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.

> 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 $t$-statistic reflects precisely this:

Assesses “trustworthiness”

Large difference: Trustworthy

Small variability: Trustworthy

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

Many measurements: Trustworthy
Tools of classical inference

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

Or expressed in GLM lingo

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

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

Large difference: Trustworthy
Small variability: Trustworthy
Many measurements: Trustworthy
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

\[ c^T \hat{\beta} = 1.17 \]
\[ t = \frac{c^T \hat{\beta}}{\sqrt{\sigma^2} \sqrt{c^T (X^T X)^{-1} c}} \]
\[ \sigma^2 = 0.71 \]
Tools of classical inference

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

We might then get these data:

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

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

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

\[
\sigma^2 = 0.71
\]

Constant
Tools of classical inference

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

or we could have gotten these

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

Constant

\[ \sigma^2 = 1.28 \]

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

\[ t = -0.51 \]
Tools of classical inference

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

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

$$\sigma^2 = 1.01$$

$$c^T \hat{\beta} = 0.31$$

maybe 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

or perhaps these

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

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

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

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

\[ t = -1.66 \]

\[ c^T \hat{\beta} = -0.69 \]
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 \( H_0: \overline{x}_1 = \overline{x}_2 \), \( H_1: \overline{x}_1 > \overline{x}_2 \)

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

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

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

\[ t_8 = 2.64 \]

\[ t_8 = 2.64^* \]

There is \(~1.46\%\) 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*}
\text{\( H_0 \) is true} & \quad \text{\( H_0 \) is false} \\
\text{True state of affairs}
\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?

\[
\begin{align*}
\text{H}_0 \text{ is true} & \quad \text{True state of affairs} \\
\text{H}_0 \text{ is false} & \quad \text{Our decision}
\end{align*}
\]

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

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>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

\[
\begin{array}{c|c}
H_0 \text{ is true} & \text{We don’t reject } H_0 \\
\hline
H_0 \text{ is false} & \text{We reject } H_0 \\
\end{array}
\]
**False positives/negatives**

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

- 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>False positive</td>
<td></td>
</tr>
<tr>
<td>$H_0$ is false</td>
<td>False negative</td>
<td></td>
</tr>
</tbody>
</table>
**False positives/negatives**

\[
\begin{align*}
H_0 \text{ is true} & \quad \{ \text{False positive / Type I error} \\
H_0 \text{ is false} & \quad \{ \text{False negative / Type II error}
\end{align*}
\]

**True state of affairs**

\[
\begin{align*}
\text{We don’t reject } H_0 & \quad \{ \text{Our decision} \\
\text{We reject } H_0 & \quad \}
\end{align*}
\]
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.
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
Italians doing maths: The Bonferroni correction

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

5255 voxels

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

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

Observed values in the z-map

Largest observed value

Bonferroni threshold

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?

Family-wise error