Multiple comparison problem

DISCOS SPM course, CRC, Liège, 2009

Contents

- Recap & Introduction
- Inference & multiple comparison
- « Take home » message

Voxel by voxel statistics...

- model specification
- parameter estimation
- hypothesis test
- statistic
- statistic image or SPM

Recap & Introduction

- Inference & multiple comparison
- « Take home » message

DISCOS SPM course, CRC, Liège, 2009

Multiple comparison problem

Based on slides from: T. Nichols

Recap & Introduction

- Inference & multiple comparison
- « Take home » message

Voxel by voxel statistics...

- model specification
- parameter estimation
- hypothesis test
- statistic
- statistic image or SPM
General Linear Model (in SPM)

- Auditory words every 20s
  - Sampled every TR = 1.7s
  - Design matrix, $X$
    - Orthogonalised

  Gamma functions $\gamma_i(u)$ of peristimulus time $u$

SPM{F}

Inference at a single voxel

- $\alpha = p(t>|u|)$
  - p-value: probability of getting a value of $t$ at least as extreme as $u$. If $\alpha$ is small we reject the null hypothesis $H$.

Contents

- Recap & Introduction
  - Inference & multiple comparison
    - Single/multiple voxel inference
    - Family wise error rate (FWER)
    - False Discovery rate (FDR)
    - SPM results
- « Take home » message

Hypothesis Testing

- Test statistic $T$
  - $t$ observed realization of $T$
- $\alpha$ level
  - Acceptable false positive rate
    - Level $\alpha = P(T>|u_\alpha|)$
    - Threshold $u_\alpha$ controls false positive rate
- P-value
  - Assessment of $t$ assuming $H_\alpha$
    - $P(T>|u_\alpha|)$
      - Prob. of obtaining stat. as large or larger in a new experiment
    - $P(Null|Data)$ not $P(Data|Null)$
**Inference at a single voxel**

NULL hypothesis, H: activation is zero

\[ \alpha = p(t > u | H) \]

We can choose \( u \) to ensure a voxel-wise significance level of \( \alpha \).

This is called an 'uncorrected' p-value, for reasons we'll see later.

We can then plot a map of above threshold voxels.

**What we'd like**

- Don't threshold, model the signal!
  - Signal location?
    - Estimates and CI's on \((x,y,z)\) location
  - Signal magnitude?
    - CI's on % change
  - Spatial extent?
    - Estimates and CI's on activation volume
    - Robust to choice of cluster definition

...but this requires an explicit spatial model

**What we need**

- Need an explicit spatial model
- No routine spatial modeling methods exist
  - High-dimensional mixture modeling problem
  - Activations don’t look like Gaussian blobs
  - Need realistic shapes, sparse representation
    - Some work by Hartvig et al., Penny et al.

**Real-life inference: What we get**

- **Signal location**
  - Local maximum – no inference
  - Center-of-mass – no inference
    - Sensitive to blob-defining-threshold
- **Signal magnitude**
  - Local maximum intensity – P-values (& CI's)
- **Spatial extent**
  - Cluster volume – P-value, no CI’s
    - Sensitive to blob-defining-threshold
**Voxel-level Inference**

- Retain voxels above $\alpha$-level threshold $u_\alpha$
- 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 $u_{\text{clus}}$
  - Retain clusters larger than $\alpha$-level threshold $k_\alpha$

**Cluster-level Inference**

- Typically better sensitivity
- Worse spatial specificity
  - The null hypothesis of entire cluster is rejected
  - Only means that one or more of voxels in cluster active

**Set-level Inference**

- Count number of blobs $c$
  - Minimum blob size $k$
- Worst spatial specificity
  - Only can reject global null hypothesis

Here $c = 1$; only 1 cluster larger than $k$.
### Sensitivity and Specificity

<table>
<thead>
<tr>
<th>ACTION</th>
<th>Don’t Reject</th>
<th>Reject</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUTH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H True (o)</td>
<td>TN</td>
<td>FP</td>
</tr>
<tr>
<td>H False (x)</td>
<td>FN</td>
<td>TP</td>
</tr>
</tbody>
</table>

Sensitivity = \( \frac{TP}{TP+FN} = \beta \)
Specificity = \( \frac{TN}{TN+FP} = 1 - \alpha \)

\( FP \) = Type I error or ‘error’
\( FN \) = Type II error
\( \alpha = \frac{p-value}{FP rate/error rate/significance level} \)
\( \beta = \text{power} \)

### Inference for Images

<table>
<thead>
<tr>
<th>Source</th>
<th>Signal</th>
<th>Noise</th>
<th>Signal+Noise</th>
</tr>
</thead>
</table>

#### fMRI Multiple Comparisons 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!**

#### Multiple comparison problem

Use of ‘uncorrected’ p-value, \( \alpha=0.1 \)

Percentage of Null Pixels that are False Positives

Using an ‘uncorrected’ p-value of 0.1 will lead us to conclude on average that 10% of voxels are active when they are not.

This is clearly undesirable: multiple comparison problem. To correct for this we can define a null hypothesis for images of statistics.
Multiple Comparisons Problem

- Which of 100,000 voxels are sig.?
  - \( \alpha = 0.05 \Rightarrow 5,000 \) false positive voxels
- Which of (random number, say) 100 clusters significant?
  - \( \alpha = 0.05 \Rightarrow 5 \) false positives clusters

Assessing Statistic Images

Where’s the signal?

<table>
<thead>
<tr>
<th>High Threshold</th>
<th>Med. Threshold</th>
<th>Low Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t &gt; 5.5 )</td>
<td>( t &gt; 3.5 )</td>
<td>( t &gt; 0.5 )</td>
</tr>
</tbody>
</table>

Good Specificity
- Poor Power
  - (risk of false negatives)

Poor Specificity
  - (risk of false positives)
  - Good Power

Multiple comparisons...

- Threshold at \( p \)?
  - expect \( (100 \times p)\% \) by chance
- Surprise?
  - extreme voxel values
    - voxel level inference
  - big suprathreshold clusters
    - cluster level inference
  - many suprathreshold clusters
    - set level inference
- Power & localisation
  - sensitivity
  - spatial specificity

Solutions for Multiple Comparison Problem

- A MCP Solution must control “False Positives”
  - How to measure multiple false positives?
- Familywise Error Rate (FWER)
  - Chance of any false positives
  - Controlled by Bonferroni, Random Field Methods, non-parametric method (SnPM).
- False Discovery Rate (FDR)
  - Proportion of false positives among rejected tests
Contents

- Recap & Introduction
- Inference & multiple comparison
  - Single/multiple voxel inference
  - Family wise error rate (FWER)
    - Bonferroni correction/Random Field Theory
    - Non-parametric approach
  - False Discovery rate (FDR)
- SPM results
- « Take home » message

Family-wise Null Hypothesis

FAMILY-WISE NULL HYPOTHESIS:

- Family of hypotheses
  - $H_k \in \Omega = \{1, \ldots, K\}$
  - $H_\Omega = H_1 \cap H_2 \cdots \cap H_k \cap H_K$

If we reject a voxel null hypothesis at any voxel, we reject the family-wise null hypothesis.

A FP anywhere gives a Family Wise Error (FWE)

$\text{family-wise error rate} = \text{corrected} \ p\text{-value}$

FWE MCP Solutions: Controlling FWE w/ Max

- FWE & distribution of maximum
  - $\text{FWE} = P(\text{FWE})$
  - $= P(T, \geq u | H_1)$
  - $= P_{\max}(T, \geq u | H_1)$

- $100(1-\alpha)%$ile of max dist controls FWE
  - $\text{FWE} = P_{\max}(T, \geq u | H_1) = \alpha$
  - where
    - $u = F^{-1}_{\max}(1-\alpha)$

Multiple comparison problem...

Example: Experiment with 100,000 « voxels » and 40 d.f.

- type I error $\alpha = 0.05$ (5% risk) $\Rightarrow t_{\alpha} = 1.68$
- 100,000 $t$ values $\Rightarrow 5000 t$ values $> 1.68$ just by chance!

Familywise Error I test, $\rho^{\text{FWE}}$:

- find threshold $\xi$ such that, in a family of 100,000 $t$ statistics, only 5% probability of one or more $t$ values above that threshold

Bonferroni correction:

- simple method to find the new threshold

Random field theory:

- more accurate for functional imaging
Given
• a family of N independent voxels and
• a voxel-wise error rate \( v \)

The probability that all tests are below the threshold, i.e. that \( H_0 \) is true :  \( (1 - v)^N \)
The Family-Wise Error rate (FWE) or ‘corrected’ error rate \( \alpha \) is
\[ \alpha = 1 - (1-v)^N \]
\[ \approx Nv \quad \text{(for small } v) \]
Therefore, to ensure a particular FWE we choose
\[ v = \frac{\alpha}{N} \]
A Bonferroni correction is appropriate for independent tests.

Experiment with \( N = 100,000 \) « voxels » and 40 d.f.
– \( v \) = unknown corrected probability threshold,
– find \( v \) such that family-wise error rate \( \alpha = 0.05 \)

Bonferroni correction:
– probability that all tests are below the threshold,
– Use \( v = \frac{\alpha}{N} \)
– Here \( v = 0.05/100000 = 0.0000005 \)
\( \Rightarrow \) threshold \( t = 5.77 \)

Interpretation:
Bonferroni procedure gives a corrected \( p \) value,
i.e. for a \( t \) statistics = 5.77,
– uncorrect \( p \) value = 0.0000005
– corrected \( p \) value = 0.05

Use of ‘uncorrected’ \( p \)-value, \( \alpha = 0.1 \)

Use of ‘corrected’ \( p \)-value, \( \alpha = 0.1 \)

\( v = \frac{\alpha}{n} \) where \( n \) is the number of independent observations.
"Bonferroni" correction & smoothed observations

100 by 100 voxels, with a $z$ value. 10000 independent measures. Fix the $\text{PFWE} = 0.05$, $z$ threshold? Bonferroni?

$\nu = 0.05/10000 = 0.000005 \\
\Rightarrow \text{threshold } z = 4.42$

Random Field Theory

- Consider a statistic image as a lattice representation of a continuous random field
- Use results from continuous random field theory

Lattice representation

Euler Characteristic (EC)

Topological measure
- Threshold an image at $u$
- Excursion set $A_u$
- $\chi(A_u) = \# \text{blobs} - \# \text{holes}$
- At high $u$, $\chi(A_u) = \# \text{blobs}$

Reject $H^0$ if Euler characteristic non-zero
$\alpha = \Pr(\chi(A_u) > 0)$

Expected Euler characteristic
$= \nu$-value
(at high $u$)
$\alpha = \mathbb{E}[\chi(A_u)]$

Euler characteristic (EC) ≈ # blobs in a thresholded image.
(True only for high threshold)

$EC = \text{function of}$
- threshold used
- number of resels

where resels (= resolution elements) ~ number of independent observations
$\Rightarrow \mathbb{E}[EC] \approx \text{PFWE}$
Euler characteristic...

Threshold z-map at 2.50

EC = 3

Threshold z-map at 2.75

EC = 1

Euler characteristic (EC) \approx \# blocks in a thresholded image.
(True only for high threshold)

EC = function of:
• threshold used
• number of resels

where resels (« resolution elements ») \sim number of independent observations

\Rightarrow \mathbb{E}[EC] \approx \rho^{PFWE}

For a threshold \( z_t \)
at 2.50, \( \mathbb{E}[EC] = 1.9 \)
at 2.75, \( \mathbb{E}[EC] = 1.1 \) instead of 3 and 1 as in example

Expected Euler characteristic...

\mathbb{E}[\chi(A_u)] \approx \lambda(\Omega) \sqrt{A_u(u^2 - 1) \exp(-u^2/2) / (2\pi)^2}

- \Omega \rightarrow large search region \Omega \in \mathbb{R}^d
- \lambda(\Omega) \rightarrow volume
- \nu(\Omega) \rightarrow smoothness
- A_u \rightarrow excursion set \ A_u = \{ x \in \mathbb{R}^d : Z(x) > u \}
- Z(\Omega) \rightarrow Gaussian random field \ Z(\mathbb{R}^d)

+ Multivariate normal finite dimensional distributions
+ continuous
+ strictly stationary
+ marginal \( \mathcal{N}(0,1) \)
+ continuously differentiable
+ twice differentiable at 0
+ Gaussian ACF
(at least near local maxima)

Unified Theory

\alpha = \sum R_d(\Omega) \rho_d(u)

\mathbb{R}_d(\Omega), \text{ RESEL count depends on:}
• the search region
  - how big, how smooth, what shape?

\rho_d(u): \text{ EC density depends on:}
• the field \ (\text{eg. Gaussian, t})
• the threshold, u

Worsley et al. (1996), HBM
Unified Theory

- General form for expected Euler characteristic
  - $\chi_f$, $f$, & $\Omega$ fields
  - Restricted search regions
  - $\alpha = \sum R_\Omega(\Omega) \rho_\Omega(u)$

R$^\Omega(\Omega)$, $d$-dimensional Minkowski functional of $\Omega$

- $R_\Omega(\Omega)$: Euler characteristic of $\Omega$
- $R_\Omega(\Omega)$: voxel count
- $R_\Omega(\Omega)$: voxel surface area
- $R_\Omega(\Omega)$: voxel volume

Estimated component fields

- Estimate component fields
- Parameters $\hat{\beta}$
- Errors $\hat{\delta}$
- Each row is an estimated component field
- Estimated variance

Random Field Theory

- RESELS
  - Resolution Elements
  - 1 RESEL = FWHM $\times$ FWHM $\times$ FWHM
  - RESEL Count $R$
    - $R = \lambda \Omega = (4\lambda \log 2)^{3/2} \Omega$
    - Volume of search region in units of smoothness
  - E.g.: 10 voxels, 2.5 FWHM

- Beware RESEL misinterpretation
  - RESEL are not “number of independent ‘things’ in the image”

Gaussian PDF

- $\Sigma$ - kernel var/cov matrix
- ACF: $\Sigma = \Gamma / \Sigma$
  - $\Gamma = \Omega / \Omega$
  - $\Sigma = \Omega / \Omega$
- Full Width at Half Maximum
  - Resel dimensions $\ell_x \times \ell_y \times \ell_z$
  - $R_\Omega(\Omega) = 2\log 2 / (\ell_x \ell_y \ell_z)$
- Point Response Function (PRF)
- Approximate the peak of the covariance function with a Gaussian

Smoothness, PRF, resels...

- Smoothness $\lambda$
  - Variance-covariance matrix of partial derivatives
  - Gaussian PDF
  - Gaussian PRF
  - Resolution Element (RESEL)
  - Voxel dimensions $\ell_x \times \ell_y \times \ell_z$
  - $\Omega = 2\log 2 / (\ell_x \ell_y \ell_z)$
- Approximate the peak of the covariance function with a Gaussian
**RFT Assumptions**

- Model fit & assumptions
  - valid distributional results
- Multivariate normality
  - of component images
- Covariance function of component images must be
  - Can be nonstationary
  - Twice differentiable

**Random Field Intuition**

- Corrected $P$-value for voxel value $t$
  \[ P = P(\max T > t) \approx E(\chi^2) \approx \frac{\lambda(\Omega)}{\Lambda^{1/2}} t^2 \exp(-t^2/2) \]

- Statistic value $t$ increases
  - $P$ decreases (but only for large $t$)
- Search volume increases (bigger $\Omega$)
  - $P$ increases (more severe MCP)
- Smoothness increases (roughness $\Lambda^{1/2}$ decreases)
  - $P$ decreases (less severe MCP)

**Small Volume Correction**

$SVC = \text{correction for multiple comparison in a user's defined volume ‘of interest’}$. Shape and size of volume become important for small or oddly shaped volume!

Example of $SVC$ (900 voxels):
- Compact volume: samples from maximum 16 resels
- Spread volume: sample from up to 36 resels
  \[ \Rightarrow \text{threshold higher for spread volume than compact volume} \]

**Resel Counts for Brain Structures**

<table>
<thead>
<tr>
<th>Structure</th>
<th>$V$ (voxels)</th>
<th>$R_0$ (resels)</th>
<th>$R_1$ (resels)</th>
<th>$R_2$ (resels)</th>
<th>$R_3$ (resels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FWHM=20mm</td>
<td>(1) Threshold depends on Search Volume (2) Surface area makes a large contribution</td>
<td>FWHM=20mm</td>
<td>FWHM=20mm</td>
<td>FWHM=20mm</td>
<td>FWHM=20mm</td>
</tr>
</tbody>
</table>
Summary

• We should correct for multiple comparisons
  – We can use Random Field Theory (RFT) or other methods
RFT requires
  – a good lattice approximation to underlying multivariate Gaussian fields,
  – that these fields are continuous with a twice differentiable correlation function
To a first approximation, RFT is a Bonferroni correction using RESELS.
We only need to correct for the volume of interest.
Depending on nature of signal we can trade-off anatomical specificity for signal sensitivity with the use of cluster-level inference.

Nonparametric Permutation Test

• Parametric methods
  – Assume distribution of statistic under null hypothesis
• Nonparametric methods
  – Use data to find distribution of statistic under null hypothesis
  – Any statistic!

Permutation Test : Toy Example

• Data from V1 voxel in visual stim. experiment
  A: Active, flashing checkerboard   B: Baseline, fixation
  6 blocks, ABABAB     Just consider block averages...
  • Null hypothesis $H_0$
    – No experimental effect, A & B labels arbitrary
  • Statistic
    – Mean difference
Permutation Test: Toy Example

• Under $H_0$
  – Consider all equivalent relabelings
  – Compute all possible statistic values
  – Find 95%ile of permutation distribution

<table>
<thead>
<tr>
<th></th>
<th>AAABBB</th>
<th>ABABAB</th>
<th>BAAABB</th>
<th>BABBA</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.82</td>
<td>9.45</td>
<td>-1.48</td>
<td>-6.86</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-3.25</td>
<td>6.97</td>
<td>1.10</td>
<td>BAAAAB 3.15</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.38</td>
<td>BAAABB -1.38</td>
<td>BBAABA</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-3.15</td>
<td>-1.10</td>
<td>BABAAB -6.97</td>
<td>BBAABA 3.25</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6.86</td>
<td>BBBAAA 1.48</td>
<td>BABABA -9.45</td>
<td>BBAABA -4.82</td>
<td></td>
</tr>
</tbody>
</table>
**Permutation Test: Toy Example**

- **Under $H_0$**
  - Consider all equivalent relabelings
  - Compute all possible statistic values
  - Find 95%ile of permutation distribution

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AABA</td>
<td>4.82</td>
<td>ABAB</td>
<td>9.45</td>
<td>BAAABB</td>
</tr>
<tr>
<td>AABAB</td>
<td>-3.25</td>
<td>ABABA</td>
<td>6.97</td>
<td>BAAAB</td>
</tr>
<tr>
<td>ABAB</td>
<td>-0.67</td>
<td>ABAAB</td>
<td>1.38</td>
<td>ABAAB</td>
</tr>
<tr>
<td>ABBBA</td>
<td>-3.15</td>
<td>ABBBA</td>
<td>-1.10</td>
<td>BABA</td>
</tr>
<tr>
<td>BABBA</td>
<td>6.86</td>
<td>ABBBA</td>
<td>1.48</td>
<td>BABBA</td>
</tr>
<tr>
<td>BBBBB</td>
<td>-4.82</td>
<td>BBBBB</td>
<td></td>
<td>BBBBB</td>
</tr>
</tbody>
</table>

**Controlling FWER: Permutation Test**

- **Parametric methods**
  - Assume distribution of max statistic under null hypothesis
- **Nonparametric methods**
  - Use data to find distribution of max statistic under null hypothesis
  - Again, any max statistic!

**Permutation Test & Exchangeability**

- Exchangeability is fundamental
  - Def: Distribution of the data unperturbed by permutation
  - Under $H_0$, exchangeability justifies permuting data
  - Allows us to build permutation distribution
- Subjects are exchangeable
  - Under $H_0$, each subject’s A/B labels can be flipped
- Are fMRI scans exchangeable under $H_0$?
  - If no signal, can we permute over time?

**Permutation Test & Exchangeability**

- fMRI scans are not exchangeable
  - Permuting disrupts order, temporal autocorrelation
- **Intrasubject fMRI permutation test**
  - Must decorrelate data, model before permuting
  - What is correlation structure?
    - Usually must use parametric model of correlation
    - E.g. Use wavelets to decorrelate
      - Bullmore et al 2001, HBM 12:61-78
- **Intersubject fMRI permutation test**
  - Create difference image for each subject
  - For each permutation, flip sign of some subjects
Permutation Test: Example

- **fMRI Study of Working Memory**
  - 12 subjects, block design  Marshuetz et al (2000)
  - Item Recognition
    - **Active**: View five letters, 2s pause, view probe letter, respond
    - **Baseline**: View XXXXX, 2s pause, view Y or N, respond
- **Second Level RFX**
  - Difference image, A-B constructed for each subject
  - One sample, smoothed variance t test

- **Permute!**
  - $2^{12} = 4,096$ ways to flip 12 A/B labels
  - For each, note maximum of $t$ image

---

**Permutation Distribution**

Thresholded $t$

- Maximum

**Does this Generalize?**

RFT vs Bonf. vs Perm.

<table>
<thead>
<tr>
<th></th>
<th>t Threshold (0.05 Corrected)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>df</td>
</tr>
<tr>
<td>Verbal Fluency</td>
<td>4</td>
</tr>
<tr>
<td>Location Switching</td>
<td>9</td>
</tr>
<tr>
<td>Task Switching</td>
<td>9</td>
</tr>
<tr>
<td>Faces: Main Effect</td>
<td>11</td>
</tr>
<tr>
<td>Faces: Interaction</td>
<td>11</td>
</tr>
<tr>
<td>Item Recognition</td>
<td>11</td>
</tr>
<tr>
<td>Visual Motion</td>
<td>11</td>
</tr>
<tr>
<td>Emotional Pictures</td>
<td>12</td>
</tr>
<tr>
<td>Pain: Warning</td>
<td>22</td>
</tr>
<tr>
<td>Pain: Anticipation</td>
<td>22</td>
</tr>
</tbody>
</table>
### RFT vs Bonf. vs Perm.

<table>
<thead>
<tr>
<th></th>
<th>No. Significant Voxels (0.05 Corrected)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>df</td>
</tr>
<tr>
<td>Verbal Fluency</td>
<td>4</td>
</tr>
<tr>
<td>Location Switching</td>
<td>9</td>
</tr>
<tr>
<td>Task Switching</td>
<td>9</td>
</tr>
<tr>
<td>Faces: Main Effect</td>
<td>11</td>
</tr>
<tr>
<td>Faces: Interaction</td>
<td>11</td>
</tr>
<tr>
<td>Item Recognition</td>
<td>11</td>
</tr>
<tr>
<td>Visual Motion</td>
<td>11</td>
</tr>
<tr>
<td>Emotional Pictures</td>
<td>12</td>
</tr>
<tr>
<td>Pain: Warning</td>
<td>22</td>
</tr>
<tr>
<td>Pain: Anticipation</td>
<td>22</td>
</tr>
</tbody>
</table>

### Performance Summary

- **Bonferroni**
  - Not adaptive to smoothness
  - Not so conservative for low smoothness
- **Random Field**
  - Adaptive
  - Conservative for low smoothness & df
- **Permutation**
  - Adaptive (Exact)

### "Old" Conclusions

- t random field results conservative for
  - Low df & smoothness
  - 9 df & ≤12 voxel FWHM; 19 df & < 10 voxel FWHM
- Bonferroni surprisingly satisfactory for low smoothness
- Nonparametric methods perform well overall

### Contents

- Recap & Introduction
- **Inference & multiple comparison**
  - Single/multiple voxel inference
  - Family wise error rate (FWER)
  - False Discovery rate (FDR)
  - SPM results
- "Take home" message
**False Discovery Rate**

<table>
<thead>
<tr>
<th>TRUTH</th>
<th>ACTION</th>
<th>Don’t Reject</th>
<th>Reject</th>
</tr>
</thead>
<tbody>
<tr>
<td>H True (o)</td>
<td>TN</td>
<td>FP</td>
<td></td>
</tr>
<tr>
<td>H False (x)</td>
<td>FN</td>
<td>TP</td>
<td></td>
</tr>
</tbody>
</table>

\[
FDR = \frac{FP}{FP + TP} \\
\alpha = \frac{FP}{FP + TN}
\]

Eg. t-scores from regions that truly do and do not activate

At \( u1 \)
- \( FDR = \frac{3}{17} = 18\% \)
- \( \alpha = \frac{3}{10} = 30\% \)

At \( u2 \)
- \( FDR = \frac{1}{12} = 8\% \)
- \( \alpha = \frac{1}{10} = 10\% \)

**Illustration:**
- Noise
- Signal
- Signal + Noise

**Benjamini & Hochberg Procedure**

- Select desired limit \( q \) on \( \text{E}(\text{FDR}) \)
- Order \( p \)-values, \( p(1) \leq p(2) \leq \ldots \leq p(V) \)
- Let \( r \) be largest \( i \) such that
  \[
  p(i) \leq \frac{i}{V} \times q
  \]
- Reject all hypotheses corresponding to \( p(1) \), \( \ldots \), \( p(r) \).

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\%

Control of Familywise Error Rate at 10%
- Occurrence of Familywise Error
- \( \text{FWE} \)

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

\( \text{JRSS-B (1995) 57:289-300} \)
**Benjamini & Hochberg: Varying Signal Extent**

- Signal Intensity: 3.0
- Signal Extent: 9.5
- Noise Smoothness: 3.0

$p = 0.001628 \quad z = 2.94$

---

**Benjamini & Hochberg: Varying Signal Extent**

- Signal Intensity: 3.0
- Signal Extent: 16.5
- Noise Smoothness: 3.0

$p = 0.007157 \quad z = 2.45$

---

**Benjamini & Hochberg: Varying Signal Extent**

- Signal Intensity: 3.0
- Signal Extent: 25.0
- Noise Smoothness: 3.0

$p = 0.019274 \quad z = 2.07$

---

**Benjamini & Hochberg: Properties**

- **Adaptive**
  - Larger the signal, the lower the threshold
  - Larger the signal, the more false positives
    - False positives constant as fraction of rejected tests
    - Not a problem with imaging’s sparse signals
- **Smoothness OK**
  - Smoothing introduces positive correlations
**Contents**

- Recap & Introduction
- Inference & multiple comparison
  - Single/multiple voxel inference
  - Family wise error rate (FWER)
  - False Discovery rate (FDR)
  - SPM results
- « Take home » message

**Levels of inference...**

- **voxel-level**
  - $P(c \geq 1 \mid n \geq 0, t \geq 4.37) = 0.048$ (corrected)
  - $P(t \geq 4.37) = 1 - \Phi(4.37) < 0.001$ (uncorrected)

- **set-level**
  - $P(c \geq 1 \mid n \geq 82, t \geq 3.09) = 0.029$ (corrected)
  - $P(n \geq 82 \mid t \geq 3.09) = 0.019$ (uncorrected)

- **omnibus**
  - $P(c \geq 3 \mid n \geq 0, t \geq 3.09) = 0.031$

- **cluster-level**
  - $P(c \geq 1 \mid n \geq 82, t \geq 3.09) = 0.019$ (corrected)
  - $P(t \geq 3.09) = 1 - \Phi(3.09) < 0.001$ (uncorrected)

**Parameters**

- $a = 3.09$
- $k = 12$ voxels
- $s = 32$ voxels
- $FWHM = 4.7$ voxels
- $D = 3$

**SPM results...**
Conclusions: FWER vs FDR

- Must account for multiplicity
  - Otherwise have a “fishing expedition”
- FWER
  - Very specific, less sensitive
- FDR
  - Less specific, more sensitive
  - Trouble with cluster inference...
More Power to Ya!

Statistical Power
- the probability of rejecting the null hypothesis when it is actually false
- "if there’s an effect, how likely are you to find it"?

Effect size
- bigger effects, more power
  - e.g., MT localizer (moving rings - stationary runs) -- 1 run is usually enough
  - looking for activation during imagined motion might require many more runs

Sample size
- larger n, more power
  - more subjects - longer runs - more runs

Signal:Noise Ratio
- better SNR, more power
  - stronger, cleaner magnet - more focal coil - fewer artifacts - more filtering