Inference
how surprising is your statistic? (thresholding)

But ... can I trust it?
Outline

• Null-hypothesis and Null-distribution
• Multiple comparisons and Family-wise error
• Gaussian Random Field theory
  • Voxel-wise tests
  • Cluster-based inference
• Enhanced Clusters
• Randomise, slow and reliable
• FDR - False Discovery Rate
• Extra: examples of different inferences
Outline

• Null-hypothesis and Null-distribution
• Multiple comparisons and Family-wise error
• Gaussian Random Field theory
  • Voxel-wise tests
  • Cluster-based inference
• Enhanced Clusters
• Randomise, slow and reliable
• FDR - False Discovery Rate
• Extra: examples of different inferences
The task of classical inference

- Given some data we want to know if (e.g.) a mean is different from zero or if two means are different.

> 0 ?  
Different?
Tools of classical inference

1. A null-hypothesis

Typically the opposite of what we actually “hope”, e.g.

There is **no** effect of treatment: $\mu = 0$

There is **no** difference between groups: $\mu_1 = \mu_2$
Tools of classical inference

1. A null-hypothesis
2. A test-statistic

Assesses “trustworthiness”

Trustworthy

Dodgy
Tools of classical inference

1. A null-hypothesis
2. A test-statistic

A test-statistic reflects precisely this:

A null-hypothesis

Assesses “trustworthiness”

Many measurements: 
Trustworthy

A t-statistic reflects precisely this:

Large difference: 
Trustworthy

Small variability: 
Trustworthy
Tools of classical inference

1. A null-hypothesis

2. A test-statistic

Or expressed in GLM lingo

\[
\begin{align*}
\hat{\beta}_1 & = \frac{\bar{x}_1}{\bar{x}_2} \\
\hat{\beta}_2 & = \frac{\bar{x}_1}{\bar{x}_2}
\end{align*}
\]

Large difference: Trustworthy

\[
t = \frac{c^T \hat{\beta}}{\sqrt{\sigma^2} \sqrt{c^T (X^T X)^{-1} c}}
\]

Small variability: Trustworthy

Many measurements: Trustworthy

\[
\bar{x}_1 - \bar{x}_2
\]
Tools of classical inference

1. A null-hypothesis
2. A test-statistic
3. A null-distribution

Let us assume there is no difference, i.e. the null-hypothesis is true.

We might then get these data

\[ c^T \hat{\beta} = 1.17 \]

\[ c^T \hat{\beta} \]

\[ t = 2.19 \]

\[ t = \frac{c^T \hat{\beta}}{\sqrt{\sigma^2 \sqrt{c^T (X^T X)^{-1} c}}} \]

\[ \sigma^2 = 0.71 \]

Constant
Tools of classical inference

1. A null-hypothesis
2. A test-statistic
3. A null-distribution

We might then get these data

$$c^T \hat{\beta} = 1.17$$

$$t = \frac{c^T \hat{\beta}}{\sqrt{\sigma^2} \sqrt{c^T (X^T X)^{-1} c}}$$

$$\sigma^2 = 0.71$$

$$t = 2.19$$
Tools of classical inference

1. A null-hypothesis
2. A test-statistic
3. A null-distribution

\[ t = \frac{c^T \hat{\beta}}{\sqrt{\sigma^2} \sqrt{c^T (X^T X)^{-1} c}} \]

or we could have gotten these

\[ \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + e \]

\[ t = -0.51 \]

\[ c^T \hat{\beta} = -0.37 \]

\[ \sigma^2 = 1.28 \]
Tools of classical inference

1. A null-hypothesis
2. A test-statistic
3. A null-distribution

\[ t = \frac{c^T \hat{\beta}}{\sqrt{\sigma^2} \sqrt{c^T (X^T X)^{-1} c}} \]

\[ \sigma^2 = 1.01 \]

\[ c^T \hat{\beta} = 0.31 \]

\[ c^T \hat{\beta} = 0.49 \]

\[ \beta_1 \quad \beta_2 \]

\[ = \]

\[ + e \]
Tools of classical inference

1. A null-hypothesis
2. A test-statistic
3. A null-distribution

\[ t = \frac{c^T \hat{\beta}}{\sqrt{\sigma^2 \sqrt{c^T (X^T X)^{-1} c}}} \]

or perhaps these

\[ c^T \hat{\beta} = 1.22 \]

\[ \sigma^2 = 0.78 \]
Tools of classical inference

1. A null-hypothesis
2. A test-statistic
3. A null-distribution

\[ t = \frac{c^T \hat{\beta}}{\sqrt{\sigma^2} \sqrt{c^T (X^T X)^{-1} c}} \]

Constant

\[ \sigma^2 = 0.44 \]

\[ c^T \hat{\beta} = -0.69 \]
Tools of classical inference

1. A null-hypothesis
2. A test-statistic
3. A null-distribution

And if we do this til the cows come home
Tools of classical inference

1. A null-hypothesis
2. A test-statistic
3. A null-distribution

So, why is this helpful?
Tools of classical inference

1. A null-hypothesis
2. A test-statistic
3. A null-distribution

Well, it for example tells us that in \( \sim 1\% \) of the cases \( t > 3.00 \), even when the null-hypothesis is true.
Tools of classical inference

1. A null-hypothesis
2. A test-statistic
3. A null-distribution

Or that in ~5% of the cases $t > 1.99$.

When the null-hypothesis is true.
Tools of classical inference

1. A null-hypothesis
2. A test-statistic
3. A null-distribution

And best of all: This distribution is known i.e. one can calculate it. Much as one can calculate sine or cosine
Tools of classical inference

1. A null-hypothesis
2. A test-statistic
3. A null-distribution

And best of all: This distribution is known i.e. one can calculate it. Much as one can calculate sine or cosine

Provided that $e \sim N(0, \sigma^2)$
An example experiment

1. A null-hypothesis
   \( H_0: \bar{x}_1 = \bar{x}_2 \), \( H_1: \bar{x}_1 > \bar{x}_2 \)

2. A test-statistic

3. A null-distribution

So, with these tools let us do an experiment
An example experiment

1. A null-hypothesis
   \[ H_0: \bar{x}_1 = \bar{x}_2, \quad H_1: \bar{x}_1 > \bar{x}_2 \]

2. A test-statistic
   \[ t_8 = 2.64 \]

3. A null-distribution

So, with these tools let us do an experiment

\[
\begin{pmatrix}
\beta_1 \\
\beta_2
\end{pmatrix}
= 
\begin{pmatrix}
1 \\
2
\end{pmatrix} \cdot
\frac{c^T \hat{\beta}}{\sqrt{\sigma^2} \sqrt{c^T (X^T X)^{-1} c}}
= 
\frac{1.53}{\sqrt{0.85 \cdot 0.4}}
= 2.64
\]
An example experiment

1. A null-hypothesis
2. A test-statistic
3. A null-distribution

So, with these tools let us do an experiment

\[ H_0: \bar{x}_1 = \bar{x}_2, \ H_1: \bar{x}_1 > \bar{x}_2 \]

\[ t_8 = 2.64 \]

If the null-hypothesis is true, we would expect to have a \( \sim 1.46\% \) chance of finding a t-value this large or larger.
An example experiment

1. A null-hypothesis
2. A test-statistic
3. A null-distribution

So, with these tools let us do an experiment

$$H_0: \bar{x}_1 = \bar{x}_2 \ , \ H_1: \bar{x}_1 > \bar{x}_2$$

$$t_8 = 2.64$$

$$t_8 = 2.64^*$$

There is \( \sim 1.46\% \) risk that we reject the null-hypothesis (i.e. claim we found something) when the null is actually true. We can live with that (well, I can).
False positives/negatives

• I am sure you have all heard about “false positives” and “false negatives”.
• But what does that actually mean?
False positives/negatives

• I am sure you have all heard about “false positives” and “false negatives”.
• But what does that actually mean?
• We want to perform an experiment and as part of that we define a null-hypothesis, e.g. $H_0 : \mu = 0$
• Now what can happen?
False positives/negatives

- I am sure you have all heard about “false positives” and “false negatives”.
- But what does that actually mean?
- We want to perform an experiment and as part of that we define a null-hypothesis, e.g. \( H_0 : \mu = 0 \)
- Now what can happen?

\[
\begin{align*}
H_0 \text{ is true} & \quad \text{True state of affairs} \\
H_0 \text{ is false} & \\
\end{align*}
\]
False positives/negatives

• I am sure you have all heard about “false positives” and “false negatives”.
• But what does that actually mean?
• We want to perform an experiment and as part of that we define a null-hypothesis, e.g. \( H_0 : \mu = 0 \)
• Now what can happen?

\[
\begin{align*}
\text{H}_0 \text{ is true} & \quad \text{We don’t reject } \text{H}_0 \\
\text{H}_0 \text{ is false} & \quad \text{We reject } \text{H}_0
\end{align*}
\]

\{ True state of affairs \} \{ Our decision \}
**False positives/negatives**

\[
\begin{align*}
H_0 \text{ is true} & \quad \{ \text{True state of affairs} \\
H_0 \text{ is false} & \quad \{ \\
\text{We don’t reject } H_0 & \\
\text{We reject } H_0 & \quad \{ \text{Our decision} \\
\end{align*}
\]

<table>
<thead>
<tr>
<th>( H_0 \text{ is true} )</th>
<th>( H_0 \text{ is false} )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>We don’t reject } H_0</strong></td>
<td><strong>We reject } H_0</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>We don’t reject } H_0</th>
<th>We reject } H_0</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 \text{ is true} )</td>
<td>( H_0 \text{ is false} )</td>
</tr>
</tbody>
</table>
### False positives/negatives

**True state of affairs**

- $H_0$ is true
- $H_0$ is false

**Our decision**

- We don’t reject $H_0$
- We reject $H_0$

<table>
<thead>
<tr>
<th>$H_0$ is true</th>
<th>We don’t reject $H_0$</th>
<th>We reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>😊</td>
<td></td>
</tr>
<tr>
<td>$H_0$ is false</td>
<td></td>
<td>😊</td>
</tr>
</tbody>
</table>
False positives/negatives

\( H_0 \) is true \( \{ \) True state of affairs
\( H_0 \) is false \( \} \)

We don’t reject \( H_0 \) \( \} \) Our decision
We reject \( H_0 \) \( \} \)

<table>
<thead>
<tr>
<th>( H_0 ) is true</th>
<th>We don’t reject ( H_0 )</th>
<th>We reject ( H_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>False positive</td>
<td></td>
<td>False negative</td>
</tr>
<tr>
<td>False negative</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### False positives/negatives

<table>
<thead>
<tr>
<th></th>
<th>We don’t reject $H_0$</th>
<th>We reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ is true</td>
<td>False positive</td>
<td>False negative</td>
</tr>
<tr>
<td></td>
<td>Type I error</td>
<td>Type II error</td>
</tr>
</tbody>
</table>

- **True state of affairs**:
  - $H_0$ is true
  - $H_0$ is false

- **Our decision**:
  - We don’t reject $H_0$
  - We reject $H_0$
Outline

- Null-hypothesis and Null-distribution
- Multiple comparisons and Family-wise error
- Gaussian Random Field theory
  - Voxel-wise tests
  - Cluster-based inference
- Enhanced Clusters
- Randomise, slow and reliable
- FDR - False Discovery Rate
- Extra: examples of different inferences
Multiple Comparisons

• In neuroimaging we typically perform many tests as part of a study
What happens when we apply this to imaging data?

Z-map where each voxel $\sim N$. Null-hypothesis true everywhere, i.e. NO ACTIVATIONS

16 clusters
288 voxels
$\sim 5.5\%$ of the voxels

That’s a LOT of false positives
Italians doing maths: The Bonferroni correction

Bonferroni says threshold at $\alpha$ divided by # of tests

5255 voxels
$0.05/5255 \approx 10^{-5}$

z-map thresholded at 5.65

No false positives. Hurrah for Italy!
But ... doesn’t 5.65 sound very high?

Largest observed value

Bonferroni threshold

Observed values in the z-map

Too lenient

Too harsh

0.05
1.64
10^{-5}
5.65

So what do we want then?
Let’s say we perform a series of identical studies. Each z-map is the end result of a study.

Let us further say that the null-hypothesis is true. We want to threshold the data so that only once in 20 studies do we find a voxel above this threshold. But how do we find such a threshold?
Outline

• Null-hypothesis and Null-distribution
• Multiple comparisons and Family-wise error

• Gaussian Random Field theory
  • Voxel-wise tests
  • Cluster-based inference
• Enhanced Clusters
• Randomise, slow and reliable
• FDR - False Discovery Rate
• Extra: examples of different inferences
Maximum \( z \)

- When we want to control “family-wise error”, what do we in practice want?
  - If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
  - And if we reject anything, we will definitely reject the most “extreme” value (\( \max(z) \)) in the brain.

\[
\max(z) = 5.16
\]
Maximum $z$

- When we want to control “family-wise error”, what do we in practice want?
- If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
- And if we reject anything, we will definitely reject the most “extreme” value in the brain.

$\text{max}(z) = 6.84$
Maximum $z$

- When we want to control “family-wise error”, what do we in practice want?
- If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
- And if we reject anything, we will definitely reject the most “extreme” value in the brain.

$\max(z) = 5.93$
Maximum z

- When we want to control “family-wise error”, what do we in practice want?
  - If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
  - And if we reject anything, we will definitely reject the most “extreme” value in the brain.

\[
\text{max}(z) = 4.62
\]
Maximum $z$

- When we want to control “family-wise error”, what do we in practice want?
- If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
- And if we reject anything, we will definitely reject the most “extreme” value in the brain.
Maximum z

- When we want to control “family-wise error”, what do we in practice want?
- If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
- And if we reject anything, we will definitely reject the most “extreme” value in the brain.

Etc…
Maximum \( z \)

- When we want to control “family-wise error”, what do we in practice want?
- If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
- And if we reject anything, we will definitely reject the most “extreme” value in the brain.

This is the distribution we want to use for our FWE control.
Maximum $z$

- When we want to control “family-wise error”, what do we in practice want?
- If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
- And if we reject anything, we will definitely reject the most “extreme” value in the brain.

This is the distribution we want to use for our FWE control.
But there is no known expression for it! 😞
But let’s pretend, and see if we can at least get a feel for how it behaves.

I get this Max-z distribution

5255 voxels, FWHM 3mm
Maximum $t$

Let us see if we can at least get a feel for how it behaves.

Then what would the distribution look like for this experiment?

Think about it for a second.

114 voxels, FWHM 3mm
Maximum $z$

Let us see if we can at least get a feel for how it behaves.

114 voxels, FWHM 3mm

What implications would this have for our threshold?
Let us see if we can at least get a feel for how it behaves.

What about this experiment then?

Think about it for a second.

5255 voxels, FWHM 6mm
Maximum $z$

Let us see if we can at least get a feel for how it behaves.

5255 voxels, FWHM 6mm

And what are the implications for our threshold?
Maximum $z$

So, the Maximum t-distribution, and hence the threshold for Family Wise Error control depends on:

- # of voxels
- Smoothness
The “RESEL” combines these two factors?

- RESEL stands for Resolution Element

- It is a bunch of voxels equal to the same size as the FWHM of the smoothness of the image

See [http://imaging.mrc-cbu.cam.ac.uk/imaging/PrinciplesRandomFields](http://imaging.mrc-cbu.cam.ac.uk/imaging/PrinciplesRandomFields) for a nice tutorial
The “RESEL” combines these two factors?

- RESEL stands for RESolution Element

- It is a bunch of voxels equal to the same size as the FWHM of the smoothness of the image

- Number of RESELS is similar to, but NOT equal to, the number of independent observations in an image

See [http://imaging.mrc-cbu.cam.ac.uk/imaging/PrinciplesRandomFields](http://imaging.mrc-cbu.cam.ac.uk/imaging/PrinciplesRandomFields) for a nice tutorial
The “RESEL” combines these two factors?

So, we know there is some function of the number of Resels, $R$, that describes the Max-z distribution.

But it doesn’t help us because we don’t know how to calculate $f(R)$.

😊
The “RESEL” combines these two factors?

So, we know there is some function of the number of Resels, $R$, that describes the Max-z distribution:

$$f(R)$$

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

$$EC(R) \Rightarrow u(R)$$

This approximation is derived from Gaussian Random Field (GRF) Theory.

see [http://imaging.mrc-cbu.cam.ac.uk/imaging/PrinciplesRandomFields](http://imaging.mrc-cbu.cam.ac.uk/imaging/PrinciplesRandomFields) for a nice tutorial.
GRF for voxel-wise tests

• Needs a null-hypothesis and a test-statistic.
• Calculates a (z) threshold based on the the # of RESELS in the volume.

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

- Hinges on strict assumptions about the distribution of the data (~N(0,Λ))
- Is too conservative when df is small
- Is too lenient when volume is small (SVC)
Outline

• Null-hypothesis and Null-distribution

• Multiple comparisons and Family-wise error

• Gaussian Random Field theory
  • Voxel-wise tests
  • Cluster-based inference

• Enhanced Clusters

• Randomise, slow and reliable

• FDR - False Discovery Rate

• Extra: examples of different inferences
Spatial extent: another way to be surprised

This far we have talked about voxel-based tests

We say: Look! A z-value of 7. That is so surprising (under the null-hypothesis) that I will have to reject it. (Though we are of course secretly delighted to do so)
Spatial extent: another way to be surprised

But sometimes our data just aren’t that surprising.

Nothing surprising here! The largest $z$-value is $\sim 4$. We cannot reject the null-hypothesis, and we are **devastated**.
Spatial extent: another way to be surprised

So we threshold the z-map at 2.3 (arbitrary threshold) and look at the spatial extent of clusters.

We say: Look at that whopper! 301 connected voxels all with z-values > 2.3. That is really surprising (under the null-hypothesis). I will have to reject it.
As with the z-values we need a “null-distribution”. What would that look like in this case?

Let’s say we have acquired some data
Distribution of Max Cluster Size

If we reject any cluster we will reject the largest. So what we want is the distribution of the largest cluster, under the null-hypothesis.

Threshold the z-map at 2.3 (arbitrary)
If we reject any cluster we will reject the largest. So what we want is the distribution of the largest cluster, under the null-hypothesis.

Locate the largest cluster anywhere in the brain.
If we reject any cluster we will reject the largest. So what we want is the distribution of the largest cluster, under the null-hypothesis.

And record how large it is.
If we reject any cluster we will reject the largest. So what we want is the distribution of the largest cluster, under the null-hypothesis.

And do the same for another experiment...
If we reject any cluster we will reject the largest. So what we want is the distribution of the largest cluster, under the null-hypothesis.
Distribution of Max Cluster Size

If we reject any cluster we will reject the largest. So what we want is the distribution of the largest cluster, under the null-hypothesis.

Until we have...
If we reject any cluster we will reject the largest. So what we want is the distribution of the largest cluster, under the null-hypothesis. If we find a cluster larger than 76 voxels we reject the null-hypothesis. And this (76) is the level we want to threshold at.
Distribution of Max Cluster Size

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

But what else does $f(R)$ depend on?

![Graph showing the distribution of Max Cluster Size](image)
Distribution of Max Cluster Size

So, just as was the case for the $z$-values, we now have a distribution $f(R)$ that allows us to calculate a Family Wise threshold $u(R)$ pertaining to cluster size. $f(R)$ depends crucially on the initial “cluster-forming” threshold?

$z = 2.3$
Distribution of Max Cluster Size

So, just as was the case for the z-values, we now have a distribution \( f(R) \) that allows us to calculate a Family Wise threshold \( u(R) \) pertaining to cluster size.

\( f(R) \) depends crucially on the initial “cluster-forming” threshold?  

\( u = 76 \)

\( z = 2.3 \)
Distribution of Max Cluster Size

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

$f(R)$ depends crucially on the initial “cluster-forming” threshold?

$u = 49$

$z = 2.7$
Distribution of Max Cluster Size

So, just as was the case for the $z$-values, we now have a distribution $f(R)$ that allows us to calculate a Family Wise threshold $u(R)$ pertaining to cluster size. $f(R)$ depends crucially on the initial “cluster-forming” threshold?

$u = 25$

$z = 3.1$
Distribution of Max Cluster Size

Hence the distribution for the cluster size should really be written $f(R, z)$ and the same for $u(R, z)$.

$z = 3.1$

$u = 25$

$z = 2.7$

$z = 2.3$

$u = 49$

$u = 76$

And as before these distributions are approximations based on Gaussian Random Field Theory.
Pick the right threshold

Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund\textsuperscript{a,b,c,1}, Thomas E. Nichols\textsuperscript{d,e}, and Hans Knutsson\textsuperscript{a,c}

\textsuperscript{a}Division of Medical Informatics, Department of Biomedical Engineering, Linköping University, S-581 85 Linköping, Sweden; \textsuperscript{b}Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University, S-581 83 Linköping, Sweden; \textsuperscript{c}Center for Medical Image Science and Visualization, Linköping University, S-581 83 Linköping, Sweden; \textsuperscript{d}Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom; and \textsuperscript{e}WMG, University of Warwick, Coventry CV4 7AL, United Kingdom

Edited by Emery N. Brown, Massachusetts General Hospital, Boston, MA, and approved May 17, 2016 (received for review February 12, 2016)

\begin{equation}
z = 2.3
\end{equation}

\begin{equation}
z = 3.1
\end{equation}

Always use an initial cluster-forming threshold of \textbf{3.1} or higher!
Instead of resel-based correction, we can do clustering:

Threshold at (arbitrary!) z level
Instead of resel-based correction, we can do clustering:

- **Threshold at (arbitrary!) z level**
  - Form clusters from surviving voxels.
  - Calculate the size threshold $u(R,z)$.
  - Any cluster larger than $u$ “survives” and we reject the null-hypothesis for that.
How do we choose the (arbitrary!) z-threshold?

This is arbitrary and a trade-off
How do we choose the (arbitrary!) z-threshold?

This is arbitrary and a trade-off

1. **Low threshold** - can violate RFT assumptions, but can detect clusters with large spatial extent and low z
How do we choose the (arbitrary!) z-threshold?

This is arbitrary and a trade-off

1. **Low threshold** - can violate RFT assumptions, but can detect clusters with large spatial extent and low z

2. **High threshold** - gives more power to clusters with small spatial extent and high z
How do we choose the (arbitrary!) z-threshold?

This is arbitrary and a trade-off

1. **Low threshold** - can violate RFT assumptions, but can detect clusters with large spatial extent and low $z$

2. **High threshold** - gives more power to clusters with small spatial extent and high $z$

Tends to be more sensitive than voxel-wise corrected testing

Results depend on extent of spatial smoothing in pre-processing
GRF for cluster-wise tests

- 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

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

- Hinges on strict assumptions about the distribution of the data ($\sim N(0,\Lambda)$)
- Inference pertains to entire cluster
- Initial threshold is arbitrary
Outline

• Null-hypothesis and Null-distribution
• Multiple comparisons and Family-wise error
• Gaussian Random Field theory
  • Voxel-wise tests
  • Cluster-based inference
• Enhanced Clusters
• Randomise, slow and reliable
• FDR - False Discovery Rate
• Extra: examples of different inferences
TFCE
Threshold-Free Cluster Enhancement
[Smith & Nichols, NeuroImage 2009]

• Cluster thresholding:
  • popular because it’s sensitive, due to its use of spatial extent
  • but the pre-smoothing extent is arbitrary
  • and so is the cluster-forming threshold
    ➡ unstable and arbitrary

• TFCE
  • integrates cluster “scores” over all possible thresholds
  • output at each voxel is measure of local cluster-like support
  • similar sensitivity to optimal cluster-thresholding, but stable and non-arbitrary

The TFCE value at point p is given by the sum, over the shaded area, of the score from each contributing incremental section:

$$\text{TFCE}(p) = \sum_h e(h)^e \cdot h^H$$
Qualitative example

Original signal
TFCE enhancement
TFCE for FSL-VBM

- TFCE cluster-based (red)
- TFCE voxel-based (blue)

Z=22
Z=48
Y=-16

p (corrected)

0.003
0.05
TFCE for TBSS

controls > schizophrenics

$p<0.05$ corrected for multiple comparisons across space, using randomise

cluster-based: cluster-forming threshold = 2 or 3

TFCE
Outline

• Null-hypothesis and Null-distribution
• Multiple comparisons and Family-wise error
• Gaussian Random Field theory
  • Voxel-wise tests
  • Cluster-based inference
• Enhanced Clusters
  • Randomise, slow and reliable
• FDR - False Discovery Rate
• Extra: examples of different inferences
Randomise, slow and reliable.

Impressive as the GRF based thresholding is, there are situations where we can’t use it.

We may want to use a test-statistic for which the distribution is unknown. **Example:** The “TFC enhanced” $t$-statistic. (A highly non-linear spatial filter)

Our data may not be normally distributed. Then our $t$-values will, paradoxically, not be $t$-distributed. *(what are the chances...)* **Example:** VBM-style data (data whose value is probability of a certain tissue-type)

We want to restrict our analysis to a particular (small and irregularly shaped) sulcus to increase our sensitivity when we have a prior spatial hypothesis.
Example: VBM-style analysis

- Our data is segmented grey matter maps
- A voxel is either grey matter, or not.

\[
\begin{bmatrix}
\beta_1 \\
\beta_2
\end{bmatrix}
= \begin{bmatrix}
0.4 \\
0.6
\end{bmatrix}
\]

\(~ N? ~\)
Oh dear! What now?

- We could use Monte-Carlo to simulate the distributions. As I did for these slides.
  - But, hinges on lots of assumptions about the data
- We could permute the data itself.

We have performed an experiment

And calculated a statistic, e.g. a $t$-value

\[ t = 2.27 \]

If the null-hypothesis is true, there is no difference between the groups. That means we should be able to “re-label” the individual points without changing anything.
We could Monte-Carlo simulate the distributions. As I did for these slides.

- But, hinges on lots of assumptions about the data

We could permute the data itself.

Oh dear! What now?

One re-labelling

$t = 0.67$

Let’s start collecting them
Oh dear! What now?

- We could Monte-Carlo simulate the distributions. As I did for these slides.
  - But, hinges on lots of assumptions about the data
- We could permute the data itself.

Second re-labelling

$t = 1.97$

And another one
We could Monte-Carlo simulate the distributions. As I did for these slides.

- But, hinges on lots of assumptions about the data

We could permute the data itself.

Of the 5000 re-labellings, only 90 had a $t$-value $> 2.27$ (the original labelling).

I.e. there is only a $\sim 1.8\%$ ($90/5000$) chance of obtaining a value $> 2.27$ if there is no difference between the groups

C.f. $p(x \geq 2.27) = 1.79\%$ for $t_{18}$

5000 re-labellings. Phew!
And now we can do the wacky stats.

We compared activation by painful stimuli in two groups of 5 subjects each.

This is what we got

Very intriguing activation. $t_8 = 4.65$

Prof. ran to write to Science. But, did she jump the gun?
And now we can do the wacky stats.

We compared activation by painful stimuli in two groups of 5 subjects each.

This is what we got:

Very intriguing activation. $t_8 = 4.65$

Prof. ran to write to Science. But, did she jump the gun?

Group 1

Group 2

2nd level model

Our group difference map

max($t$) = 4.65
And now we can do the wacky stats.

We compared activation by painful stimuli in two groups of 5 subjects each.

This is what we got

Very intriguing activation. $t_8 = 4.65$

Prof. ran to write to Science. But, did she jump the gun?

max($t$) = 8.23
And now we can do the wacky stats.

We compared activation by painful stimuli in two groups of 5 subjects each.

This is what we got

Very intriguing activation. $t_8 = 4.65$

Prof. ran to write to Science. But, did she jump the gun?

Group 1

Group 2

2nd Permutation

2nd permuted map

max($t$) = 5.43
And now we can do the wacky stats.

We compared activation by painful stimuli in two groups of 5 subjects each.

This is what we got

Very intriguing activation. $t_8 = 4.65$

Prof. ran to write to Science. But, did she jump the gun?

max($t$) = 5.84
And now we can do the wacky stats.

We compared activation by painful stimuli in two groups of 5 subjects each.

This is what we got

Original labelling

Very intriguing activation. $t_8 = 4.65$

Prof. ran to write to Science. But, did she jump the gun?

3925 permutations yielded higher max(t)-value than original labelling.

We cannot reject the null-hypothesis.
Permutations for dummies

- Needs a null-hypothesis and a test-statistic.

  + Builds its own “null-distribution” from your single data-set.
  + No assumptions about the data.
  + Can use any test-statistic, e.g. $\max(t)$, $\max(n_{\text{cluster}})$ etc.
  + Can use “classical” statistics (e.g. $t$-test) when data have strange distribution.

- Need to ensure exchangeability.
- Don’t hold your breath.
Outline

• Null-hypothesis and Null-distribution
• Multiple comparisons and Family-wise error
• Gaussian Random Field theory
  • Voxel-wise tests
  • Cluster-based inference
• Enhanced Clusters
• Randomise, slow and reliable
• FDR - False Discovery Rate
• Extra: examples of different inferences
False Discovery Rate

• FDR: False Discovery Rate
A “new” way to look at inference.

• Uncorrected (for multiple-comparisons):
  • Is equivalent to saying: “I am happy to nearly always say something silly about my experiments”.

• Family-Wise Error (FWE):
  • Is equivalent to saying: “I am happy to say something silly about 5% of my experiments”.

• False Discovery Rate
  • Is equivalent to saying: “I am happy if 5% of what I say about each experiment is silly”.
False Discovery Rate

- **FDR: False Discovery Rate**
  A “new” way to look at inference.

- **Uncorrected (for multiple-comparisons):**
  - Is equivalent to saying: “I am happy to nearly always say something silly about my experiments”.
  - On average, 5% of all voxels are false positives

- **Family-Wise Error (FWE):**
  - Is equivalent to saying: “I am happy to say something silly about 5% of my experiments”.
  - On average, 5% of all experiments have one or more false positive voxels

- **False Discovery Rate**
  - Is equivalent to saying: “I am happy if 5% of what I say about each experiment is silly”.
  - On average, 5% of significant voxels are false positives
Little imaging demonstration.

Noise

Signal

Signal + Noise
uncorrected voxelwise control of FP rate at 10%

percentage of all null pixels that are False Positives

closest of FamilyWise Error rate at 10%

occurrence of FamilyWise Error

control of False Discovery Rate at 10%

percentage of activated (reported) pixels that are False Positives
FDR for dummies

• Makes assumptions about how errors are distributed (like GRT).

• Used to calculate a threshold.

• Threshold such that X% of super-threshold (reported) voxels are false positives.

• Threshold depends on the data. May for example be very different for [1 0] and [0 1] in the same study.
Outline

• Null-hypothesis and Null-distribution
• Multiple comparisons and Family-wise error
• Gaussian Random Field theory
  • Voxel-wise tests
  • Cluster-based inference
• Enhanced Clusters
• Randomise, slow and reliable
• FDR - False Discovery Rate
  • Extra: examples of different inferences
A little example...

We performed an experiment and analyzed it using feat. This is the t-map we got. Is there anything significant in there?
A little example...

We performed an experiment and analyzed it using feat. This is the t-map we got. Is there anything significant in there?

We start by using GRF theory to calculate the threshold for voxel-base FWE error 0.05.

\[ T = 4.27 \]
A little example...

We performed an experiment and analyzed it using feat. This is the $t$-map we got. Is there anything significant in there?

And then threshold at $T = 4.27$

And we find three blobs. Remember that if there had been no activations anywhere there is only a 5% risk we would have seen even one blob.
A little example...

But we are a little surprised that we see so little in the occipital cortex since our contrast had a visual component.

So we threshold, arbitrarily, at 2.3.
A little example...

But we are a little surprised that we see so little in the occipital cortex since our contrast had a visual component.

And use cluster-based GRF theory to calculate the FWE p-value for each cluster:

\[ p(k \geq 40) = 0.014 \]
A little example...

But we are a little surprised that we see so little in the occipital cortex since our contrast had a visual component.

And use cluster-based GRF theory to calculate the FWE $p$-value for each cluster:

$$p(k \geq 3) = 0.73$$
A little example...

But we are a little surprised that we see so little in the occipital cortex since our contrast had a visual component.

And use cluster-based GRF theory to calculate the FWE p-value for each cluster.

etc ...
A little example...

But we are a little surprised that we see so little in the occipital cortex since our contrast had a visual component.

And use cluster-based GRF theory to calculate the FWE p-value for each cluster.

And the winners are ...
A little example...

Let us now, just for fun, try another arbitrary threshold.
Say 3.5
A little example...

Let us now, just for fun, try another arbitrary threshold. Say 3.5

And calculate cluster-based FWE corrected $p$-values
A little example...

Let us now, just for fun, try another arbitrary threshold. Say 3.5

And calculate cluster-based FWE corrected $p$-values

And the winners this time are ...
A little example...

So depending on exactly how we perform our inference testing we can get any of these

- Voxel-based
  
  FWE 0.05

- Cluster-based
  
  Initial threshold 2.3
  
  FWE 0.05

- Cluster-based
  
  Initial threshold 3.5
  
  FWE 0.05

But they look quite different!
A little example...

Let us finish by calculating also the FDR threshold for 0.05

$$T_{FDR} = 2.67$$

And that looks different too. How do we bring some order to all this?
Since it was fake...

“Magnitude” 1

“Magnitude” 0.05

We know the “true” activations
Since it was fake...

- Detected both clusters
- Poor characterization of “large” activation
- No false positives
Since it was fake...

- Detected both clusters
- Much better characterization of “large” activation
- One false positive. But that was just a fluke

Cluster-based
Initial threshold 2.3
FWE 0.05
Since it was fake...

- Detected both clusters
- So-so characterization of “large” activation
- No false positives

Cluster-based
Initial threshold 3.5
FWE 0.05
Since it was fake...

- Detected both clusters
- Quite good characterization of “large” activation
- Two false positive blobs
- 8.8% of all reported voxels are false positives