Inference

how surprising is your statistic? (thresholding)

But ... can I trust it?
Outline

• Null-hypothesis and Null-distribution
• Multiple comparisons and Family-wise error
• Different ways of being surprised
  • Voxel-wise inference (Maximum z)
  • Cluster-wise inference (Maximum size)
• Parametric vs non-parametric tests
• Enhanced clusters
• FDR - False Discovery Rate
Outline

• Null-hypothesis and Null-distribution
• Multiple comparisons and Family-wise error
• Different ways of being surprised
  • Voxel-wise inference (Maximum $z$)
  • Cluster-wise inference (Maximum size)
• Parametric vs non-parametric tests
• Enhanced clusters
• FDR - False Discovery Rate
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
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

Untrustworthy
Tools of classical inference

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

Assesses “trustworthiness”

A $t$-statistic reflects precisely this

$$t = \sqrt{n} \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\sigma^2}}$$

- Large difference: Trustworthy
- Small variability: Trustworthy

- Many measurements: Trustworthy
Tools of classical inference

1. A null-hypothesis

2. A test-statistic

Or expressed in GLM lingo

Large difference: Trustworthy

Small variability: Trustworthy

Many measurements: Trustworthy

\[
\begin{bmatrix}
\hat{\beta}_1 \\
\hat{\beta}_2
\end{bmatrix} = \begin{bmatrix}
\bar{x}_1 \\
\bar{x}_2
\end{bmatrix}
\]

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

\[
\bar{x}_1 - \bar{x}_2
\]
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

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

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

\[
sigma^2 = 0.71
\]

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

\[
t = 2.19
\]
Tools of classical inference

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

We might then get these data:

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

- \[ t = 2.19 \]
- \[ c^T \hat{\beta} = 1.17 \]
- \[ \sigma^2 = 0.71 \]
- Constant
Tools of classical inference

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

or we could have gotten these

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

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

\[ \sigma^2 = 1.28 \]

Constant
Tools of classical inference

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

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

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

\[ \sigma^2 = 1.01 \]

\[ t = 0.49 \]
Tools of classical inference

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

\[ t = \frac{c^T \hat{\beta}}{\sqrt{\sigma^2} \sqrt{c^T (X^TX)^{-1} c}} = \frac{1.22}{\sqrt{0.78}} \]

or perhaps these

\[ = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + e \]
Tools of classical inference

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

\[ t = \frac{c^T \hat{\beta}}{\sqrt{\sigma^2} \sqrt{c^T (X^T X)^{-1} c}} = \frac{-0.69}{\sqrt{0.44}} \]

\[ \sigma^2 \] = 0.44

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

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

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

And if we do this many many many many many times…
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 ~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 \(\sim 5\%\) 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
   \[ H_0: \overline{x}_1 = \overline{x}_2, \ H_1: \overline{x}_1 > \overline{x}_2 \]
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

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

\[
t = \frac{c^T \hat{\beta}}{\sqrt{\sigma^2} \sqrt{c^T (X^T X)^{-1} c}} = \frac{1.53}{\sqrt{0.85} \sqrt{0.4}} = 2.64
\]
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

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

\[ t_8 = 2.64 \quad t_8 = 2.64^* \]

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.
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?

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

\[
\begin{align*}
H_0 & \text{ is true } \\
H_0 & \text{ is false }
\end{align*}
\]  True state of affairs

\[
\begin{align*}
\text{We don’t reject } H_0 & \\
\text{We reject } H_0
\end{align*}
\]  Our decision
### False positives/negatives

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

### True state of affairs

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

### Our decision

<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></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>
</tr>
</thead>
<tbody>
<tr>
<td>☺</td>
<td></td>
</tr>
<tr>
<td>$H_0$ is false</td>
<td>We reject $H_0$</td>
</tr>
<tr>
<td></td>
<td>☺</td>
</tr>
</tbody>
</table>
# False positives/negatives

\[
\begin{align*}
H_0 \text{ is true} & \quad \{ \text{True state of affairs} \} \\
H_0 \text{ is false} & \\
\{ \text{Our decision} \} \\
\end{align*}
\]

- We don’t reject \( H_0 \)
- We reject \( H_0 \)

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

\[ H_0 \text{ is true} \} \rightarrow \text{True state of affairs} \]

\[ H_0 \text{ is false} \} \rightarrow \text{Our decision} \]

We don’t reject \( H_0 \)

We reject \( H_0 \)

<table>
<thead>
<tr>
<th>( H_0 \text{ is true} )</th>
<th>( H_0 \text{ is false} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>We don’t reject ( H_0 )</td>
<td>False positive \nType I error</td>
</tr>
<tr>
<td>We reject ( H_0 )</td>
<td>False negative \nType II error</td>
</tr>
</tbody>
</table>
Outline

- Null-hypothesis and Null-distribution

- Multiple comparisons and Family-wise error

- Different ways of being surprised
  - Voxel-wise inference (Maximum z)
  - Cluster-wise inference (Maximum size)

- Parametric vs non-parametric tests

- Enhanced clusters

- FDR - False Discovery Rate
Multiple Comparisons

- In neuroimaging we typically perform many tests as part of a study.

Different here?  Maybe here?  Or here?

...
What happens when we apply this to imaging data?

z-map where each voxel \( \sim N \).
Null-hypothesis true everywhere, i.e. NO ACTIVATIONS

z-map thresholded at 1.64

16 clusters
288 voxels
\( \sim 5.5\% \) of the voxels

That’s a LOT of false positives
The strict approach: 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!
But ... doesn’t 5.65 sound very high?

Largest observed value

Bonferroni threshold

Observed values in the z-map

Too lenient

0.05

1.64

Too harsh

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
- Different ways of being surprised
  - Voxel-wise inference (Maximum z)
  - Cluster-wise inference (Maximum size)
- Parametric vs non-parametric tests
- Enhanced clusters
- FDR - False Discovery Rate
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.
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.

$\max(z) = 6.84$
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.

\[ \text{max}(z) = 5.93 \]
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.

$max(z) = 4.62$
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.

$max(z) = 7.36$
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.

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! 😞
Outline

• Null-hypothesis and Null-distribution
• Multiple comparisons and Family-wise error
• Different ways of being surprised
  • Voxel-wise inference (Maximum z)
  • Cluster-wise inference (Maximum size)
• Parametric vs non-parametric tests
• Enhanced clusters
• FDR - False Discovery Rate
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 ~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 \( f \) that allows us to calculate a Family Wise threshold \( u \) pertaining to cluster size.

But what does \( f \) and \( u \) crucially depend on?
Distribution of Max Cluster Size

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

\[ f \] 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 $f$ that allows us to calculate a Family Wise threshold $u$ pertaining to cluster size. $f$ depends crucially on the initial “cluster-forming” threshold?

$u = 49$

$z = 2.7$
So, just as was the case for the z-values, we now have a distribution $f$ that allows us to calculate a Family Wise threshold $u$ pertaining to cluster size.

$f$ 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 $f(z)$ and the same for $u(z)$

$z = 3.1$

$u = 25$

$z = 2.7$

$u = 49$

$z = 2.3$

$u = 76$

But as before we don’t have an expression for these distributions.
Outline

- Null-hypothesis and Null-distribution
- Multiple comparisons and Family-wise error
- Different ways of being surprised
  - Voxel-wise inference (Maximum z)
  - Cluster-wise inference (Maximum size)
- Parametric vs non-parametric tests
- Enhanced clusters
- FDR - False Discovery Rate
Parametric vs non-parametric

- As we described earlier, one of the great things about for example the t-test is that we know the null-distribution.

- But most distributions are not that simple.

- And errors are not always normally distributed.

Provided that $\mathbf{e} \sim N(0,\sigma^2)$
Example: VBM-style analysis

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

Group #1 (FSL Course Tutors)

\[
\begin{bmatrix}
\beta_1 \\
\beta_2
\end{bmatrix}
= \begin{bmatrix}
\mathbf{1} \\
\mathbf{1}
\end{bmatrix}
\]

Group #2 (FSL Course Attendees)

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

\(\text{hist}(e)\)

Ok!

\(\sim N?\)
Parametric vs non-parametric

- There are approximations to the Max-z and Max-size statistics

- These are valid under certain sets of assumptions

- But can be a problem when applied outside of that set of assumptions

- Search area “large relative to boundary”
- “High enough” cluster forming threshold
- Normal distributed errors

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

Anders Eklund*, Thomas E. Nichols*, and Han Knutsson*
Parametric vs non-parametric

- Those approximations were based on Gaussian Random Field Theory, and was an impressive body of work.

- They served us fantastically well at a time when we had little choice.

- But the we’ve moved towards non-parametric testing.
Parametric vs non-parametric
A simple permutation test

- We can permute the data itself to create a distribution that we can use to test our statistic.
  - Makes very few assumptions about the data
  - Works for any test statistic

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.
A simple permutation test

- We can permute the data itself to create a distribution that we can use to test our statistic.
  - Makes very few assumptions about the data
  - Works for any test statistic

One re-labelling

$t$-value after re-labelling

$t = 0.67$

Let's start collecting them
A simple permutation test

- We can permute the data itself to create a distribution that we can use to test our statistic.
  + Makes very few assumptions about the data
  + Works for any test statistic

Second re-labelling

\[ t = 1.97 \]

Original labelling

And another one
A simple permutation test

• We can permute the data itself to create a distribution that we can use to test our statistic.
  
  + Makes very few assumptions about the data
  
  + Works for any test statistic

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

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

i.e. \( p(x \geq 2.27) = 1.79\% \) for \( t_{18} \)
And we can use this for any statistic

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 Nature Neuro. But, did they jump the gun?
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 Nature Neuro. **But**, did they jump the gun?

And we can use this for any statistic.

**Group 1**

**Group 2**

2nd level model

Our group difference map

\[ \text{max}(t) = 4.65 \]
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 Nature Neuro. **But**, did they jump the gun?

And we can use this for any statistic.
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 Nature Neuro. But, did they jump the gun?

And we can use this for any statistic.
And we can use this for any statistic

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 Nature Neuro. But, did they jump the gun?

max($t$) = 5.84

Group 1

Group 2

3rd Permutation 3rd permuted map
And we can use this for any statistic

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 Nature Neuro. But, did they jump the gun?

3925 permutations yielded higher max(t)-value than original labelling. We cannot reject the null-hypothesis.

5000 permutations
But beware the "exchangeability"

- When we swap the labels of two data-points we need to make sure that they are "exchangeable"

- "Exchangeable" means that the covariance matrix of the noise/error after model fitting isn’t changed by a permutation (will show examples of this)
1st level fMRI data is not exchangeable

- You may, or may not, have seen this slide in the 1st level GLM talk.

\[
\begin{align*}
\mathbf{y} &= \mathbf{X} \mathbf{\beta} + \mathbf{e} \\
\mathbf{X} &= \begin{bmatrix} x_1 \ x_2 \end{bmatrix} \\
\mathbf{\beta} &= \begin{bmatrix} \beta_1 \\
\beta_2 \end{bmatrix} \\
\mathbf{e} &\sim N(\mathbf{0}, \Sigma)
\end{align*}
\]

This time we will look more closely at this part.

This is the (potentially) problematic covariance matrix.
1st level fMRI data is not exchangeable

- One important component of noise in fMRI consists of physiological/neuronal events convolved by the HRF
1st level fMRI data is not exchangeable

- One important component of noise in fMRI consists of physiological/neuronal events convolved by the HRF

If we sample this every 20 seconds it no longer looks “smooth”
First level fMRI data is not exchangeable

- One important component of noise in fMRI consists of physiological/neuronal events convolved by the HRF

\[ e \sim N(0, \sigma^2 I) \]
First level fMRI data is not exchangeable

- One important component of noise in fMRI consists of physiological/neuronal events convolved by the HRF

But that is not a realistic TR. What about every 3 seconds?
1st level fMRI data is not exchangeable

- One important component of noise in fMRI consists of physiological/neuronal events convolved by the HRF.
1st level fMRI data is not exchangeable

- Let us now return to our model again

\[ Y = X \beta + \epsilon \]

Regressor, Explanatory Variable (EV)

Regression parameters, Effect sizes

\[ Y = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \epsilon \]

- The model consists of our regressors \( X \) and the noise model

- All permutations must result in “equivalent models”

- Let us now see what happens if we swap two data-points (points 5 and 10)
1st level fMRI data is not exchangeable

- One important component of noise in fMRI consists of physiological/neuronal events convolved by the HRF.

“Point” 10 now covaries with points 4 and 6.

“Point 5” now covaries with points 9 and 11.

And the models are no longer equivalent.
1st level fMRI data is not exchangeable

- One important component of noise in fMRI consists of physiological/neuronal events convolved by the HRF

And for a random permutation …

And the models are no longer equivalent
Back to exchangeability

• Data-points are not “exchangeable” if swapping them means that the noise covariance-matrix ends up looking different.

• Formally “The joint distribution of the data must be unchanged by the permutations under the null-hypothesis”.

• If the noise covariance-matrix has non-zero off-diagonal elements (covariances) you need to beware.

• You typically never estimate or see the covariance-matrix. You need to “imagine it” and determine from that if there is a problem.
Examples of exchangeability:
Two groups unpaired

This is the “exchangeability group”. Here all scans are in the same group, which means any scan can be exchanged for any other.

N.B. The “group” labelling is used for completely different purposes when using FLAME/GRFT
Examples of exchangeability:
Two groups unpaired

Assumed covariance matrix

The implicit assumption here is that data from all subjects have the same uncertainty and are all independent.
Examples of exchangeability:
Two groups unpaired

Original  Perm 1  Perm 2  ...

1  6  6
2  3  1
3  7  7
4  8  4
5  5  9
6  1  5
7  2  8
8  4  3
9  9  10
10 10 2
Examples of exchangeability: Single group average

Here we model a single mean and want to know if that is different from zero

But there isn’t really anything to permute, or is there?
Examples of exchangeability: Single group average

$\text{Original} \quad t = -0.17$
Examples of exchangeability:
Single group average

First flip

\[ t = -0.09 \]
Examples of exchangeability: Single group average

$t = 1.54$
Examples of exchangeability: Single group average

Etc …
Examples of exchangeability: Single group average

And the assumptions are:

- Symmetric errors
- Errors independent
- Subjects drawn from a single population
Examples of exchangeability: Two groups paired

Here we can only exchange scans within each subject. I.e. Input 1 for Input 2, Input 3 for Input 4 etc.
Examples of exchangeability: Two groups paired

The implicit assumption here is that data from all subjects have the same uncertainty and that there is no dependence between subjects.

Assumed covariance matrix

Allowed swap
Examples of exchangeability: Two groups paired

Assumed covariance matrix

The implicit assumption here is that data from all subjects have the same uncertainty and that there is no dependence between subjects.

Disallowed swap
Examples of exchangeability: Two groups paired
Outline

• Null-hypothesis and Null-distribution
• Multiple comparisons and Family-wise error
• Different ways of being surprised
  • Voxel-wise inference (Maximum z)
  • Cluster-wise inference (Maximum size)
• Parametric vs non-parametric tests
  • Enhanced clusters
• FDR - False Discovery Rate
Clustering cookbook

Instead of resel-based correction, we can do clustering:

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

- **z stat image**
  - 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
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

Cluster-based (red)
Voxel-based (blue)

Z=22
Z=48
Y=−16

TFCE

0.003

p (corrected)

0.05

cluster-based (red)
voxel-based (blue)
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
- Different ways of being surprised
  - Voxel-wise inference (Maximum z)
  - Cluster-wise inference (Maximum size)
- Parametric vs non-parametric tests
- Enhanced clusters
- FDR - False Discovery Rate
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

control of FamilyWise Error rate at 10%
occurrence of FamilyWise Error
FWE

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.