Correction for multiple comparisons

Cyril Pernet, PhD

SBIRC/SINAPSE – University of Edinburgh

Overview

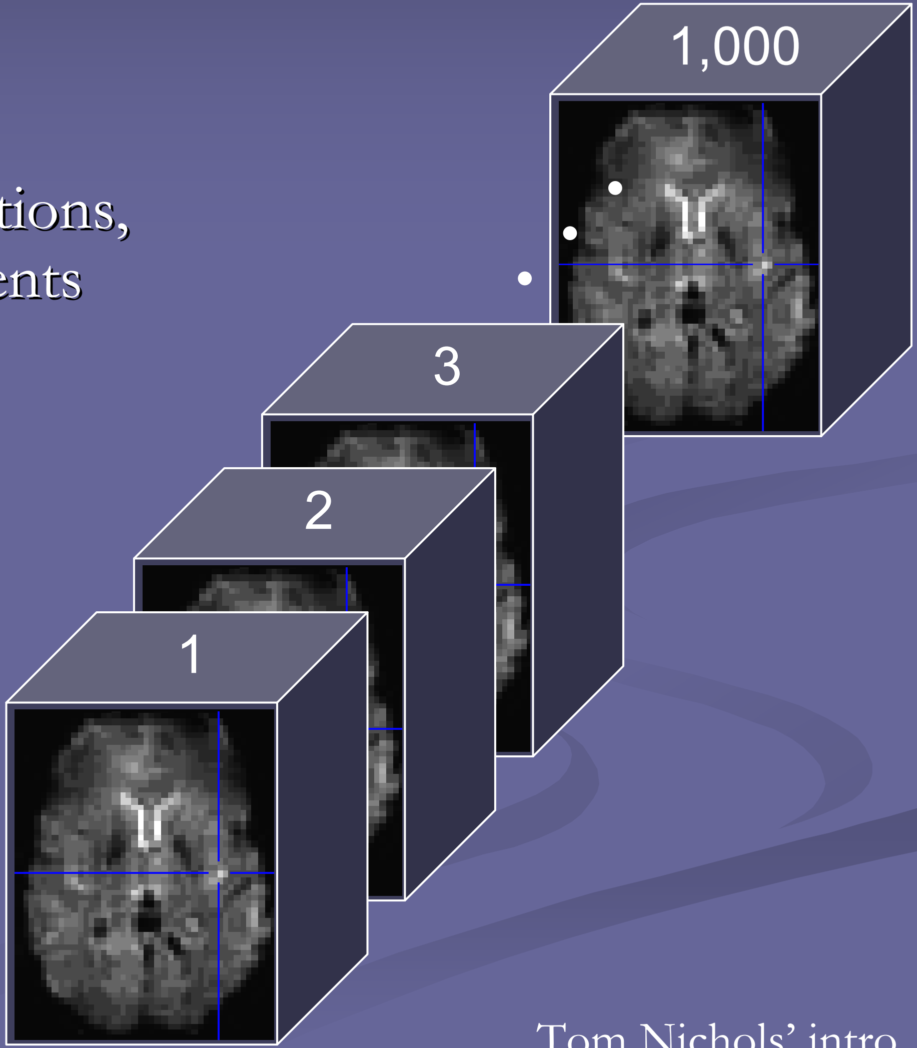
- Multiple comparisons correction procedures
- Levels of inferences (set, cluster, voxel)
- Circularity issues

Multiple comparison correction

Avoiding false positives

What 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
- Massive MCP!



Tom Nichols' intro

What Problem?

- Typical brain ~ 130000 voxels
- @ $p = .05$, it is expected = 6500 false positives!
- @ a more conservative value like $p = .001$ we still expect 130 false positives.

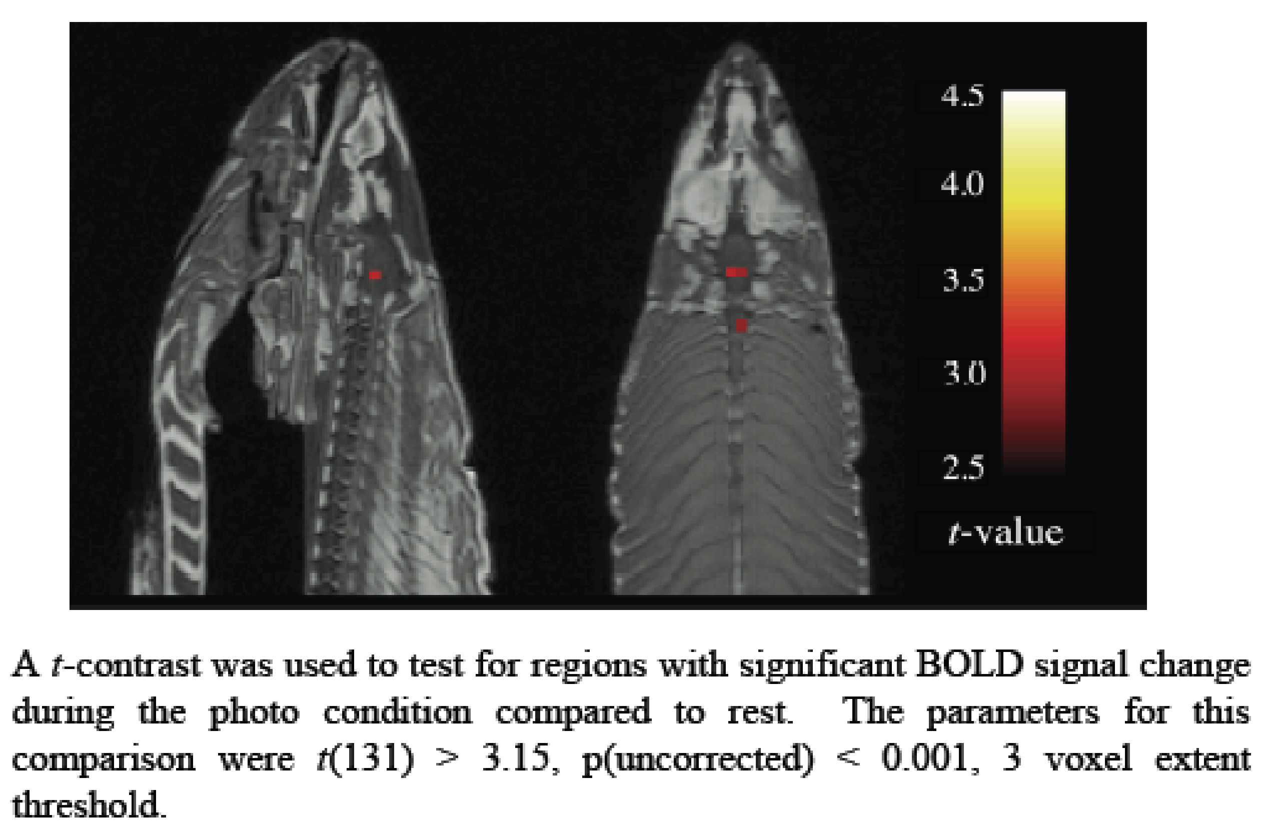
- Using extend threshold k without correction is not enough as it, by chance, can cluster as well.

What Problem?

- Bennet et al., 2009

- Task: take a decision about emotions on pictures
- Design: blocks of 12 sec activation/rest
- Analysis: standard data processing with SPM
- Subject: a dead salmon!

What Problem?



A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were *t*(131) > 3.15, p(uncorrected) < 0.001, 3 voxel extent threshold.

- The cluster was 81mm^3 ! – after multiple comparison corrections all false activations were removed.

Solutions for MCP

- Height Threshold
- Familywise Error Rate (FWER)
  - Chance of *any* false positives; Controlled by Bonferroni & Random Field Methods
- False Discovery Rate (FDR)
  - Proportion of false positives *among* rejected tests
- Bayes Statistics

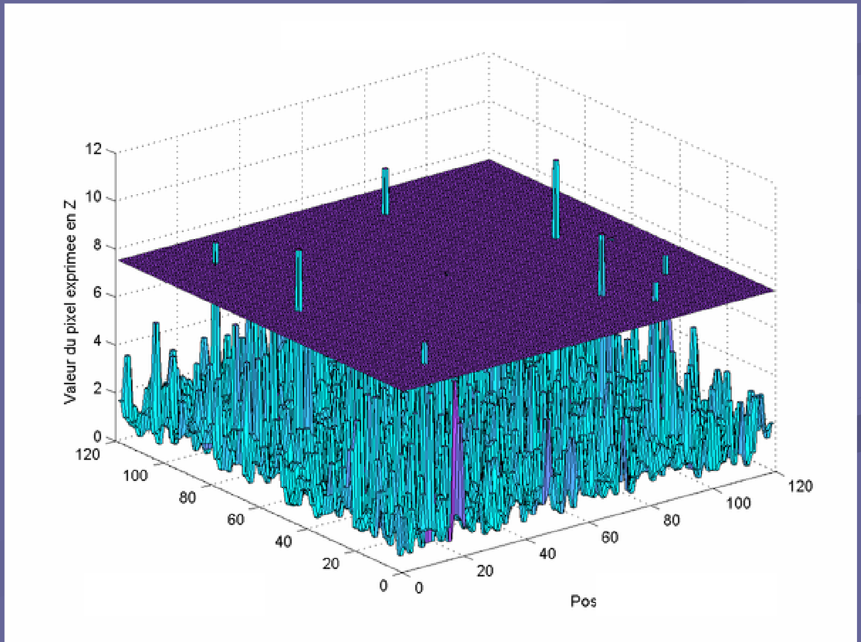
From single univariate to massive univariate

Univariate stat	Functional neuroimaging
1 observed data	Many voxels
1 statistical value	Family of statistical values
Type 1 error rate (chance to be wrong rejecting H0)	Family-wise error rate
Null hypothesis	Family-wise null hypothesis

Height Threshold

- Choose locations where a test statistic Z (T, F, ...) is large to threshold the image of Z at a height z
- The problem is how to choose this threshold z to exclude false positives with a high probability (e.g. 0.95)?

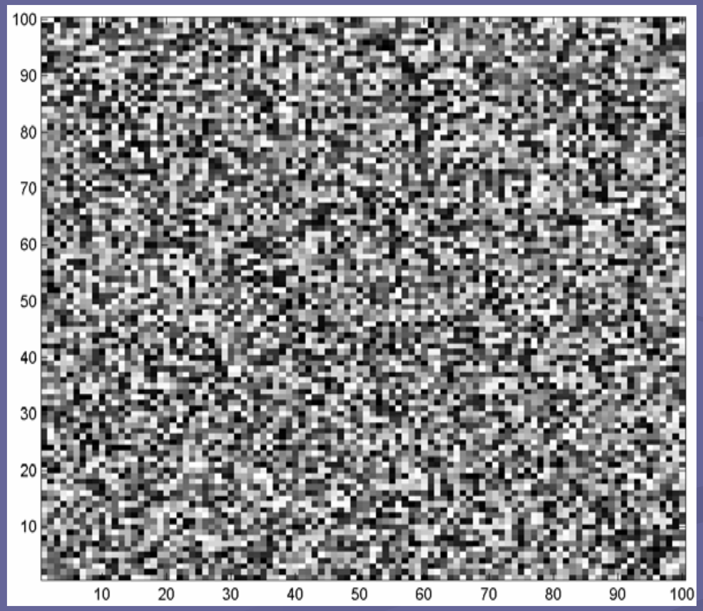
To control for family wise error on must take into account the nb of tests



Bonferroni

- 10000 Z-scores ; alpha = 5%
- alpha corrected = .000005 ; z-score = 4.42

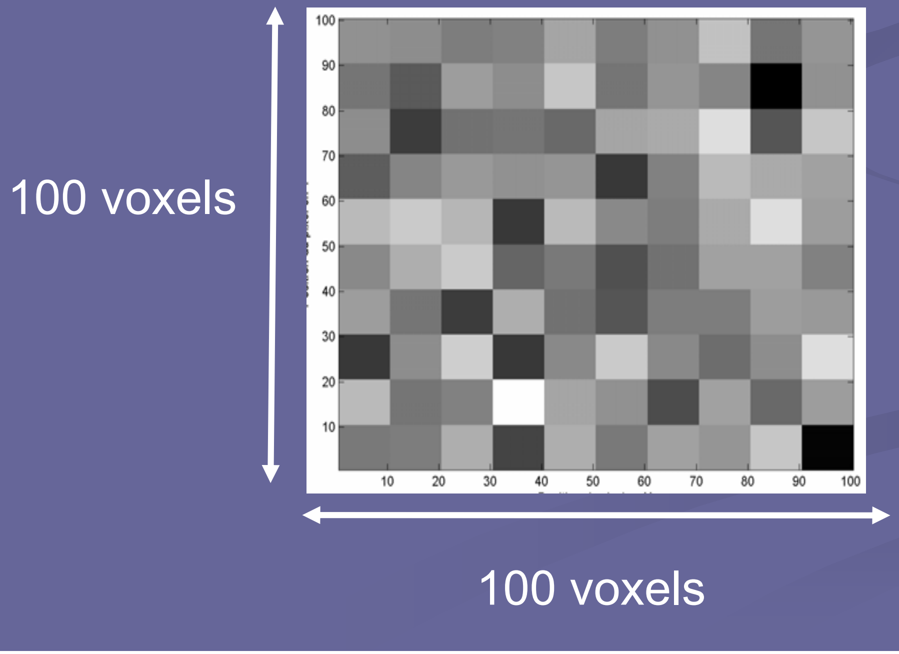
100 voxels



100 voxels

Bonferroni

- 10000 Z-scores ; alpha = 5%
- 2D homogeneous smoothing – 100 independent observations
- alpha corrected = .0005 ; z-score = 3.29

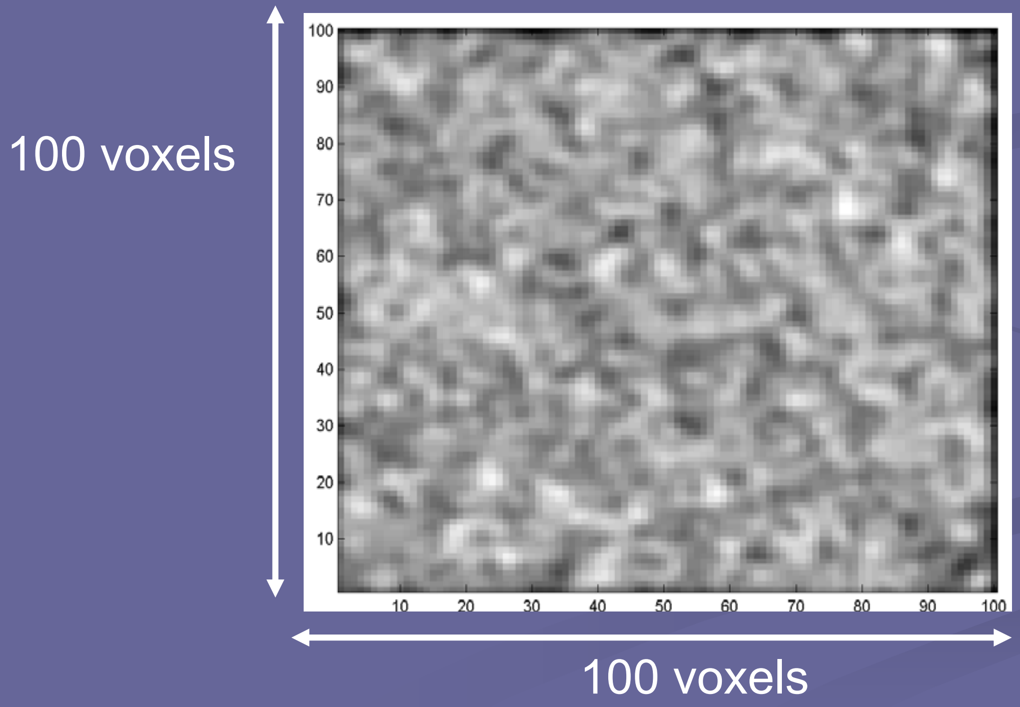


Solutions for MCP

- An important feature of neuroimaging data is that we have a family of stat values that has topological features (Bonferroni for instance consider tests as independent)
- Why considering data as a smooth lattice? (Chumbley et al., 2009 NeuroImage 44)
- fMRI/PET are projection methods of data points onto the whole space – MEEG forms continuous functions in time and are smooth by the scalp (space)
- Neural activity propagate locally through intrinsic/lateral connections and is distributed via extrinsic connections / Hemodynamic correlates are initiated by diffusing signals (e.g. NO)

Random Field Theory

- 10000 Z-scores ; alpha = 5%
- Gaussian kernel smoothing –
- How many independent observations ?



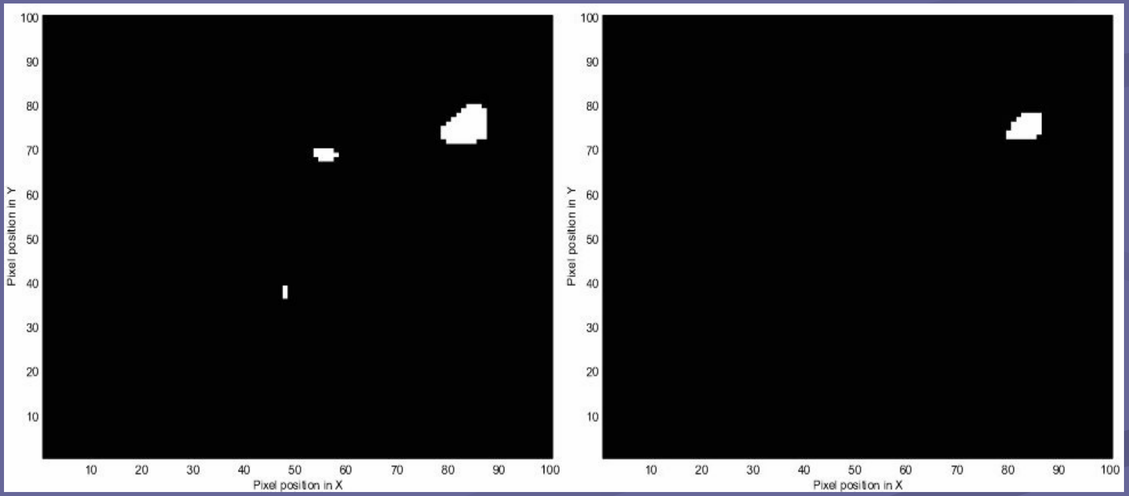
Random Field Theory

- RFT relies on theoretical results for smooth statistical maps (hence the need for smoothing), allowing to find a threshold in a set of data where it's not easy to find the number of independent variables. Uses the expected Euler characteristic (EC density)

- 1 Estimation of the smoothness = number of resel (resolution element) = f(nb voxels, FWHM)
- 2 expected Euler characteristic = number of clusters above the threshold
- 3 Calculation of the threshold

Random Field Theory

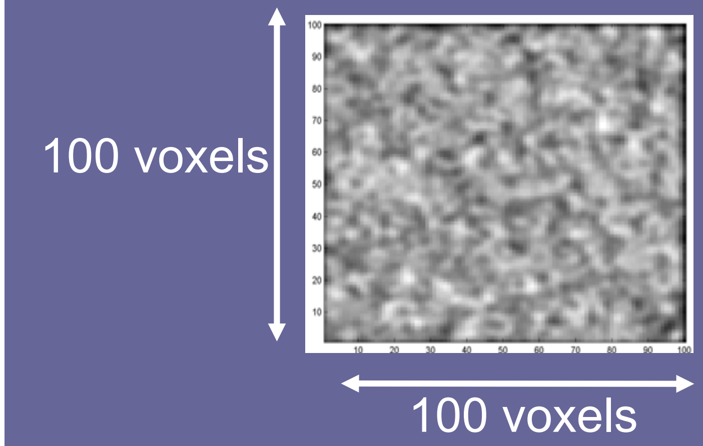
- The Euler characteristic can be seen as the number of blobs in an image after thresholding (p value that you select in SPM)
- At high threshold, EC = 0 or 1 per resel: E[EC] ≈ p^{FWE}

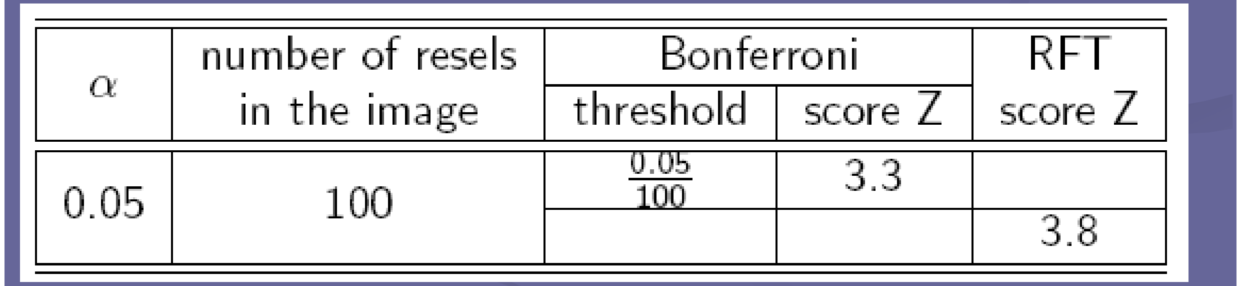


$E[EC] = R \cdot (4 \log_e 2) \cdot (2\pi)^{-2/3} \cdot Z_t \cdot e^{-1/2\, Z^2_t}$ for a 2D image, more complicated in 3D

Random Field Theory

- For 100 resels, the equation gives E[EC] = 0.049 for a threshold Z of 3.8, i.e. the probability of getting one or more blobs where Z is greater than 3.8 is 0.049





α	number of resels in the image	Bonferroni		RFT
		threshold	score Z	score Z
0.05	100	$\frac{0.05}{100}$	3.3	
				3.8

- If the resel size is much larger than the voxel size then E[EC] only depends on the nb of resels otherwise it also depends on the volume, surface and diameter of the search area (i.e. shape and volume matter)

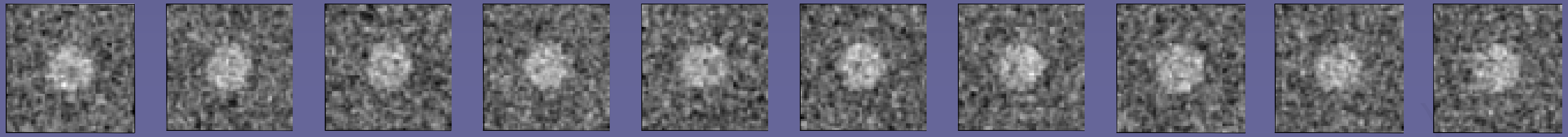
False discovery Rate

- Whereas family wise approach corrects for any false positive, the FDR approach aim at correcting among positive results only.

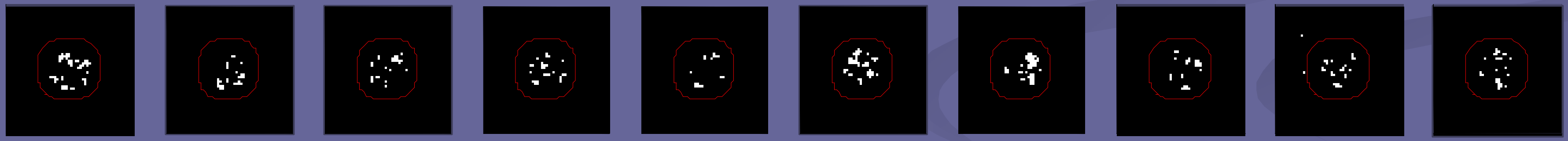
1. Run an analysis with alpha = x%
2. Sort the resulting positive data
3. Threshold to remove the false positives

False discovery Rate

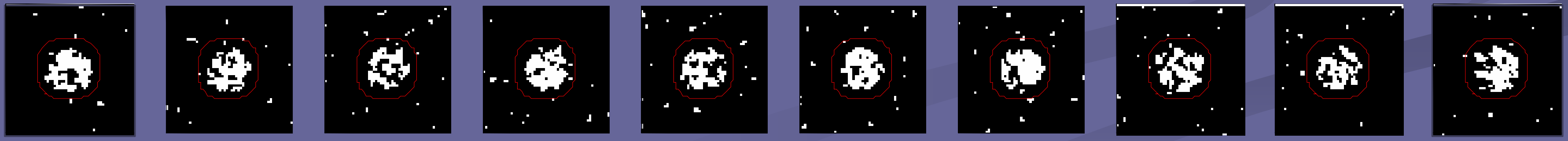
Signal+Noise



FEW correction



FDR correction

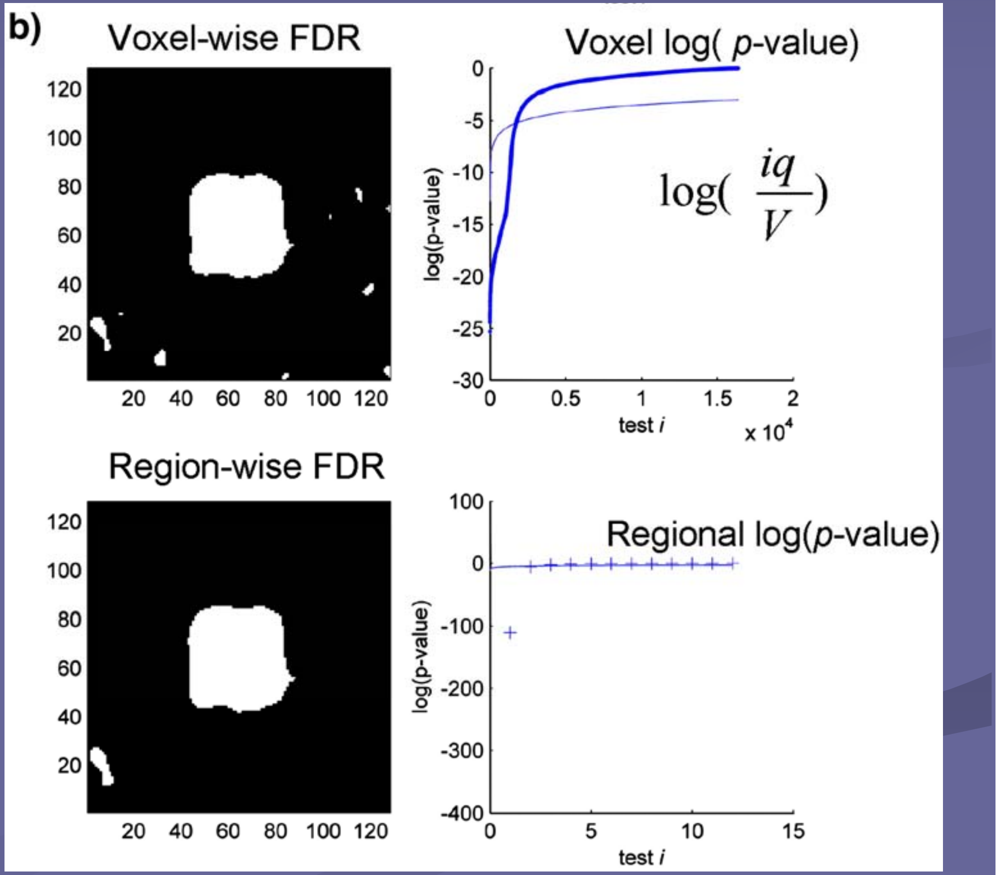


False discovery Rate

takes the spatial structure into account

Under H0 the nb of voxels per cluster is known → uncorrected p value for clusters → apply FDR on the clusters (volume-wise correction)

Assumes that the volume of each cluster is independent of the number of clusters



Levels of inference

Voxel, cluster and set

Levels of inference

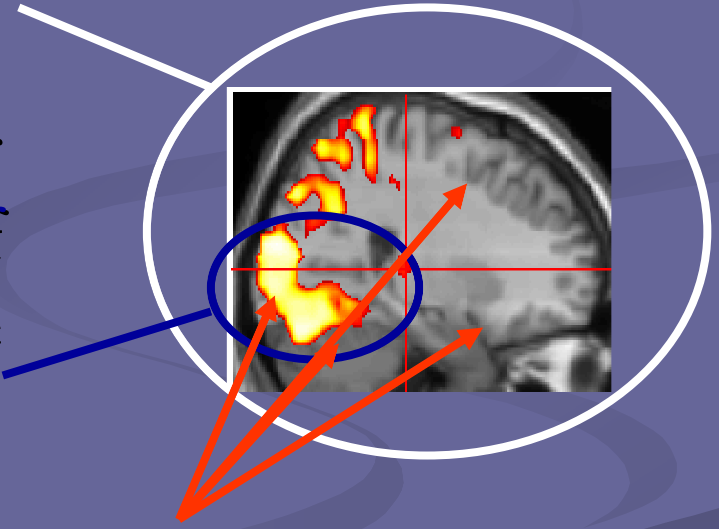
- 3 levels of inference can be considered:
- Voxel level (prob associated at each voxel)
- Cluster level (prob associated to a set of voxels)
- Set level (prob associated to a set of clusters)

- The 3 levels are nested and based on a single probability of obtaining c or more clusters (set level) with k or more voxels (cluster level) above a threshold u (voxel level): $P_w(u,k,c)$

Levels of inference

- Set level: we can reject H0 for an omnibus test, i.e. there are some significant clusters of activation in the brain.

- Cluster level: we can reject H0 for an area of a size k, i.e. a cluster of 'activated' voxels is likely to be true for a given spatial extend.



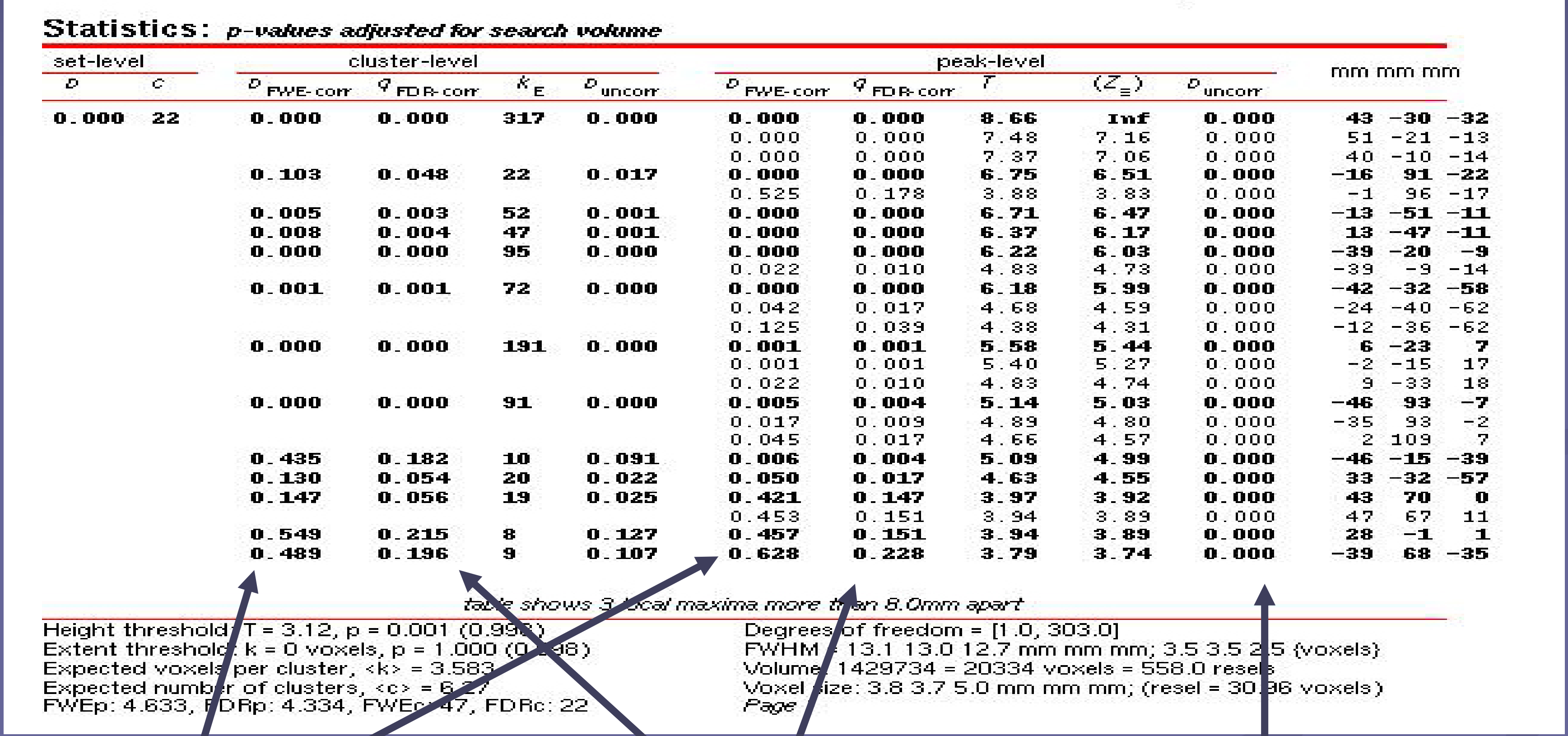
- Voxel level: we can reject H0 at each voxel, i.e. a voxel is 'activated' if exceeding a given threshold

Levels of inference

- Each level of inference is valid, but the inferences are different – e.g. a set might be enough to check that subjects activated regions selected a priori for a connectivity analysis – clusters might be good enough if hypotheses are about the use of different brain areas between groups

- Both voxel and cluster levels need to address the multiple comparison problem. If the activated region is predicted in advance, the use of corrected p values is unnecessary and inappropriately conservative – a correction for the number of predicted regions (Bonferroni) is enough

Level of inference



Statistics: *p-values adjusted for search volume*

set-level		cluster-level				peak-level					mm mm mm		
p	c	$p_{FWE\text{-}corr}$	$q_{FDR\text{-}corr}$	k_E	p_{uncorr}	$p_{FWE\text{-}corr}$	$q_{FDR\text{-}corr}$	T	$(Z_\equiv)$	p_{uncorr}			
0.000	22	0.000	0.000	317	0.000	0.000	0.000	8.66	Inf	0.000	43	-30	-32
						0.000	0.000	7.48	7.16	0.000	51	-21	-13
						0.000	0.000	7.37	7.06	0.000	40	-10	-14
		0.103	0.048	22	0.017	0.000	0.000	6.75	6.51	0.000	-16	91	-22
						0.525	0.178	3.88	3.83	0.000	-1	96	-17
		0.005	0.003	52	0.001	0.000	0.000	6.71	6.47	0.000	-13	-51	-11
		0.008	0.004	47	0.001	0.000	0.000	6.37	6.17	0.000	13	-47	-11
		0.000	0.000	95	0.000	0.000	0.000	6.22	6.03	0.000	-39	-20	-9
						0.022	0.010	4.83	4.73	0.000	-39	-9	-14
		0.001	0.001	72	0.000	0.000	0.000	6.18	5.99	0.000	-42	-32	-58
						0.042	0.017	4.68	4.59	0.000	-24	-40	-62
						0.125	0.039	4.38	4.31	0.000	-12	-36	-62
		0.000	0.000	191	0.000	0.001	0.001	5.58	5.44	0.000	6	-23	7
						0.001	0.001	5.40	5.27	0.000	-2	-15	17
						0.022	0.010	4.83	4.74	0.000	9	-33	18
		0.000	0.000	91	0.000	0.005	0.004	5.14	5.03	0.000	-46	93	-7
						0.017	0.009	4.89	4.80	0.000	-35	93	-2
						0.045	0.017	4.66	4.57	0.000	2	109	7
		0.435	0.182	10	0.091	0.006	0.004	5.09	4.99	0.000	-46	-15	-39
		0.130	0.054	20	0.022	0.050	0.017	4.63	4.55	0.000	33	-32	-57
		0.147	0.056	19	0.025	0.421	0.147	3.97	3.92	0.000	43	70	0
						0.453	0.151	3.94	3.89	0.000	47	67	11
		0.549	0.215	8	0.127	0.457	0.151	3.94	3.89	0.000	28	-1	1
		0.489	0.196	9	0.107	0.628	0.228	3.79	3.74	0.000	-39	68	-35

table shows 3 local maxima more than 8.0mm apart

Height threshold: T = 3.12, p = 0.001 (0.998)
Extent threshold: k = 0 voxels, p = 1.000 (0.998)
Expected voxels per cluster, <k> = 3.583
Expected number of clusters, <c> = 6.27
FWEp: 4.633, FDRp: 4.334, FWEc: 47, FDRc: 22

Degrees of freedom = [1.0, 303.0]
FWHM = 13.1 13.0 12.7 mm mm mm; 3.5 3.5 2.5 {voxels}
Volume: 1429734 = 20334 voxels = 558.0 resels
Voxel size: 3.8 3.7 5.0 mm mm mm; (resel = 30.96 voxels)
Page 1

RFT

Using p=.001 this creates an excursion set
Prob clusters of that size
Prob peack that height
→ after FDR correction

Uncorrected (bad)

Circularity issues in fMRI

Definition

- Refers to the problem of selecting data for analysis
- How data (areas usually) are selected, analysed and sorted is key to avoid circularity

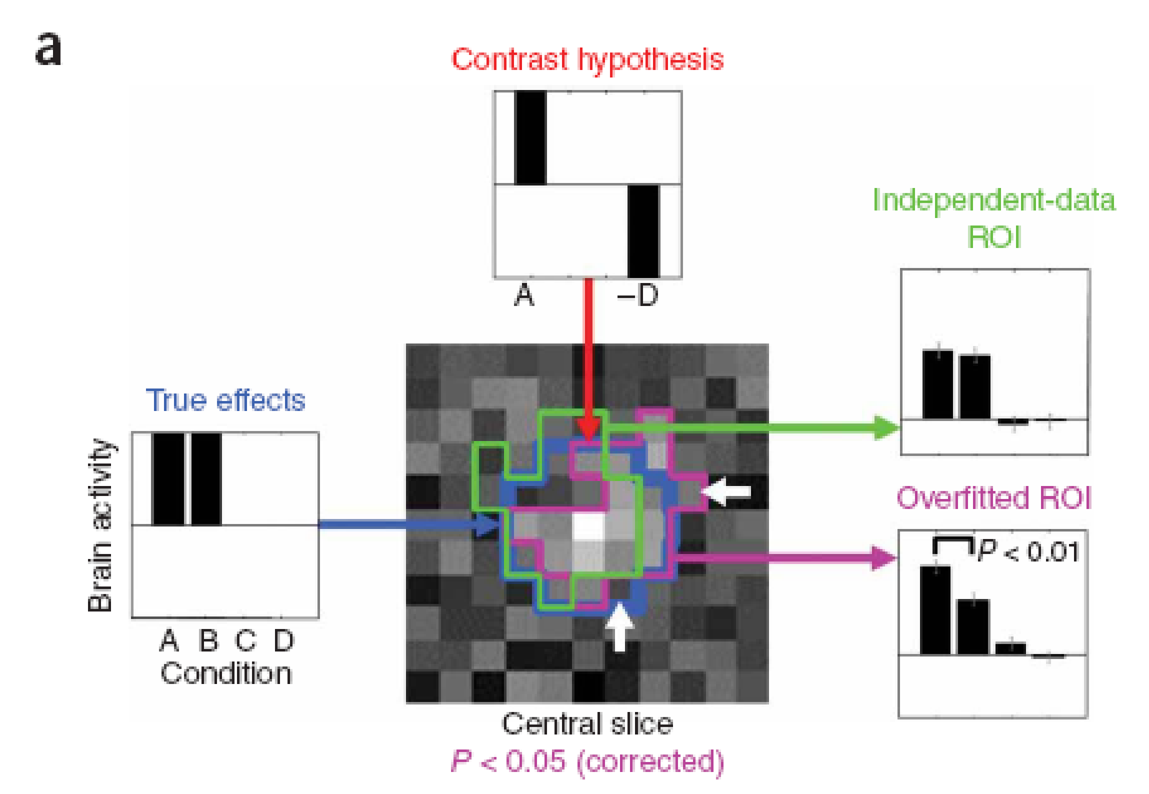
- Put forward by Vul et al. 2009, *Perspectives on Psychological Science*. 4
- Better explained in Kriegeskorte et al., 2009 *Nat. Neuroscience* 12

Circularity

- Double dipping pblm: “data are first analyzed to select a subset and then the subset is reanalyzed to obtain the results. In this context, assumptions and hypotheses determine the selection criterion and selection can, in turn, distort the results.”
- Take a gp of subjects and measures RTs, then take 2 subgroups from the same subjects and re-do some analysis?? → increases the diff.
- Take fMRI data and get activated areas, extract ROI and re-do some analyses??

Circularity

- Selection and tests must be independent – non independence create spurious effects



Circularity

- Independence of the selection and tests

1. Anatomic ROI, analysis of fMRI
2. SPM, minimal requirement is orthogonality of the contrasts (e.g. find regions using A+B>0 C=[1 1] and test A vs B C=[1 -1]) but if N_A and N_B are different there is still a bias when testing A-B (across subjects independence is ensured by $C_{selection}{}^T(X^TX)^{-1}C_{test}$)
3. Select using a subset of data, test with another one

Enough for today ☺

Thanks for your attention