Statistical Methods in functional MRI

Martin Lindquist
Department of Biostatistics
Johns Hopkins University

Lecture 7: Multiple Comparisons
04/23/13

Data Processing Pipeline

Standard Analysis
1. Fit a statistical model (e.g., the GLM) to a certain voxel in the brain.
2. Use the estimated model parameters to test for an effect of interest, i.e. \( H_0: \beta_1 = \beta_2 = 0 \).
3. Repeat steps 1 and 2 for each voxel.
4. Summarize the results in a statistical image.
5. Determine which voxels show a statistically significant effect, i.e. threshold the image.

Multiple Comparisons Problem

- Choosing an appropriate threshold is complicated by the fact we are 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 \( \alpha \) 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?
  \( \alpha = 0.05 \Rightarrow 5,000 \) 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 $i \in V = \{1, \ldots, m\}$.

• 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.

$$H_0 = \bigcap_{i \in V} H_{0i}$$

• If we reject a **single** voxel null hypothesis, $H_{0i}$, we will reject the family-wise null hypothesis.

• A false positive at any voxel gives a **Family-Wise Error (FWE)**

• Assuming $H_0$ is true, we want the probability of falsely rejecting $H_0$ to be controlled by $\alpha$, i.e.

$$P\left(\bigcup_{i \in V} \{T_i \geq u\} \mid H_0\right) \leq \alpha$$

Bonferroni Correction

• Choose the threshold so that

$$P \left( \frac{C \geq u}{H_0} \geq \frac{\alpha}{m} \right)$$

• Hence,

$$FWER = P\left( \bigcup_{i \in V} \{T_i \geq u\} \mid H_0\right)$$

$$\leq \sum_{i} P \left( C \geq u \mid H_{0i}\right)$$

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 for a FWE of 0.05, the Bonferroni correction is 0.05/10,000. This corresponds to \( u = 4.42 \). On average only 5 out of every 100 generated in this fashion will have one or more values above \( u \). No false positives.

**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.
- It is not optimal for correlated data, and most fMRI data has significant spatial correlation.

**Spatial Correlation**
- We may be able to choose a more appropriate threshold by using information about the spatial correlation in the data.
- Random field theory allows one to incorporate the 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.
  \[
  \text{FWER} = P(\text{FWE}) = P(\bigcup_i \{T_i \geq u \mid H_0\}) = P(\max T_i \geq u \mid H_0) = P(\text{max } t\text{-value exceeds } u \text{ under null})
  \]
  Choose the threshold \( u \) such that the max only exceeds it \( \alpha \% \) of the time.

**Random Field Theory**
- A random field is a set of random variables defined at every point in D-dimensional space.
- A Gaussian random field has a Gaussian distribution at every point and every collection of points.
- A Gaussian random field is defined by its mean function and covariance function.

**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.
Random Field Theory

- Consider a random field $Z(s)$ defined on $s \in \Omega \subseteq R^D$ where $D$ is the dimension of the process.

Euler Characteristic

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

Controlling the FWER

1. Link between FWER and Euler Characteristic.
   
   \[
   \text{FWER} = P(\max \{T_i \geq u \mid H_0\}) = P(\text{One or more blobs} \mid H_0) = P(\chi_u \geq 1 \mid H_0) = E(\chi_u \mid H_0)
   \]

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

3D Gaussian Random Fields

For large search regions:

\[
E(\chi_u) \approx R(4 \log 2)^{\frac{1}{2}} (u^2 - 1) e^{-u^2/2} \Phi(\frac{u}{\sqrt{2}})
\]

where

\[
R = \frac{V}{\text{FWHM} \times \text{FWHM} \times \text{FWHM}}
\]

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.

Where

\[
R = \text{Resolution Element (Resel)}
\]

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 $3-4 \times$ voxel size.
- The amount of smoothness is assumed known.
  - Estimate is biased when images not sufficiently smooth.
- Several layers of approximations.

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 that can be used to categorize it.
  - Height of a peak
  - Number of peaks
  - The volume or extent of an excursion set
- We can perform corrections on several levels.
  - Voxel-level
  - Cluster-level

Voxel-level Inference

- Retain voxels above $\alpha$-level threshold $u_\alpha$

Cluster-level Inference

- Two step-process
  - Define clusters by arbitrary threshold $u_{\alpha \text{clus}}$
  - Retain clusters larger than $\alpha$-level threshold $k_\alpha$

Cluster-level Inference

- Can use RFT to perform cluster level inference.
  - Assume clusters behave like a multidimensional Poisson point process.
  - The probability of getting one or more clusters of volume $k$ is given by:
    \[ p = 1 - \exp\{ -E(\chi_\alpha^2)P(n \geq k) \} \]
  - Here $E(\chi_\alpha^2)$ is the expected Euler Characteristic and $\alpha$ represents the volume of a cluster.

SnPM

- Statistical nonparametric mapping (Nichols & Holmes) is a nonparametric equivalent to SPM.
- It uses a permutation test, rather than random field theory, to correct for multiple comparisons.
- It allows one to avoid the assumptions needed for using random field theory.

Illustration

- Data from V1 voxel in visual stimulus experiment
  - A: Active, flashing checkerboard
  - B: Baseline, fixation
  - 6 blocks, ABABAB
  - Just consider block averages.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>A</th>
<th>B</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>103.00</td>
<td>90.48</td>
<td>99.93</td>
<td>87.83</td>
<td>99.76</td>
<td>96.06</td>
</tr>
</tbody>
</table>

- Null hypothesis $H_0$
  - No experimental effect, i.e. the A and B labels are arbitrary.
- Statistic
  - Mean difference between conditions.
**Under $H_0$**
- Consider all equivalent re-labelings.
  - Assume exchangeability, i.e. the distribution of the statistic is the same whatever the relabeling.
- Compute all possible statistic values.
- Determine the permutation distribution.
  - Each relabeling is equally likely. Hence, each statistic has equal probability.

<table>
<thead>
<tr>
<th>Label</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AABBB</td>
<td>4.82</td>
</tr>
<tr>
<td>BABAB</td>
<td>9.45</td>
</tr>
<tr>
<td>AABAB</td>
<td>-3.25</td>
</tr>
<tr>
<td>ABBBA</td>
<td>-0.67</td>
</tr>
<tr>
<td>ABBBA</td>
<td>-3.18</td>
</tr>
<tr>
<td>AABAB</td>
<td>6.86</td>
</tr>
<tr>
<td>BABBB</td>
<td>1.48</td>
</tr>
<tr>
<td>BABA</td>
<td>4.82</td>
</tr>
<tr>
<td>BABB</td>
<td>9.45</td>
</tr>
<tr>
<td>BAAB</td>
<td>3.25</td>
</tr>
<tr>
<td>BAABA</td>
<td>1.38</td>
</tr>
<tr>
<td>BAAAB</td>
<td>3.15</td>
</tr>
<tr>
<td>BAAAB</td>
<td>1.10</td>
</tr>
<tr>
<td>BAABA</td>
<td>-6.97</td>
</tr>
<tr>
<td>BAABA</td>
<td>0.67</td>
</tr>
<tr>
<td>BAAAB</td>
<td>3.25</td>
</tr>
<tr>
<td>BBAAA</td>
<td>-4.82</td>
</tr>
</tbody>
</table>

**Permutation distribution**

- Actual value = 9.45
- Permutation test don’t work very well with small sample sizes, as they tend to be conservative.

**Comments**
- Requires only the assumption of exchangeability
  - Under $H_0$, the distribution is unchanged by permutation.
  - This allows us to build the permutation distribution.
- Subjects are exchangeable
  - Under $H_0$, each subject’s A/B labels can be flipped.
  - Permutation tests useful for 2nd level analysis.
- fMRI scans not exchangeable under $H_0$
  - The problem is temporal autocorrelation.
  - Be careful performing permutation tests on individual subjects.

**Issues**
- Sample size
  - If there are $N$ possible relabelings, the smallest attainable $p$-value is $1/N$ which can be problematic at small $N$.
  - If there are too many possible relabelings it may not be feasible to compute the statistic images for all of them.
    - Randomly sample from the population of relabelings.
    - Each relabeling should be equally likely to be chosen.
- Flexibility
  - The permutation approach is free to consider any statistic and is not bound to those that have a known distributional form.

**Multiple Comparisons**
- Permutation tests can be used to control for multiple comparisons.
  - To do so all voxels need to be considered simultaneously.
    - Arguments can be extended to image-level inference by considering an appropriate max statistic.
    - Permutations carried out on the image level, i.e. entire images are relabeled.
Max statistic

By repeatedly permuting the image labels and computing the max statistic we obtain the permutation distribution needed for correcting the FWER.

Controlling FWER

• A single threshold test thresholds the statistic image at a critical value and voxels exceeding the threshold are deemed active.

• Procedure:
  – Re-label the images.
  – Compute the statistic for each voxel.
  – Compute the maximum statistic over the whole image.
  – Repeat to construct the permutation distribution of the max statistic.

Controlling FWER

• A suprathreshold cluster test thresholds the statistic image at a primary threshold, and then studies the extent of activation in a second level.

• Procedure:
  – Re-label the images.
  – Compute the statistic for each voxel.
  – Threshold at a primary threshold.
  – Compute the size of the largest cluster above the threshold.
  – Repeat to construct the permutation distribution.

Power

• Bonferroni and SPM’s GRF correction do not account accurately for spatial smoothness with the sample sizes and smoothness values typical in imaging experiments.

• Nonparametric tests more accurate and more sensitive.

False Discovery Rate

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

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

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.
Suppose we perform tests on m voxels.

\[ \begin{array}{ccc}
\text{Declared} & \text{Inactive} & \text{Active} \\
\text{Truly inactive} & U & V \\
\text{Truly active} & T & S \\
m-R & m & m-m_0 \\
\end{array} \]

U, V, T and S are unobservable random variables. R is an observable random variable.

In this notation:

- **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.
- An FDR-controlling technique guarantee controls of the FDR in the sense that FDR ≤ q.

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 \[ p(i) \leq \frac{i \cdot m \times q}{m} \]
4. Reject all hypotheses corresponding to \( p(1), \ldots, p(r) \).

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

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.

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
METHODS

Subjects: 30 adult Atlantic salmon (Salmo salar) participated in the study.

- The salmon were approximately 18 months old, weighed 1.5 kg, and were not active at the time of scanning.

Tasks: The task consisted of two conditions: an operant conditioning task. The tasks involved the presentation of a series of colored stimuli, which were associated with food rewards. The salmon were trained in operant chambers, which were housed in the filming area of the study.

DISCUSSION

Can we conclude from this data that the salmon is engaging in the perspective-taking task? Certainly not. What we can determine is that animals have the potential to engage in perspective-taking behavior, as evidenced by their ability to discriminate between different colored stimuli. However, this ability may be limited by factors such as the complexity of the task, the training regimen, and the reward system.

We suggest that using a reward-based training protocol may improve the animals' ability to engage in perspective-taking behavior. Additionally, the use of operant conditioning tasks may be useful in studying the cognitive abilities of salmon and other species.

In conclusion, our results indicate that salmon are capable of engaging in perspective-taking behavior, but further research is needed to fully understand the mechanisms underlying this behavior.