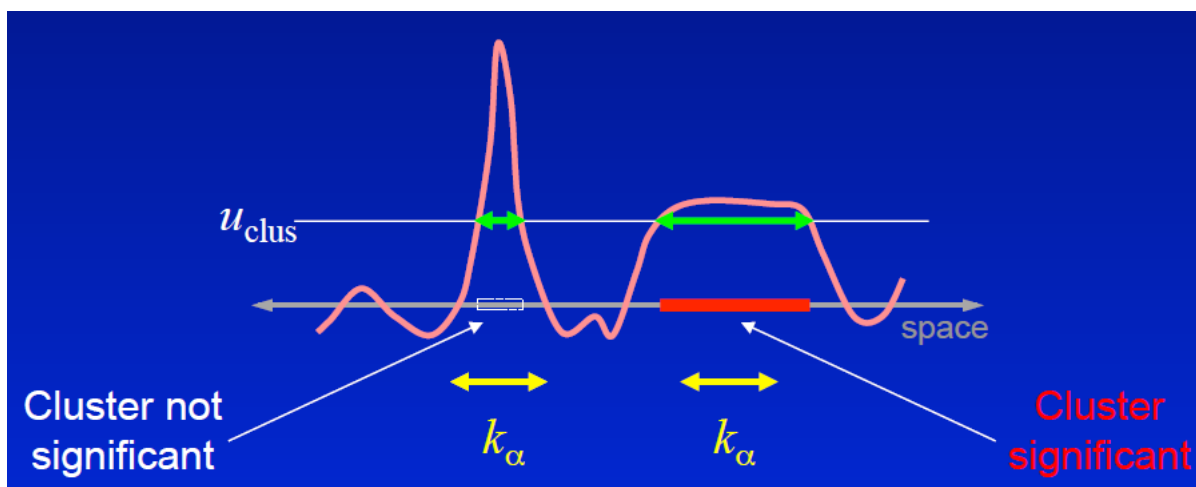
- Voxel-level
- Cluster-level
- Set-level

4.2 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

4.3 Cluster-level Inference



- Typically better sensitivity
- Worse spatial specificity
 - The null hypo of entire cluster is rejected
 - Only means that AT LEAST ONE voxels in cluster active

4.4 Set-level Inference

- Count number of blobs c
 - Minimum blob size k
- Worst spatial specificity

- Only can reject global null hypothesis
- just a global inference, same as weak control

5 Voxel-level Inference

5.1 Do you Bonferroni... or not?

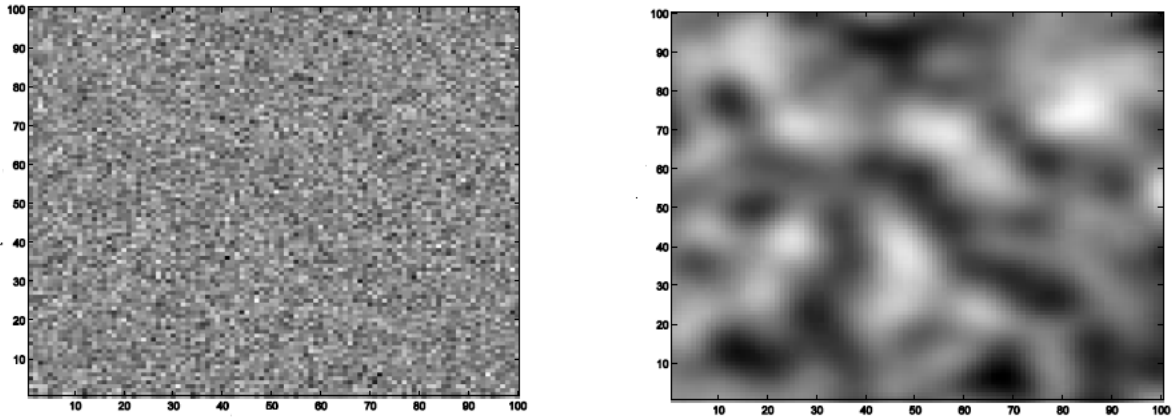
As we know, we test each hypothesis (voxel) at level: $\tilde{\alpha} = \alpha/m$

In fMRI, we use more often t -threshold (or z -threshold, F -threshold, χ^2 -threshold – similar results hold) instead of p -values and α .

$p \leq \tilde{\alpha}$ Equivalent to $t \geq t_{1-\tilde{\alpha}}$

We look for the distribution of $Max - T$ (maximum t-statistic for m test under H_0).

5.2 Smoothed Images - Autocorrelation



Intrinsic smoothness

- MRI signals are acquired in k-space (Fourier space); after projection on anatomical space, signals have continuous support
- Diffusion of vasodilatory molecules has extended spatial support

Extrinsic smoothness

- Re-sampling during preprocessing
- Deliberate additional smoothing to increase SNR
- Robustness to between-subject anatomical differences

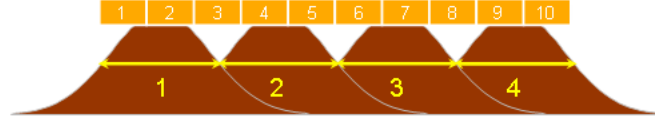
Unfortunately, the spatial correlation makes Bonferroni correction too conservative

5.3 RESEL

RESEL stands for **RE**Resolution **EL**ement

A RESEL is simply a block of pixels that is the same size as the FWHM.

Eg: 10 voxels, 2.5 FWHM, 4 RESELS



- Number of RESELS is similar to, but NOT equal to, the number of independent observations in an image
- The number of resels depends only on the number of pixels, and the FWHM
- Smoothness (FWHM) can be estimated from standardized residuals.

5.4 Max-T distribution

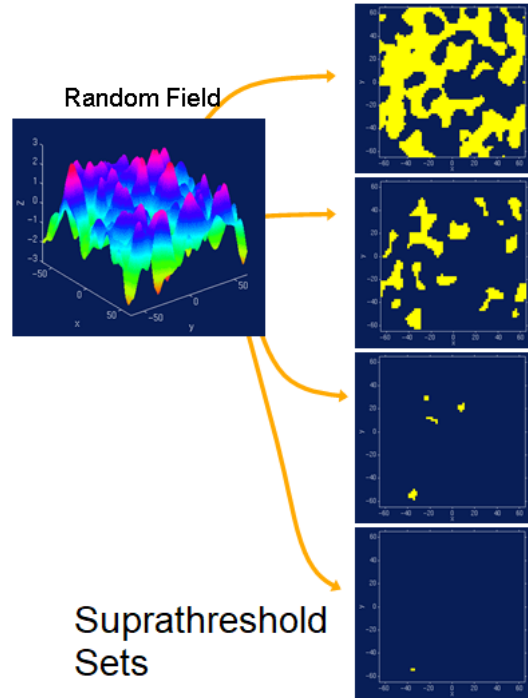
We know there is some function of the number of Resels, R , that describes the Max-t distribution

We don't know how to calculate it

But there is an approximation of the tail, and that is what matters.

This approximation is derived from Random Field Theory (RFT) Theory

5.5 Random Field Theory



Euler Characteristic χ_u can be thought of as the number of blobs in an image after thresholding.

- Topological Measure: #blobs - #holes
- At high thresholds, just counts blobs

$$\begin{aligned}
 FWER &= P(\text{Max} - t \geq u \mid H_0) \\
 &= P(\text{One or more blobs} \mid H_0) \\
 (\text{no holes}) &\approx P(\chi_u \geq 1 \mid H_0) \\
 &\leq E(\chi_u \mid H_0)
 \end{aligned}$$

e.g. for Gaussian test statistic (i.e. z , not t):

$$E(\chi_u|H_0) \approx R2\pi^{-2}W^{-3}u^2\exp(-u^2/2)$$

R is the number of resels, u is the z-score threshold

5.6 How it works

- First we estimate the smoothness (spatial correlation) of our statistical map.
- Then we use the smoothness values in the appropriate RFT equation, to give the expected EC at different thresholds.
- This allows us to calculate the threshold u at which we would expect 5% of equivalent statistical maps arising under the null hypothesis to contain at least one area above threshold.
- All hypos (voxels) with $t \geq u$ are rejected and FWER is controlled.

6 Cluster-level inference

6.1 Spatial extent - motivation

Peak extent (voxel-level)

We see a t -value of 10. It is so surprising (under the null hypothesis) that we have to reject it (i.e. t -statistic larger than u_α).

Spatial extent (cluster-level)

We threshold the t -map at $u_{clust} = 2.3$ (arbitrary threshold) and look at the spatial extent of clusters.

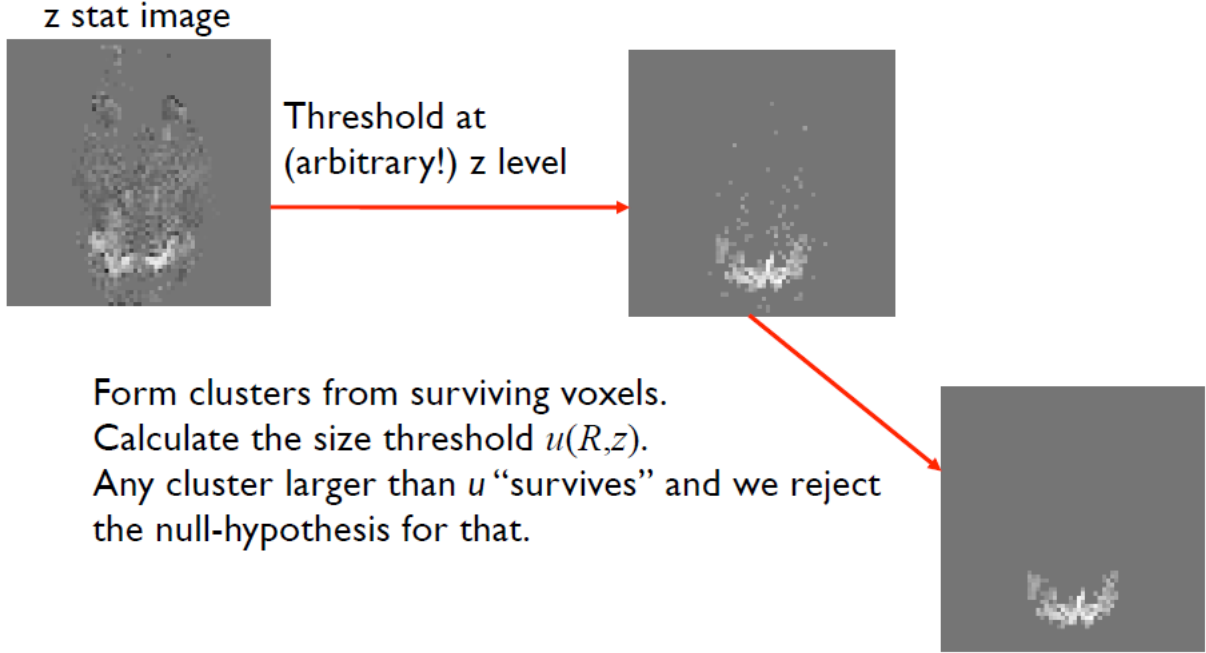
We see 300 connected voxels all with t -values $\geq u_{clust} = 2.3$.

It is so surprising (under the null hypothesis) that we have to reject it (i.e. size of the cluster larger than k_α).

6.2 Cluster-level Inference

Two step-process:

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



If we reject any cluster we will reject the largest.

We need the distribution of the largest cluster (given a threshold u_{clust}), under the null-hypothesis.

k_α is the $(1 - \alpha)$ -quantile of this distribution.

So, just as was the case for the t -values, we now have a distribution $f(R, u_{clust})$ that allows us to calculate a Family Wise threshold k_α pertaining to cluster size.

$W = |\Lambda|^{-1/(2D)} = FWHM(4\log_e 2)^{-1/2}$, where $FWHM = FWHM_x FWHM_y FWHM_z$, Λ is the covariance matrix of the field's first partial derivatives and D is the number of dimensions (i.e. $D = 3$)

At high thresholds, the number of clusters χ_u approximates the number of maxima and has been shown to have a Poisson distribution (Adler, 1981, Theorem 6.9.3, page 161):

$$P(\chi_u = c) \approx \lambda(c, E(\chi_u))$$

- the expected number of maxima $E(\chi_u)$ (i.e., clusters) is:

$$E(\chi_u) \approx R^2 \pi^{-(D+1)/2} W^{-D} u_{clust}^{D-1} \exp(-u_{clust}^2/2)$$

- Distribution of the number of voxels n in a cluster:

$$P(n \geq k) \approx \exp(-\beta k^{2/D})$$

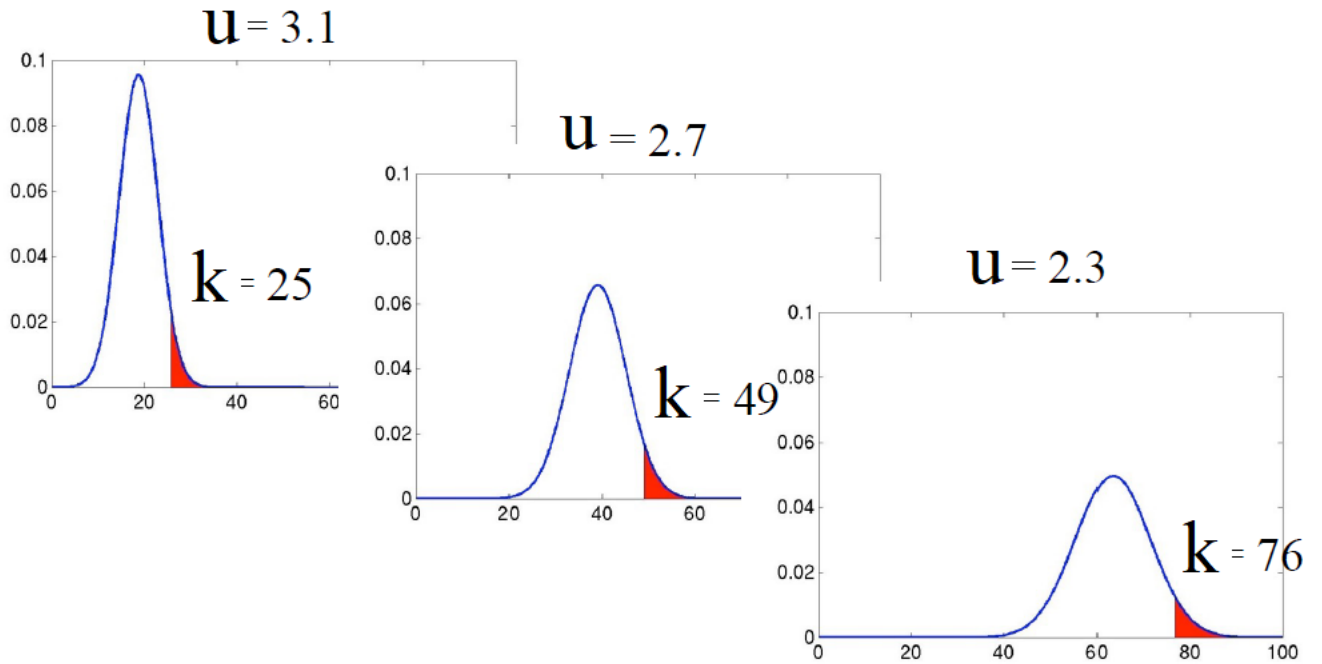
$$, \text{ where } \beta = [\Gamma(D/2 + 1)E(\chi)/(S\Phi(-u))]^{2/D}$$

The probability to observe a cluster with k or more voxles is

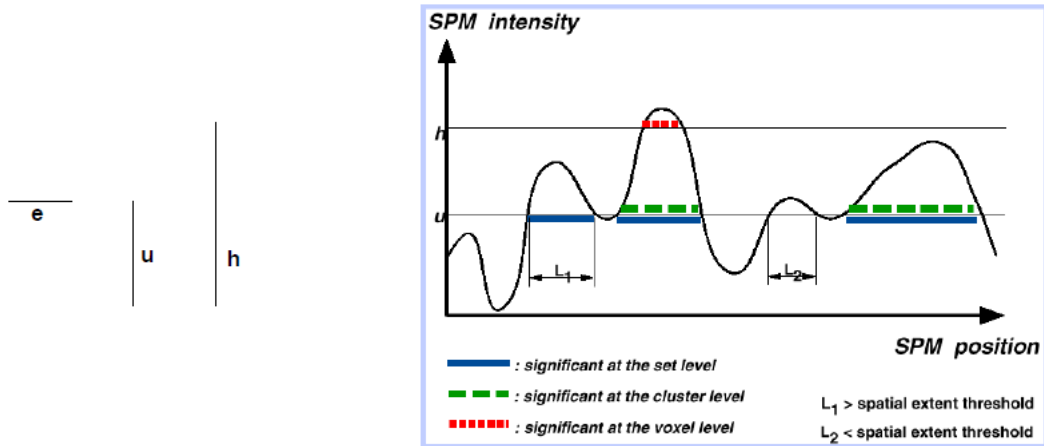
$$P(u_{clust}, k) \approx 1 - \exp(-E(\chi)P(n \geq k))$$

k_α is the k such that $P(u_{clust}, k) = \alpha$.

k_α depends on the initial *cluster-forming* threshold u_{clust} .



6.3 Voxel/Cluster-level in a glance



6.4 Remarks

- Needs a null-hypothesis, a test-statistic and an initial cluster forming threshold.
- Calculates a (size) threshold based on number of RESELS and initial (z) threshold

Pros

- Gives a (size) threshold such that the family-wise error is controlled.
- Calculates that threshold very fast.

6.5 Limitations

Limitations (1/2)

- *Sufficient smoothness*
 - FWHM smoothness $3 - 4 \times$ voxel size (Z)
 - More like $\sim 10 \times$ for low-df T images
- *Smoothness estimate* is biased when images not sufficiently smooth
- *Multivariate normality*: virtually impossible to check
- Several layers of *approximations* (e.g. Lattice Image Data \approx Continuous Random Field)
- *Stationarity* required for cluster size results

This can be solved via permutation-approach

Limitations (2/2)

- Inference pertains to entire cluster (i.e. there is at least one voxel)
 - This can be solved via All-Resolution Inference approach
- Initial threshold is arbitrary and must be chosen a priori
 - This can be solved via All-Resolution Inference approach

(All-Resolution Inference, Rosenblatt et al. <https://www.ncbi.nlm.nih.gov/pubmed/30056198>)

... **A serious problem**, no jockes:

Eklund, Nichols and Knutsson (2016) Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. PNAS