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!**
What Problem?

- Typical brain \( \sim 130000 \) voxels
- \[ p = .05 \], it is expected \( = 6500 \) false positives!
- \[ p = .001 \] 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!
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 $t(131) > 3.15$, $p(\text{uncorrected}) < 0.001$, 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

<table>
<thead>
<tr>
<th>Univariate stat</th>
<th>Functional neuroimaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 observed data</td>
<td>Many voxels</td>
</tr>
<tr>
<td>1 statistical value</td>
<td>Family of statistical values</td>
</tr>
<tr>
<td>Type 1 error rate (chance to be wrong rejecting H0)</td>
<td>Family-wise error rate</td>
</tr>
<tr>
<td>Null hypothesis</td>
<td>Family-wise null hypothesis</td>
</tr>
</tbody>
</table>
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 = 0.000005; z-score = 4.42
Bonferroni

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

100 voxels
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(n 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: \( E[EC] \approx p^{FWE} \)

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

- 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 number of voxels per cluster is known \( \rightarrow \) uncorrected p-value for clusters \( \rightarrow \) 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.
Using p=.001 this creates an excursion set
Prob clusters of that size
Prob peak that height
\(\rightarrow\) after FDR correction

**Level of inference**

<table>
<thead>
<tr>
<th>Statistics:</th>
<th>p-values adjusted for search volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>set-level</td>
<td>cluster-level</td>
</tr>
<tr>
<td>(D)</td>
<td>(c)</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>0.000</td>
<td>22</td>
</tr>
<tr>
<td>0.000</td>
<td>22</td>
</tr>
<tr>
<td>0.005</td>
<td>52</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Notes:***

- **Height threshold:** \(T = 3.12, p = 0.001 (0.993)\)
- **Extent threshold:** \(k = 0 \text{ voxels}, p = 1.000 (0.000)\)
- **Expected voxels per cluster:** \(k_o = 3.563\)
- **Expected number of clusters:** \(c = 6\)
- **FWEp:** 4.633, FDRp: 4.334, FWEp: 17, FDRp: 22

Degrees of freedom: 10, 303.0

FWHM: 13.1 13.0 12.7 mm mm mm; 3.5 3.5 3.5 (voxels)

Volume: 14287342 720334 voxels; 8568.6 (voxels)

Voxel size: 3.8 3.7 5.0 mm mm mm; voxel = 8568.6 voxels

**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

- Better explained in Kriegeskorte et al., 2009 *Nat. Neuroscience* 12
Circularity

- **Double dipping problem**: “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 $C_{\text{selection}}^T (X^TX)^{-1}C_{\text{test}}$)
  3. Select using a subset of data, test with another one
Enough for today 😊

Thanks for your attention