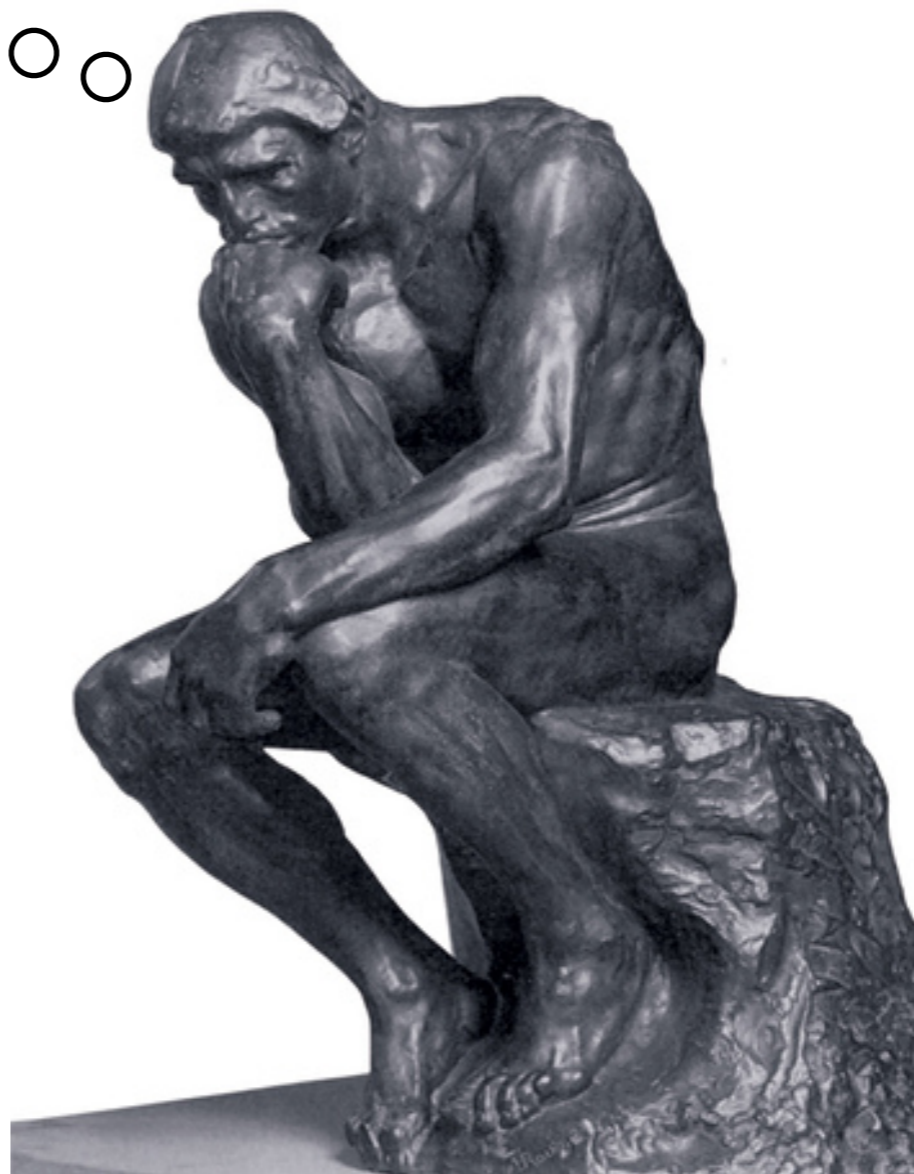
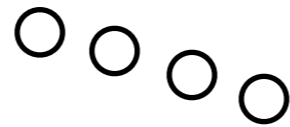




# *Inference*

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



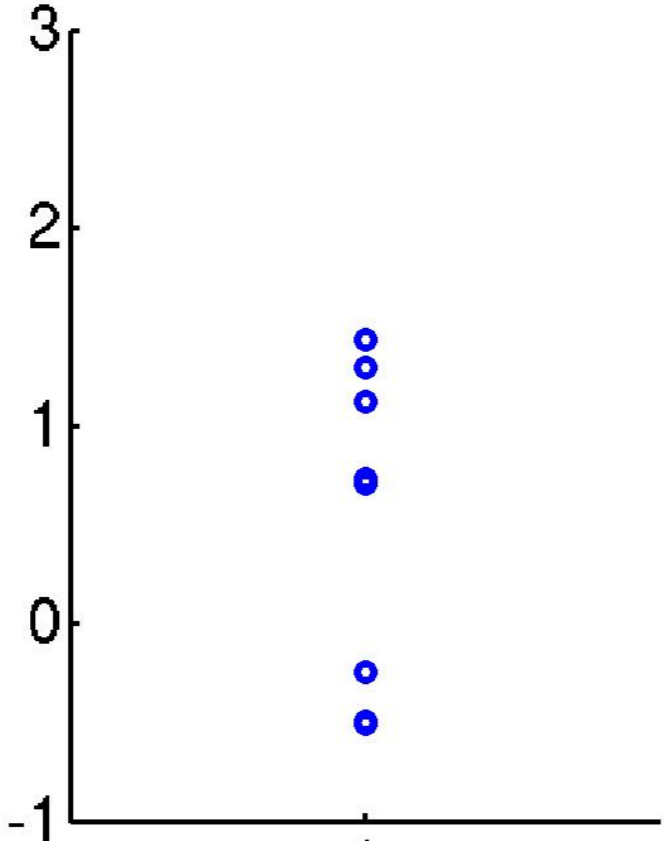
# Outline

- **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

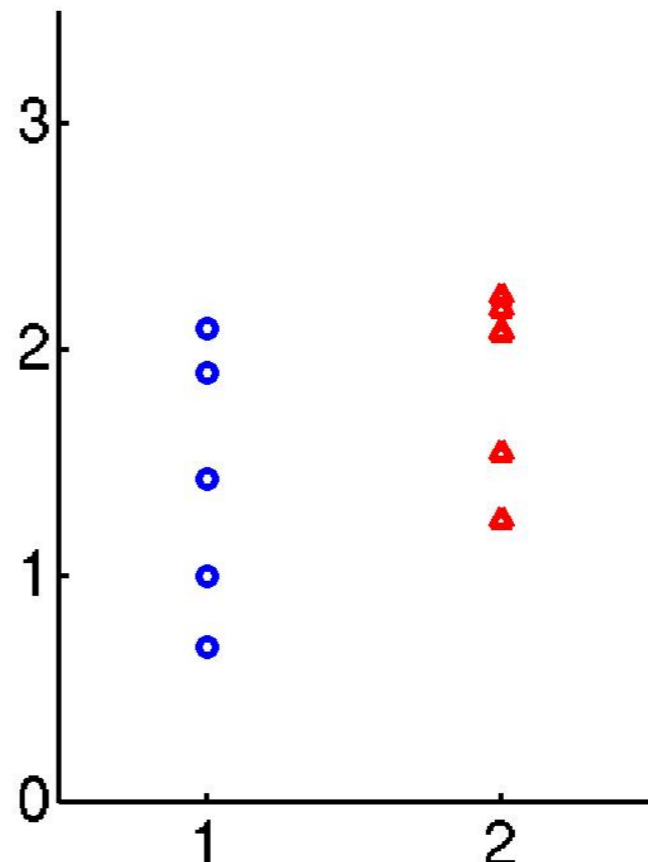


# The task of classical inference

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



> 0 ?



Different?

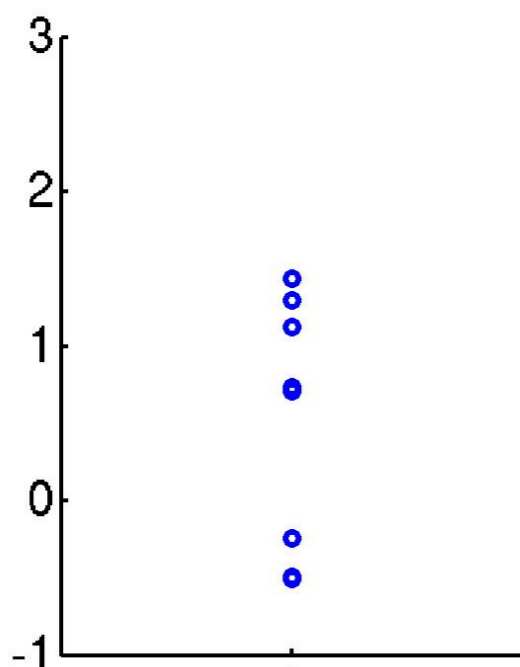


# Tools of classical inference

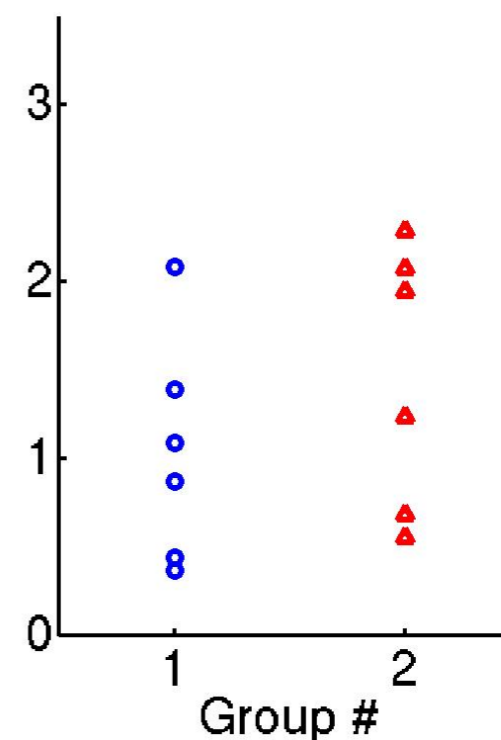
## I. A null-hypothesis

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

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



There is **no** difference between groups:  $\mu_1 = \mu_2$





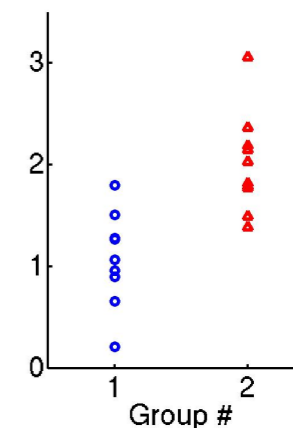
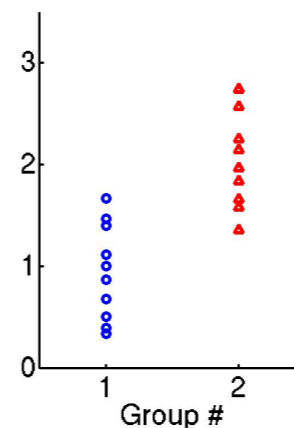
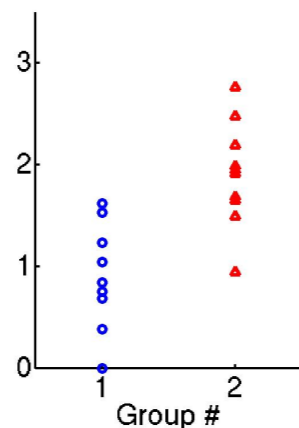
# Tools of classical inference

1. A null-hypothesis

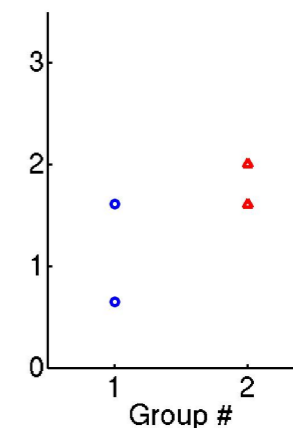
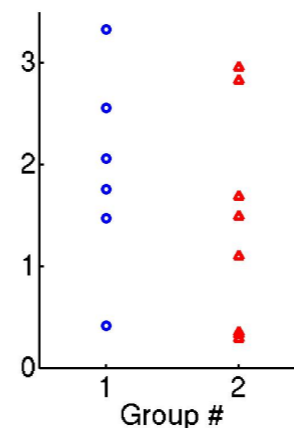
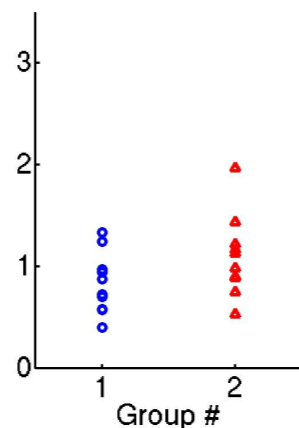
2. A test-statistic

Assesses “trustworthiness”

Trustworthy



Untrustworthy





# Tools of classical inference

1. A null-hypothesis

2. A test-statistic

Assesses “trustworthiness”

A  $t$ -statistic reflects precisely this

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

Large difference:  
Trustworthy

Many measurements:  
Trustworthy

Small variability:  
Trustworthy

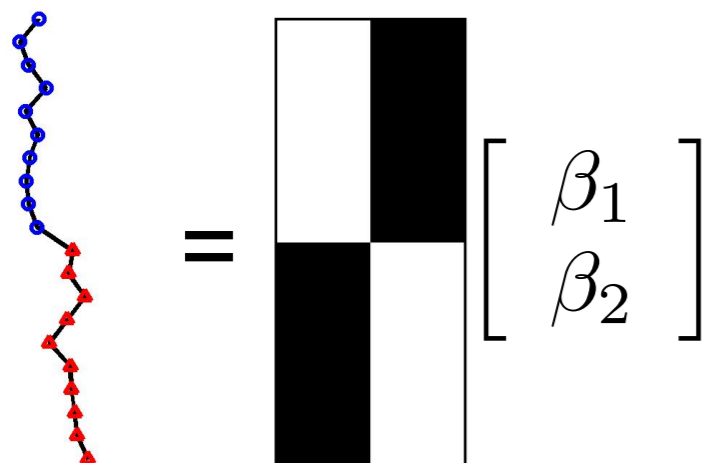


# Tools of classical inference

1. A null-hypothesis

2. A test-statistic

Or expressed in GLM lingo



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

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

Large difference: Trustworthy

Small variability: Trustworthy

Many measurements: Trustworthy

$\bar{x}_1 - \bar{x}_2$

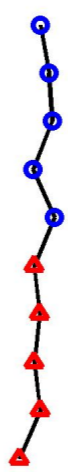




# Tools of classical inference

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

We might then get these data



$$= \begin{bmatrix} \text{white} & \text{black} \\ \text{black} & \text{white} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \mathbf{e}$$

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

$$t = 2.19$$

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

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

$\mathbf{c}^T \hat{\boldsymbol{\beta}}$

$\sigma^2 = 0.71$

Constant

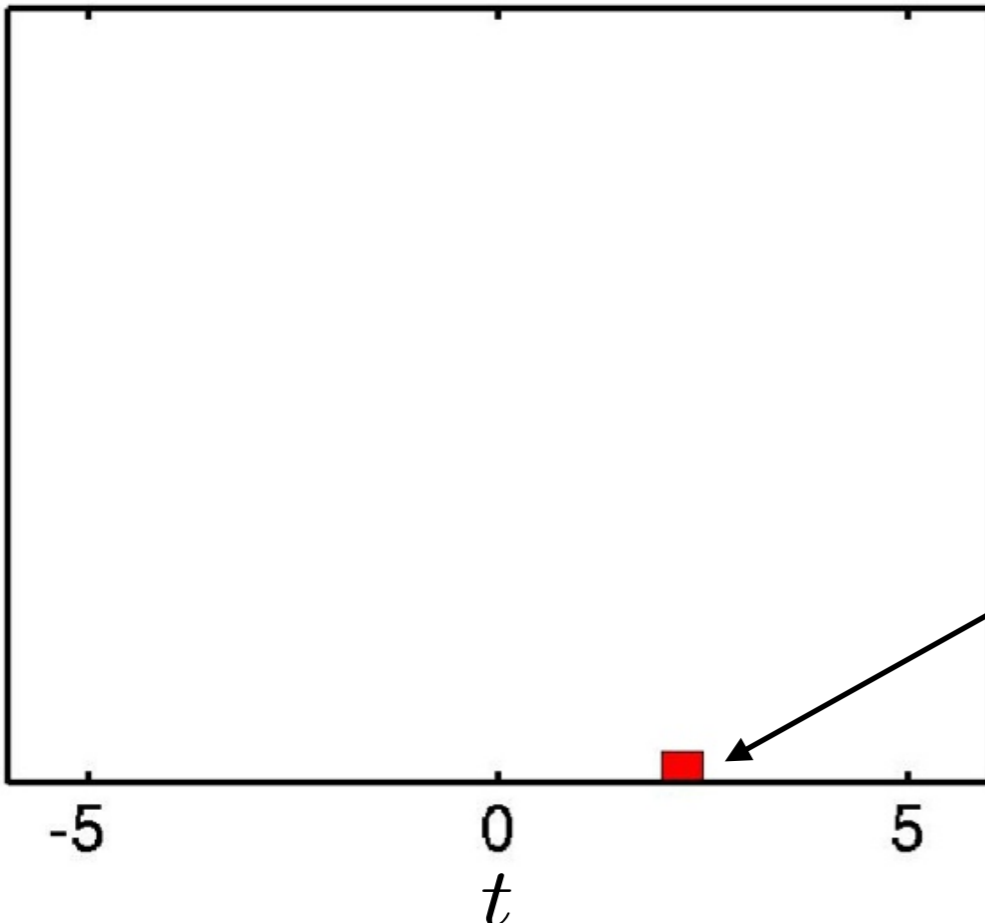


# Tools of classical inference

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

We might then get these data

$$= \begin{bmatrix} \text{white} & \text{black} \\ \text{black} & \text{white} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \mathbf{e}$$



$t = 2.19$

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

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

$\sigma^2 = 0.71$

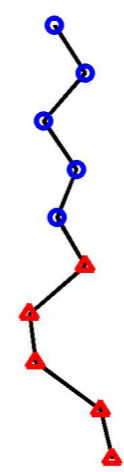
Constant



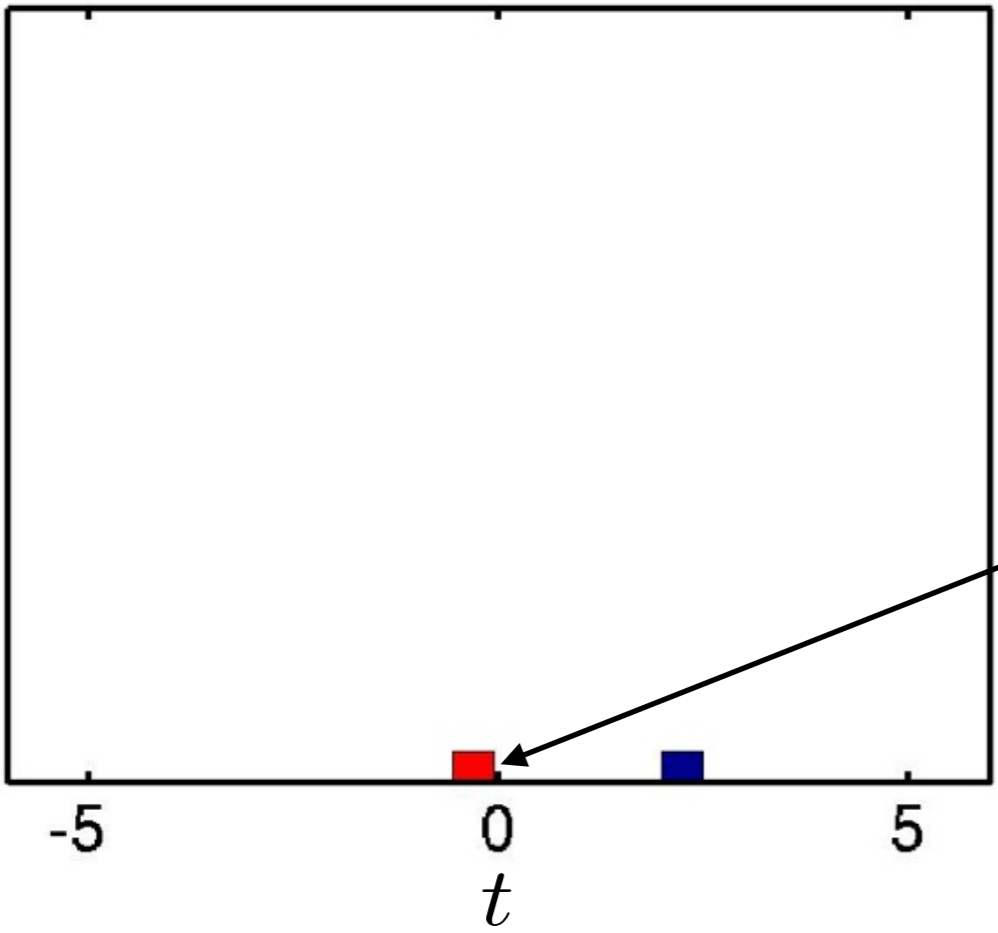
# Tools of classical inference

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

or we could have gotten these



$$= \begin{bmatrix} \text{white} & \text{black} \\ \text{black} & \text{white} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \mathbf{e}$$



$$t = -0.51$$

$$c^T \hat{\beta} = -0.37$$

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

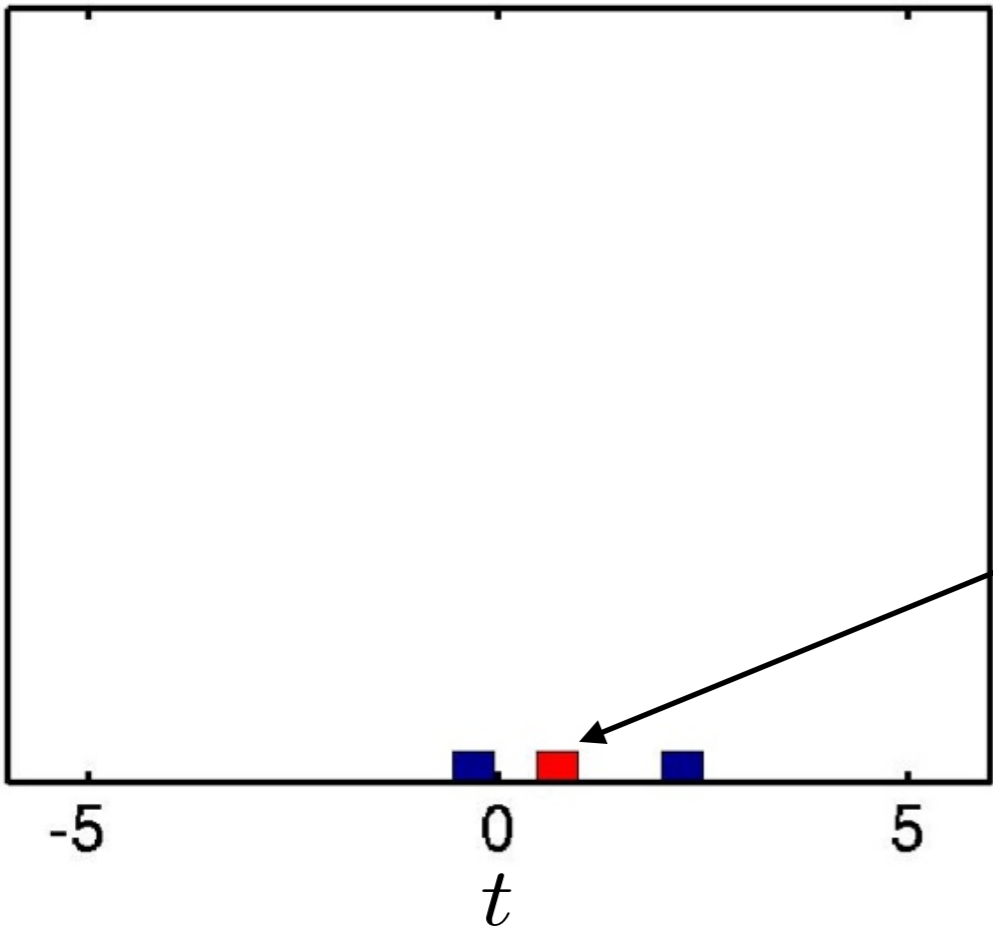
$$t = \frac{c^T \hat{\beta}}{\sqrt{\sigma^2} \sqrt{c^T (X^T X)^{-1} c}}$$

$\sigma^2 = 1.28$       Constant



# Tools of classical inference

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



maybe these

$$= \begin{bmatrix} \text{white} & \text{black} \\ \text{black} & \text{white} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \mathbf{e}$$

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

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

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

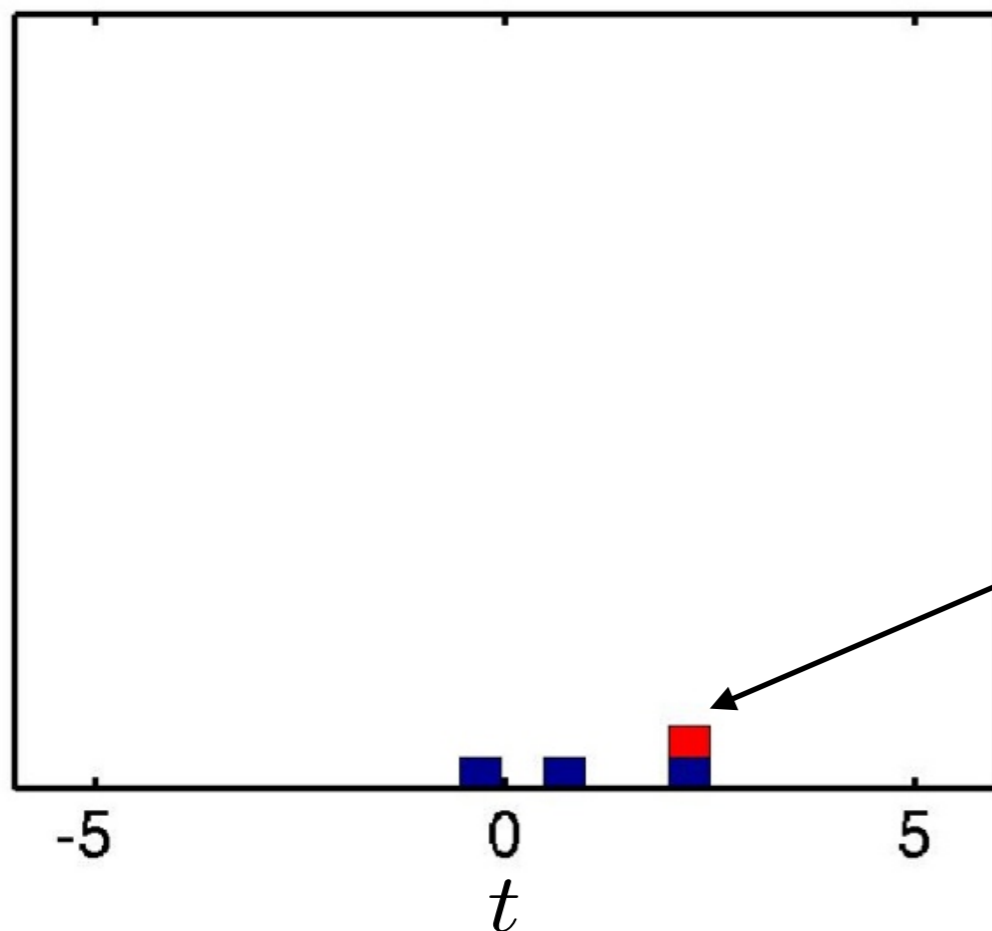
$$\sigma^2 = 1.01$$

Constant

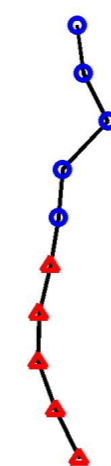


# Tools of classical inference

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



or perhaps these



$$= \begin{bmatrix} \text{white} & \text{black} \\ \text{black} & \text{white} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \mathbf{e}$$

$t = 2.19$

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

$\mathbf{c}^T \hat{\boldsymbol{\beta}} = 1.22$

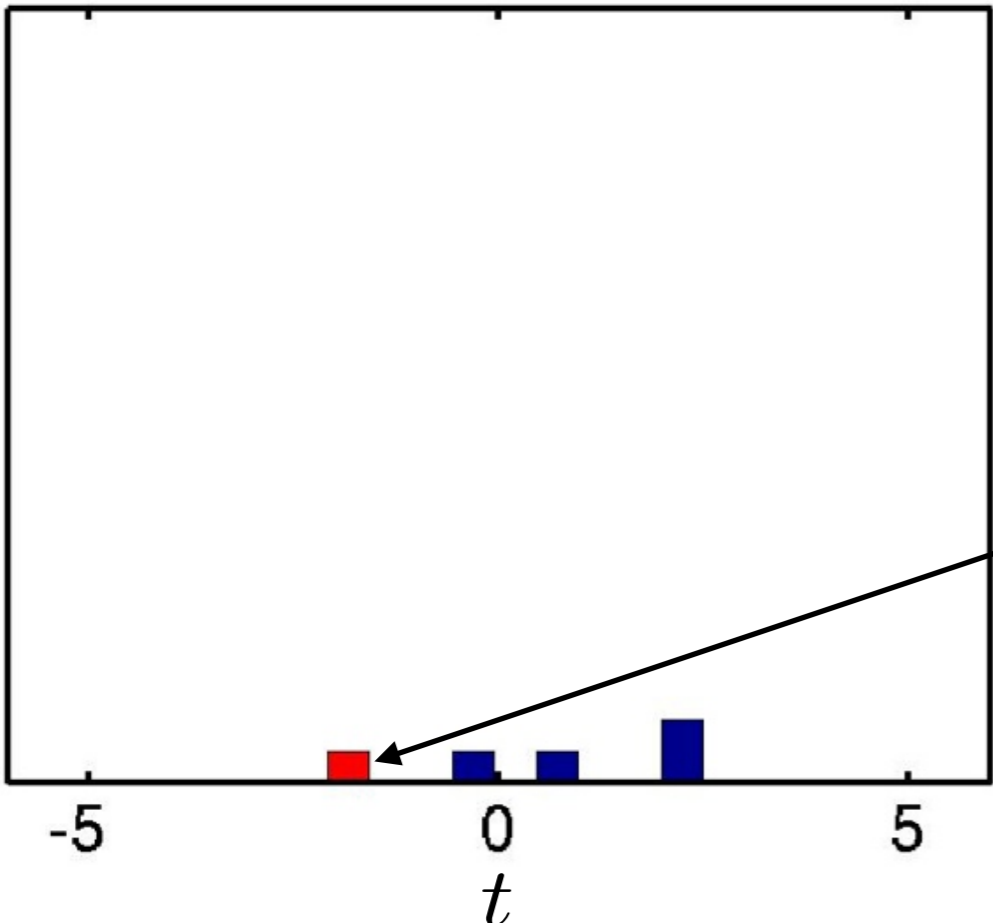
Constant

$\sigma^2 = 0.78$



# Tools of classical inference

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



$t = -1.66$

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

$c^T \hat{\beta} = -0.69$   
 $\sigma^2 = 0.44$   
 Constant

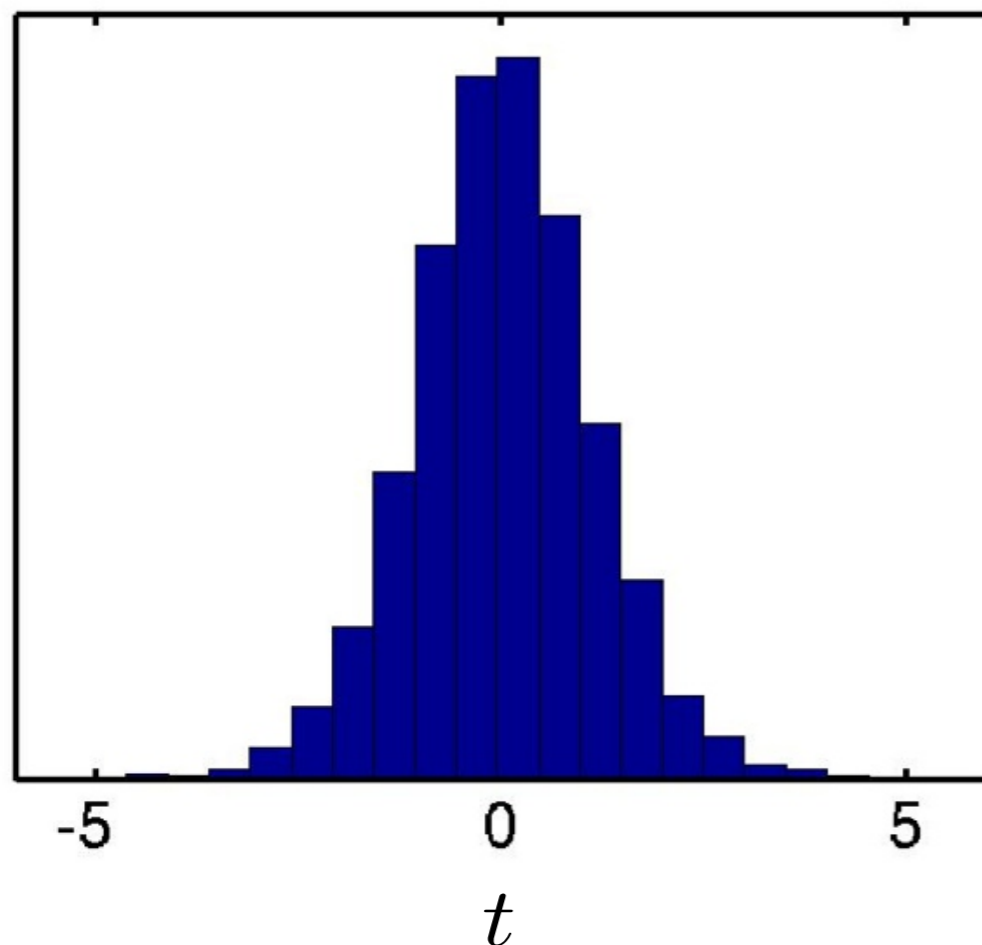
etc

$$= \begin{bmatrix} \text{white} & \text{black} \\ \text{black} & \text{white} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \mathbf{e}$$



# Tools of classical inference

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



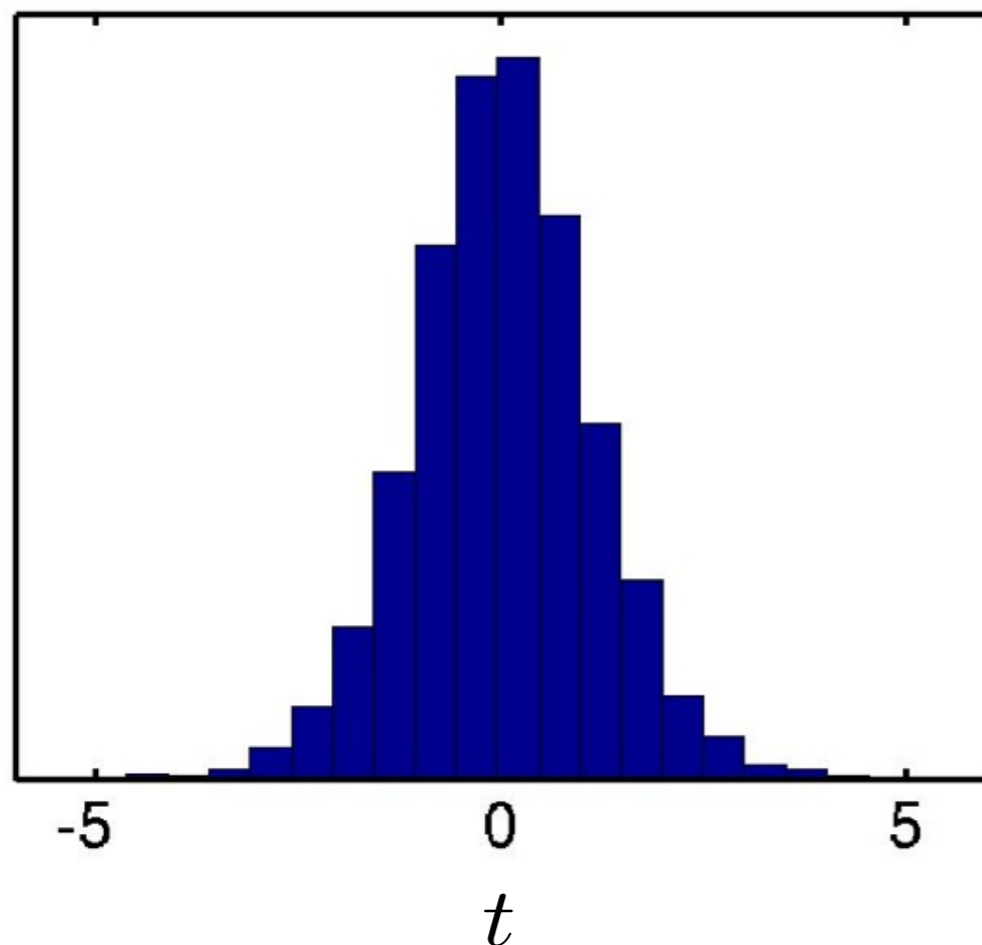
$$\begin{matrix} \circ \\ \circ \\ \circ \\ \circ \\ \circ \\ \circ \\ \circ \\ \circ \\ \circ \\ \circ \\ \circ \\ \triangle \\ \triangle \\ \triangle \\ \triangle \\ \triangle \\ \triangle \end{matrix} \begin{matrix} \circ \\ \circ \\ \circ \\ \circ \\ \circ \\ \circ \\ \circ \\ \circ \\ \circ \\ \circ \\ \circ \\ \triangle \\ \triangle \\ \triangle \\ \triangle \\ \triangle \\ \triangle \end{matrix} = \begin{matrix} \square & \square \\ \square & \square \end{matrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \mathbf{e}$$

And if we do this many many many many times...



# Tools of classical inference

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



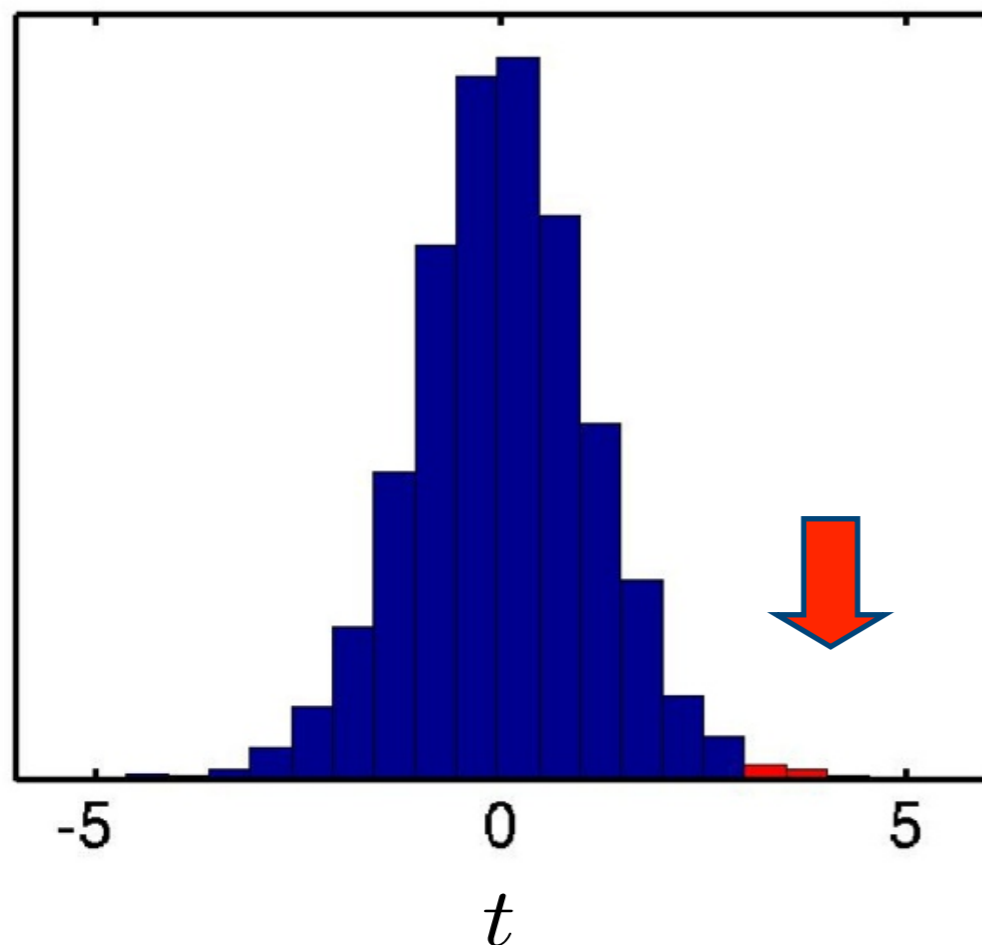
So, why is this helpful?





# Tools of classical inference

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

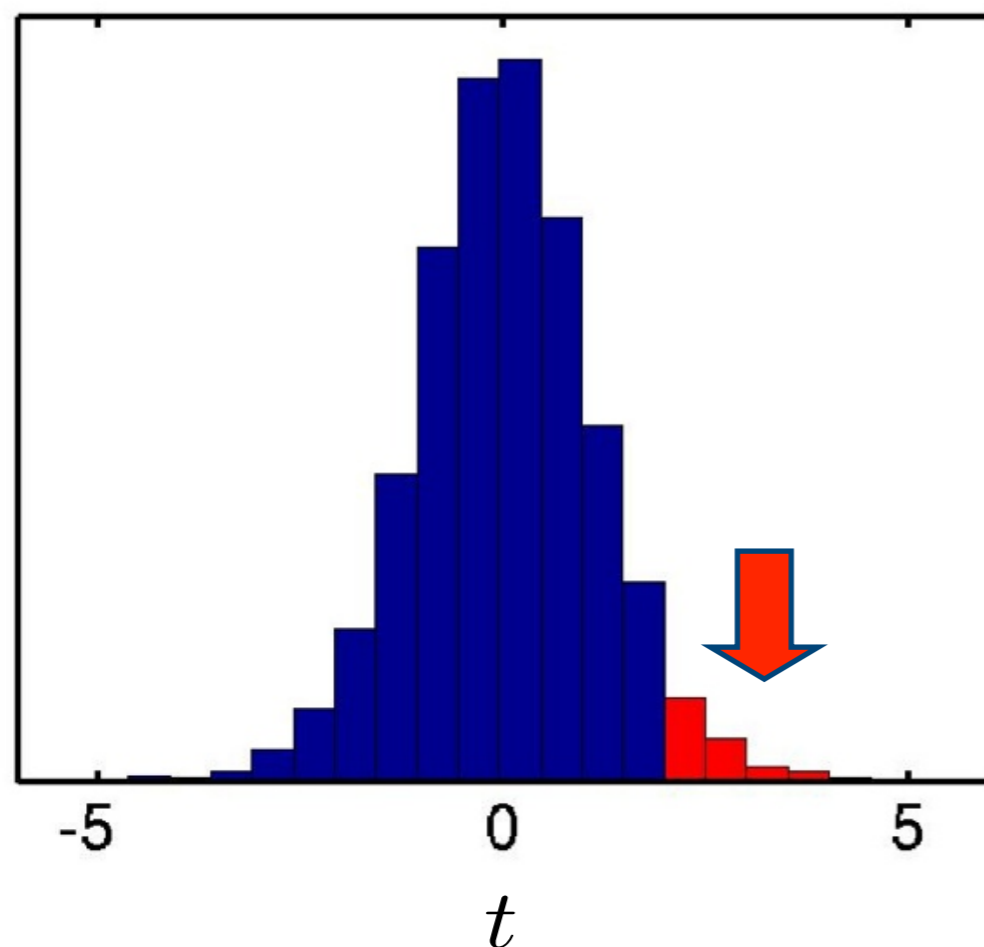


Well, it for example tells us that in  $\sim 1\%$  of the cases  $t > 3.00$ , even when the null-hypothesis is true.



# Tools of classical inference

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

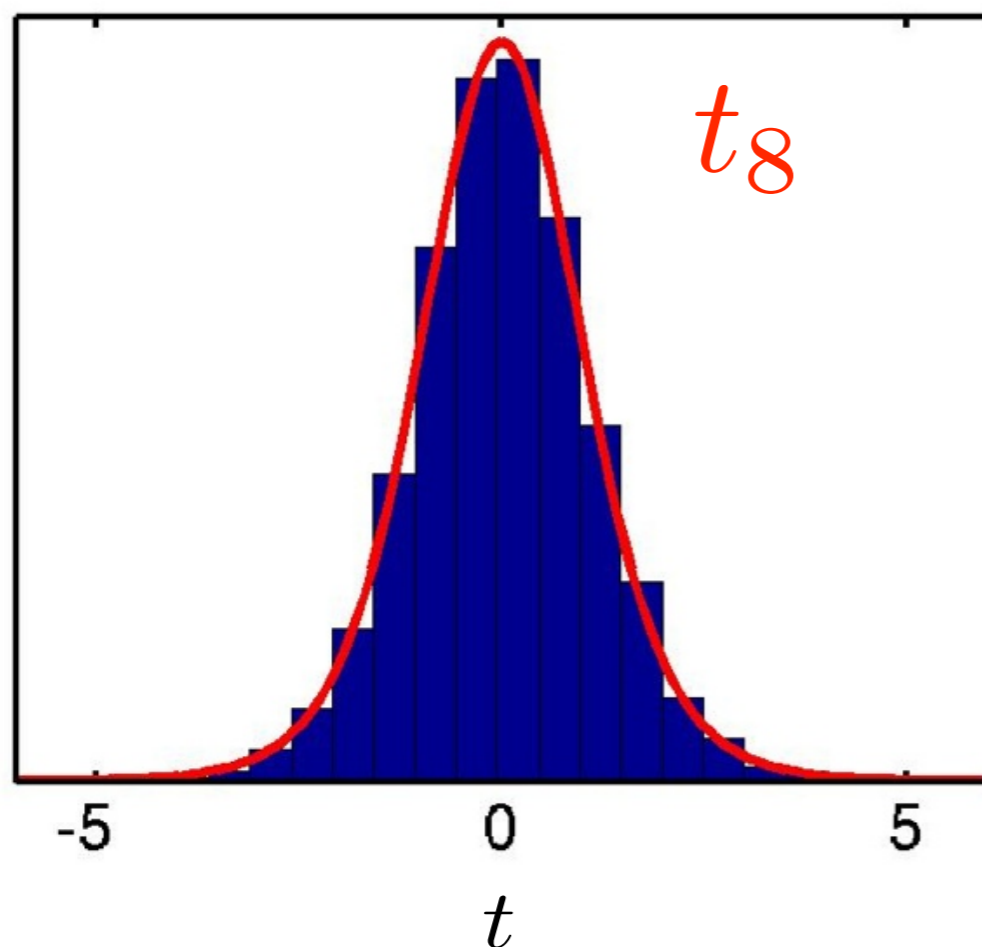


Or that in  $\sim 5\%$  of the cases  $t > 1.99$ .  
When the null-hypothesis is true.



# Tools of classical inference

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

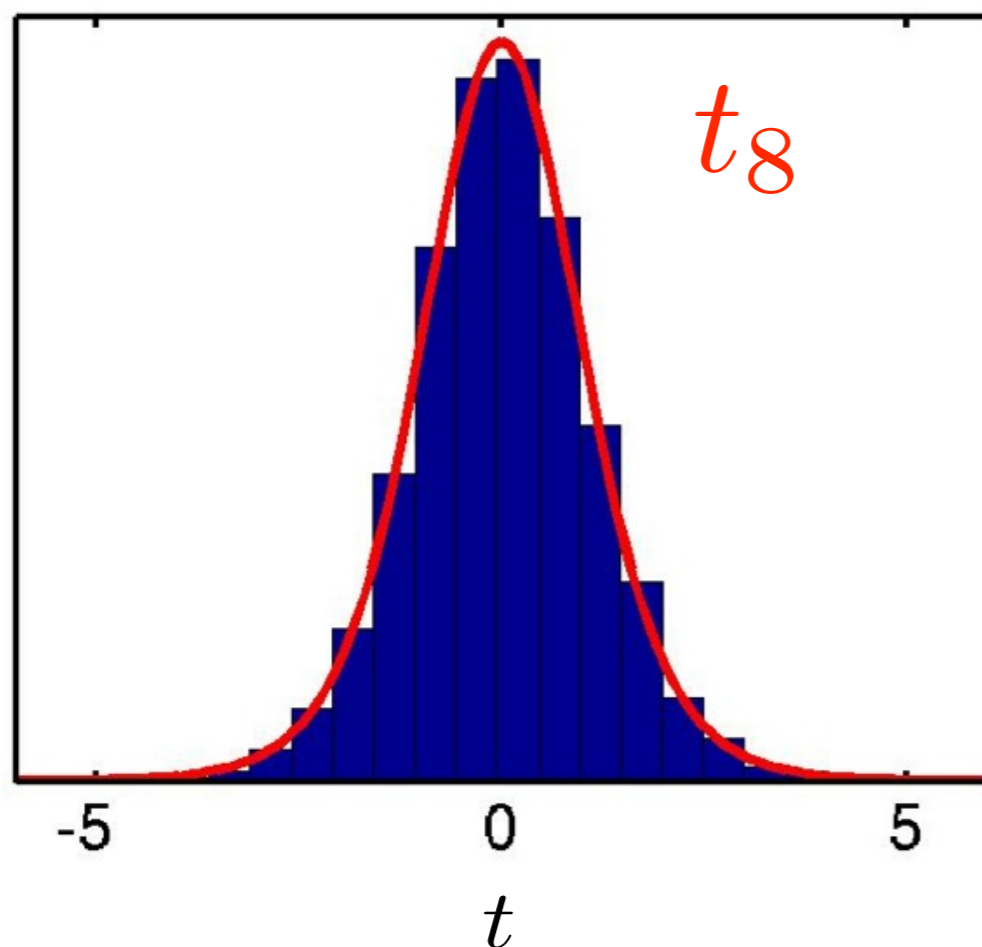


And best of all: This distribution is known *i.e.* one can calculate it.  
Much as one can calculate sine or cosine



# 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  $\mathbf{e} \sim N(0, \sigma^2)$



# 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

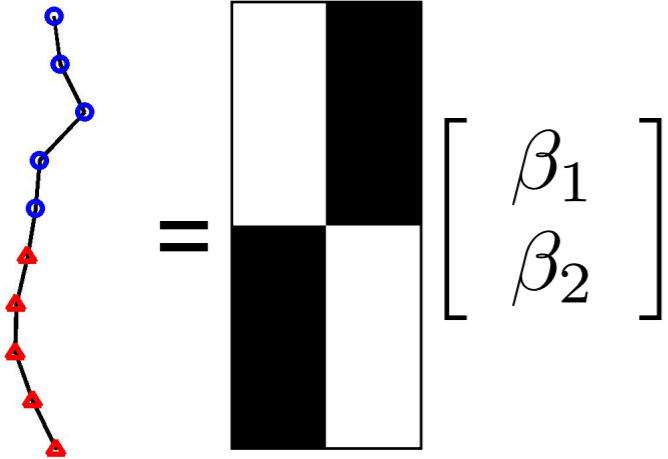
2. A test-statistic

3. A null-distribution

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

$$t_8 = 2.64$$

So, with these tools let us do an experiment



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



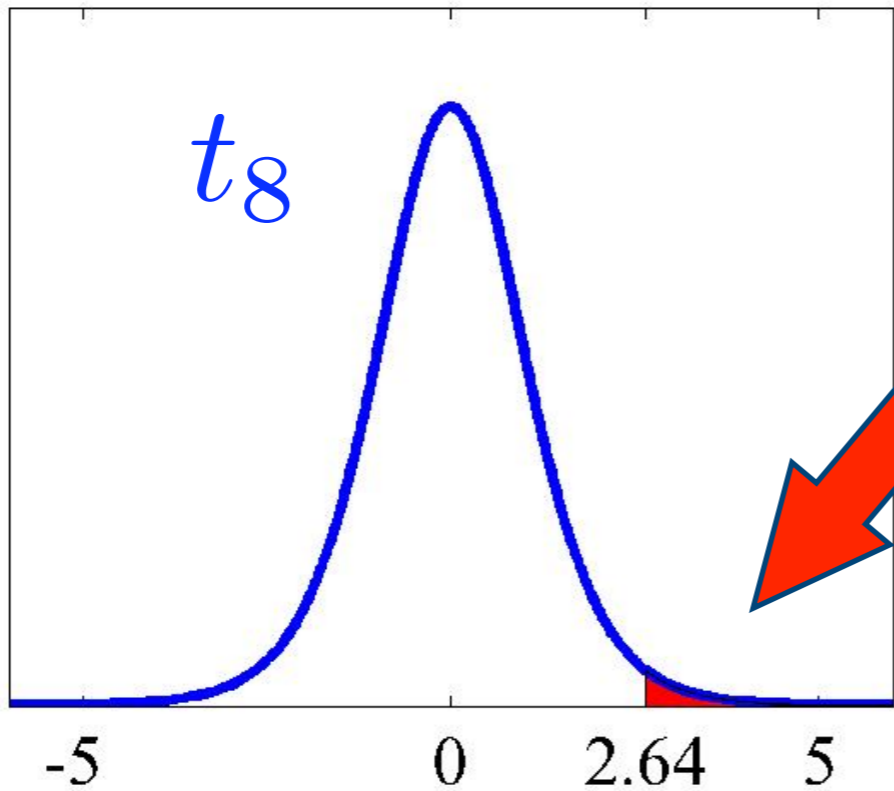
# An example experiment

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

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

$$t_8 = 2.64$$

So, with these tools let us do an experiment



If the null-hypothesis is true, we would expect to have a ~1.46% chance of finding a t-value this large or larger



# An example experiment

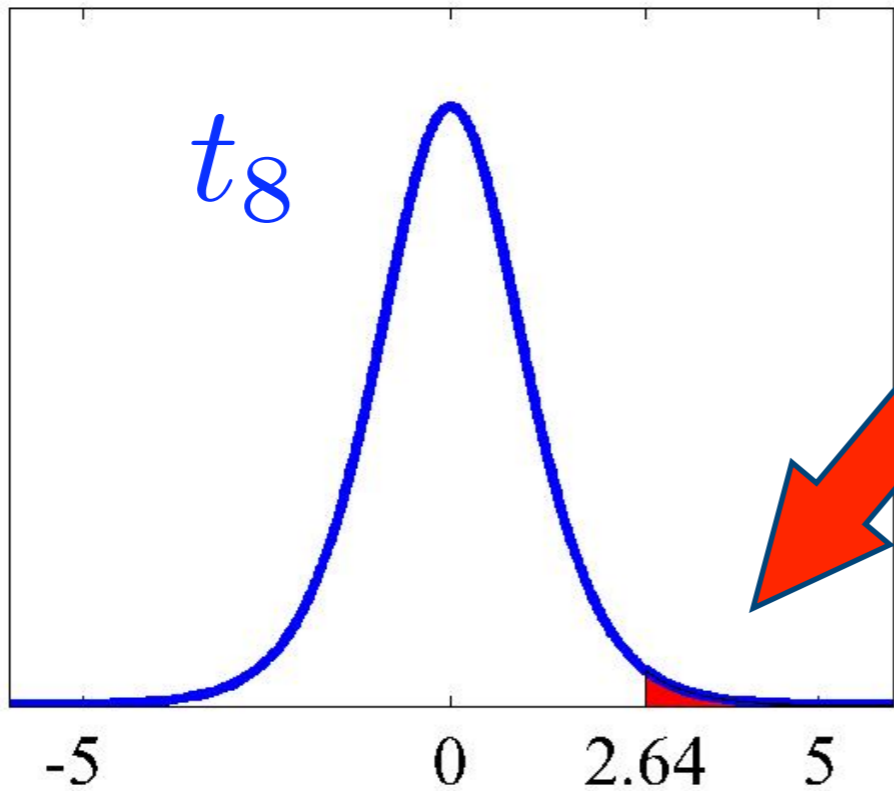
- 1. A null-hypothesis
- 2. A test-statistic
- 3. A null-distribution

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

$$t_8 = 2.64$$

$$t_8 = 2.64^*$$

So, with these tools let us do an experiment



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.





# 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?

$H_0$  is true }  
 $H_0$  is false } True state of affairs



# 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?

$H_0$  is true }  
 $H_0$  is false } True state of affairs

We don't reject  $H_0$  }  
We reject  $H_0$  } 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

We don't reject  $H_0$       We reject  $H_0$

$H_0$  is true



$H_0$  is false




# 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 $H_0$
$H_0$ is true		
$H_0$ is false		



# 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  $H_0$

$H_0$  is true

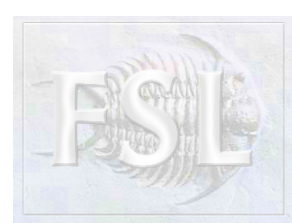


False positive

$H_0$  is false

False negative







# 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 $H_0$
$H_0$ is true		False positive Type I error
$H_0$ is false	False negative Type II error	





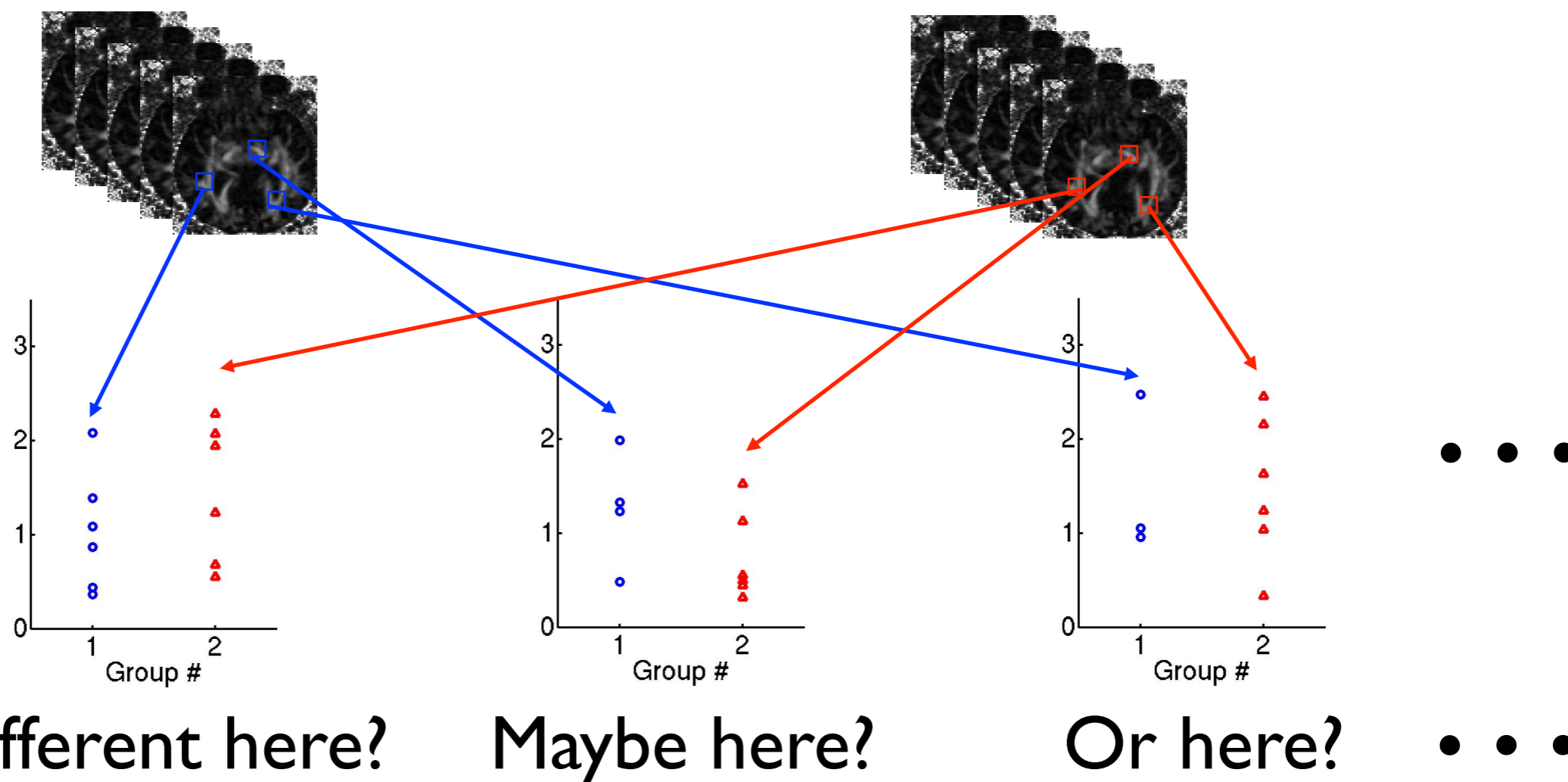
# Outline

- 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



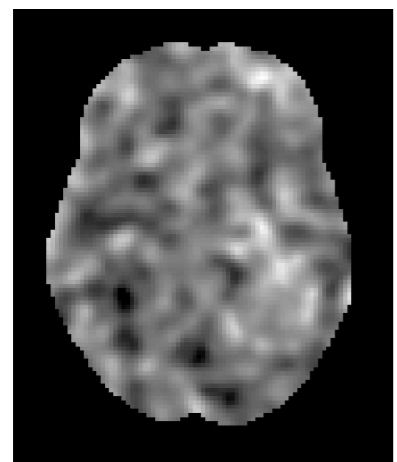
# Multiple Comparisons

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

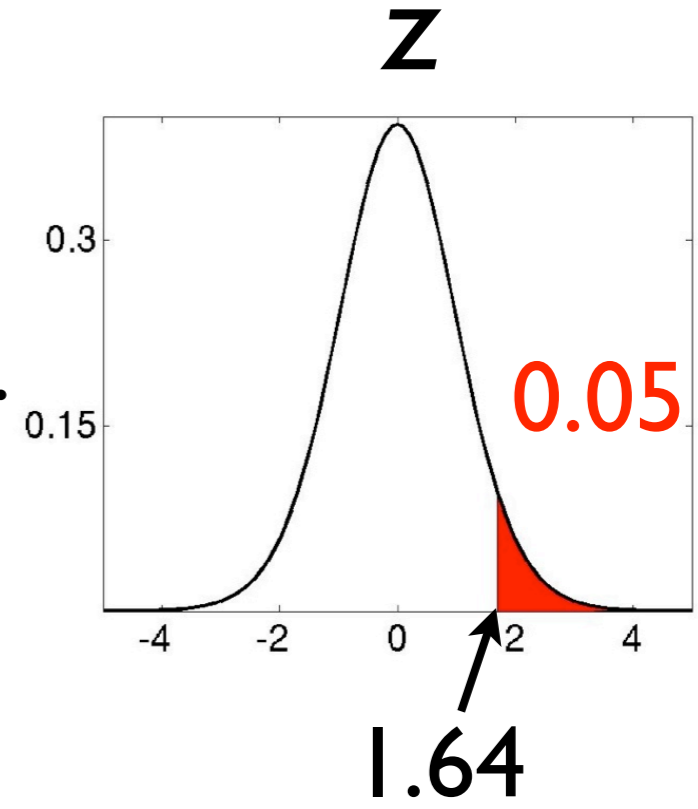




# What happens when we apply this to imaging data?



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



z-map  
thresholded at  
1.64



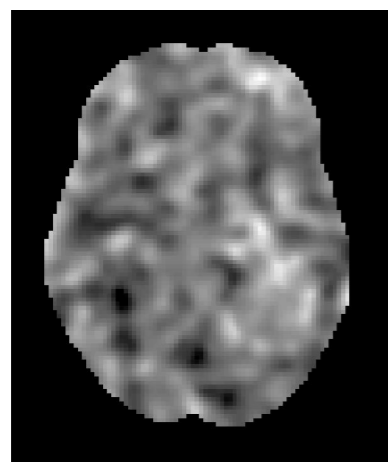
16 clusters  
288 voxels  
~5.5% of the voxels

That's a LOT of false positives



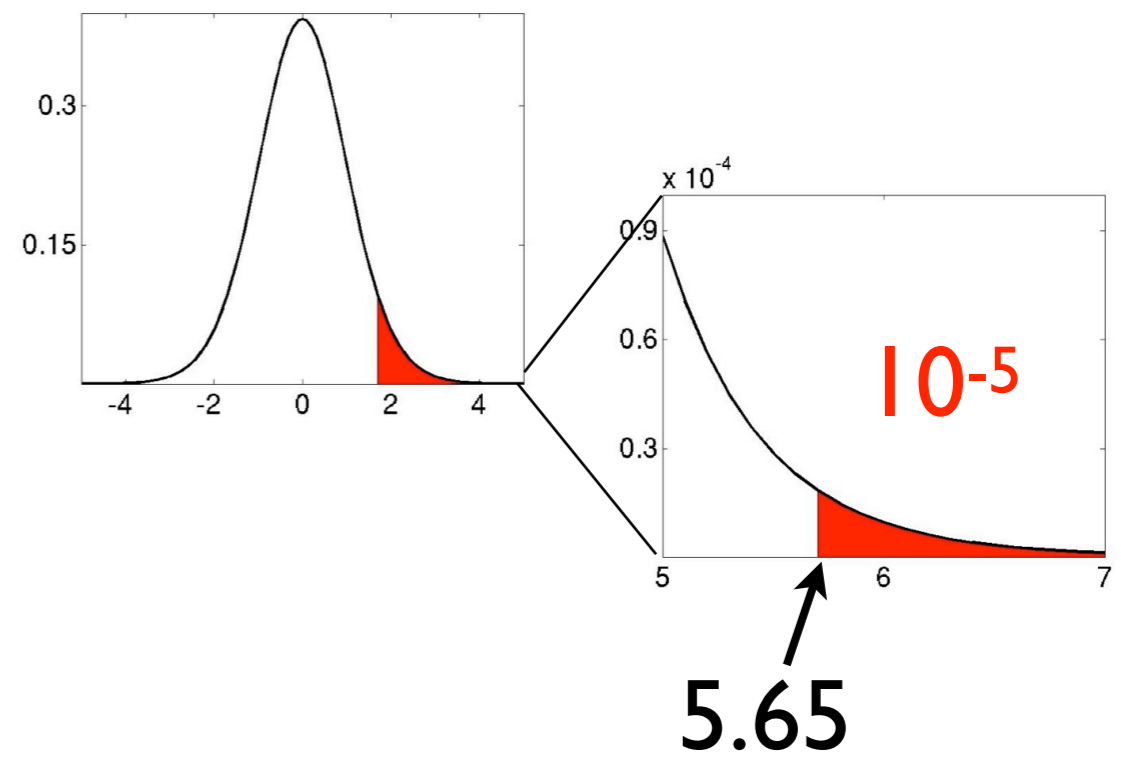
# The strict approach: Bonferroni correction

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



5255 voxels

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



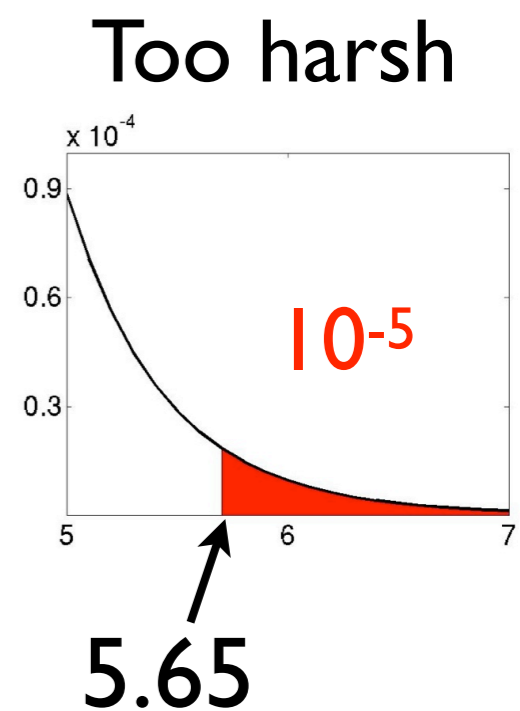
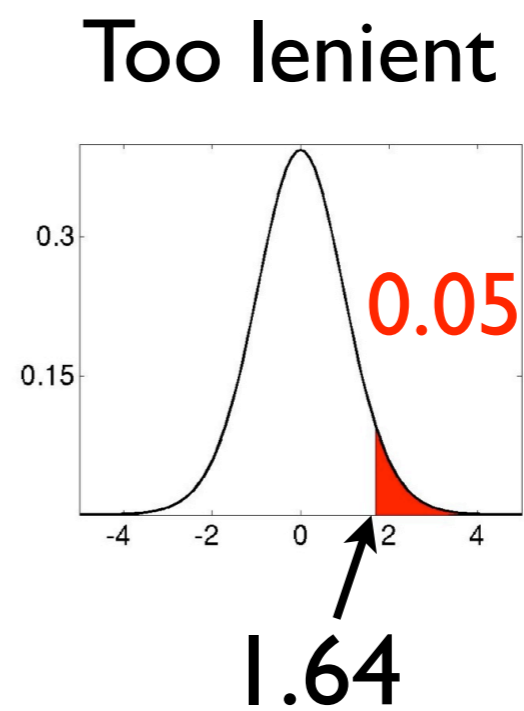
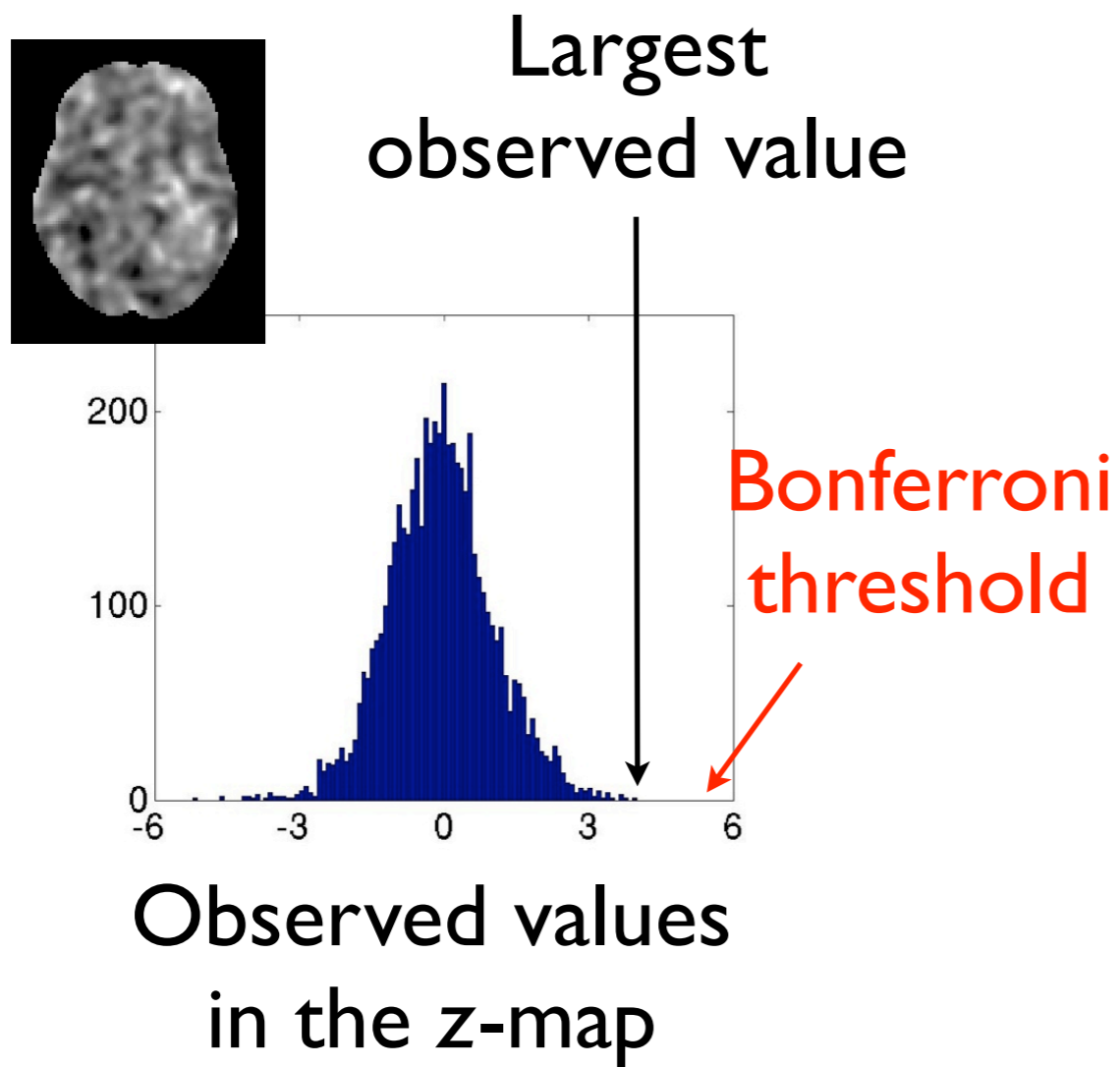
z-map  
thresholded at  
5.65



No false positives.  
Hurrah!



# But ... doesn't 5.65 sound very high?

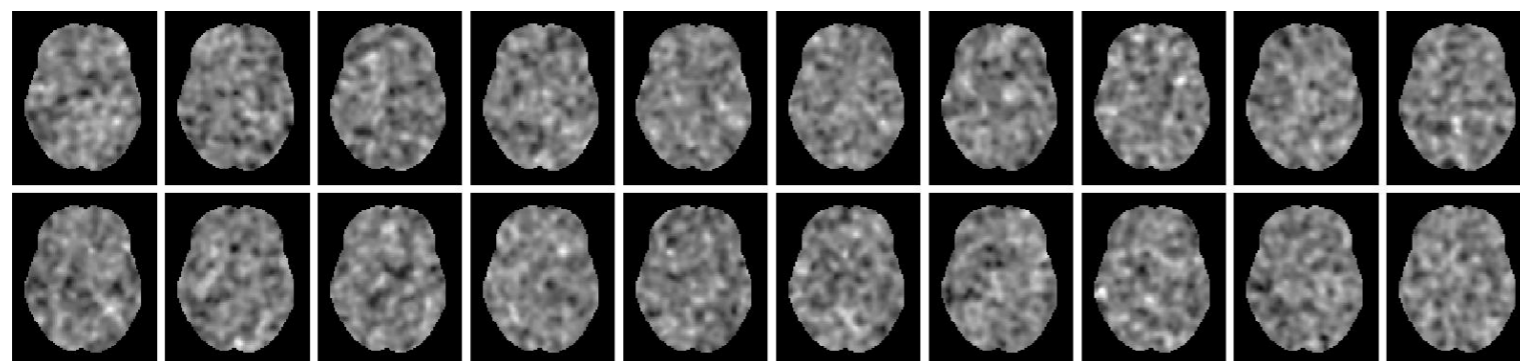


So what do we want then?



# Family-wise error

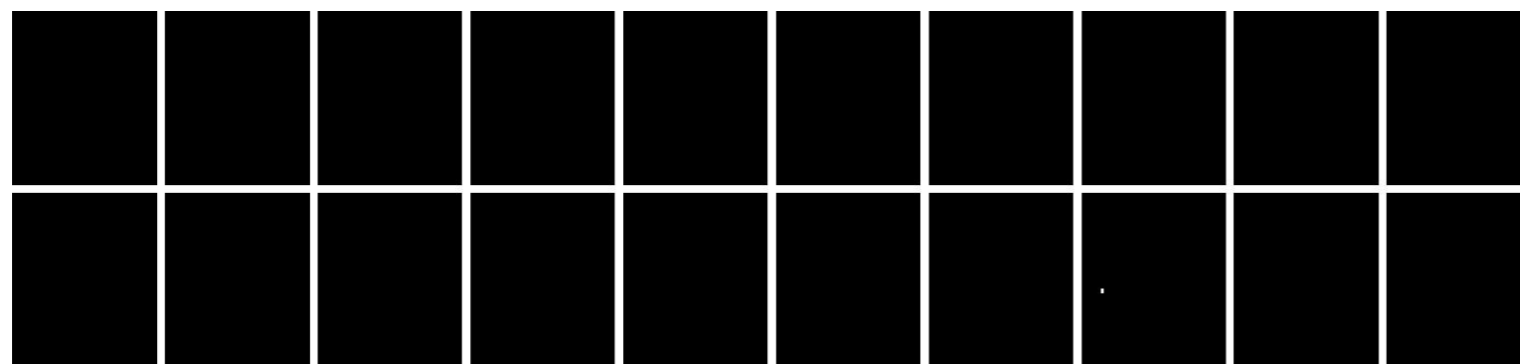
Let's say we perform a series of identical studies



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?





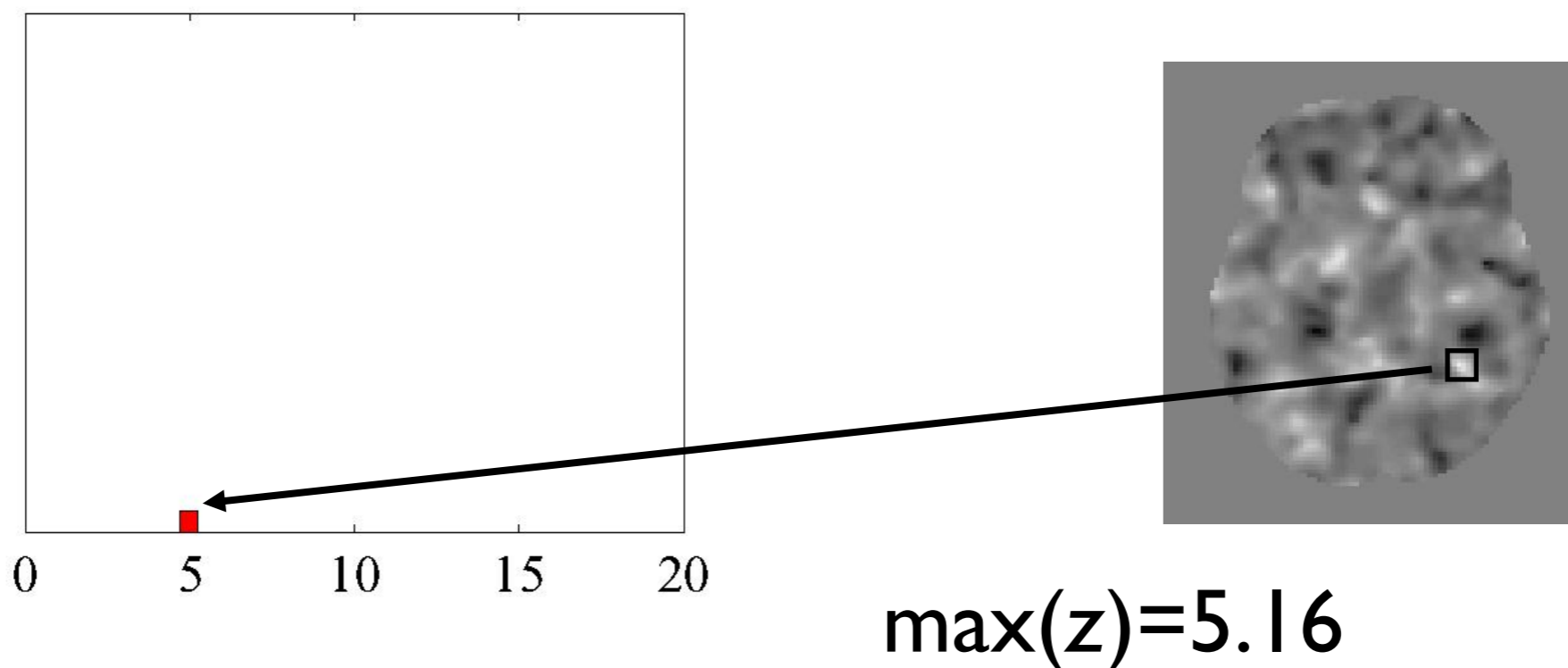
# Outline

- 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



# 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.

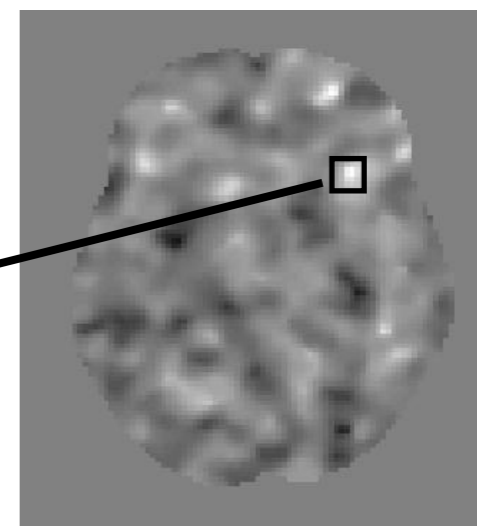
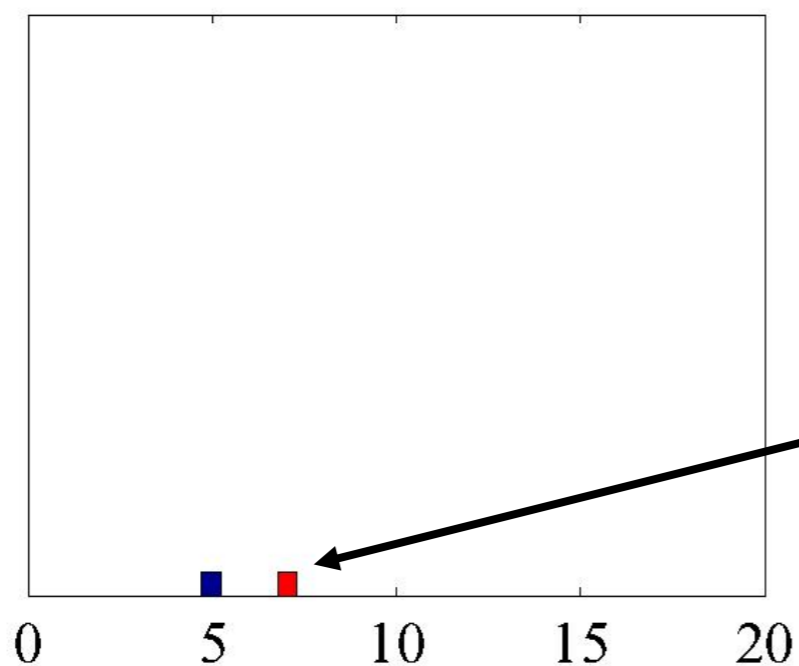






# Maximum z

- 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.

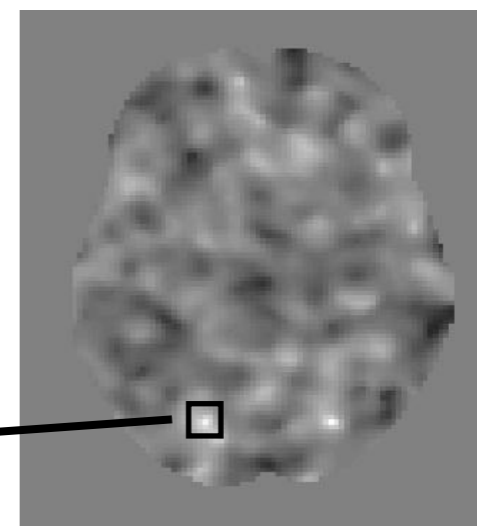
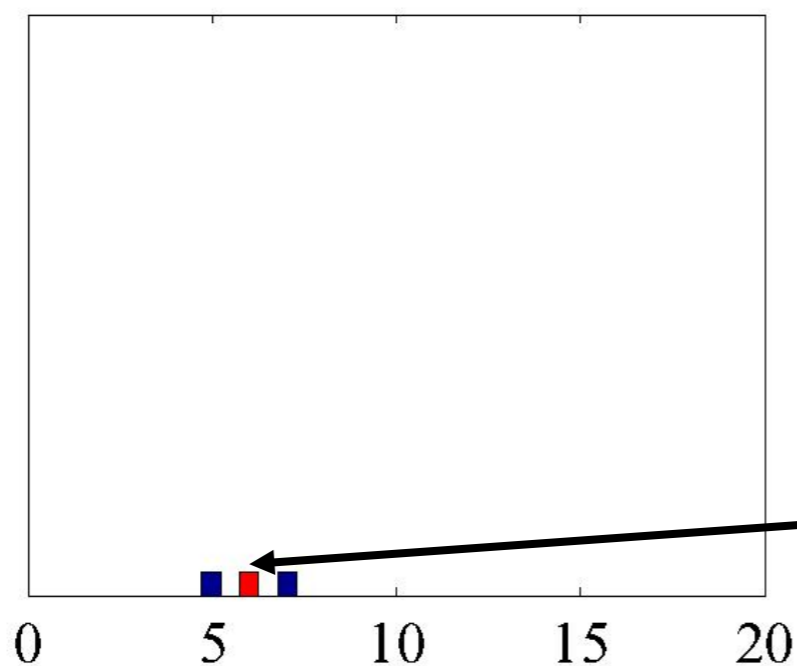


$\max(z)=6.84$



# 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.

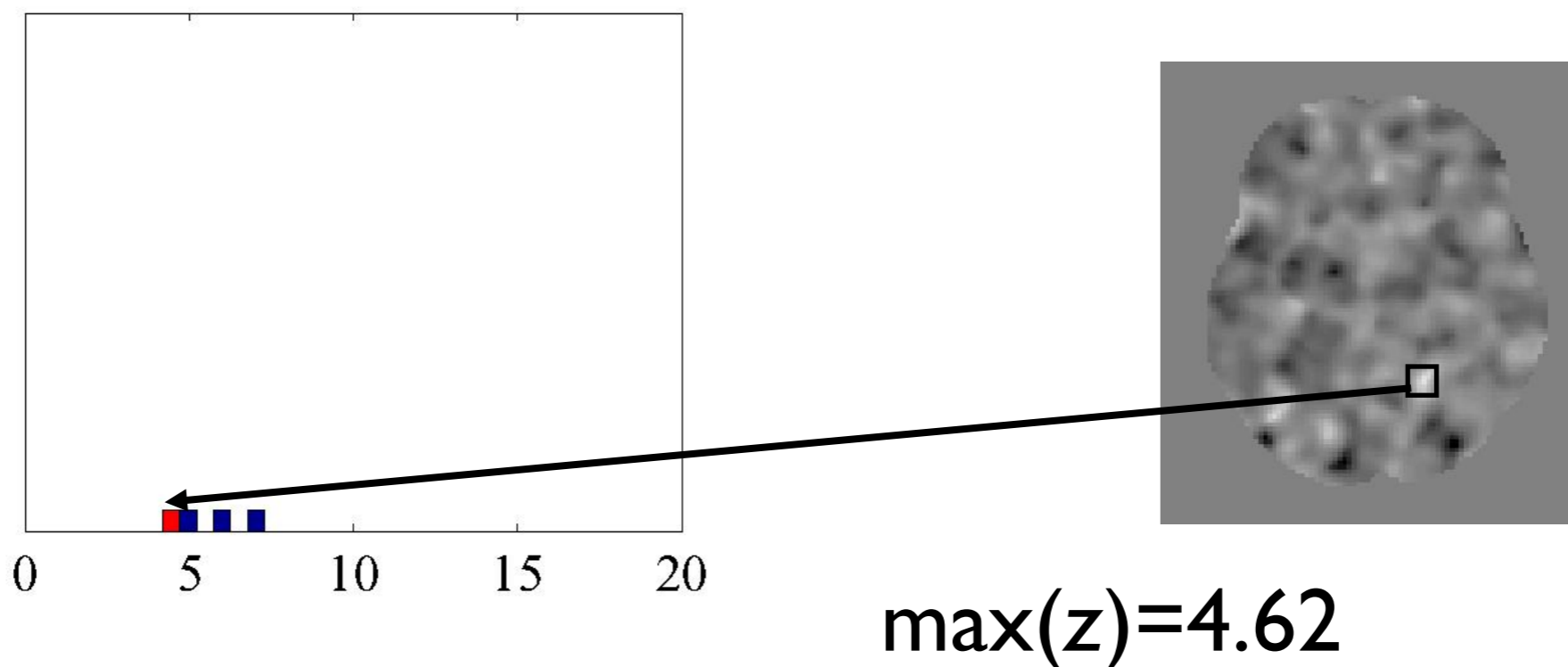


$\max(z)=5.93$



# Maximum z

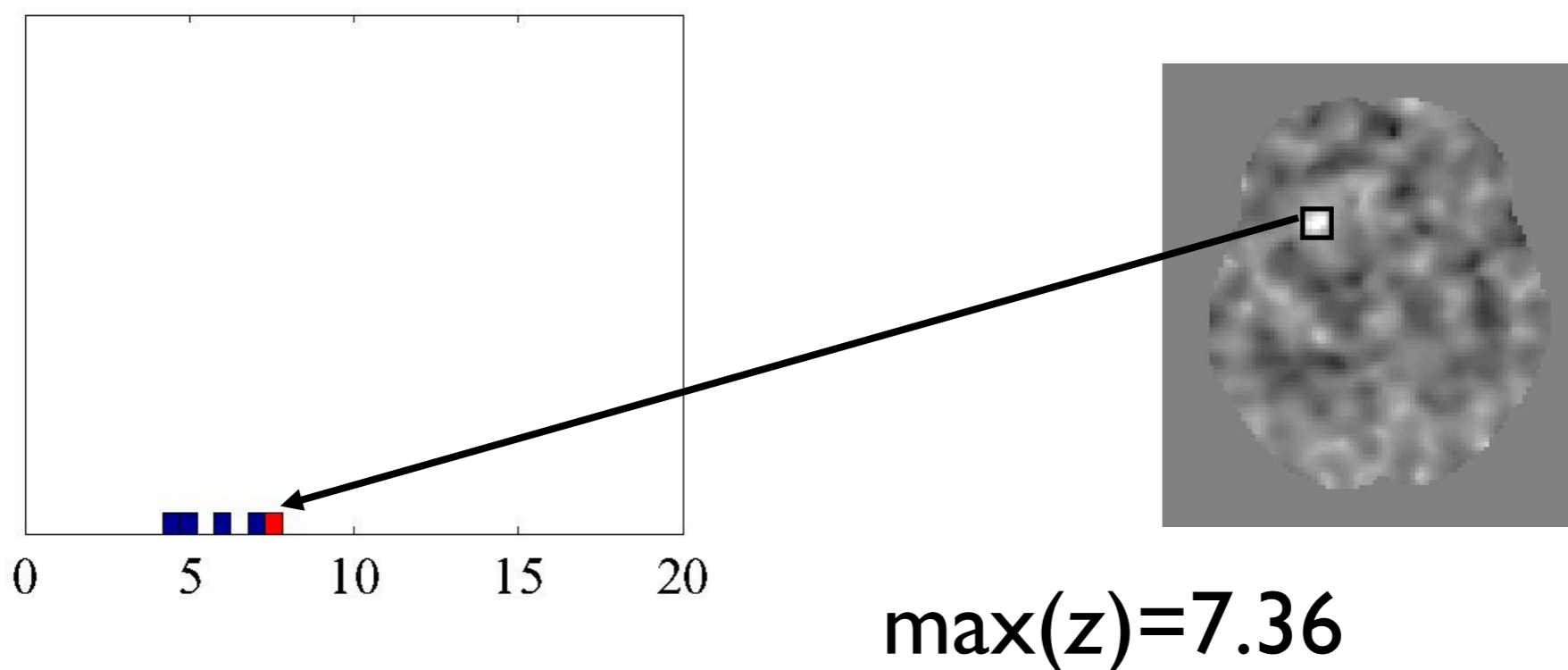
- When we want to control “family-wise error”, what do we in practice want?
- If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
- And if we reject anything, we will definitely reject the most “extreme” value in the brain.





# Maximum z

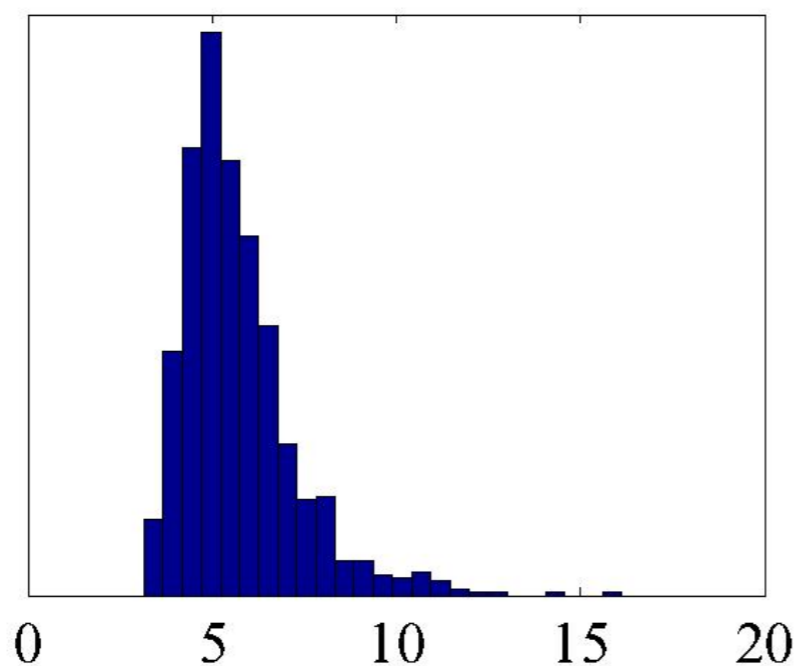
- When we want to control “family-wise error”, what do we in practice want?
- If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
- And if we reject anything, we will definitely reject the most “extreme” value in the brain.





# 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.



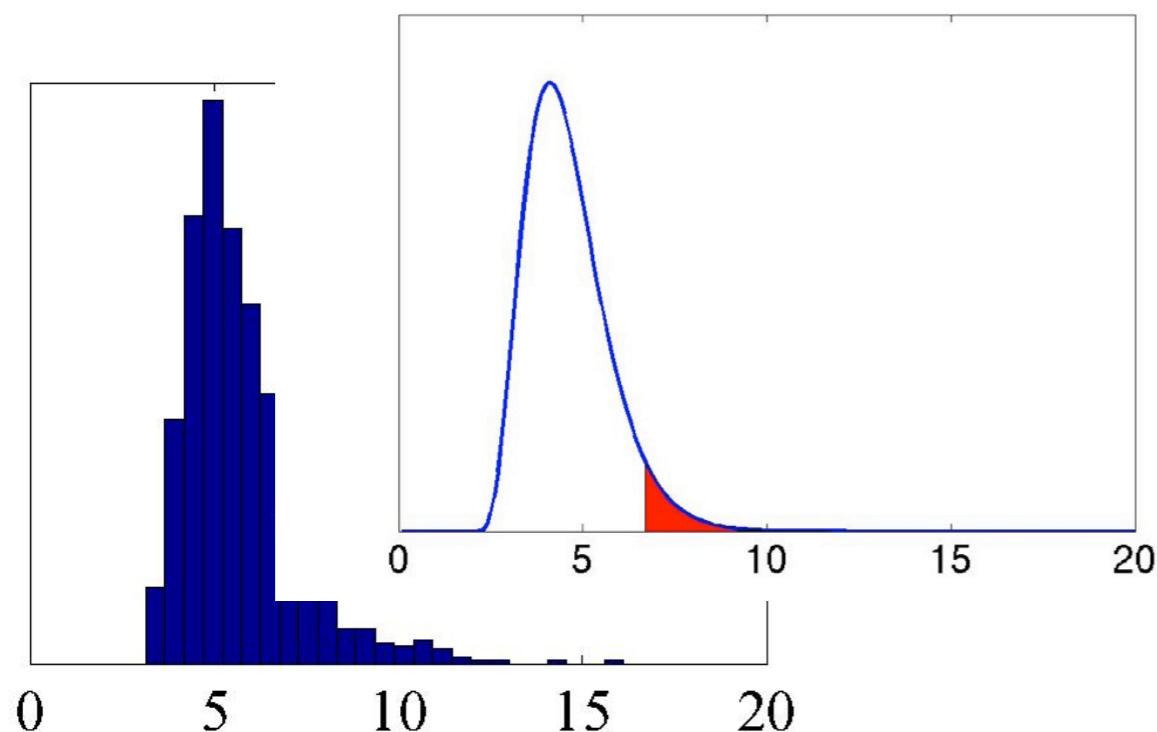
Etc...



# 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.

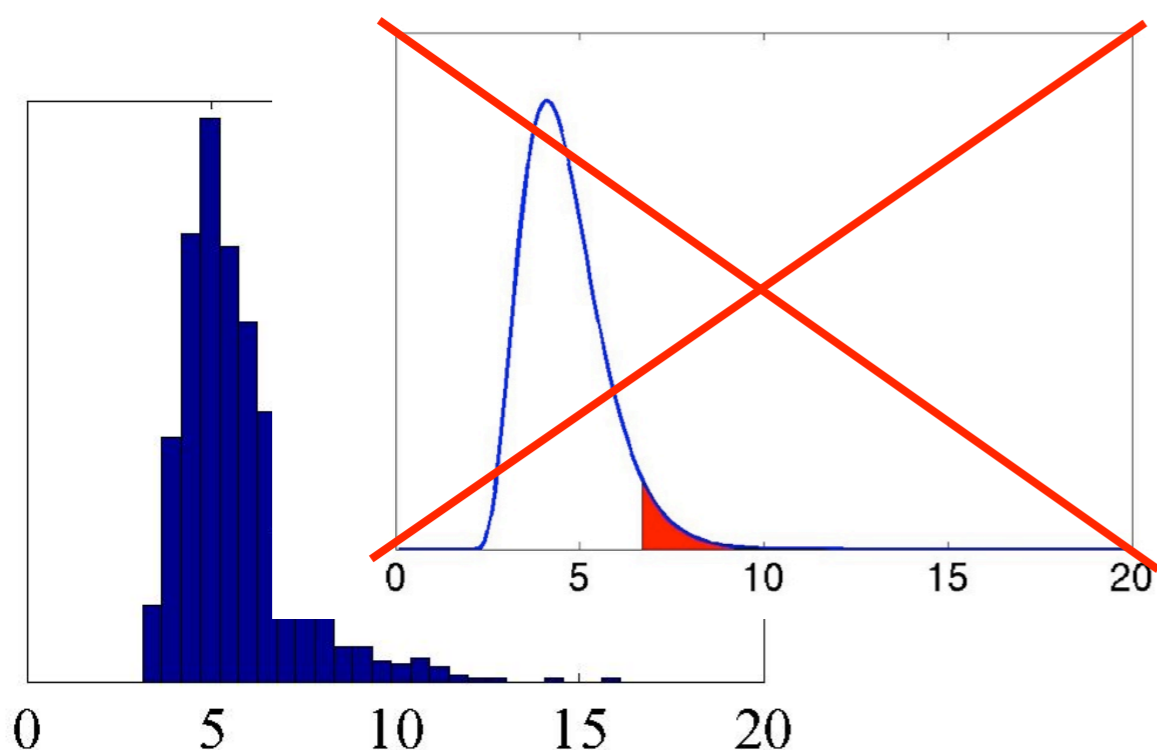
This is the distribution we want to use for our FWE control.





# 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! 😞



# Outline

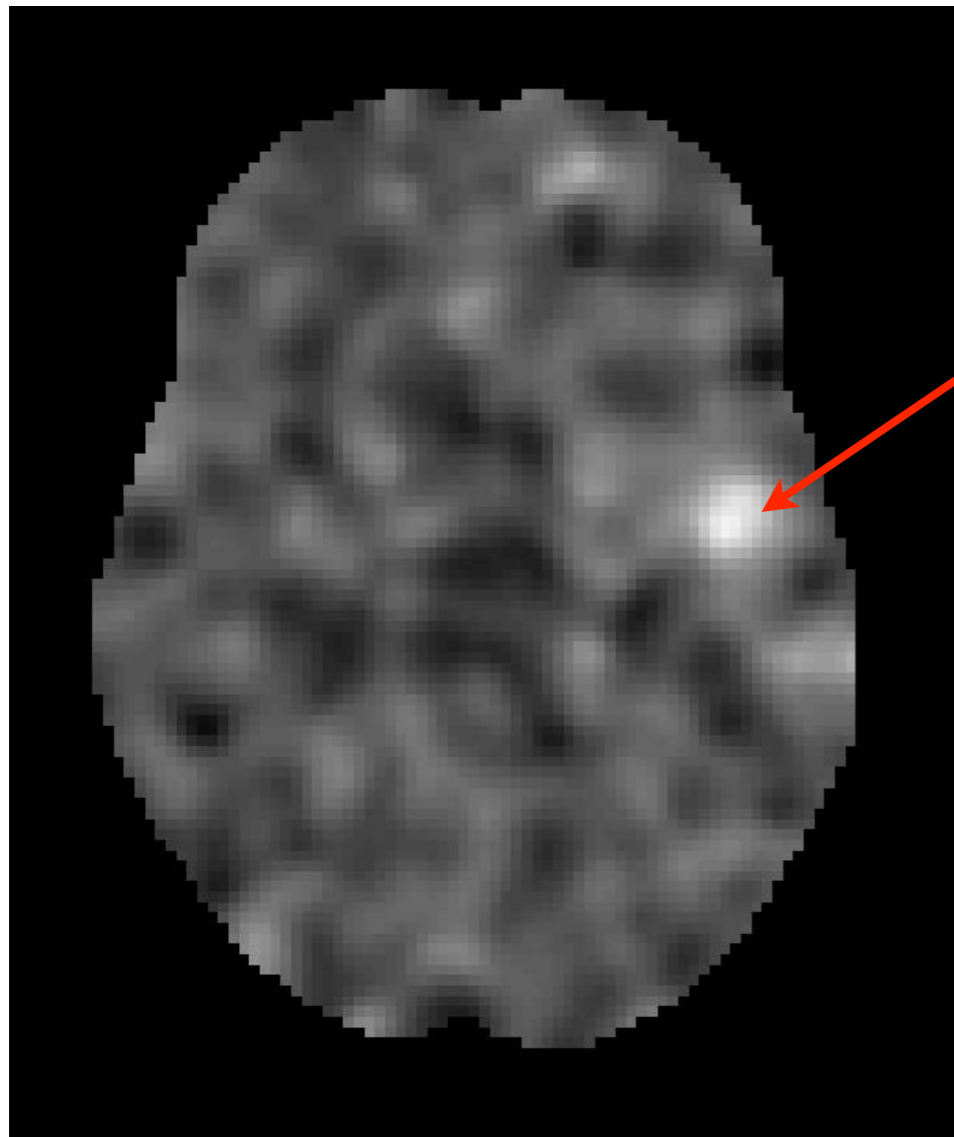
- Null-hypothesis and Null-distribution
- Multiple comparisons and Family-wise error
- **Different ways of being surprised**
  - Voxel-wise inference (Maximum  $z$ )
  - **Cluster-wise inference (Maximum size)**
- Parametric vs non-parametric tests
- Enhanced clusters
- FDR - False Discovery Rate





# Spatial extent: another way to be surprised

This far we have talked about voxel-based tests

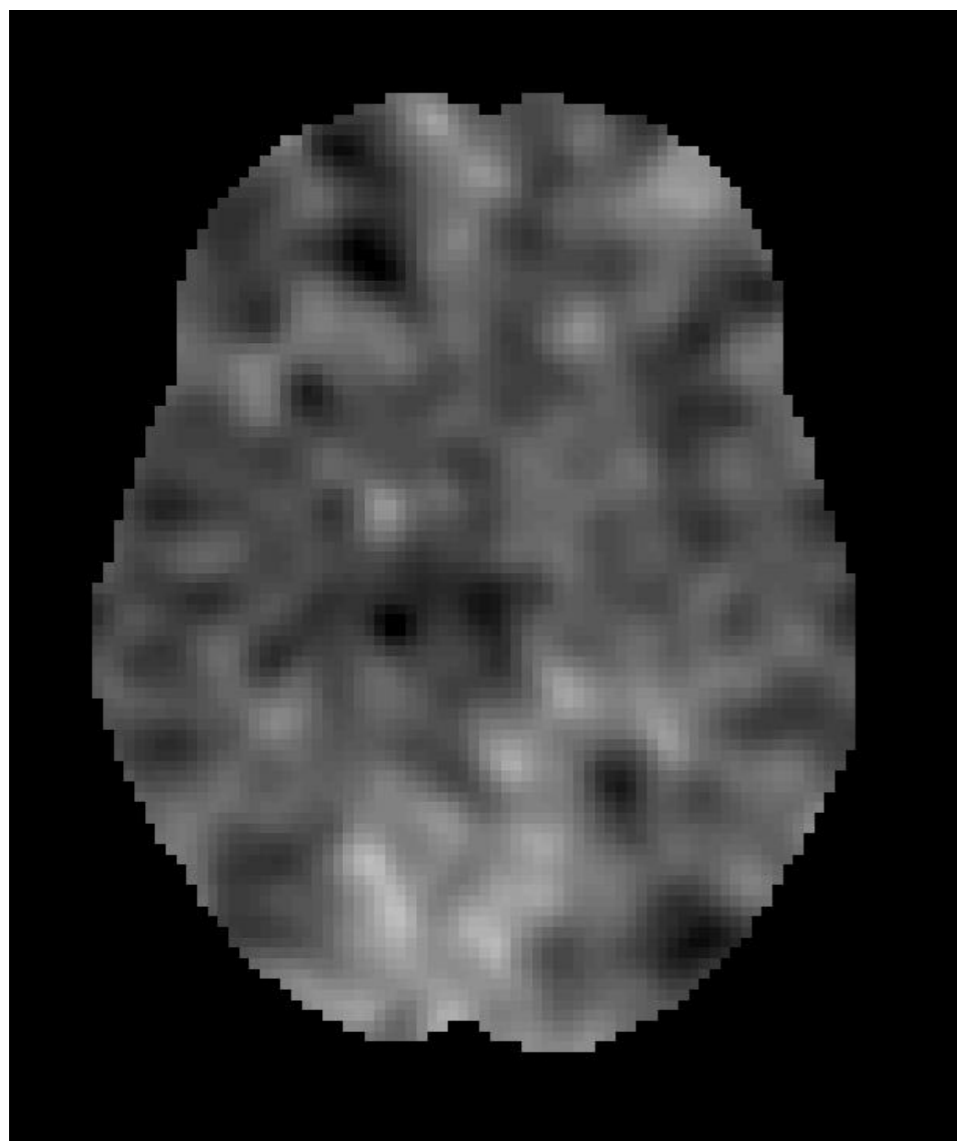


We say: Look! A z-value of 7. That is so surprising (under the null-hypothesis) that I will have to reject it. (Though we are of course secretly delighted to do so)



# Spatial extent: another way to be surprised

But sometimes our data just aren't that surprising.

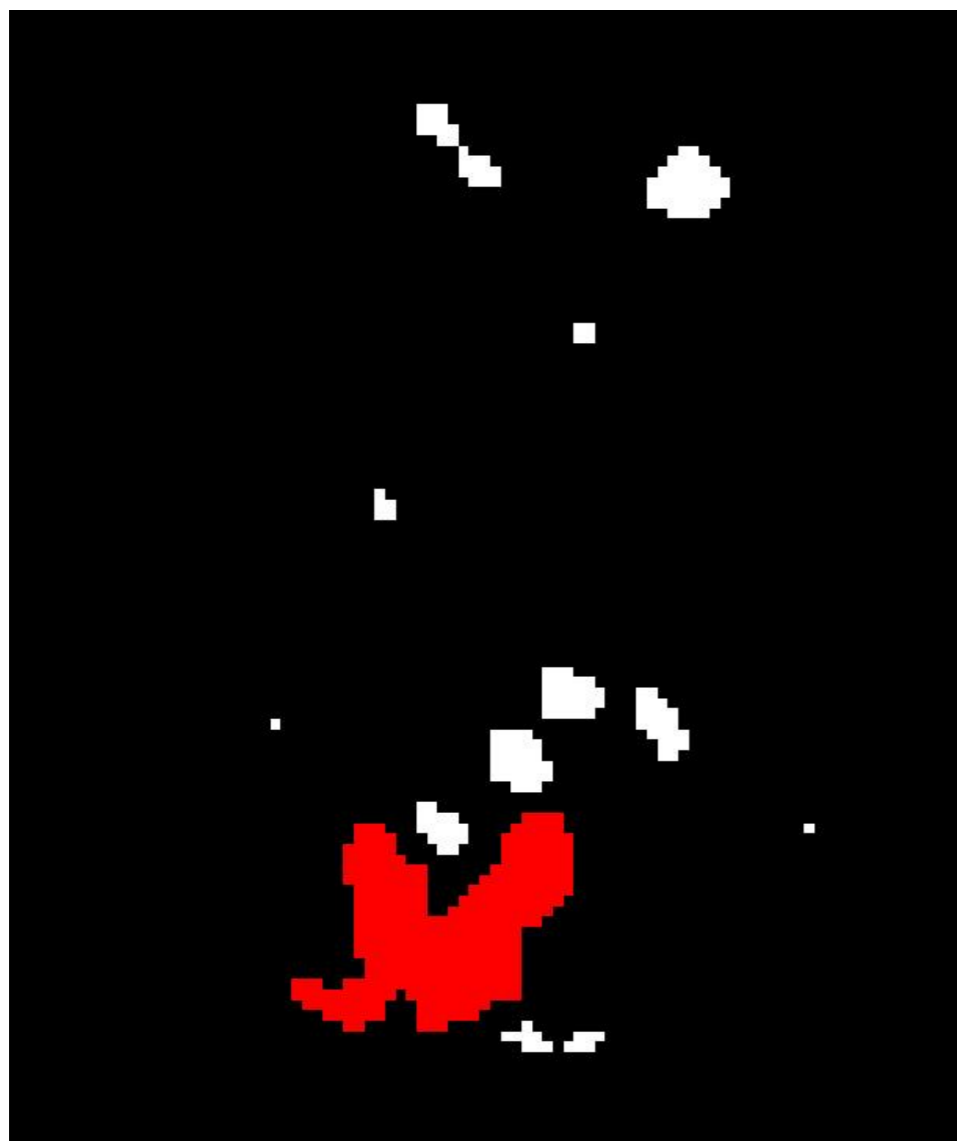


Nothing surprising here! The largest z-value is  $\sim 4$ . We cannot reject the null-hypothesis, and we are **devastated**.



# 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

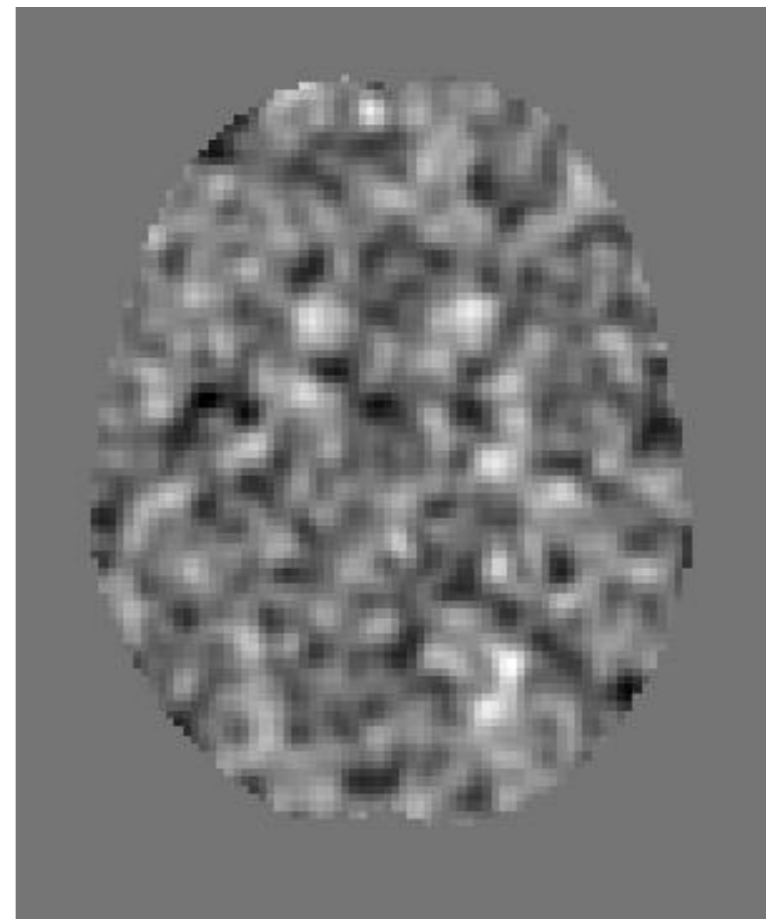


We say: Look at that **whopper!** 301 connected voxels all with z-values  $> 2.3$ . That is really surprising (under the null-hypothesis). I will have to reject it.



# Distribution of Max Cluster Size

As with the z-values we need a “null-distribution”. What would that look like in this case?

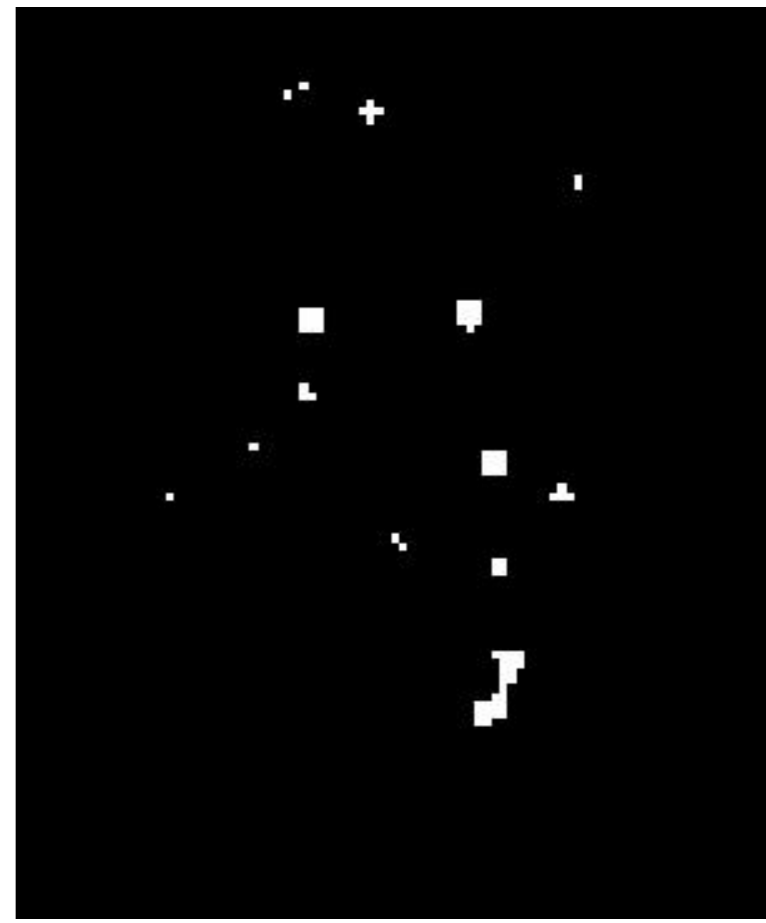


Let's say we have acquired some data



# 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.

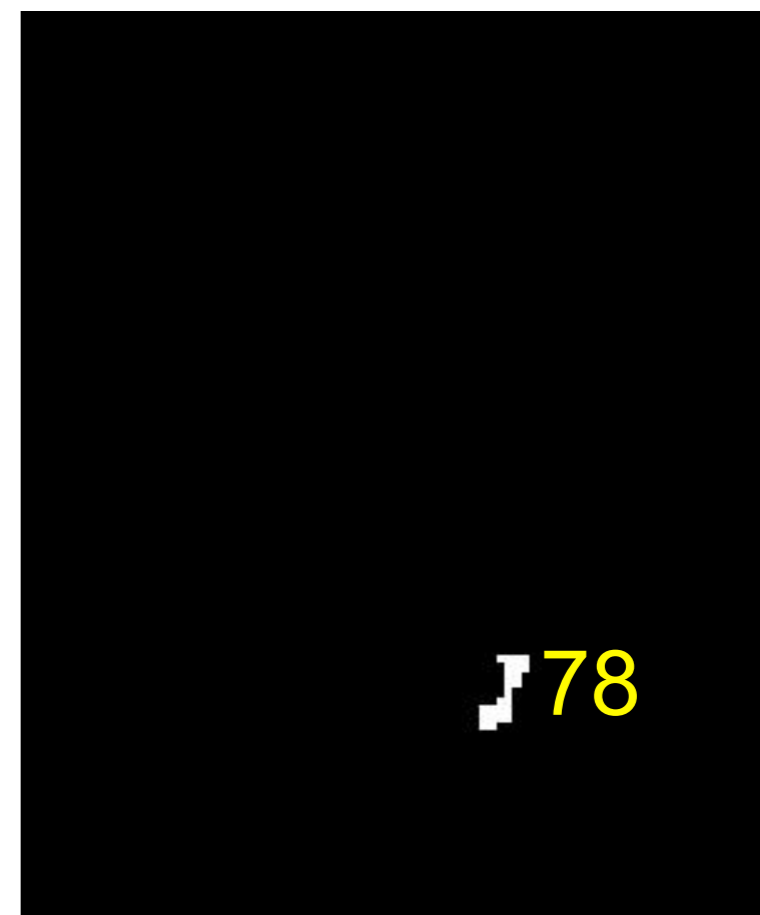


Threshold the z-map at 2.3 (arbitrary)



# 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.

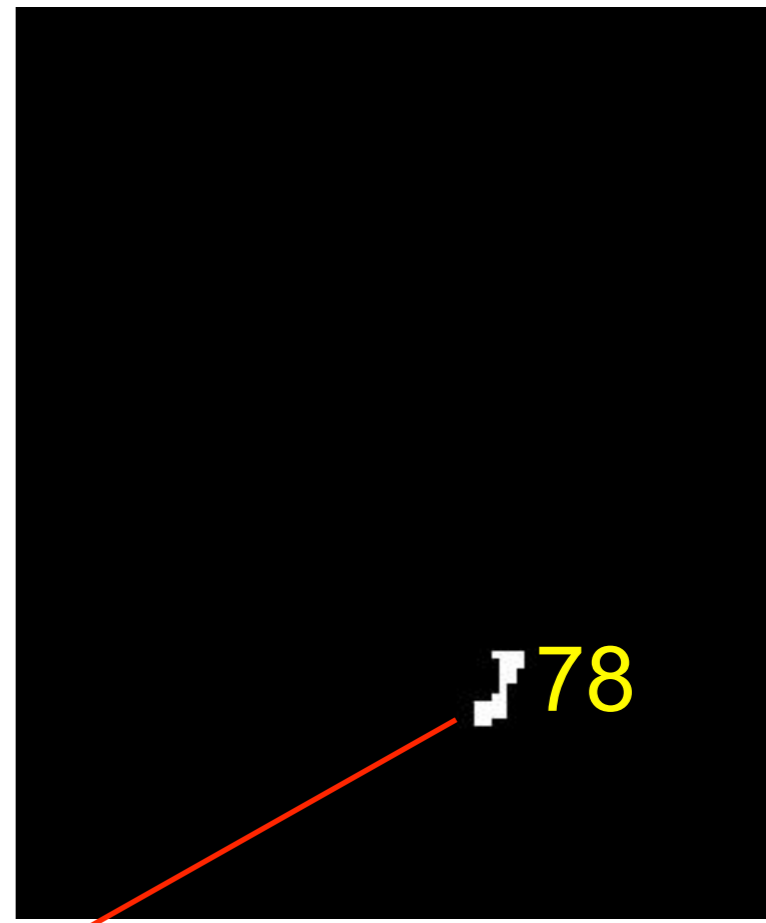
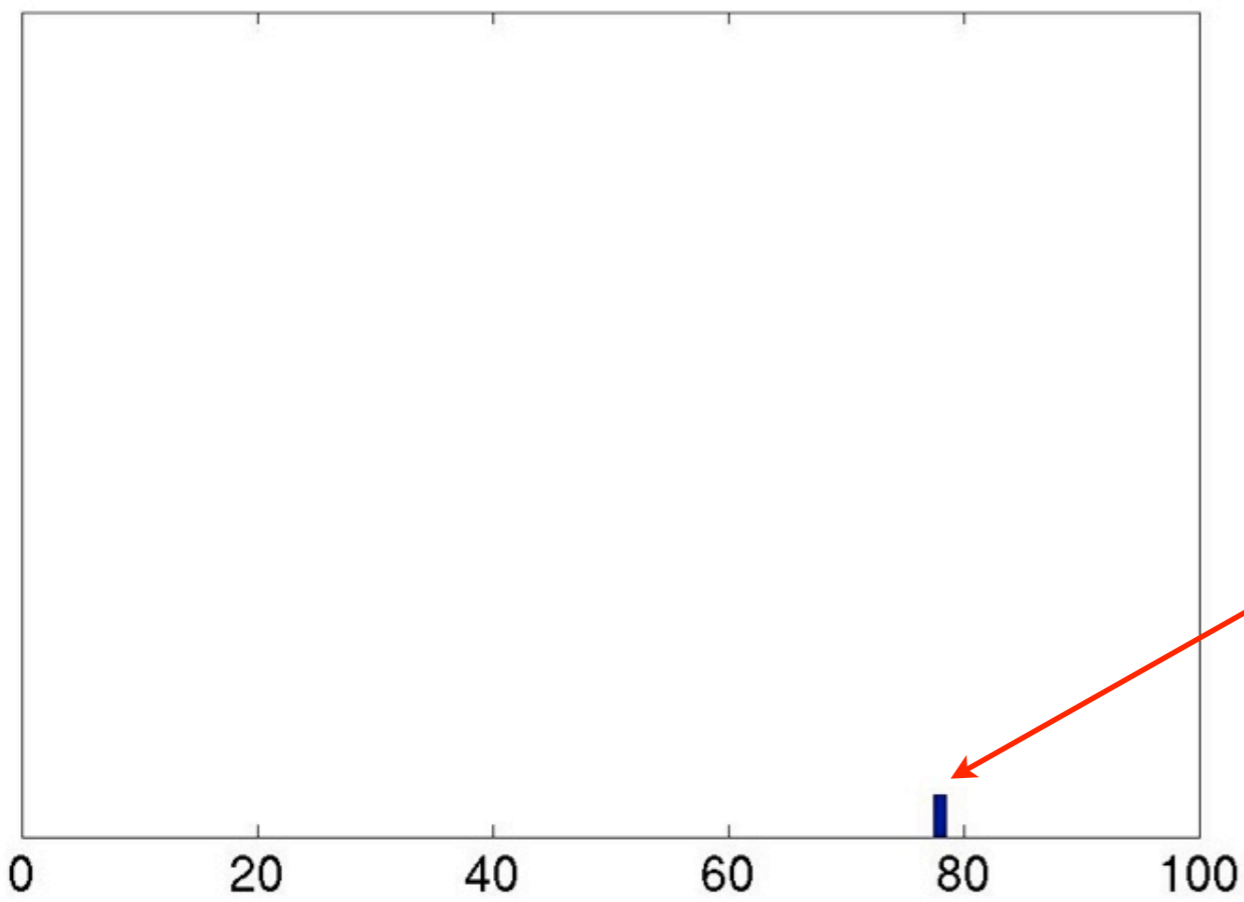


Locate the largest cluster anywhere in the brain.



# 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.

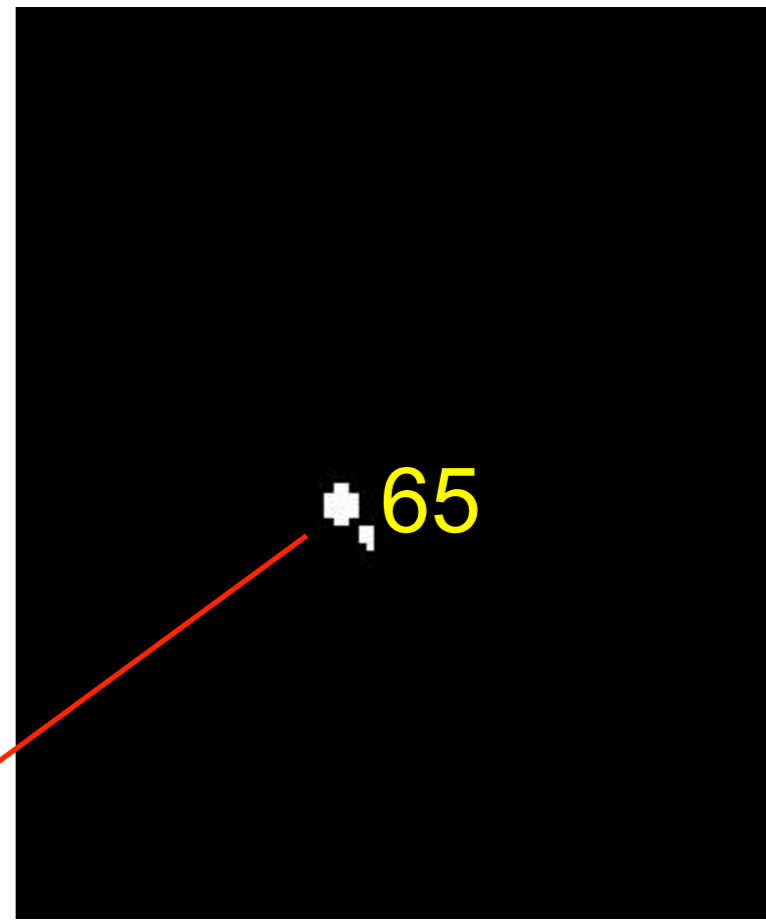
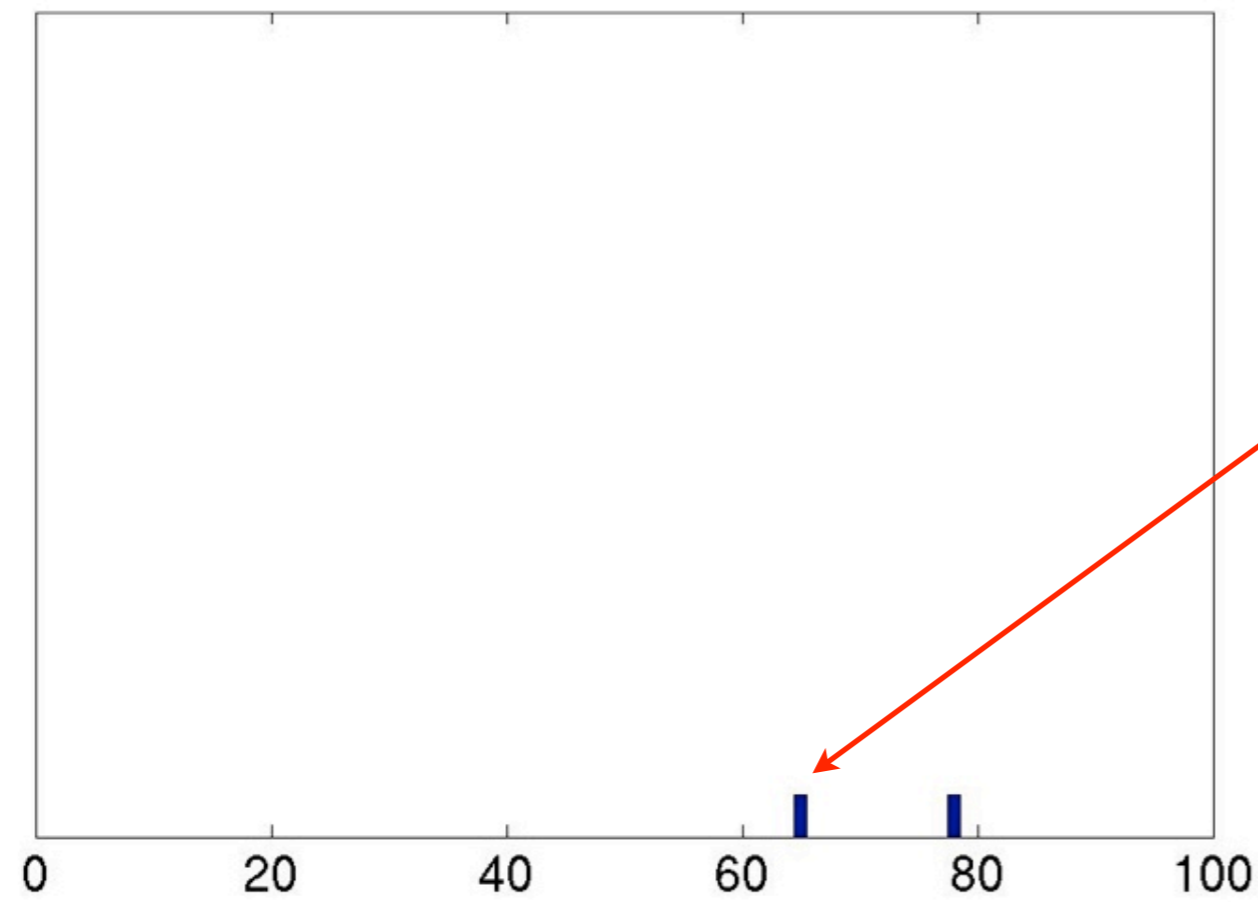


And record how large it is.



# 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.



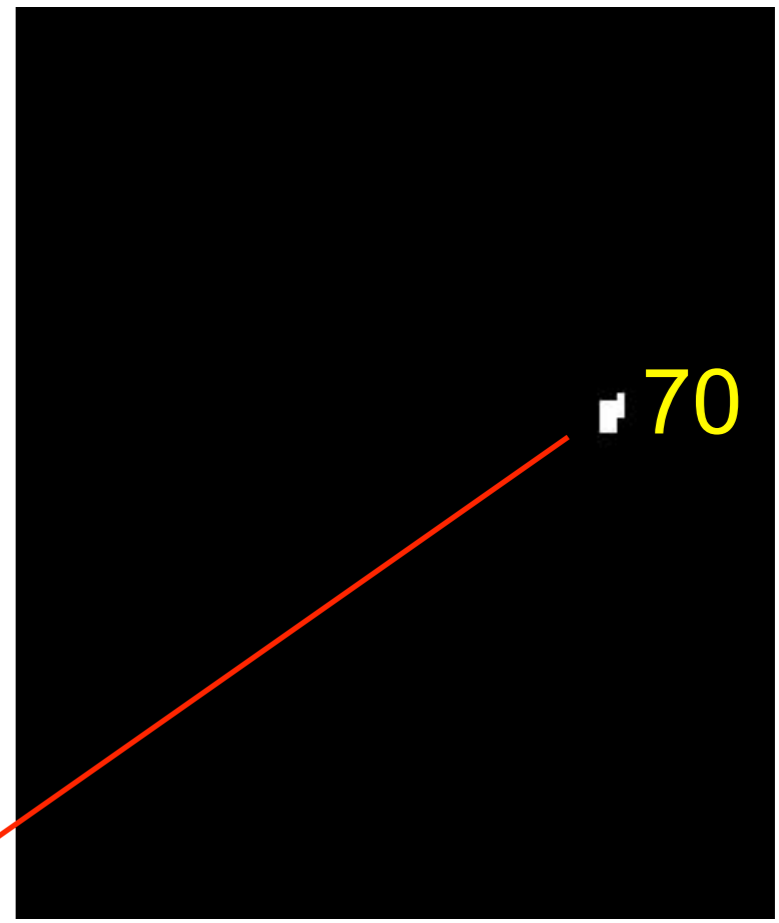
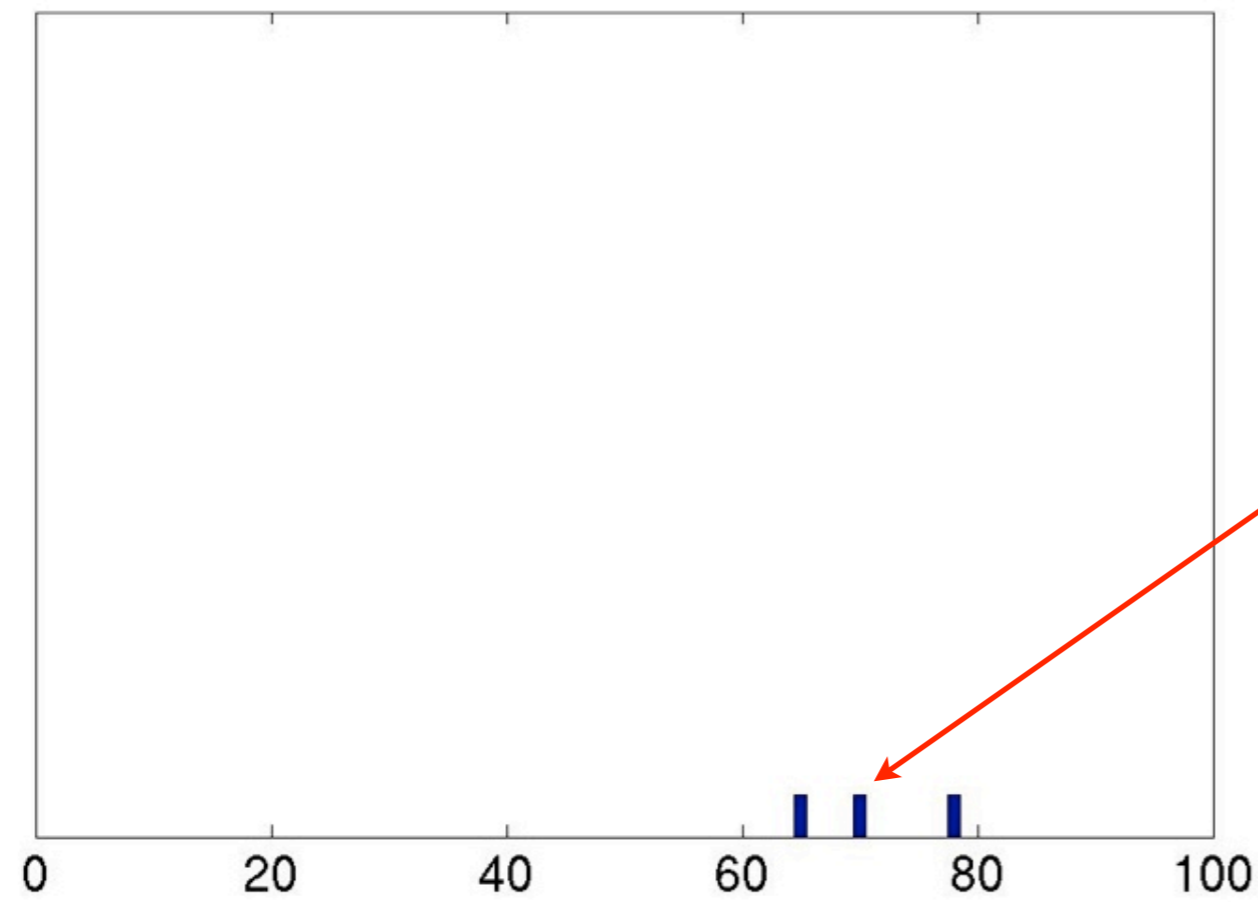
And do the same for another experiment...





# 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.

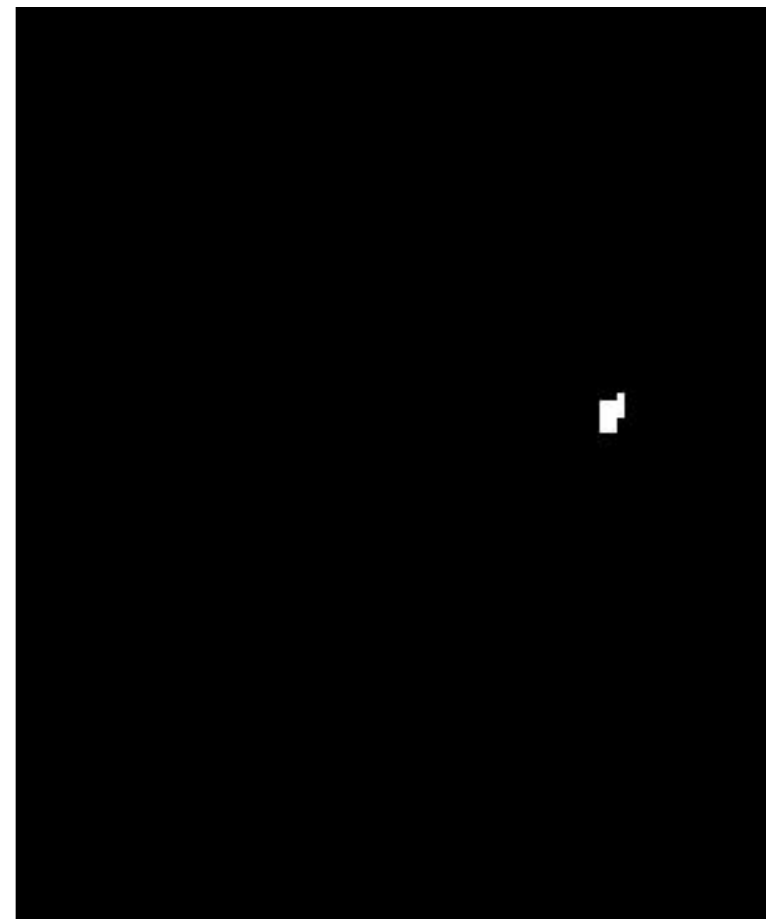
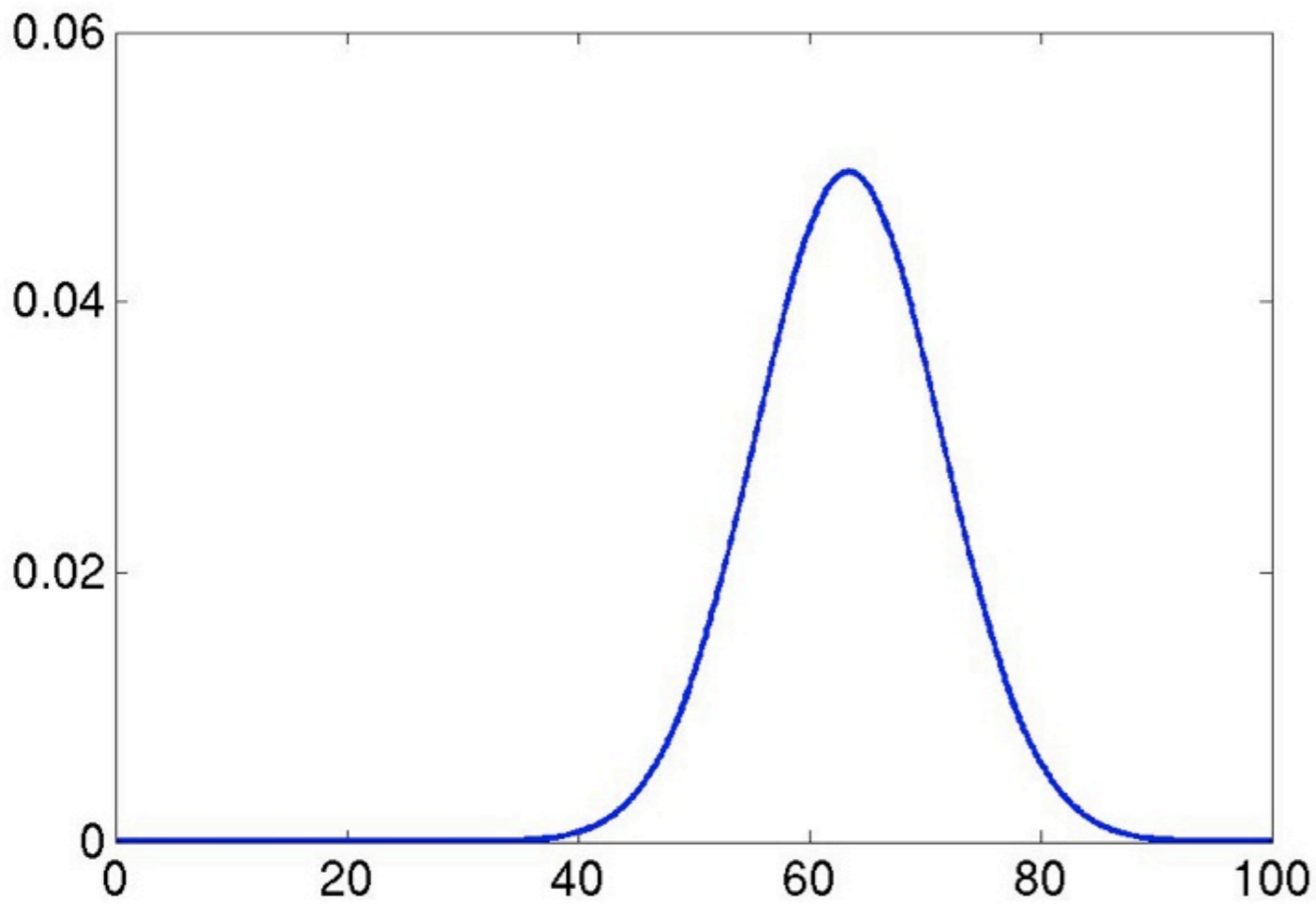


Etc ...



# 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.



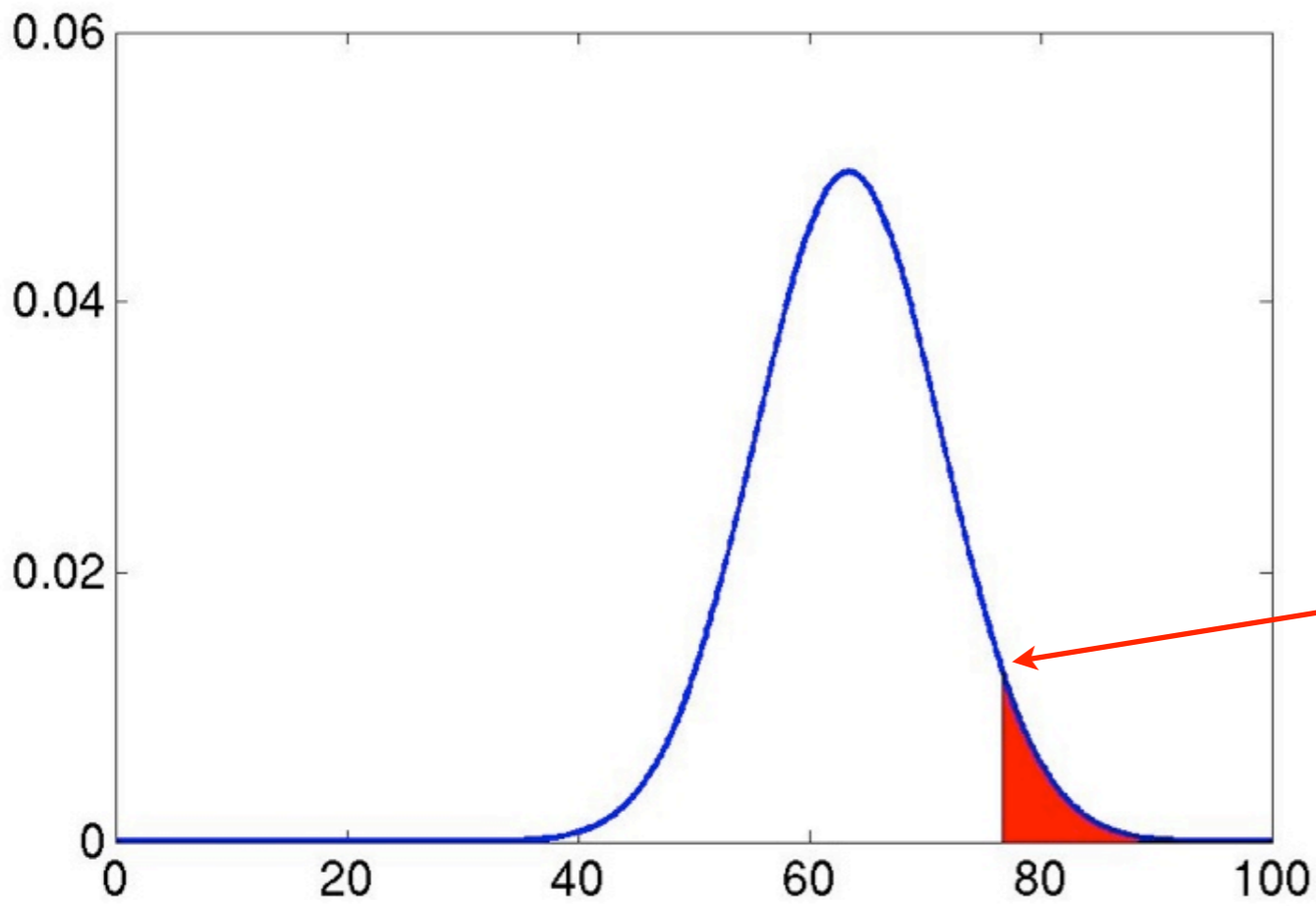
Until we have ...



# 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.

If we find a cluster larger than 76 voxels we reject the null-hypothesis.



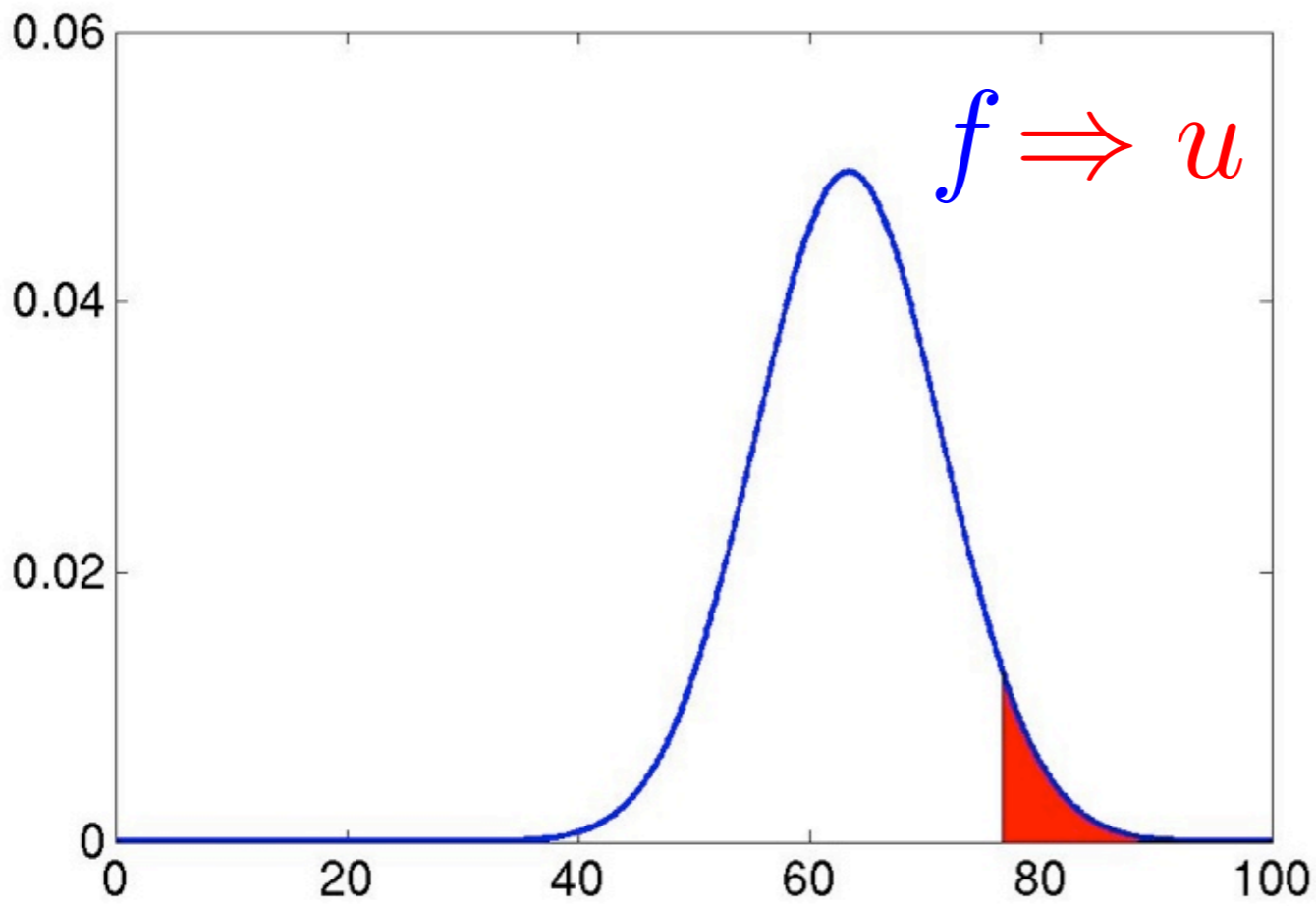
And this (76) is the level we want to threshold at



# 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?

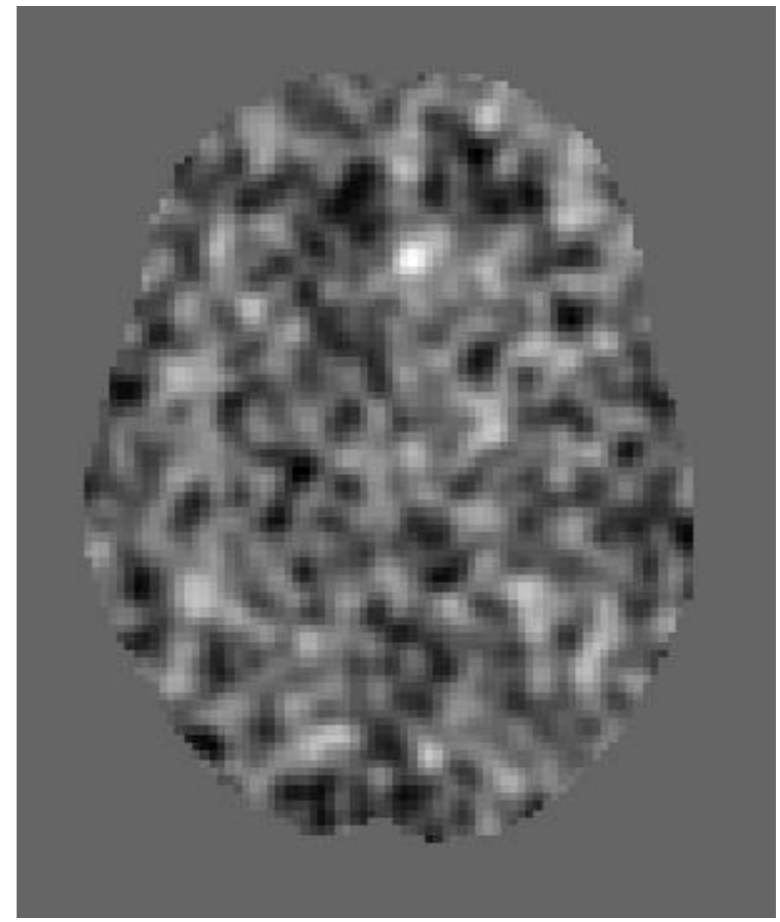




# Distribution of Max Cluster Size

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?



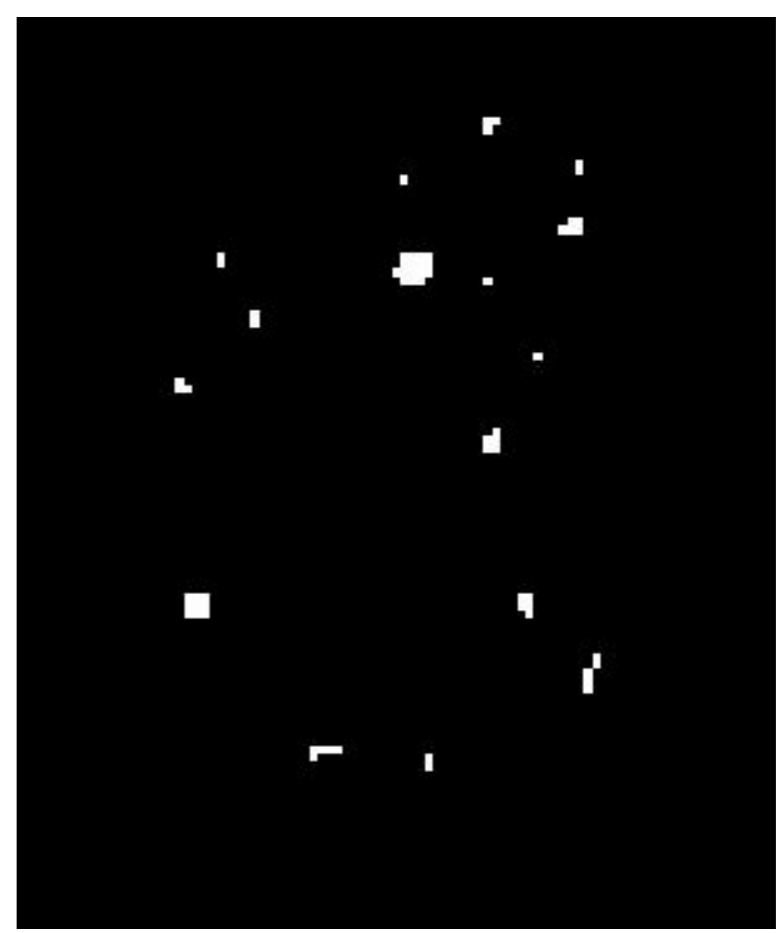
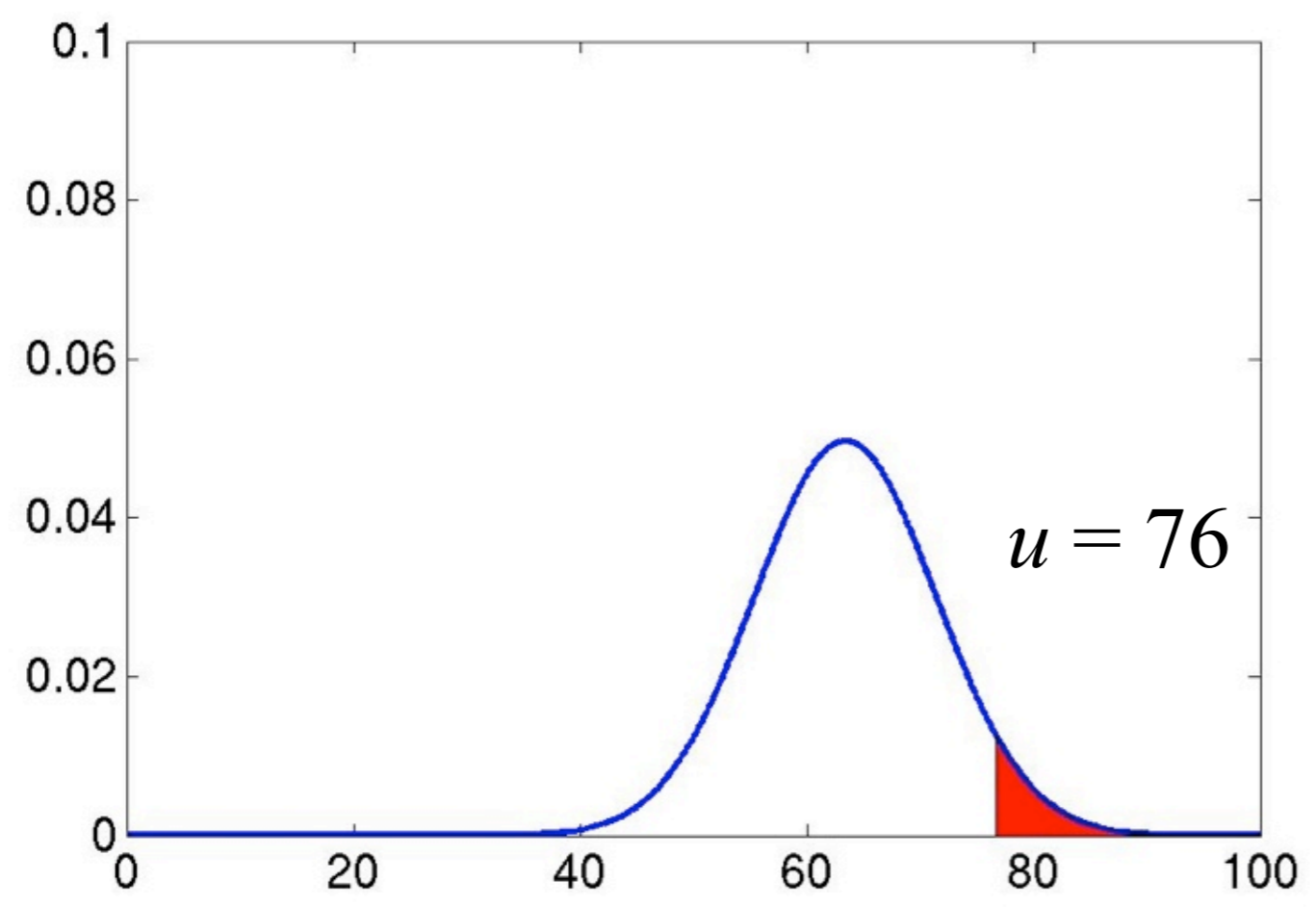
$z = 2.3$



# 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?



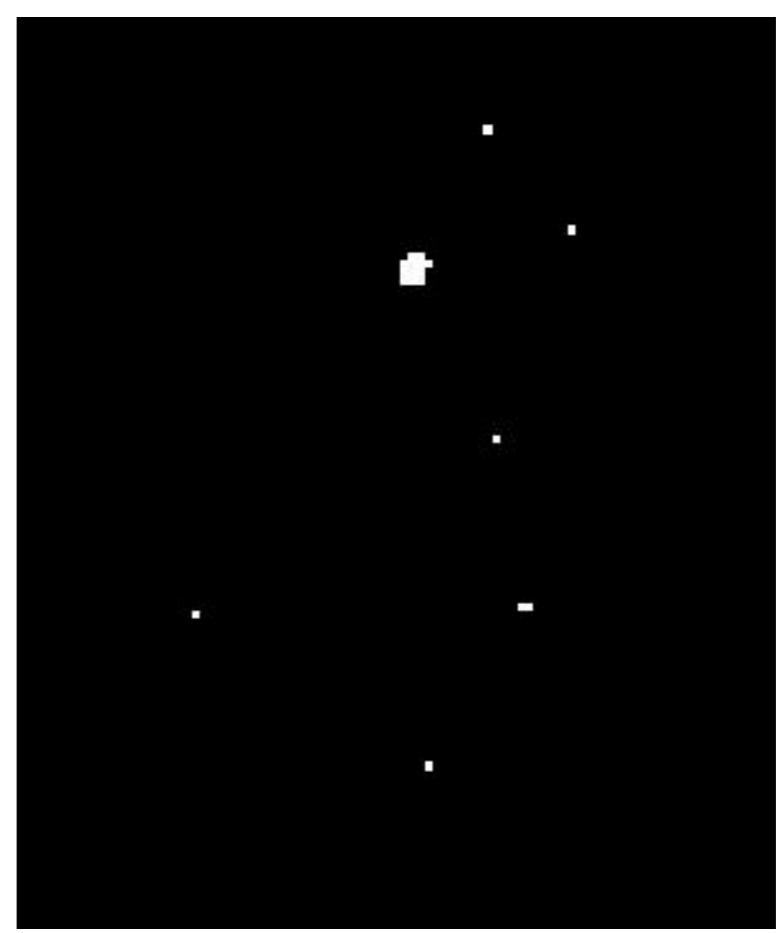
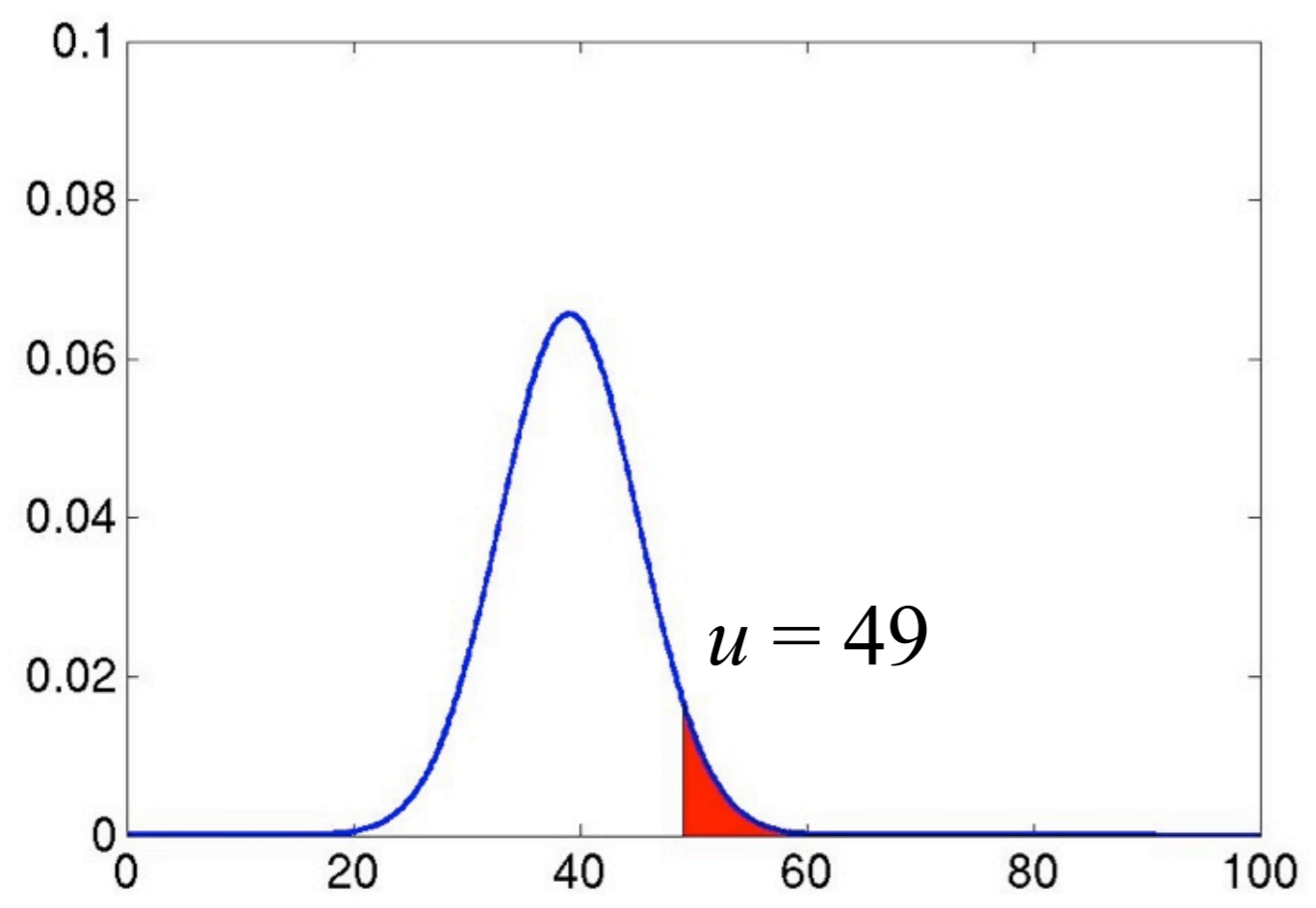
$z = 2.3$



# 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?



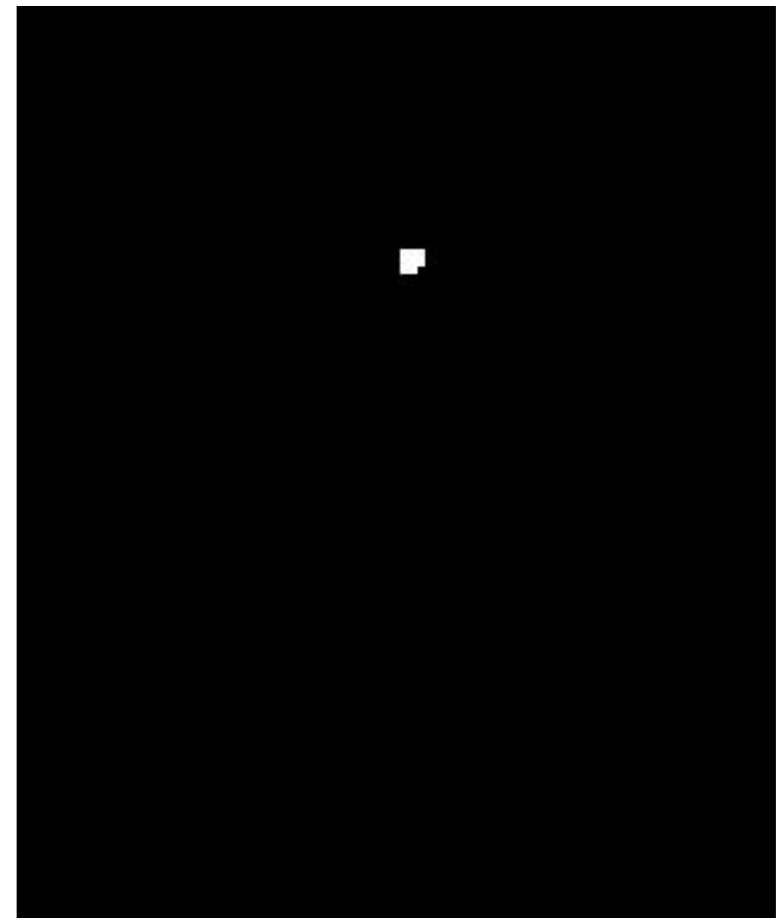
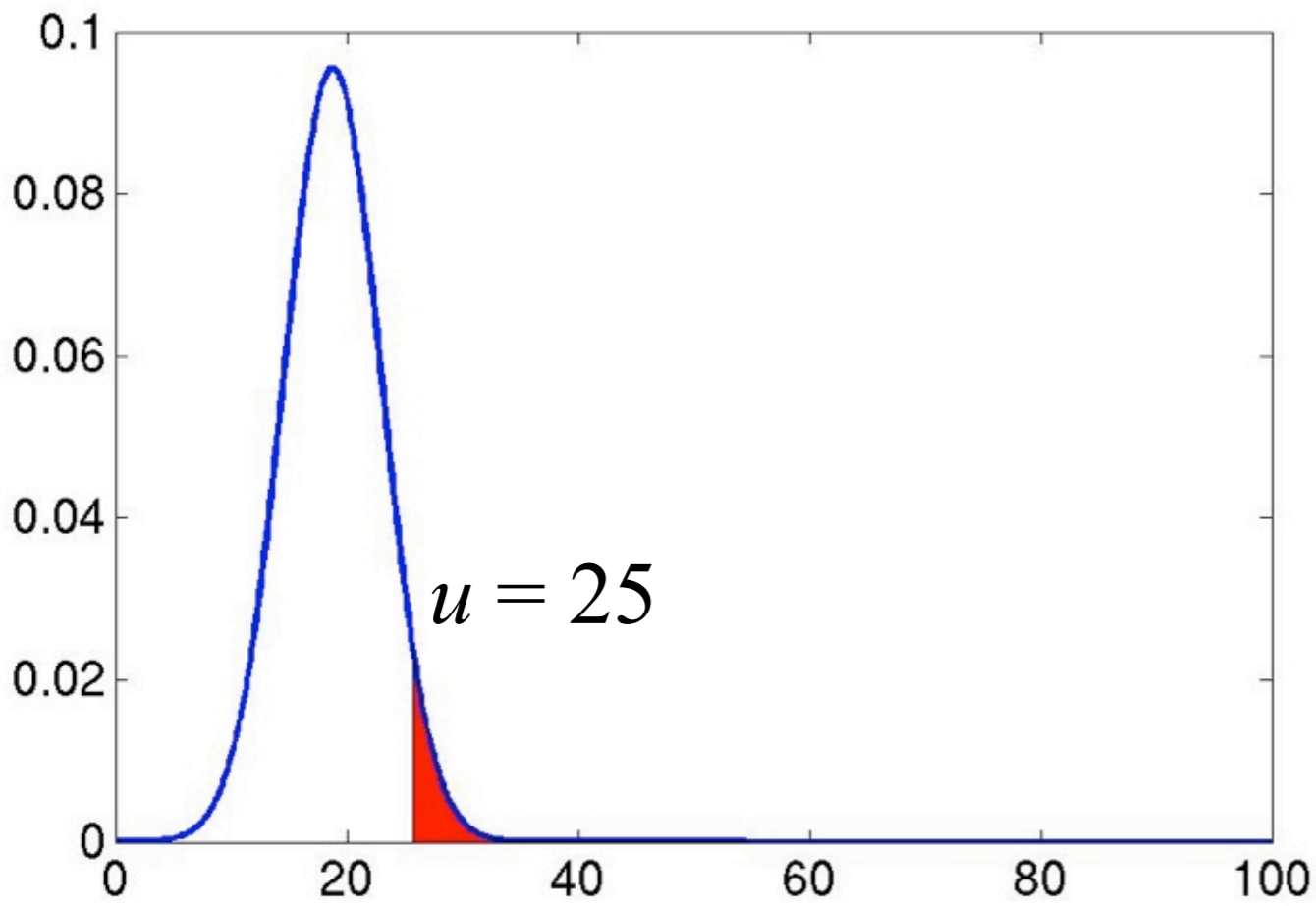
$z = 2.7$



# 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?



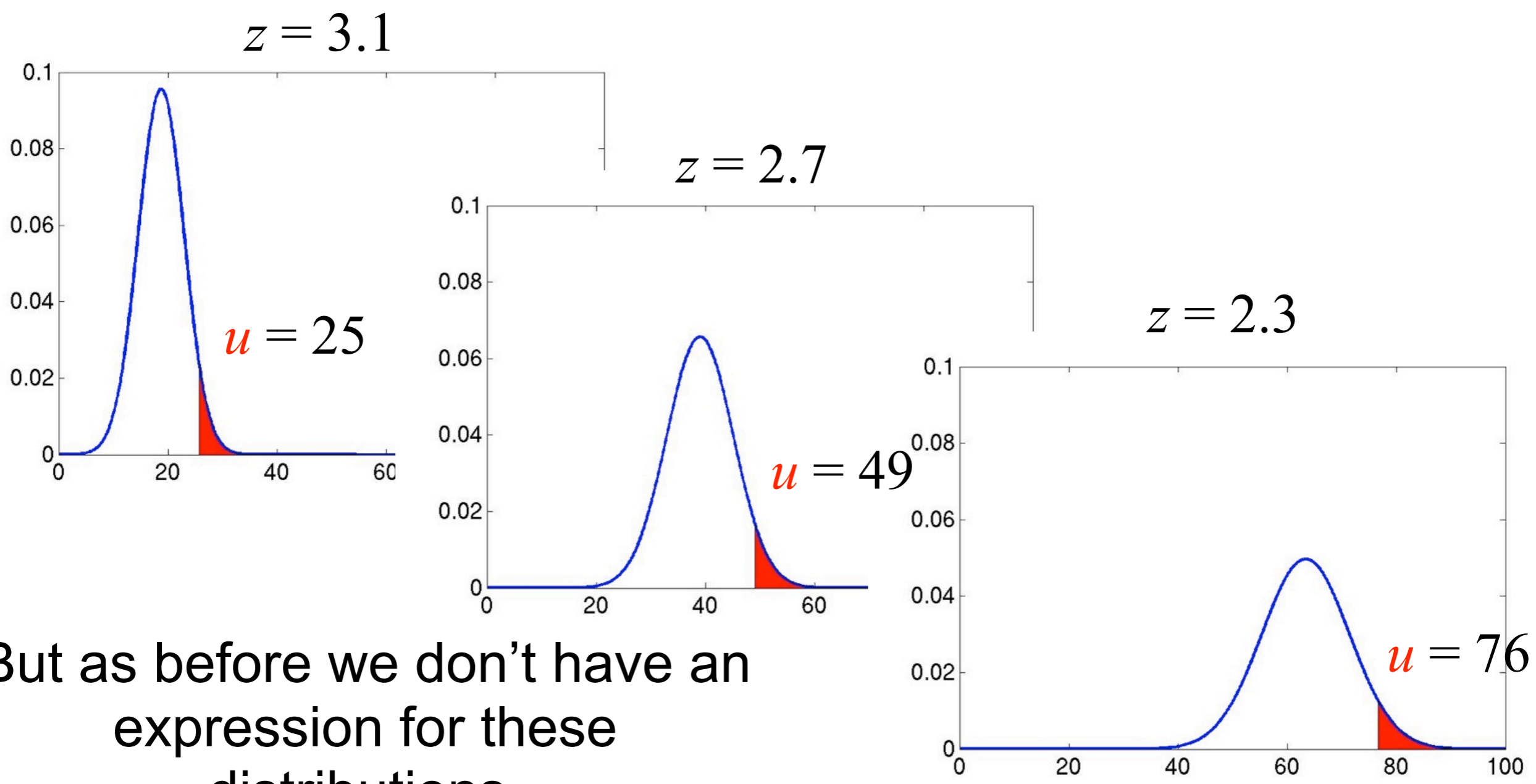
$z = 3.1$





# Distribution of Max Cluster Size

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



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



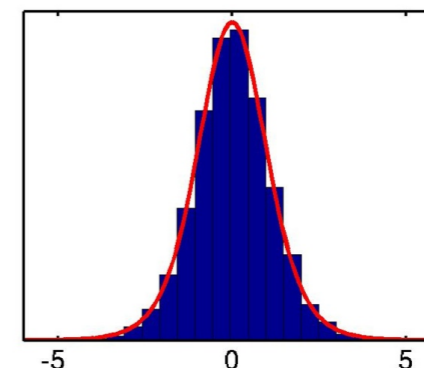
# Outline

- 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

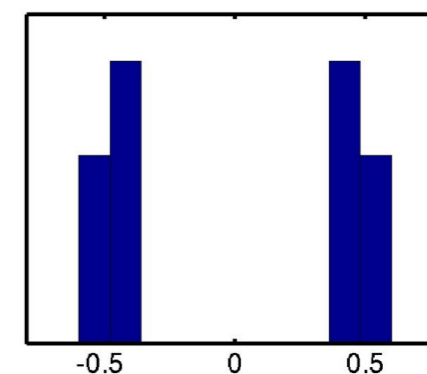
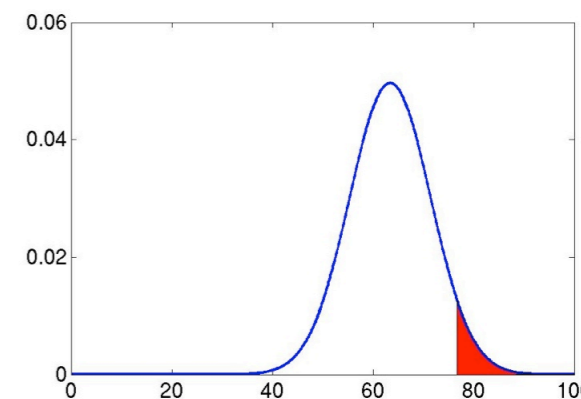
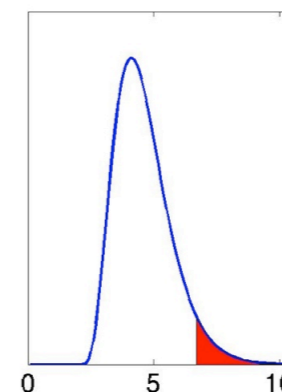


# Parametric vs non-parametric

- 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 normal-distributed



Provided that  $\mathbf{e} \sim N(0, \sigma^2)$





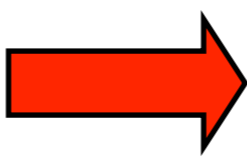
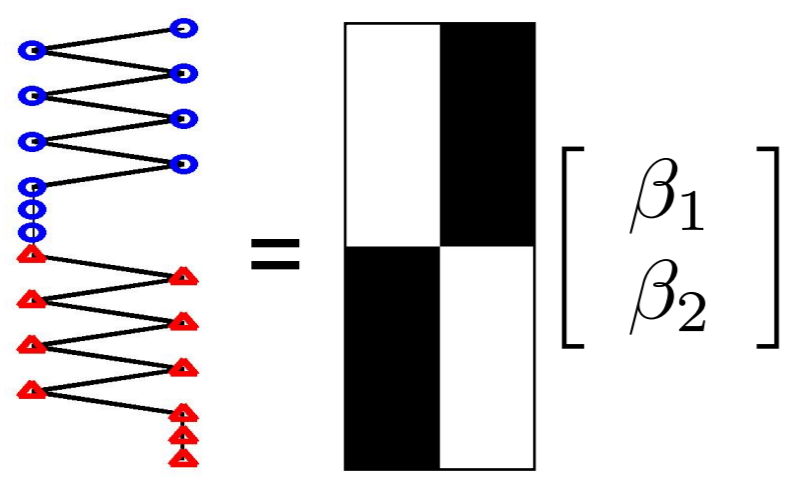
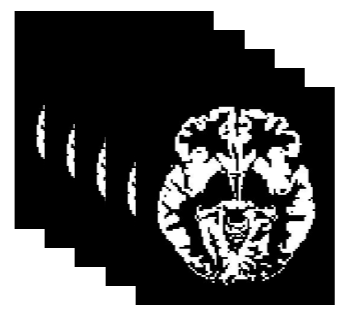
# Example: VBM-style analysis

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

Group #1  
(FSL Course Tutors)

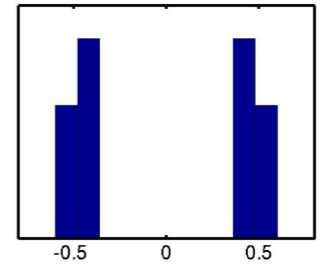


Group #2  
(FSL Course Attendees)



$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix} \text{ Ok!}$$

hist(e)

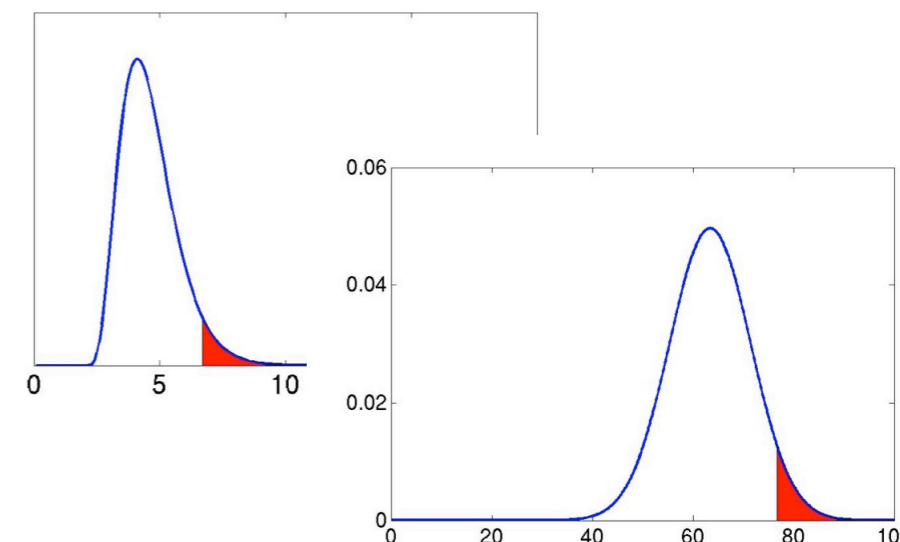


~ N?  
☹️



# 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<sup>a,b,c,1</sup>, Thomas E. Nichols<sup>d,e</sup>, and Hans Knutsson<sup>a,c</sup>

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

The most widely used task functional magnetic resonance imaging (fMRI) analyses use parametric statistical methods that depend on a variety of assumptions. In this work, we use real resting-state data and a total of 3 million random task group analyses to compute empirical familywise error rates for the fMRI software packages SPM, FSL, and AFNI. We compare these to the theoretical familywise error rates (FWE), the chance of one or more false positives, and empirically measure the FWE as the proportion of analyses that give rise to any significant results. Here, we consider both two-sample and one-sample designs. Because two groups of subjects are randomly drawn from a large group of healthy controls, the null hypothesis



# Parametric vs non-parametric

Adv. Appl. Prob. (SGSA) 33, 773-793 (2001)  
Printed in Northern Ireland  
© Applied Probability Trust 2001

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

**TESTING FOR SIGNALS WITH UNKNOWN LOCATION AND SCALE IN A  $\chi^2$  RANDOM FIELD, WITH AN APPLICATION TO FMRI**

KEITH J. WORSLEY, *McGill University*

**Abstract**

nals with unknown statistic was the 'ice',  $N$  dimensions ace is identical to a ough the emphasis ent. Two methods = 3; one based on  $r$  characteristic of the latter method result to  $\chi^2$  fields. case. In this paper in images obtained

ry; image analysis;

sion tomography (PET) : interested in detecting the signal to noise ratio, tion with a filter  $f$ . The

(1.1)

from the matched filter which states that signal

**The Geometry of Random Images**

Keith J. Worsley

The geometry in the title is not the geometry of lines and angles but the modern geometry of topology and shape. What has this to do with statistics? Some recent work has found some fascinating applications of a mixture of geometry, topology, probability, and statistics to some very pressing problems in newly emerging areas of medical imaging and astrophysics.

Where is the link? Let us begin with a quick introduction to one of the fundamental tools of topology, the Euler characteristic.

**Topology: The Euler Characteristic**

Named after Leonhard Euler (1707-1783), the most prolific mathematician of the 18th century, the Euler characteristic itself began with Euler's observation about polyhedra.

Recall that a polyhedron is a solid object bounded by plane faces, such as a cube. Euler realized that, if you count the faces ( $F$ ), edges ( $E$ ), and vertices ( $V$ ) of a polyhedron, then  $V - E + F = 2$  no matter how the polyhedron is constructed.

A cube, for example, has  $F = 6$  faces,  $E = 12$  edges and  $V = 8$  vertices (see Fig. 1a) so that  $8 - 12 + 6 = 2$ . For a solid that consists of  $P$  polyhedra, stuck together on at least one common face, the slightly more general formula becomes  $V - E + F - P = 1$ .

A little experimentation will convince you that this new formula works for all solids (see Fig. 1b)—well almost all. If the solid has a hole going right through

it, like a doughnut (see Fig. 1c), then the result no longer holds. In fact, the result is  $V - E + F - P = 0$  for any solid with just one hole.

Too bad! But this does not deter a good mathematician—far from it—it opens up vast new possibilities! What happens if there are two holes, like a figure 8 (see Fig. 1d)? Then it turns out that  $V - E + F - P = -1$ , and so on; each hole reduces  $V - E + F - P$  by 1.

So now suddenly we have a fascinating new tool. We can count the number of holes in a solid using the formula  $V - E + F - P$ ; it has the very interesting property that no matter how the solid is subdivided into polyhedra, the value of  $V - E + F - P$  is invariant. This is born the field of topology: We define the Euler characteristic (EC) of a solid as simply

$$EC = V - E + F - P$$

for any subdivision of the solid into polyhedra.

Thus the EC of a pretzel-shaped solid (Fig. 1e) is  $-2$ ; +1 for the solid part (the part you eat), and  $-3$  for each of the three holes, giving  $-2$  overall. Have we covered all possibilities? Not quite—if the solid is hollow, like a tennis ball, then surprisingly enough the EC is 2 (see Fig. 1f).

Think you've got it now? How about a solid shaped like a bicycle inner tube? Answer: The EC is 0, and if it has a puncture, then the EC is  $-1$ .

One more slight generalization, which will prove to be extremely useful for practical applications: Suppose

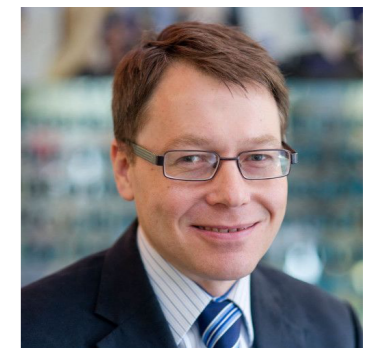
Received June 1990; revised June 1994.  
Research supported by the Natural Sciences and the Fonds pour la Formation des Chercheurs et l'Aide à la Recherche (FCR).  
AMS 1991 subject classifications. Primary 6 Key words and phrases. Euler characteristic; bility; image analysis.

27

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

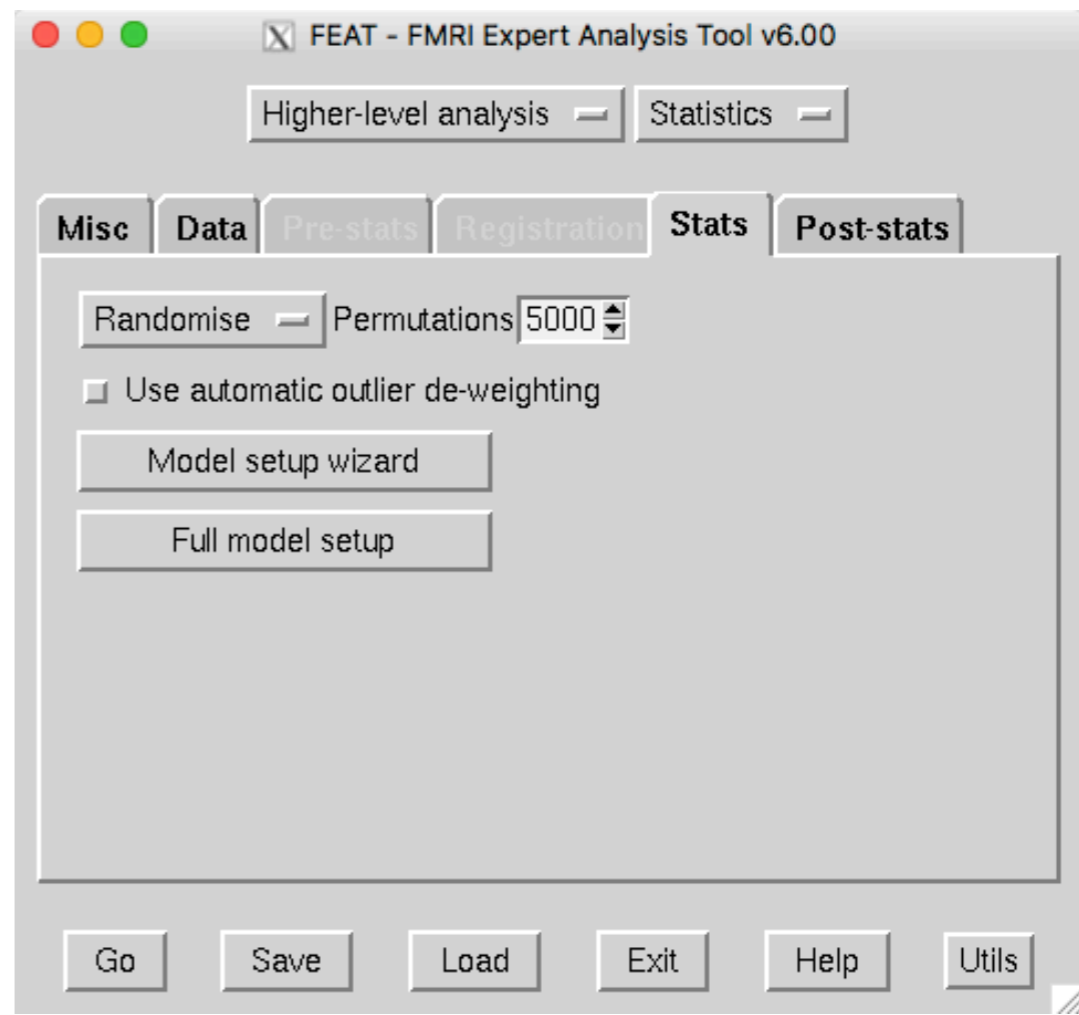
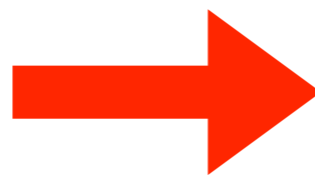
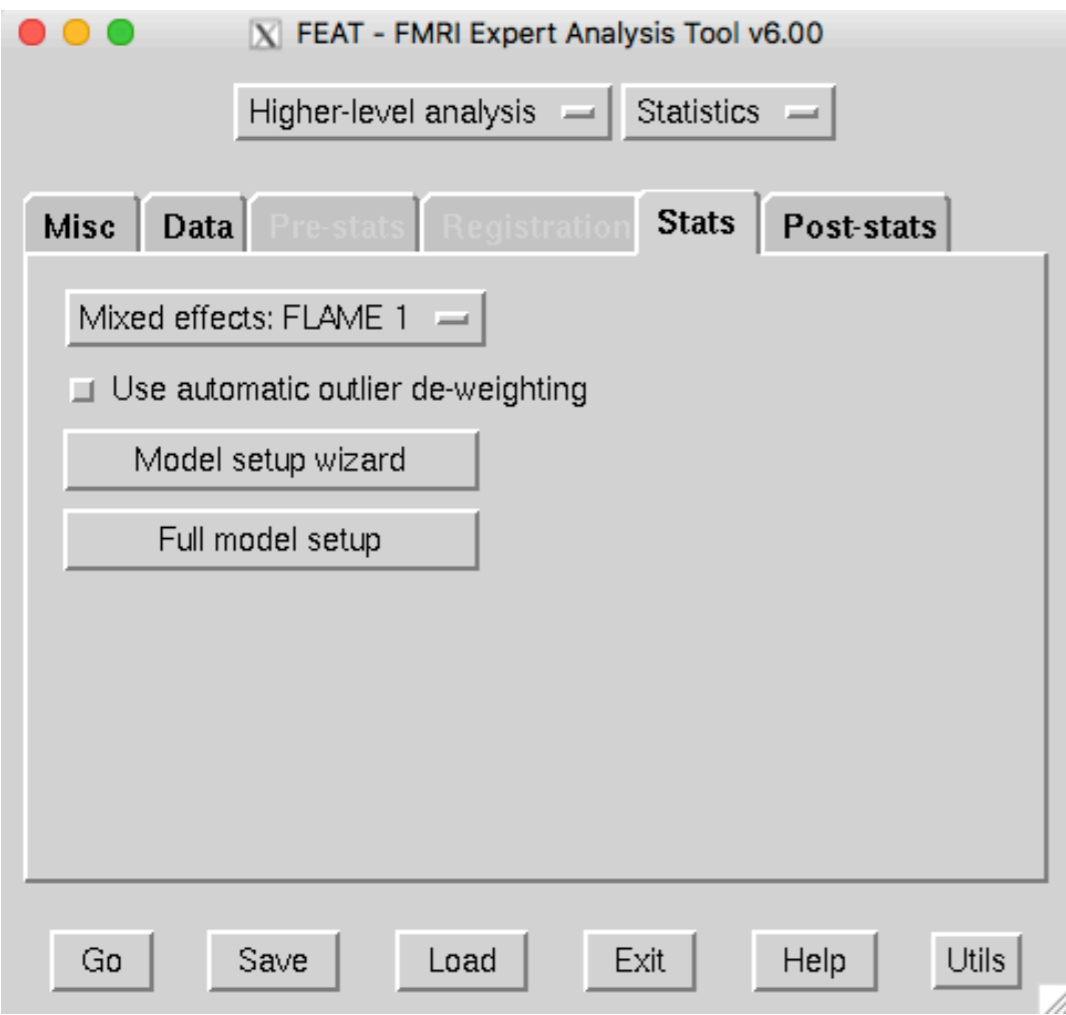


- But the we've moved towards non-parametric testing





# Parametric vs non-parametric



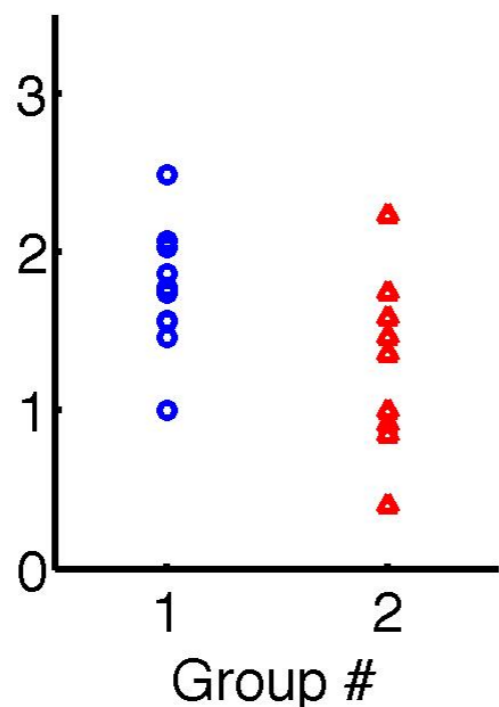




# A simple permutation test

- 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.

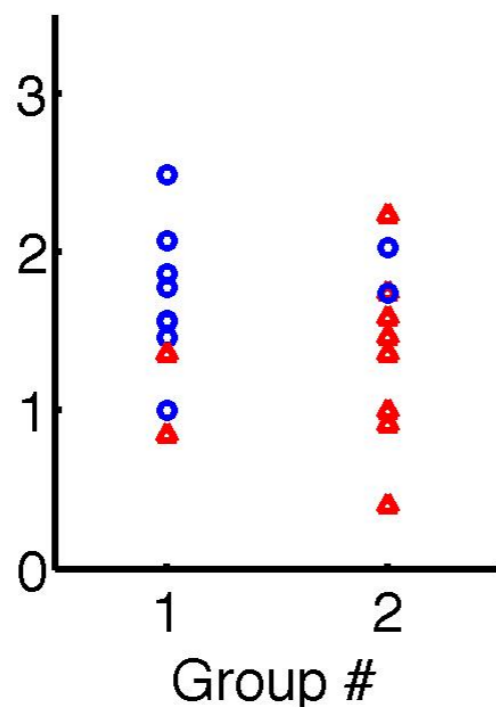




# A simple permutation test

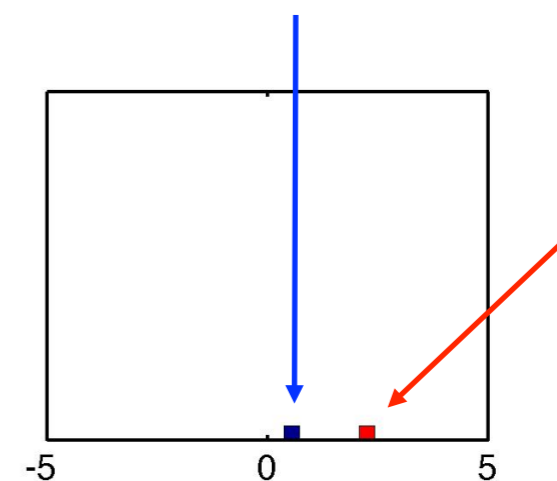
- We can permute the data itself to create a distribution that we can use to test our statistic.
- + Makes very few assumptions about the data
- + Works for any test statistic

One re-labelling



$t$ -value after re-labelling

$$t = 0.67$$



Original labelling

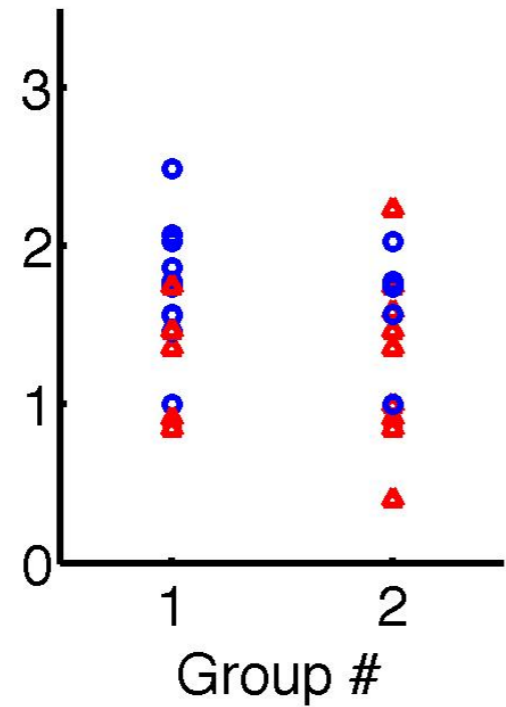
Let's start collecting them



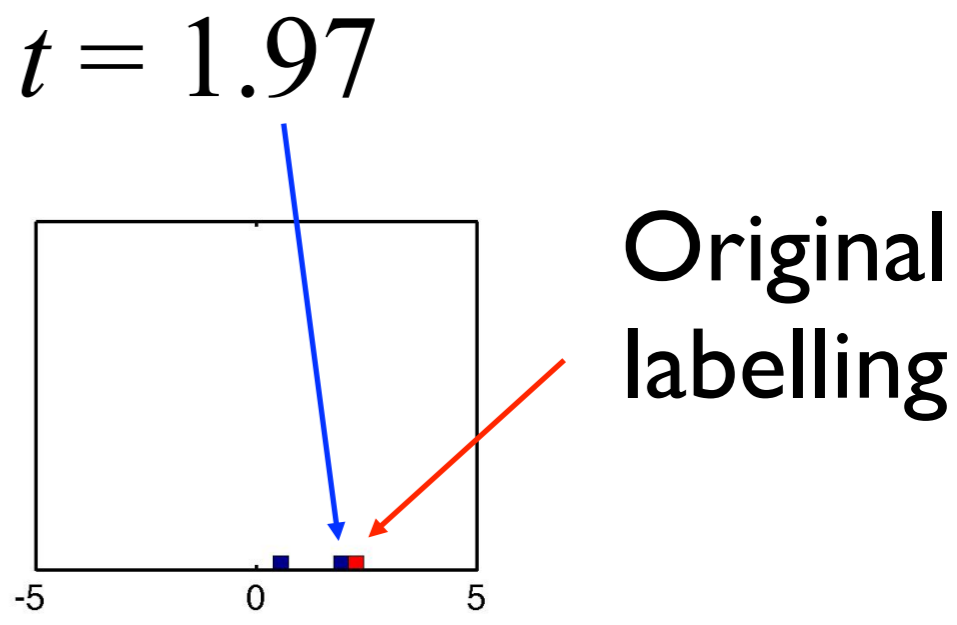
# A simple permutation test

- 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



$t$ -value after re-labelling



And another one



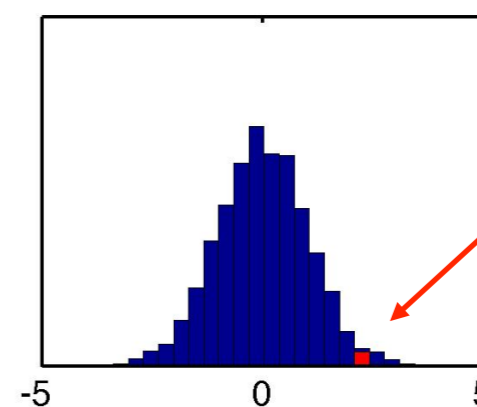
# A simple permutation test

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

$$\text{i.e. } p(x \geq 2.27) = 1.79\% \text{ for } t_{18}$$



Original labelling

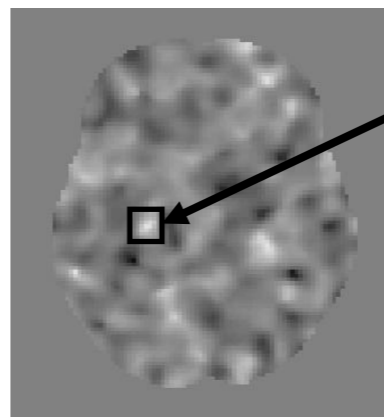
5000 re-labellings. Phew!



# 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.  $t_8 = 4.65$

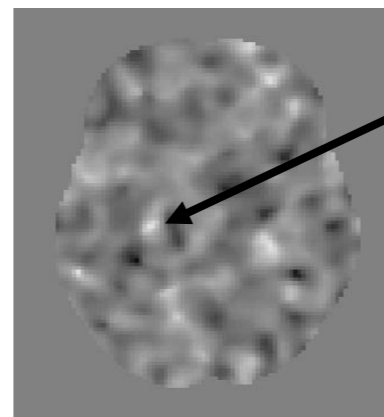
Prof. ran to write to Nature Neuro. **But**, did they jump the gun?



# And we can use this for any statistic

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

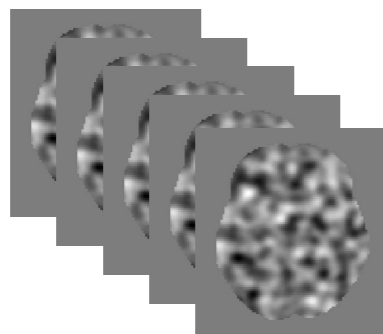
This is what we got



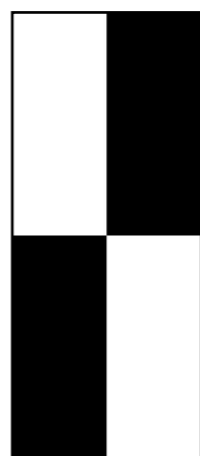
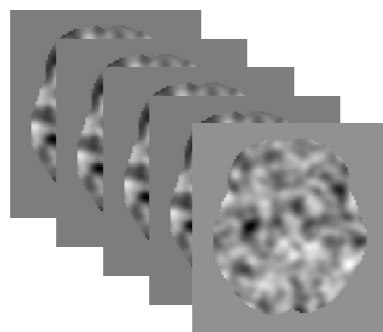
Very intriguing activation.  $t_8 = 4.65$

Prof. ran to write to Nature Neuro. **But**, did they jump the gun?

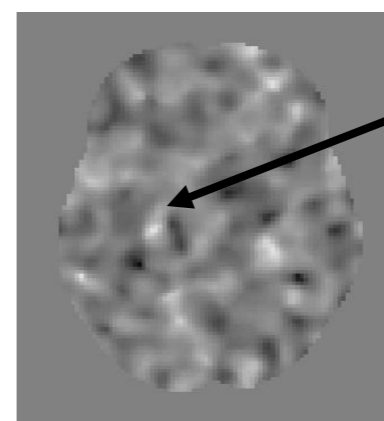
Group 1



Group 2



2nd level model



Our group difference map

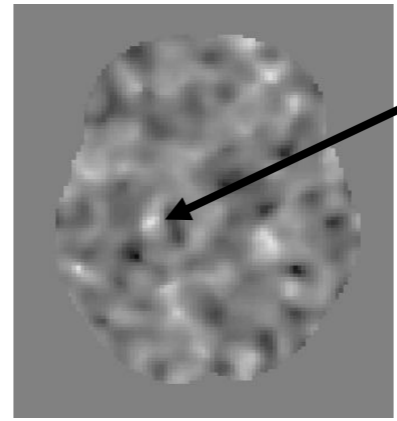
$\max(t) = 4.65$



# And we can use this for any statistic

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

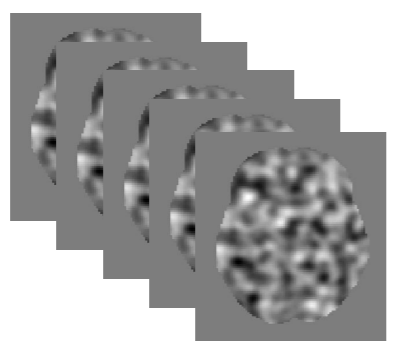
This is what we got



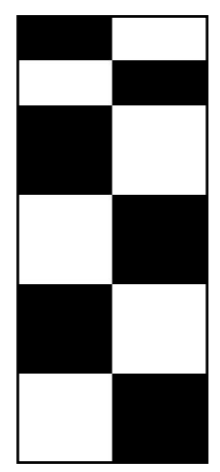
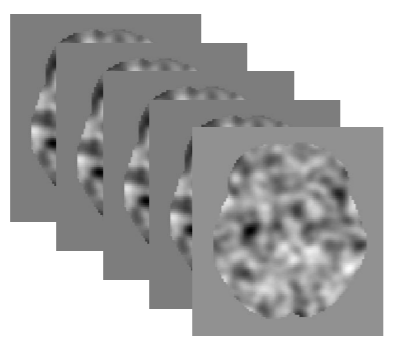
Very intriguing activation.  $t_8 = 4.65$

Prof. ran to write to Nature Neuro. **But**, did they jump the gun?

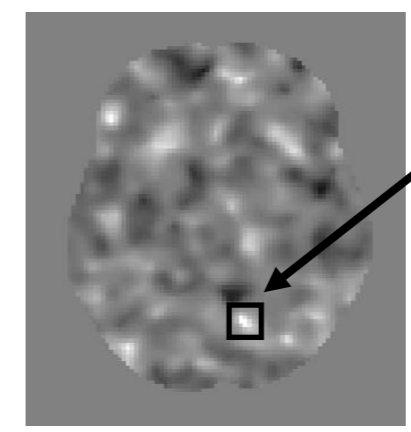
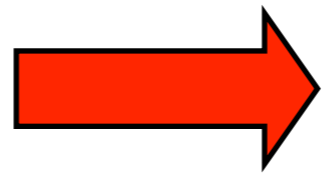
Group 1



Group 2



Permuted model



Permuted group difference map

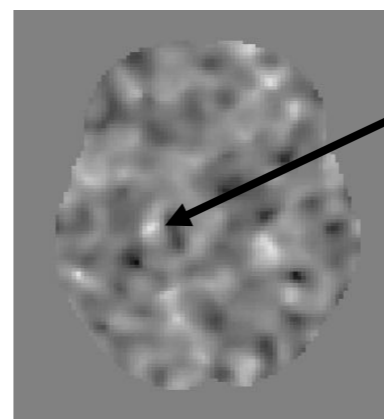
$\max(t) = 8.23$



# And we can use this for any statistic

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

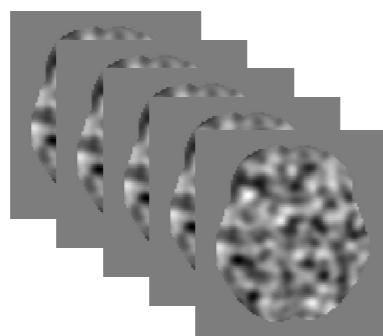
This is what we got



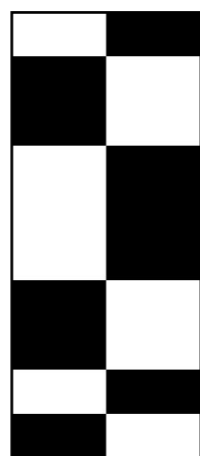
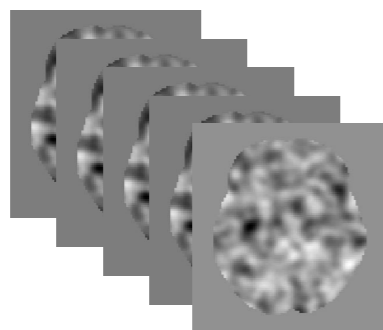
Very intriguing activation.  $t_8 = 4.65$

Prof. ran to write to Nature Neuro. **But**, did they jump the gun?

Group 1

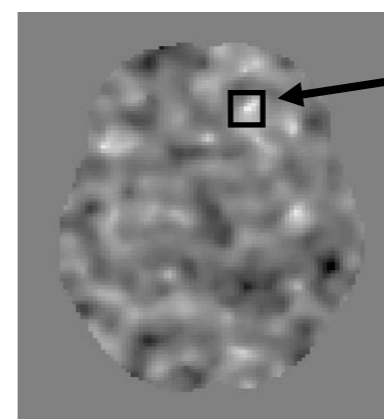


Group 2



2nd

Permutation



2nd permuted map

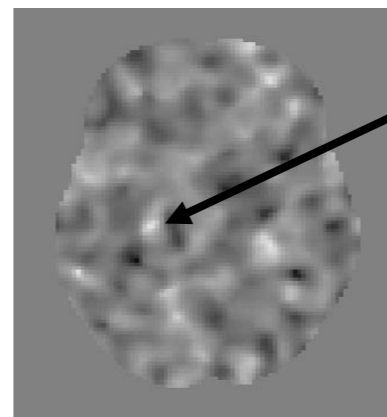
$\max(t) = 5.43$



# And we can use this for any statistic

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

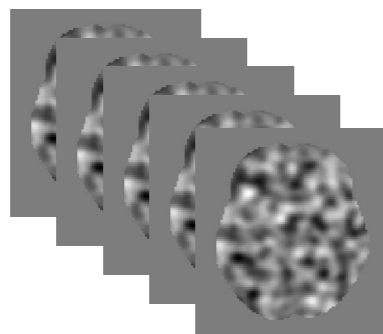
This is what we got



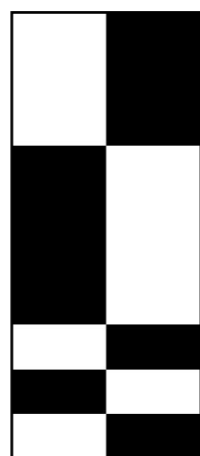
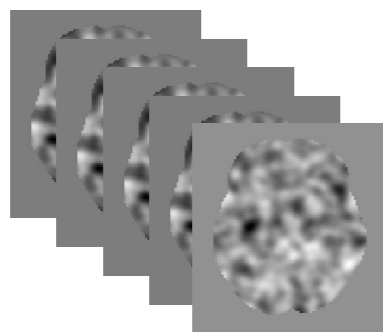
Very intriguing activation.  $t_8 = 4.65$

Prof. ran to write to Nature Neuro. **But**, did they jump the gun?

Group 1

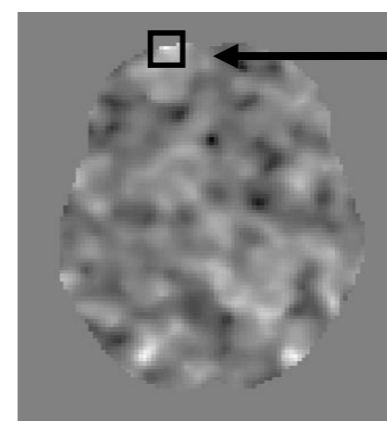
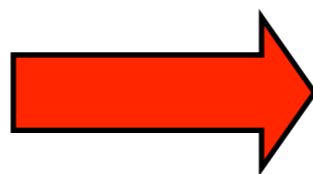


Group 2



3rd

Permutation



3rd permuted map

$\max(t) = 5.84$

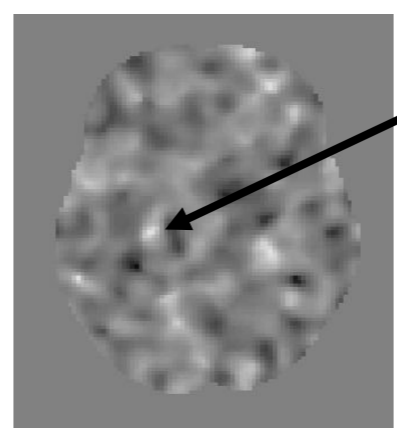




# And we can use this for any statistic

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

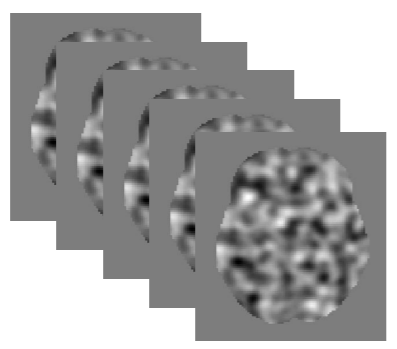
This is what we got



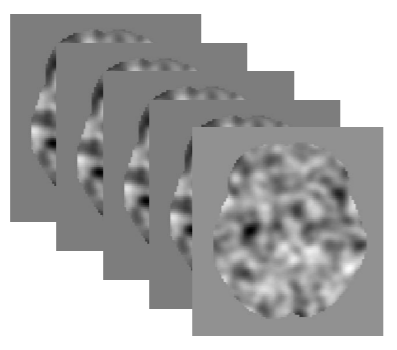
Very intriguing activation.  $t_8 = 4.65$

Prof. ran to write to Nature Neuro. **But**, did they jump the gun?

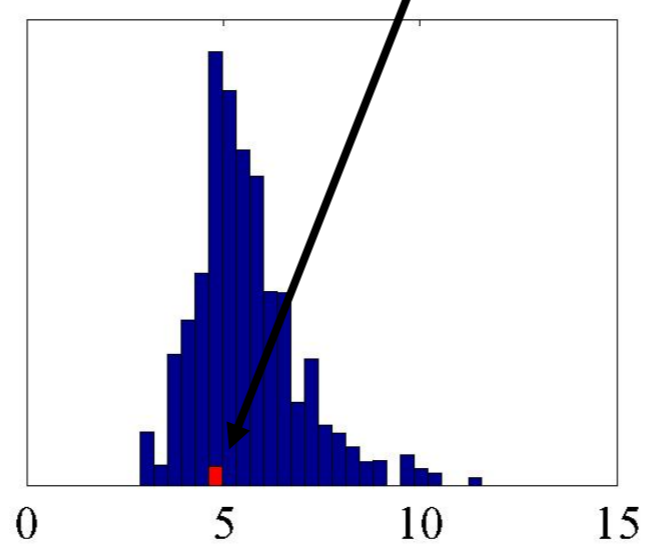
Group 1



Group 2



Original labelling



5000 permutations

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



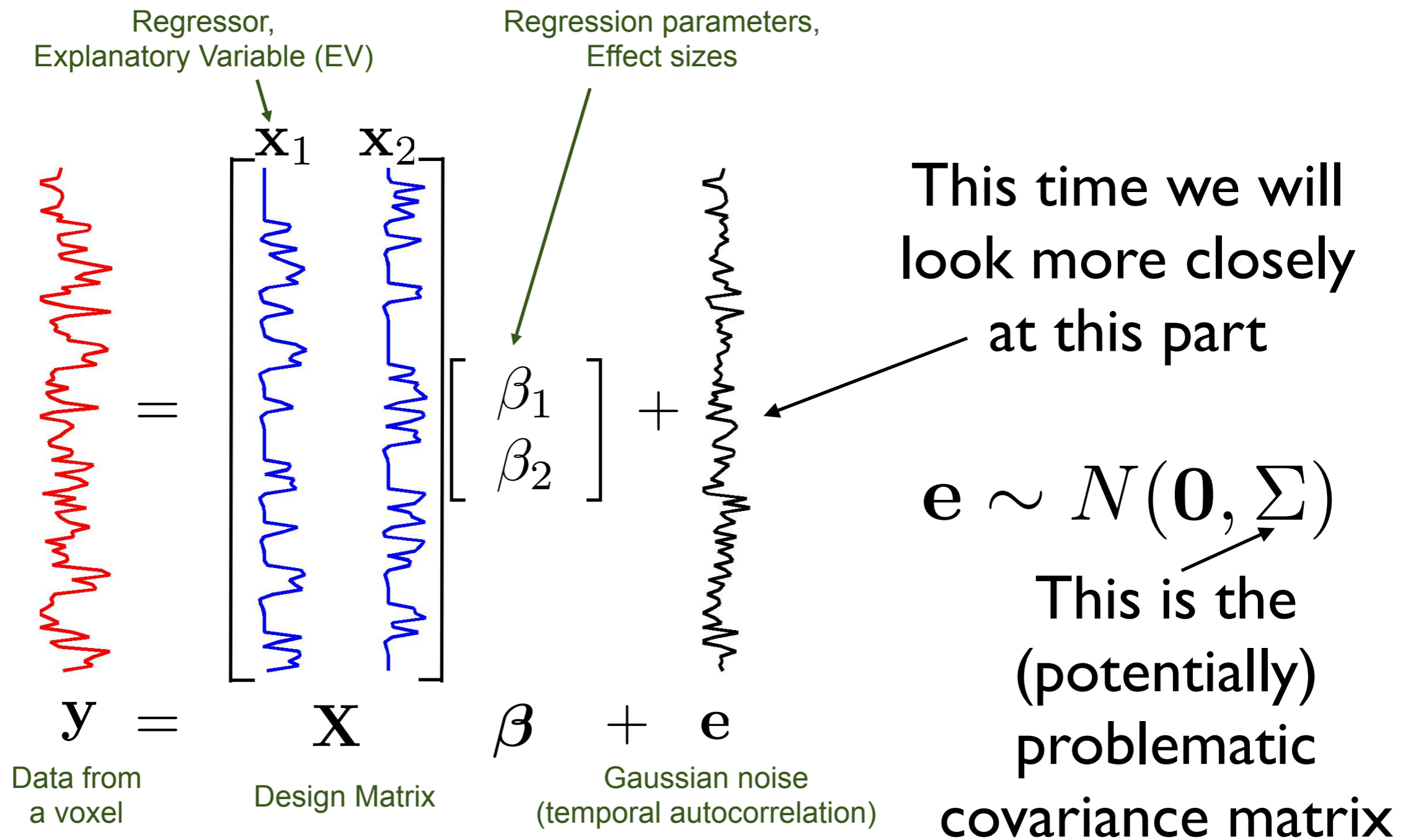
# But beware the “exchangeability”

- 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

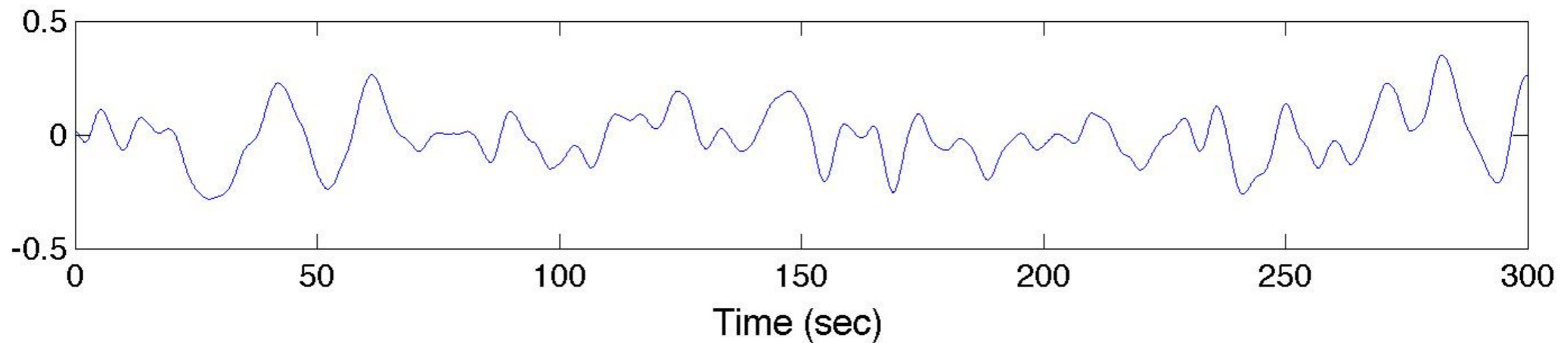
- You may, or may not, have seen this slide in the 1st level GLM talk.





# 1st level fMRI data is not exchangeable

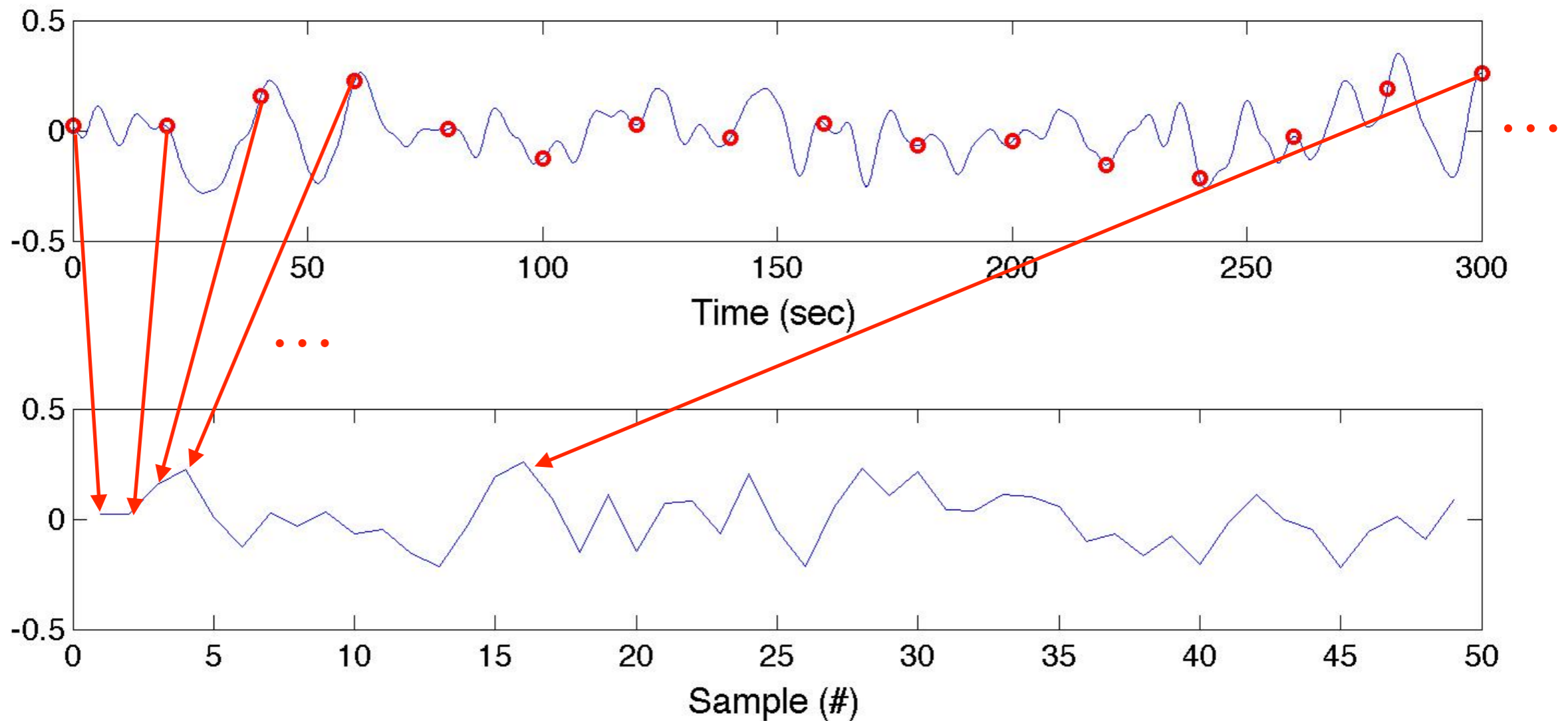
- One important component of noise in fMRI consists of physiological/neuronal events convolved by the HRF





# 1st level fMRI data is not exchangeable

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

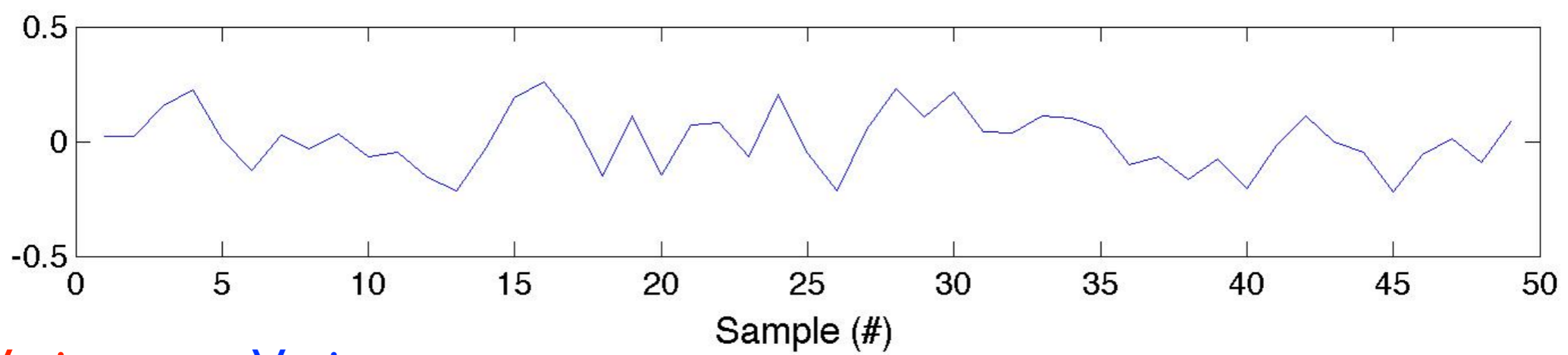


If we sample this every 20 seconds it no longer looks “smooth”



# 1st level fMRI data is not exchangeable

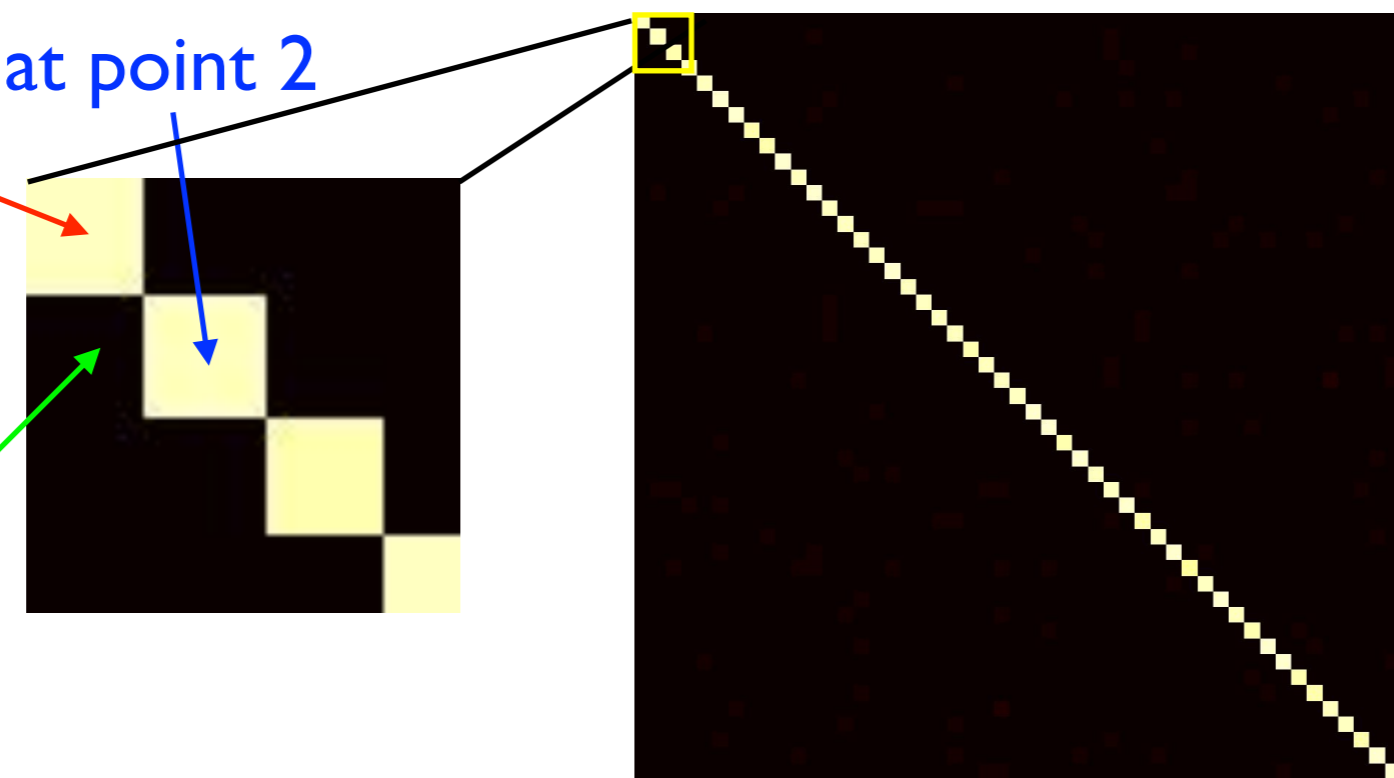
- One important component of noise in fMRI consists of physiological/neuronal events convolved by the HRF



Variance at point 1

Variance at point 2

Covariance between points 1 and 2



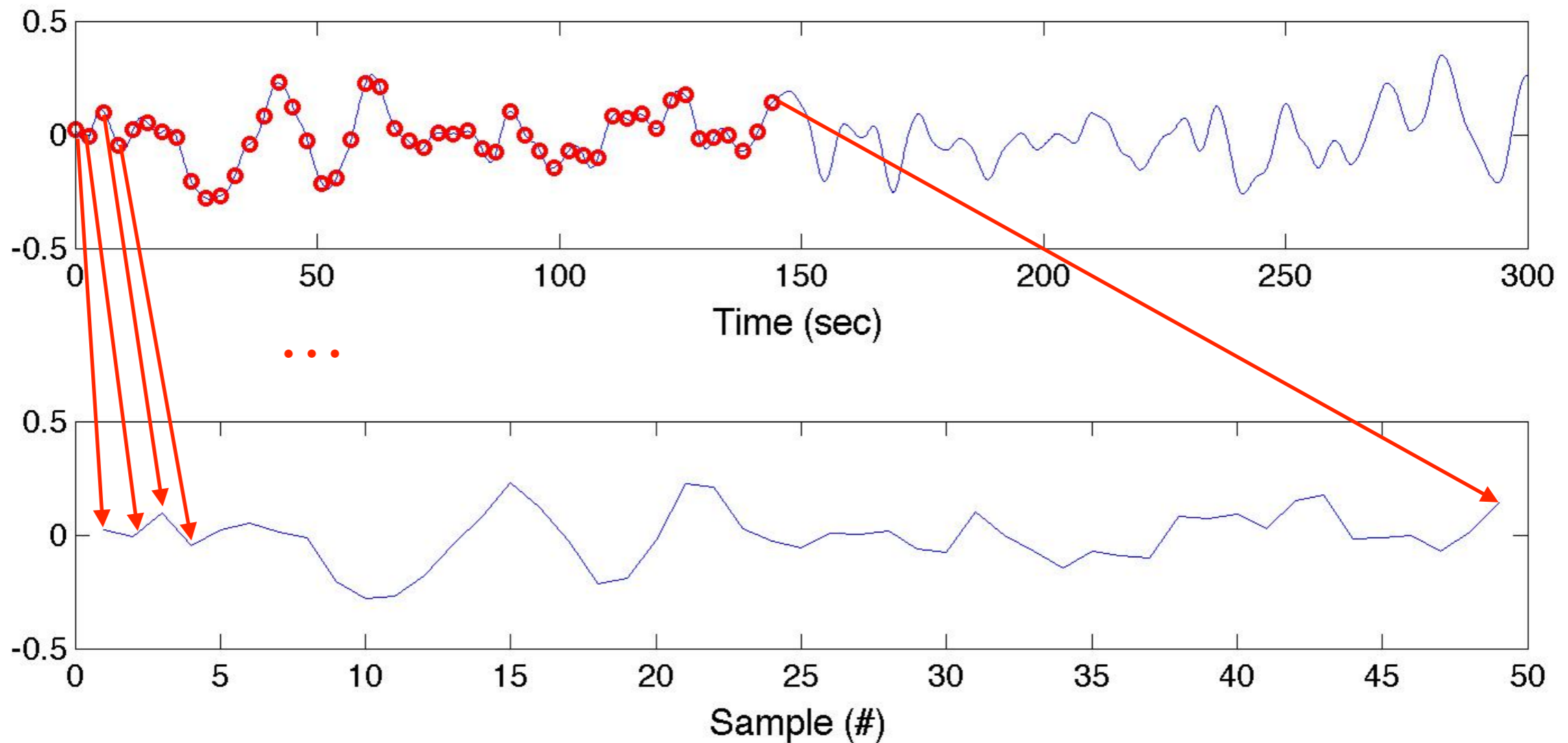
$$\mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$





# 1st level fMRI data is not exchangeable

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

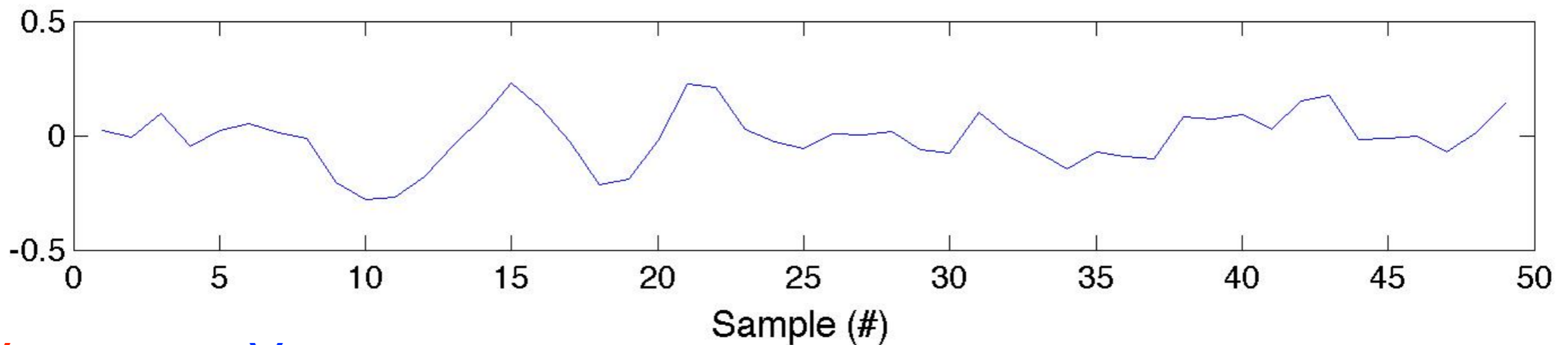


But that is not a realistic TR. What about every 3 seconds?



# 1st level fMRI data is not exchangeable

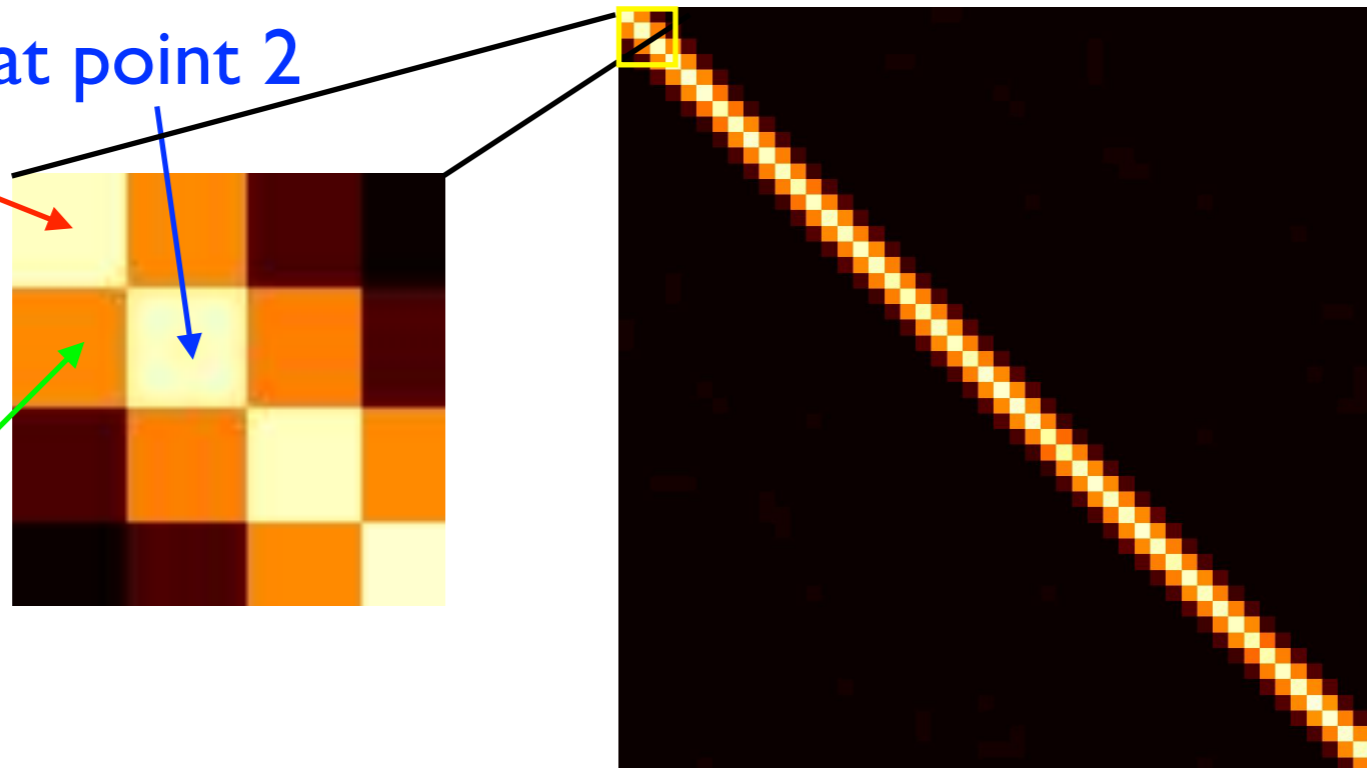
- One important component of noise in fMRI consists of physiological/neuronal events convolved by the HRF



Variance  
at point 1

Variance  
at point 2

Covariance  
between points  
1 and 2

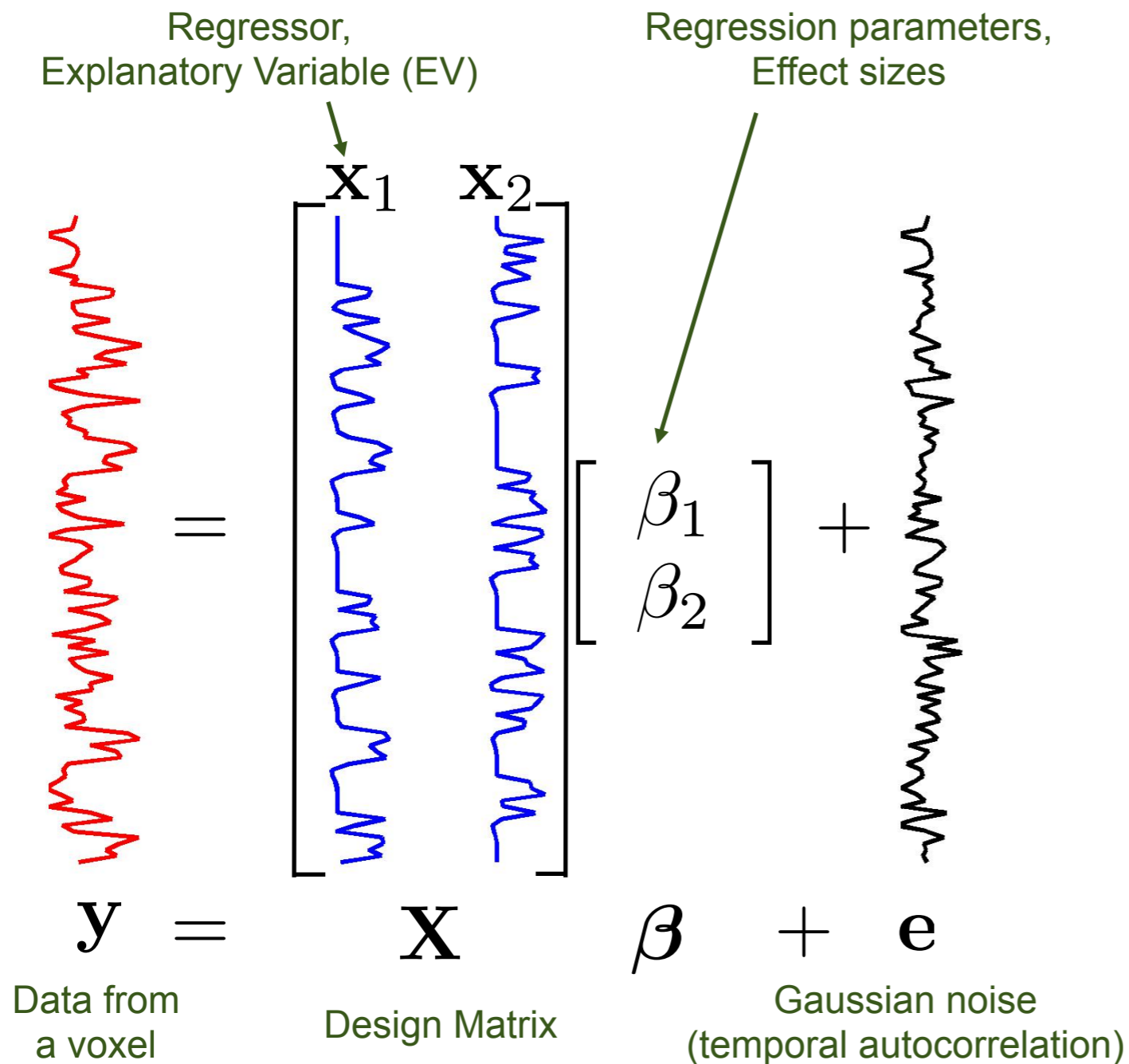






# 1st level fMRI data is not exchangeable

- Let us now return to our model again

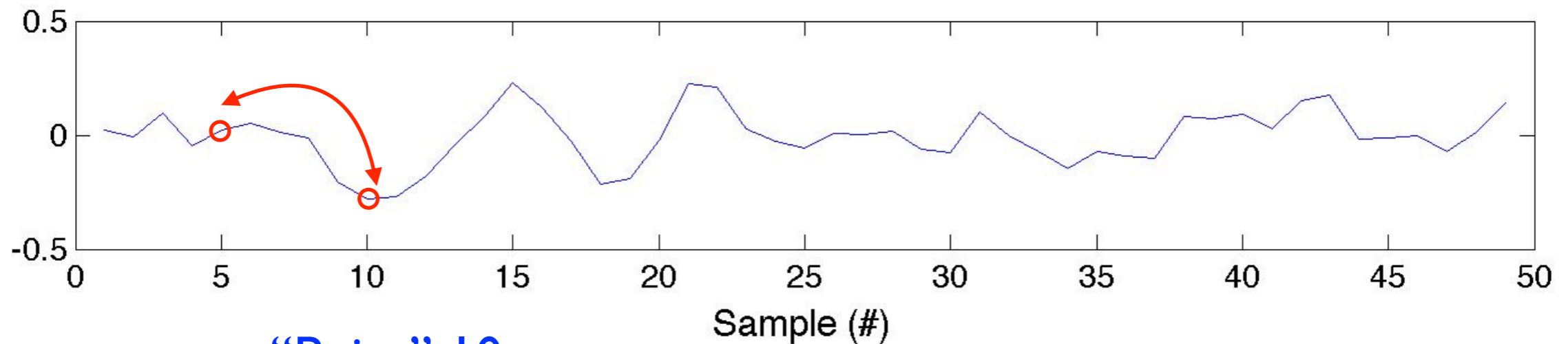


- The model consists of our regressors  $\mathbf{X}$  and the noise model
- All permutations must result in “equivalent models”
- Let us now see what happens if we swap two data-points (points 5 and 10)



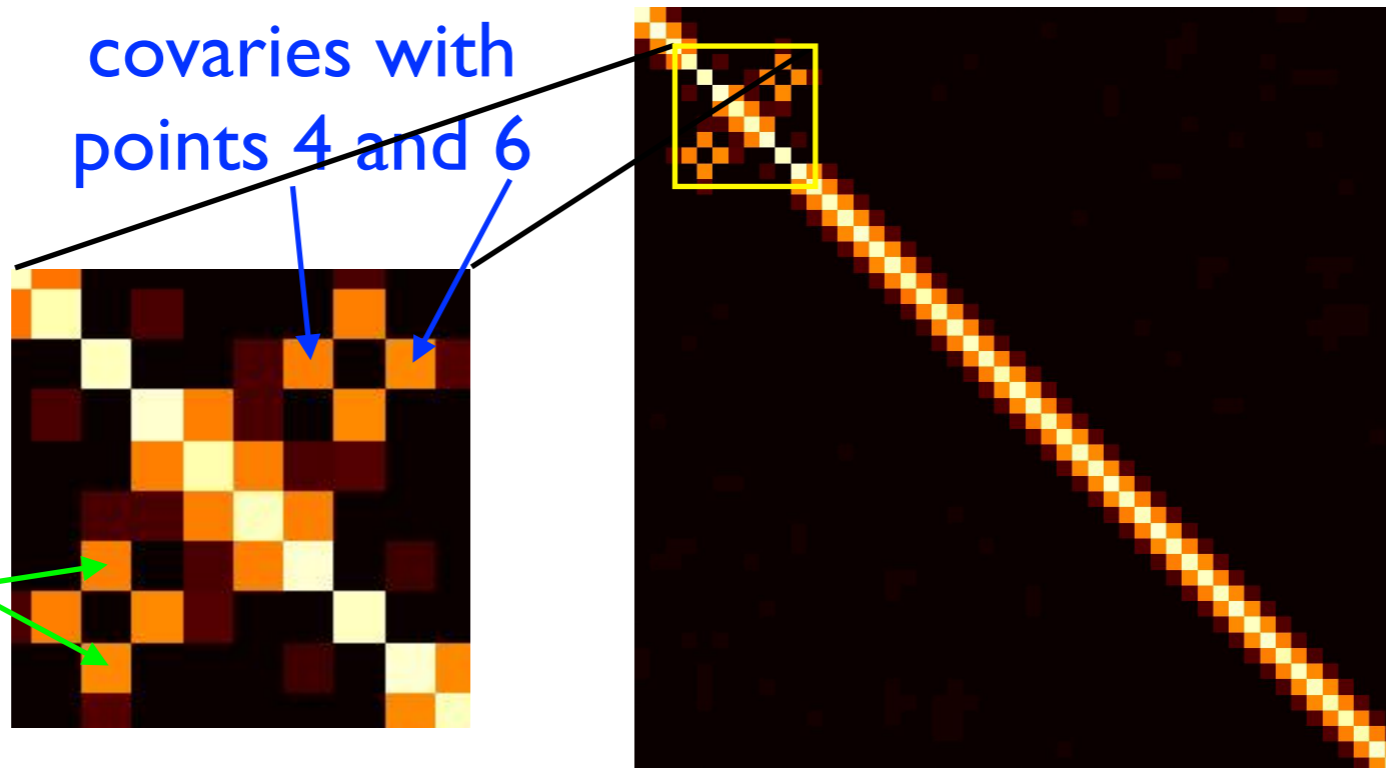
# 1st level fMRI data is not exchangeable

- 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

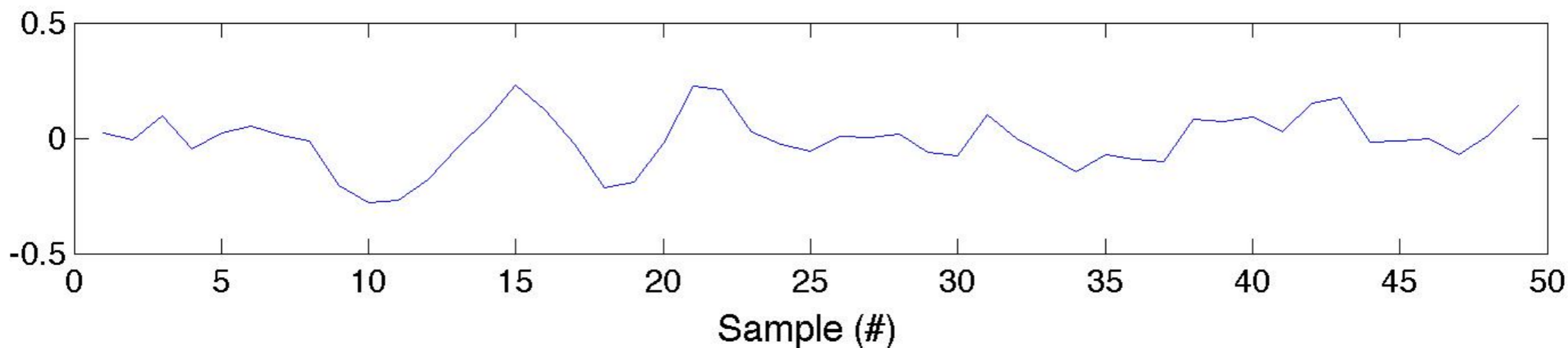


And the models are no longer equivalent

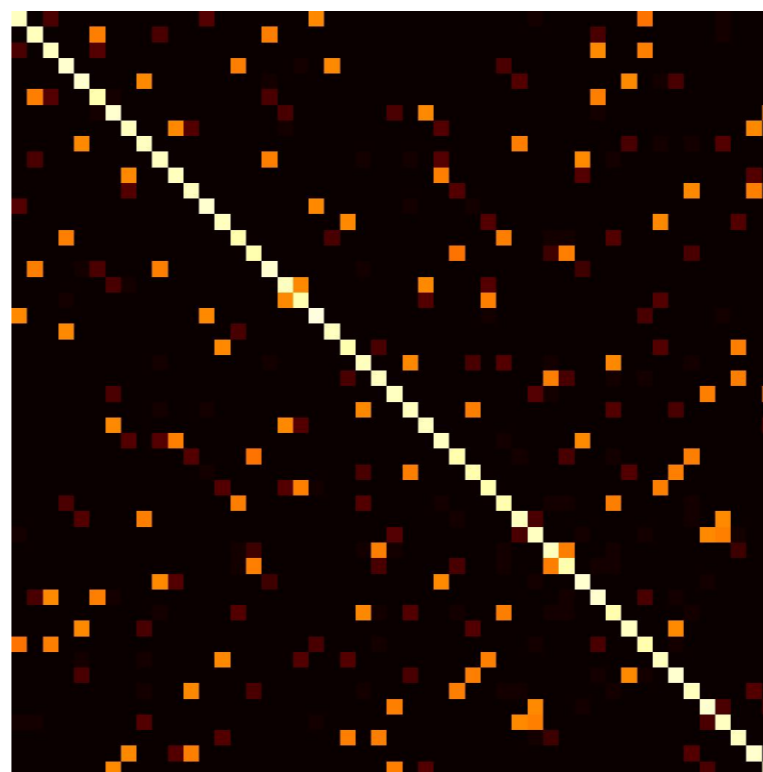


# 1st level fMRI data is not exchangeable

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



And for a random permutation ...



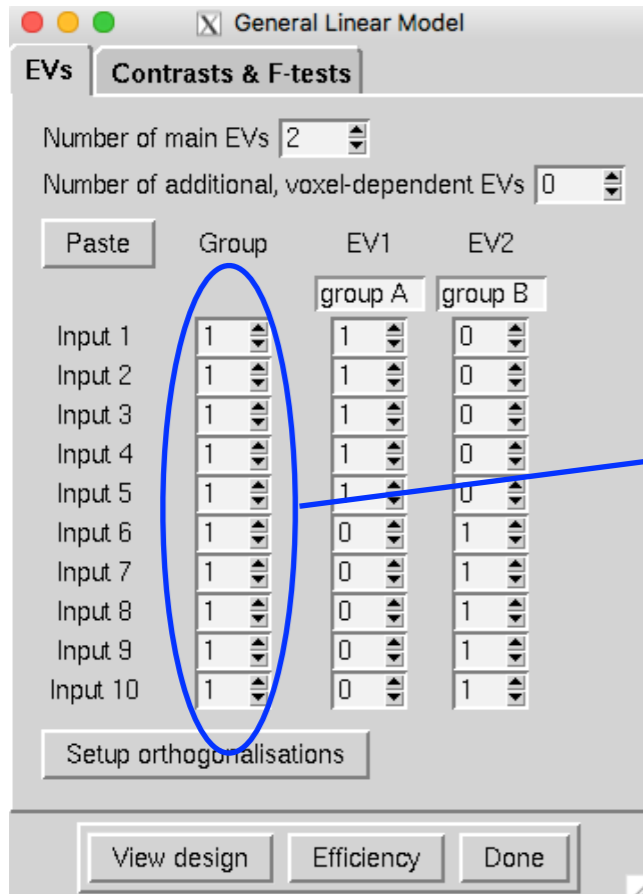
And the models are no longer equivalent



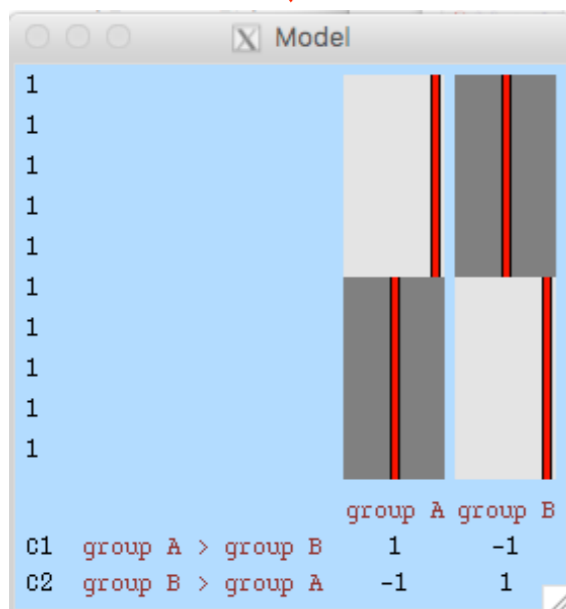
# Back to exchangeability

- 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

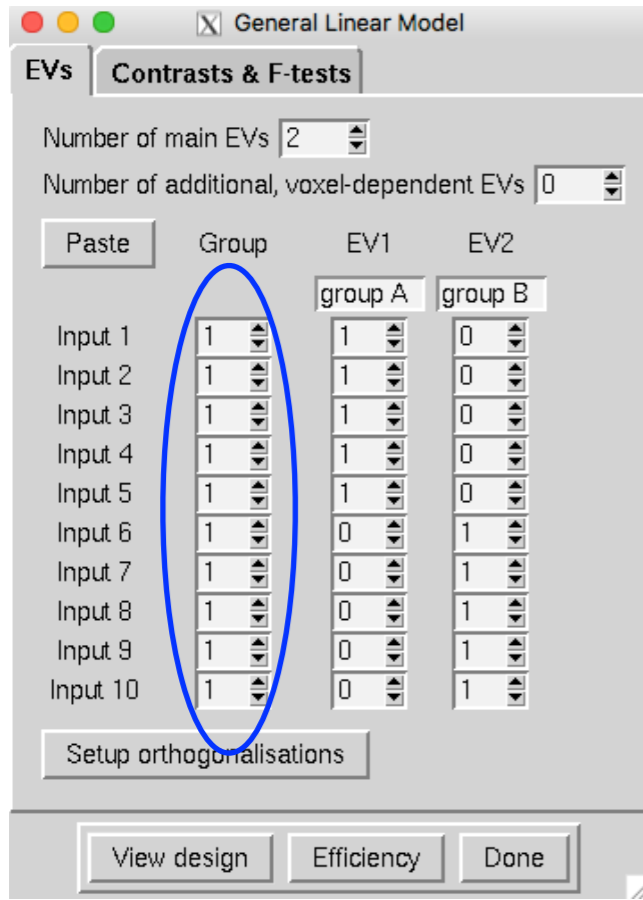


This is the “exchangeability group”. Here all scans are in the same group, which means any scan can be exchanged for any other.

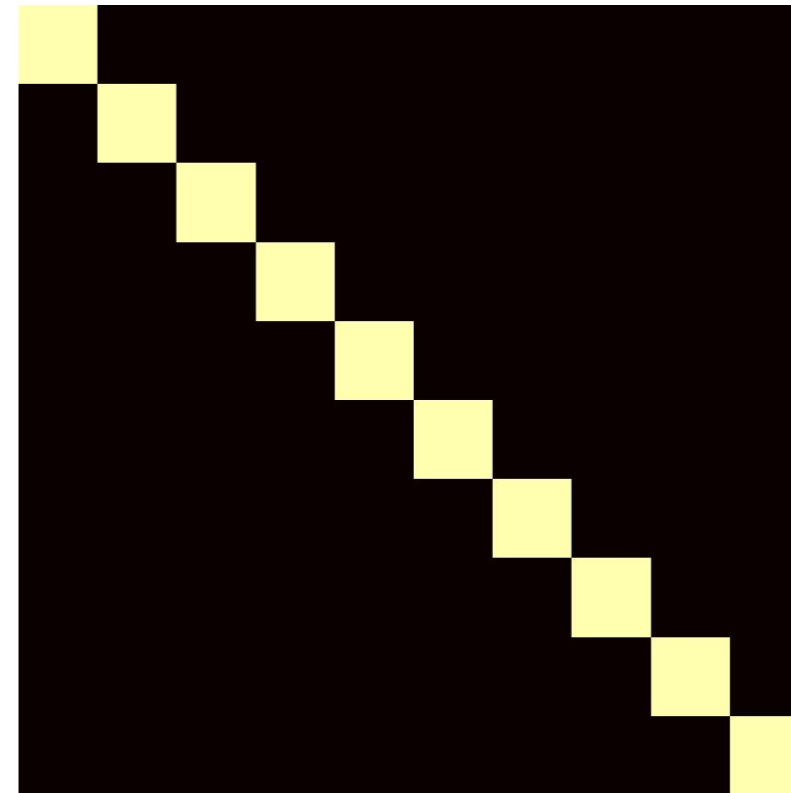


N.B. The “group” labelling is used for completely different purposes when using FLAME/GRFT

# Examples of exchangeability: Two groups unpaired



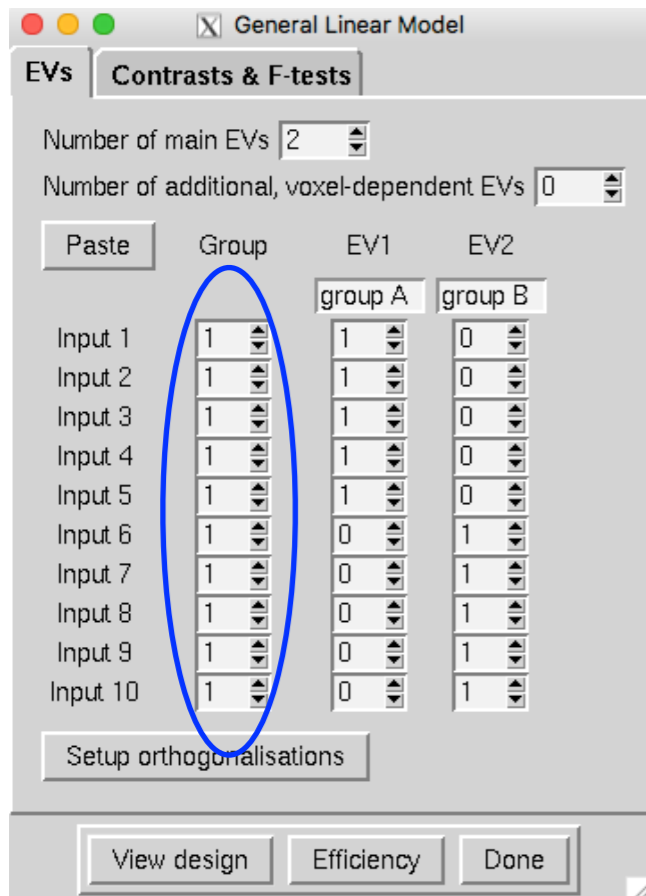
Assumed covariance matrix



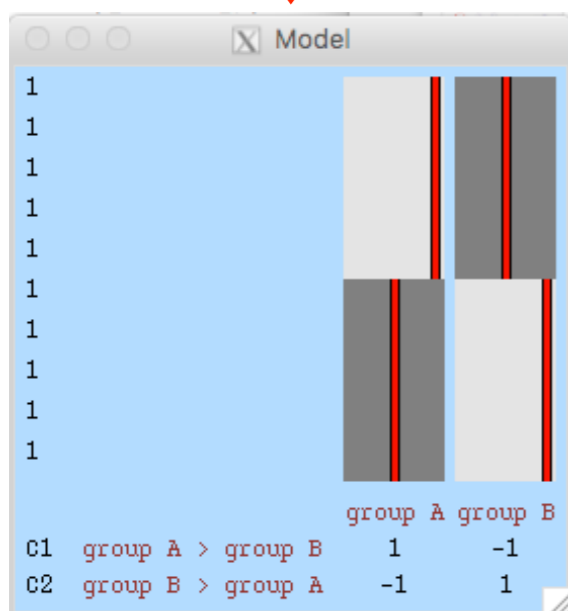
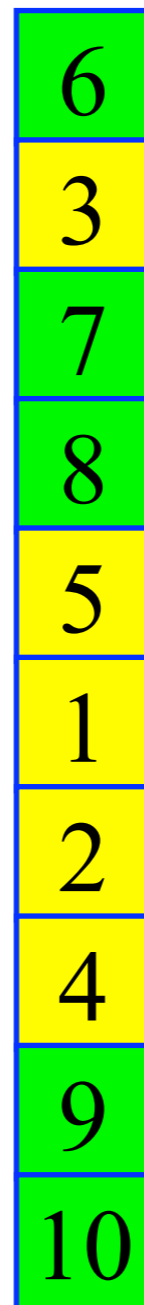
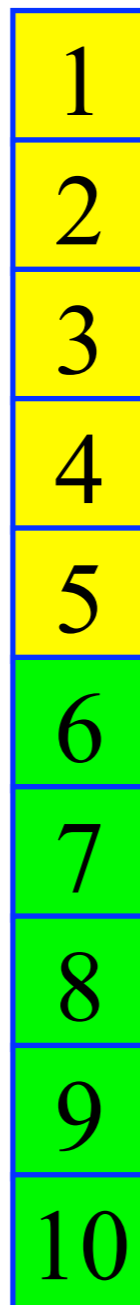
The implicit assumption here is that data from all subjects have the same uncertainty and are all independent

Model  
1  
1  
1  
1  
1  
1  
1  
1  
1  
1  
1  
1  
group A group B  
C1 group A > group B 1 -1  
C2 group B > group A -1 1

# Examples of exchangeability: Two groups unpaired



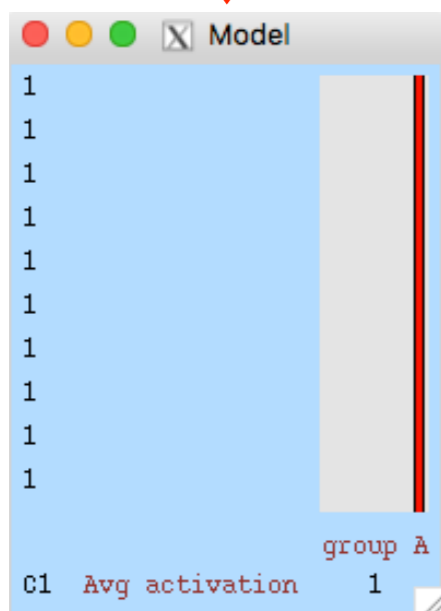
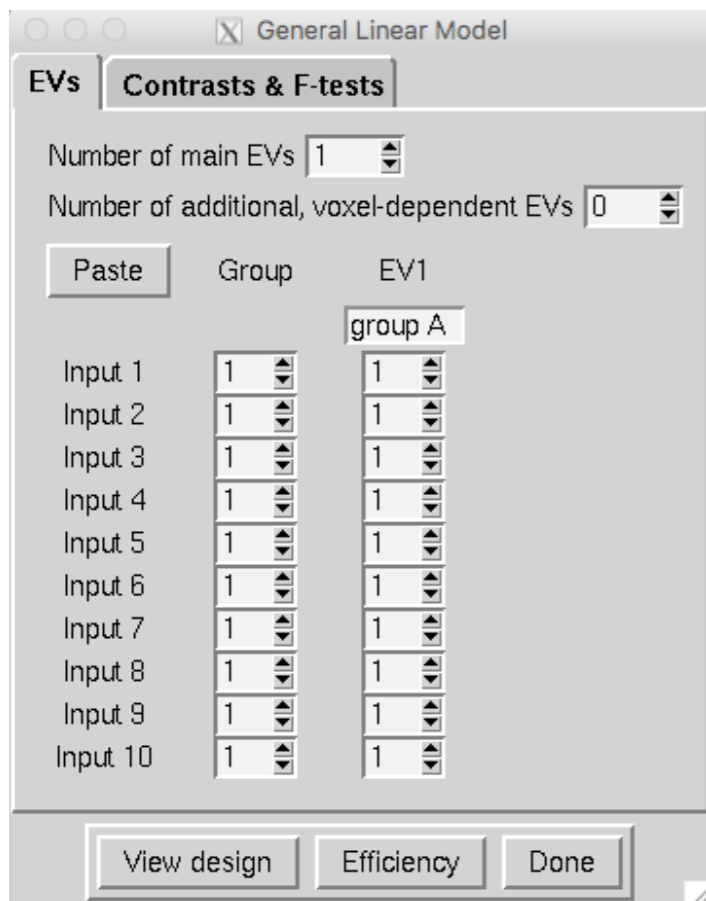
Original Perm 1 Perm 2 ...







# Examples of exchangeability: Single group average

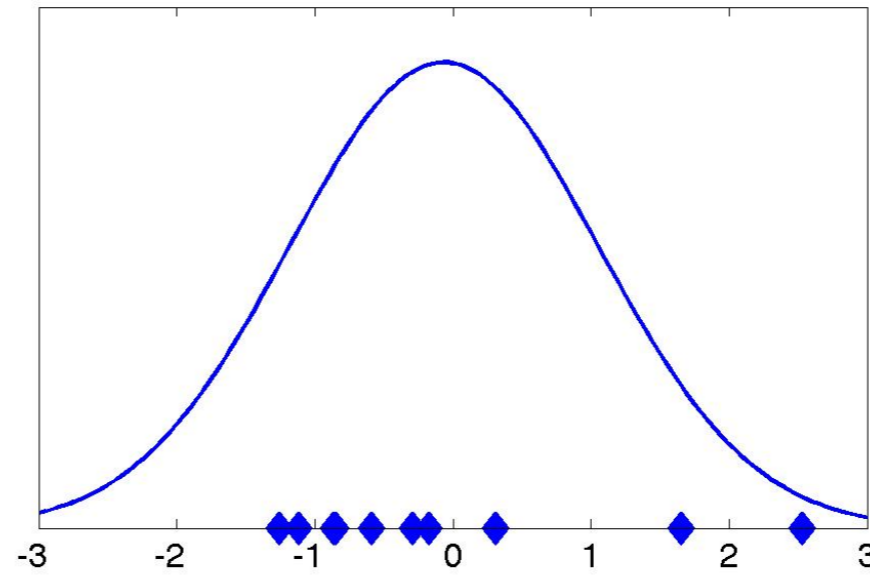
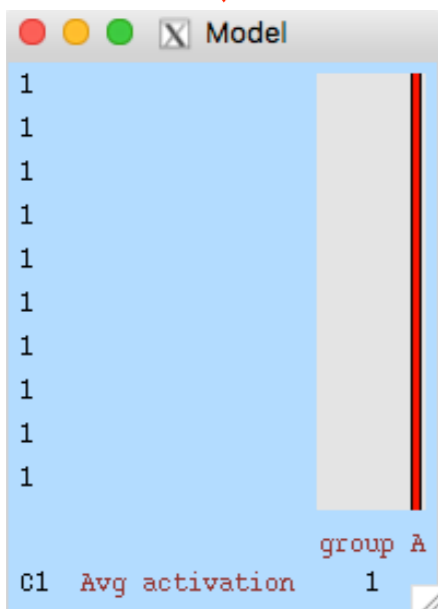
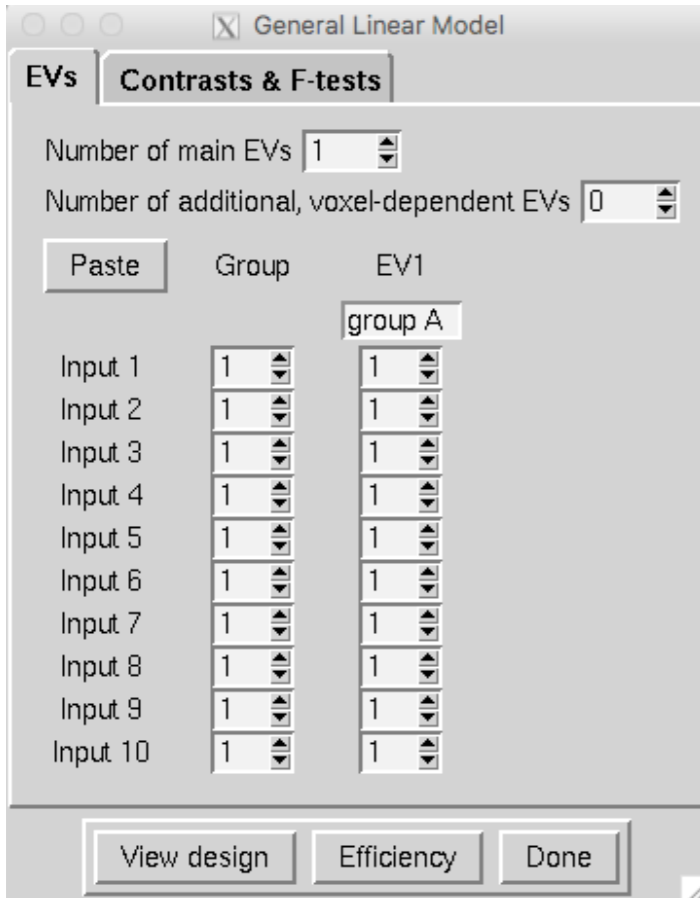


Here we model a single mean and want to know if that is different from zero

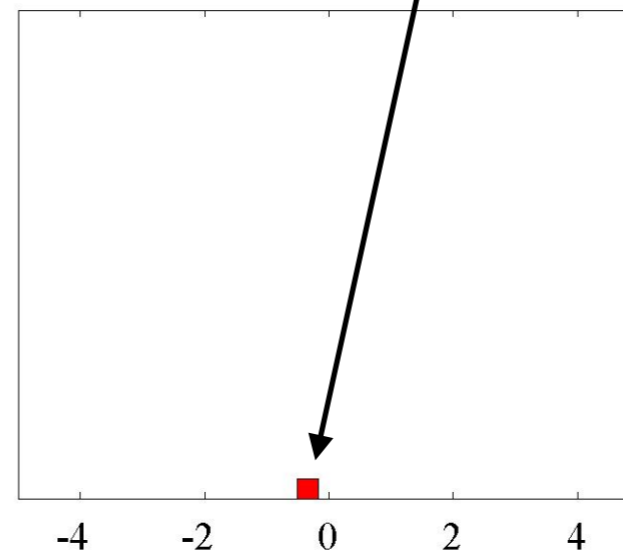
But there isn't really anything to permute, or is there?



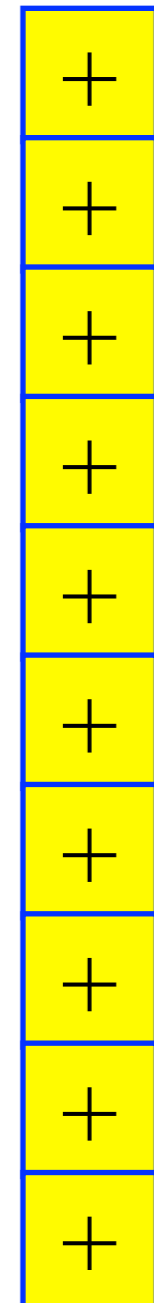
# Examples of exchangeability: Single group average



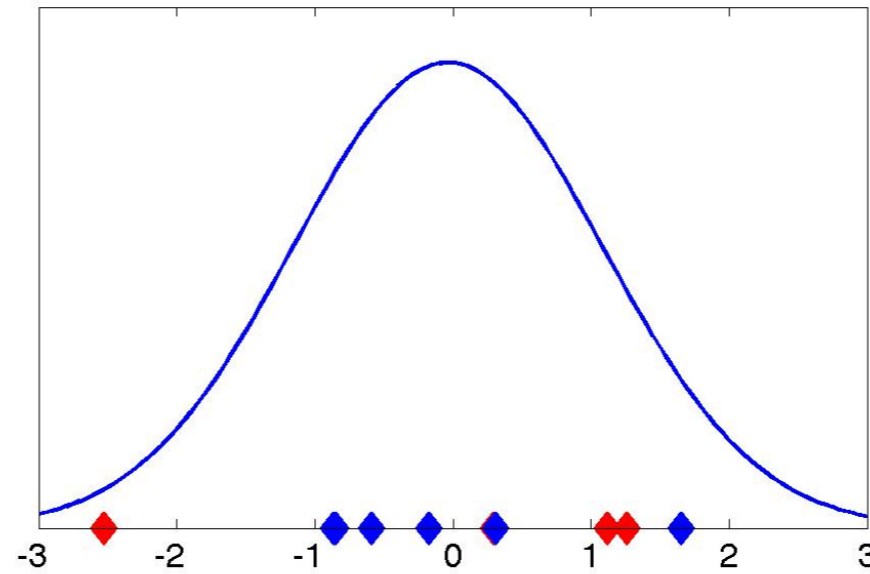
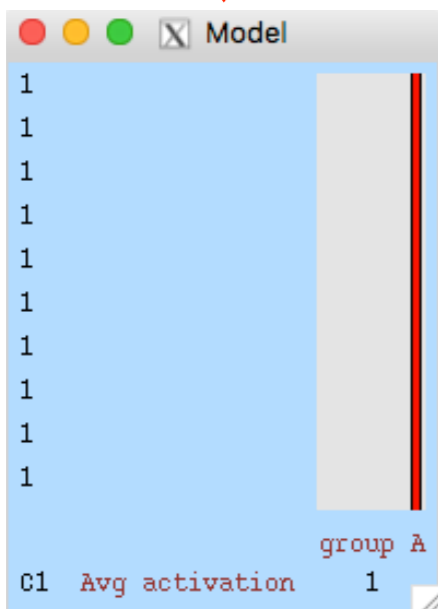
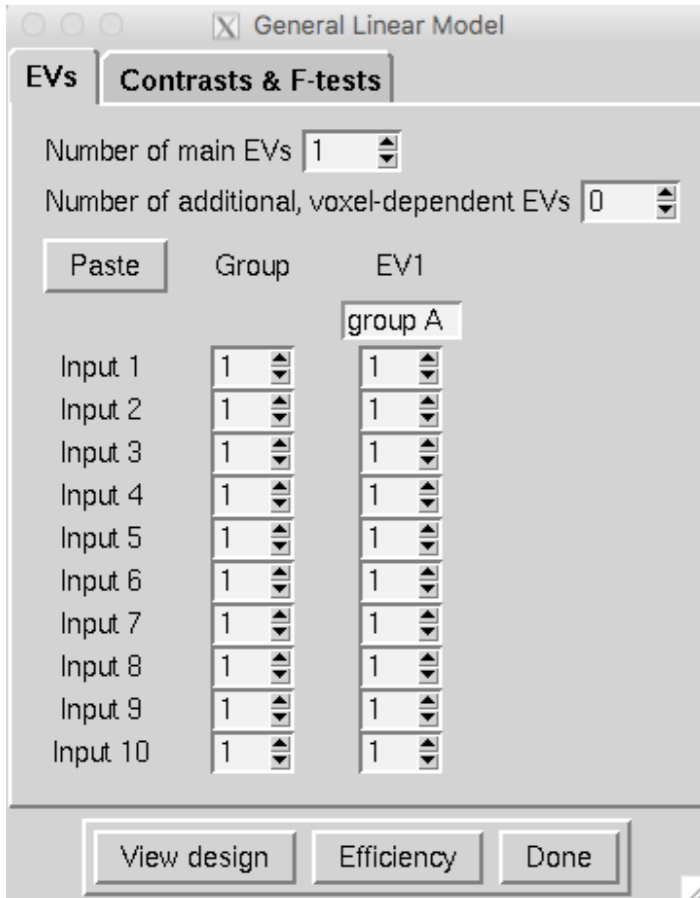
$$t = -0.17$$



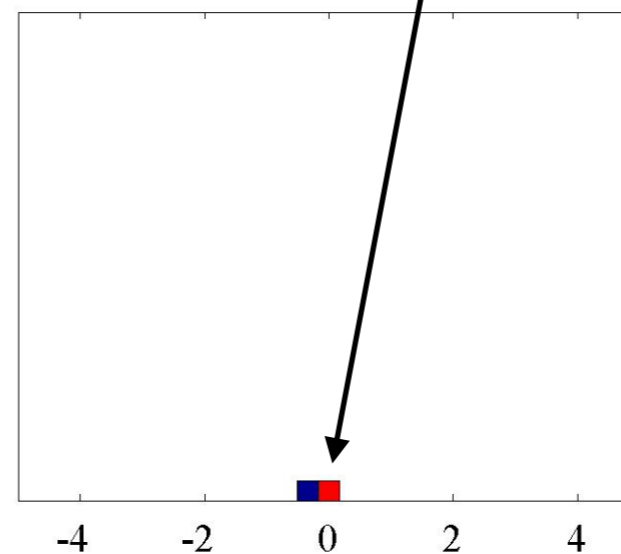
Original



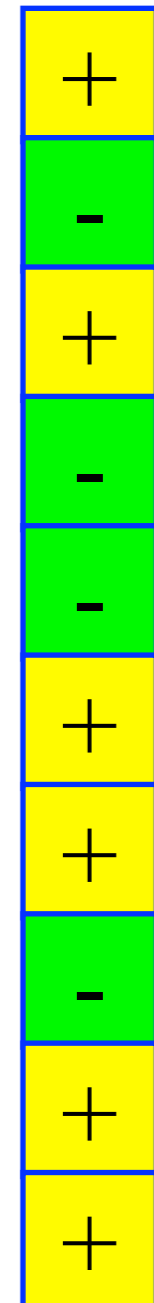
# Examples of exchangeability: Single group average



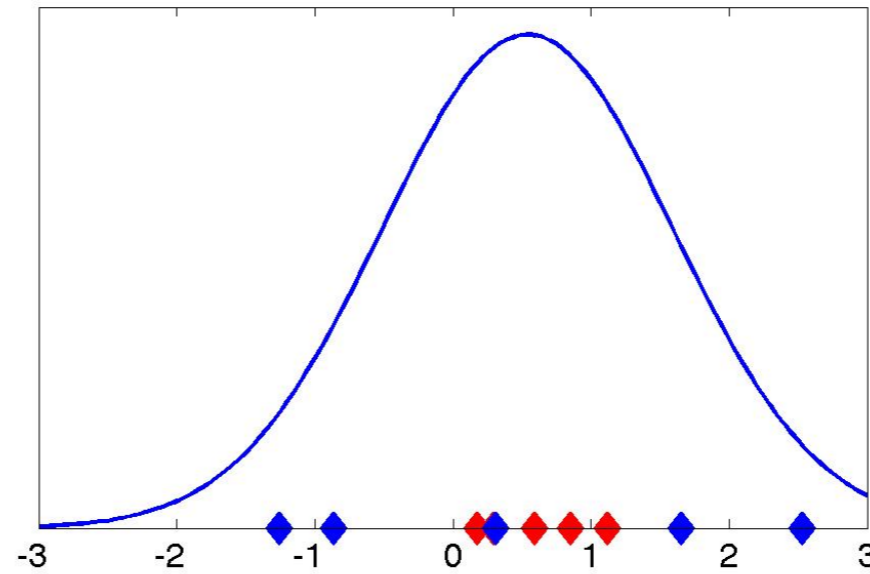
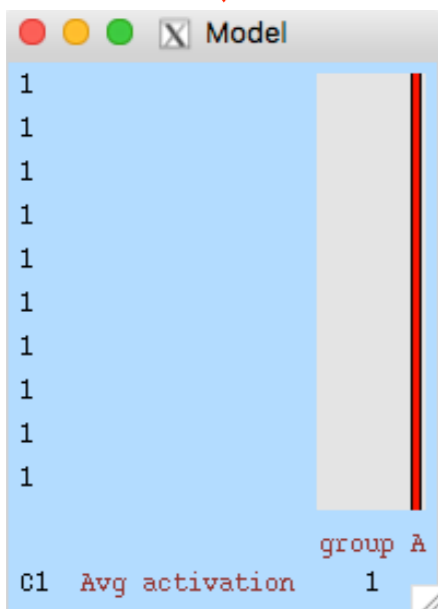
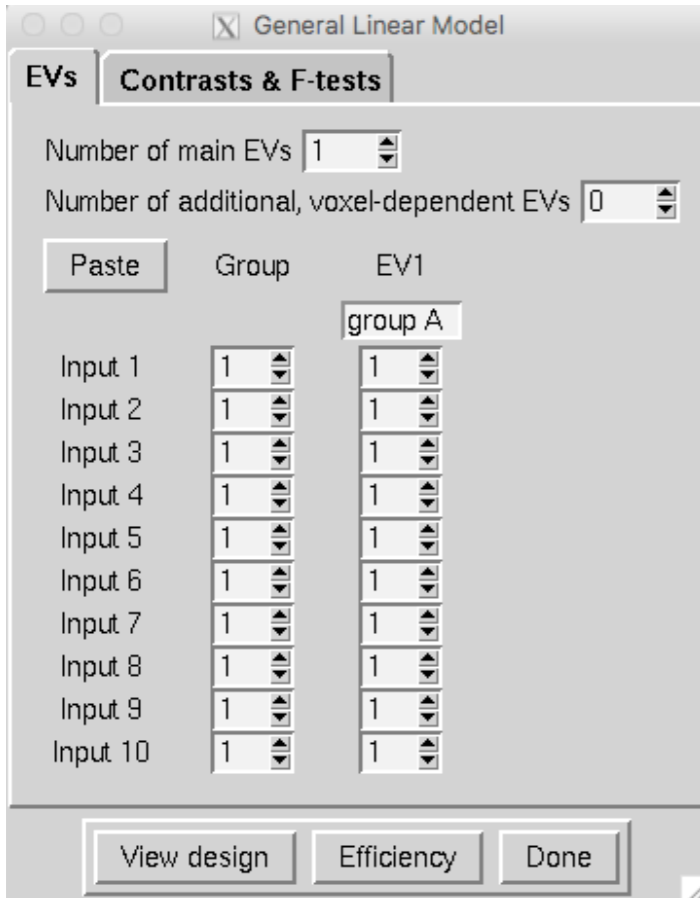
$$t = -0.09$$



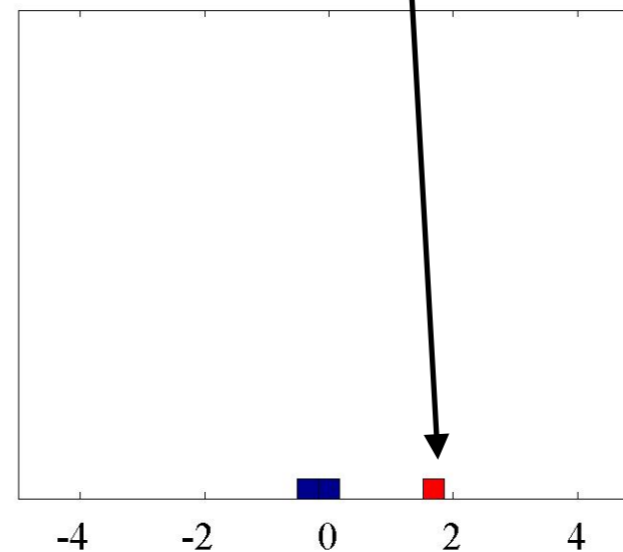
First flip



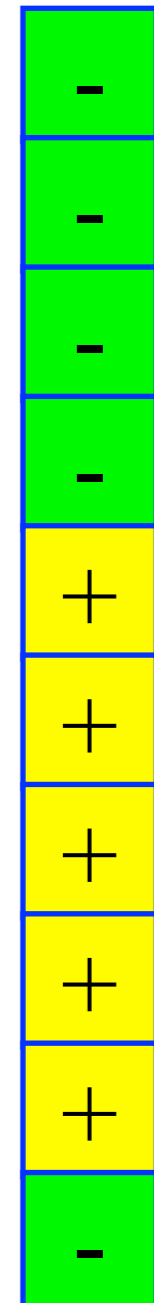
# Examples of exchangeability: Single group average



$$t = 1.54$$

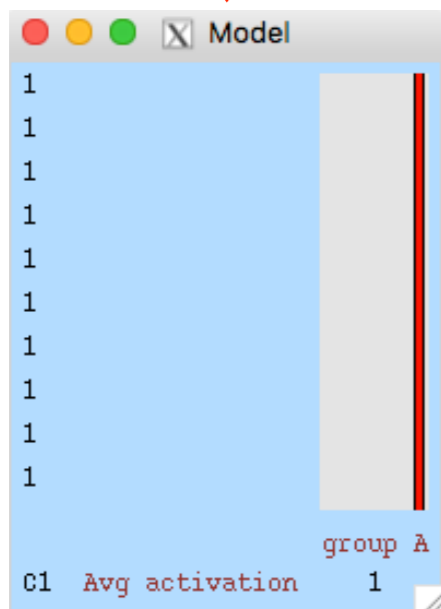
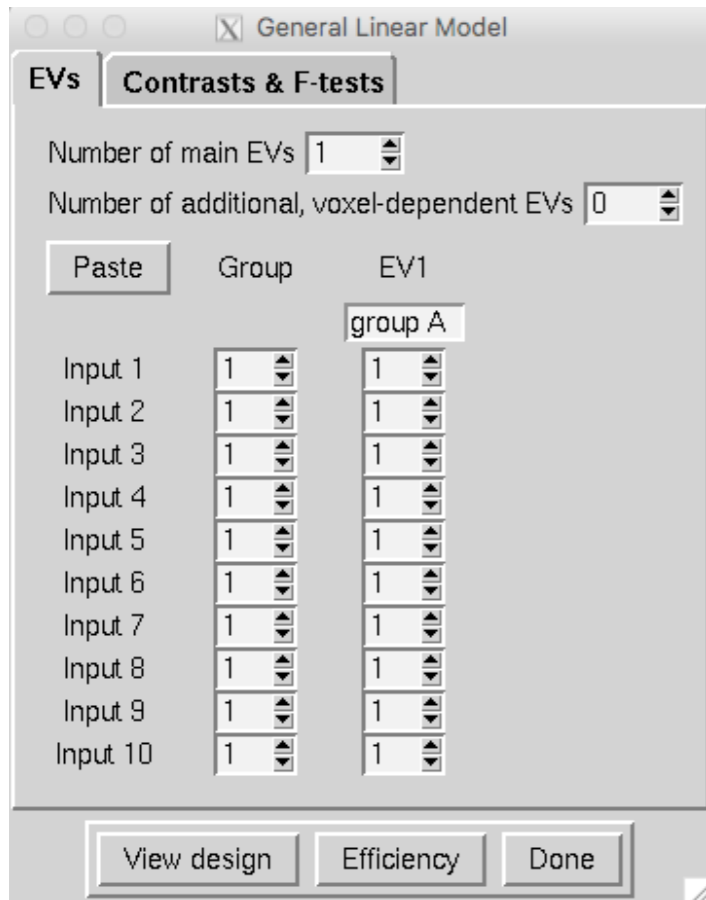


Second flip

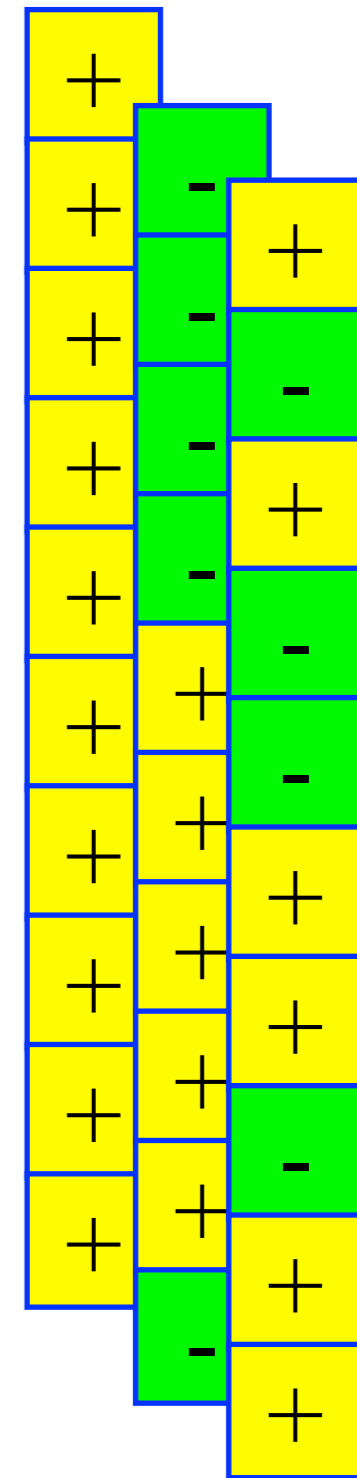
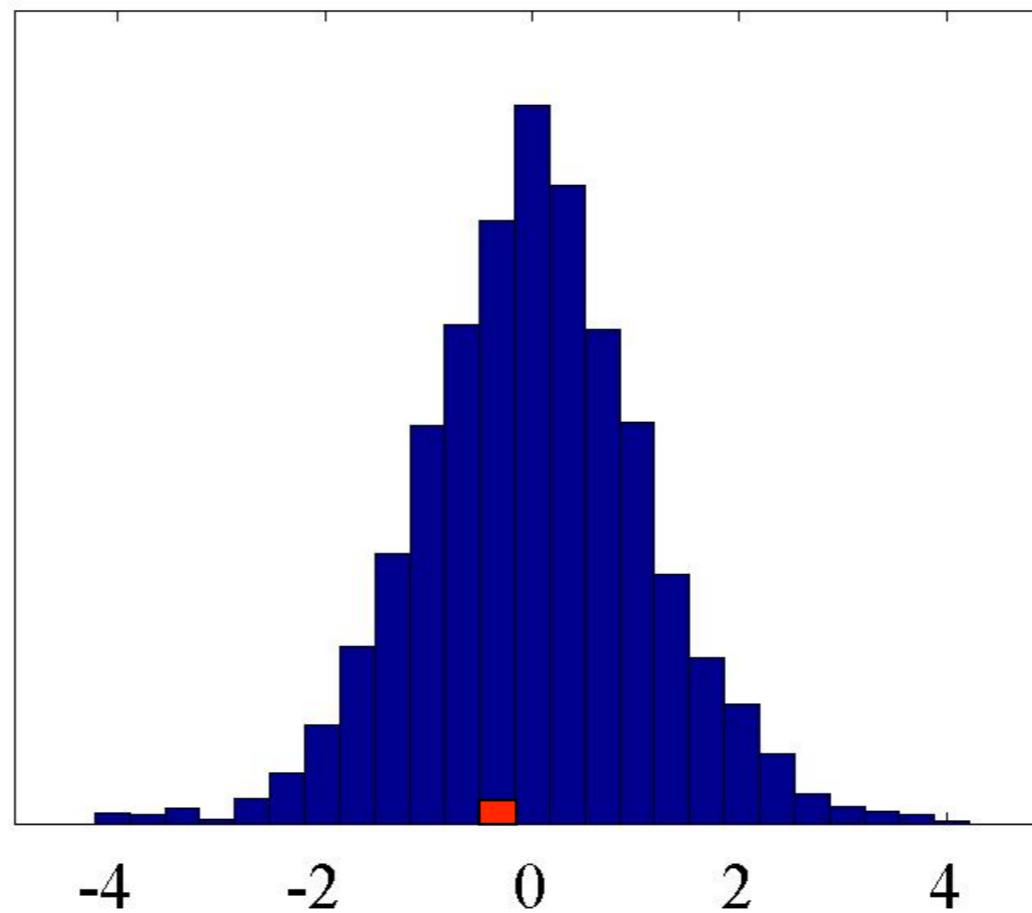




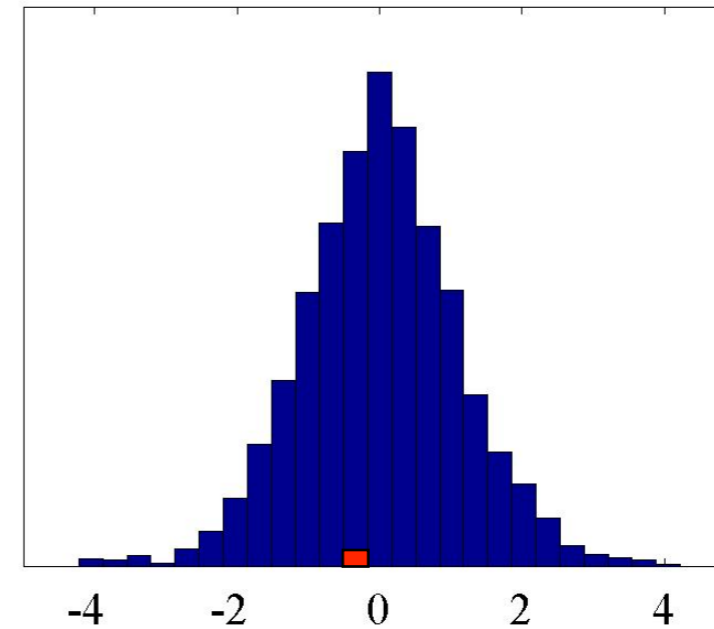
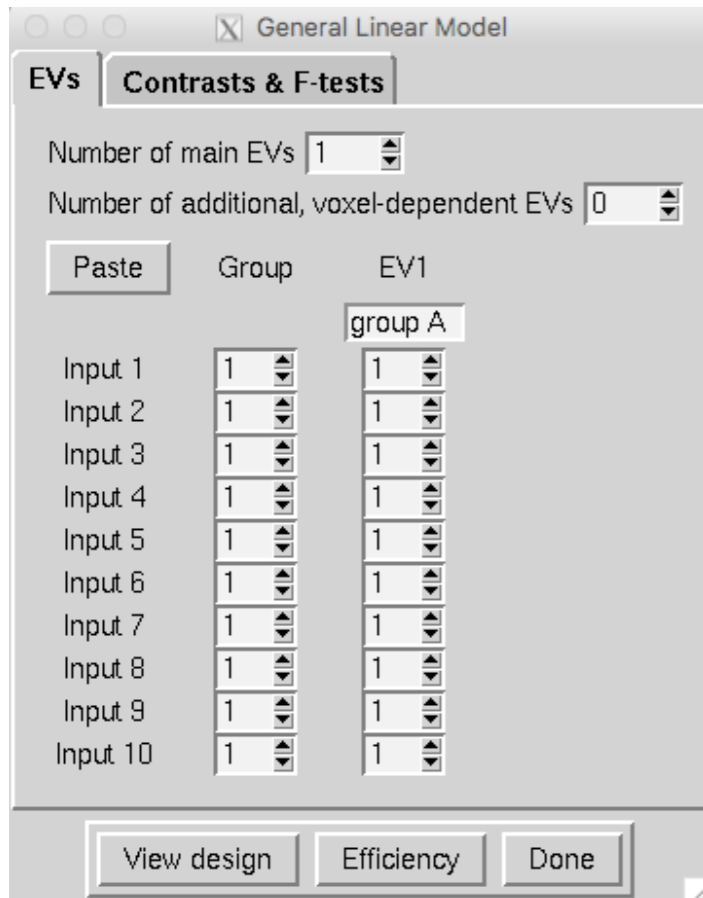
# Examples of exchangeability: Single group average



Etc ...

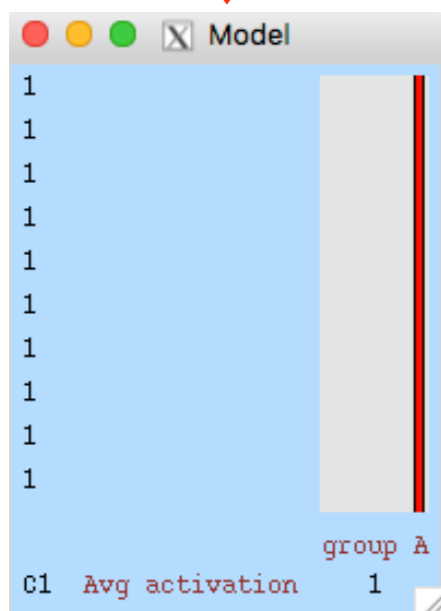


# Examples of exchangeability: Single group average



And the assumptions are:

- Symmetric errors
- Errors independent
- Subjects drawn from a single population





# Examples of exchangeability: Two groups paired

General Linear Model  
EVs | Contrasts & F-tests

Number of main EVs: 6  
Number of additional, voxel-dependent EVs: 0

Input	Group	EV1 (A>B)	EV2 (Subj 1)	EV3 (Subj 2)	EV4 (Subj 3)	EV5 (Subj 4)	EV6 (Subj 5)
Input 1	1	1	1	0	0	0	0
Input 2	1	-1	1	0	0	0	0
Input 3	2	1	0	1	0	0	0
Input 4	2	-1	0	1	0	0	0
Input 5	3	1	0	0	1	0	0
Input 6	3	-1	0	0	1	0	0
Input 7	4	1	0	0	0	1	0
Input 8	4	-1	0	0	0	1	0
Input 9	5	1	0	0	0	0	1
Input 10	5	-1	0	0	0	0	1

Buttons: Paste, Setup orthogonalisation, View design, Efficiency, Done

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

Assumed covariance matrix



General Linear Model

EVs **Contrasts & F-tests**

Number of main EVs: 6  
Number of additional, voxel-dependent EVs: 0

Group	EV1	EV2	EV3	EV4	EV5	EV6
	A>B	Subj 1	Subj 2	Subj 3	Subj 4	Subj 5
Input 1	1	1	0	0	0	0
Input 2	1	1	0	0	0	0
Input 3	2	0	1	0	0	0
Input 4	2	-1	0	1	0	0
Input 5	3	1	0	1	0	0
Input 6	3	-1	0	0	1	0
Input 7	4	1	0	0	1	0
Input 8	4	-1	0	0	1	0
Input 9	5	1	0	0	0	1
Input 10	5	-1	0	0	0	1

Setup orthogonalisations

View design Efficiency Done



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



# Examples of exchangeability: Two groups paired

Assumed covariance matrix



General Linear Model

EVs | Contrasts & F-tests

Number of main EVs: 6  
Number of additional, voxel-dependent EVs: 0

Group	EV1	EV2	EV3	EV4	EV5	EV6
	A>B	Subj 1	Subj 2	Subj 3	Subj 4	Subj 5
Input 1	1	1	0	0	0	0
Input 2	1	1	0	0	0	0
Input 3	2	0	1	0	0	0
Input 4	2	-1	0	1	0	0
Input 5	3	1	0	1	0	0
Input 6	3	-1	0	0	1	0
Input 7	4	1	0	0	1	0
Input 8	4	-1	0	0	0	1
Input 9	5	1	0	0	0	1
Input 10	5	-1	0	0	0	1

Setup orthogonalisations

View design | Efficiency | Done



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





# Examples of exchangeability: Two groups paired

General Linear Model

EVs **Contrasts & F-tests**

Number of main EVs: 6  
Number of additional, voxel-dependent EVs: 0

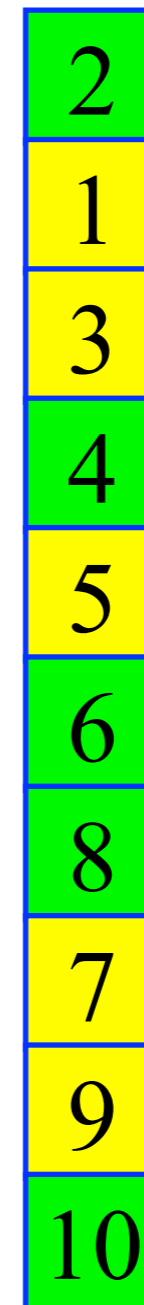
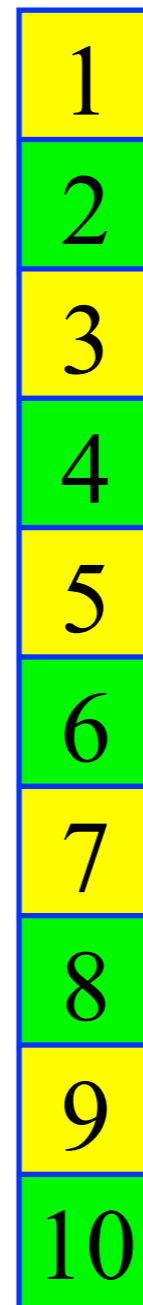
Group	EV1	EV2	EV3	EV4	EV5	EV6
	A>B	Subj 1	Subj 2	Subj 3	Subj 4	Subj 5
Input 1	1	1	0	0	0	0
Input 2	1	1	0	0	0	0
Input 3	2	0	1	0	0	0
Input 4	2	-1	0	1	0	0
Input 5	3	1	0	1	0	0
Input 6	3	-1	0	0	1	0
Input 7	4	1	0	0	1	0
Input 8	4	-1	0	0	1	0
Input 9	5	1	0	0	0	1
Input 10	5	-1	0	0	0	1

Setup orthogonalisations

View design Efficiency Done



Original Perm 1 Perm 2 ...





# 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



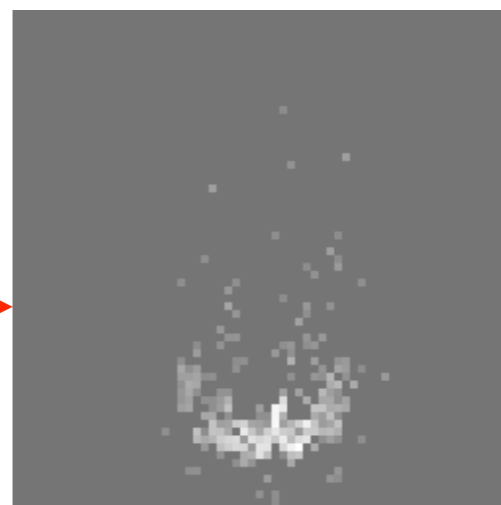
# Clustering cookbook

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

z stat image



Threshold at  
(arbitrary!) z level





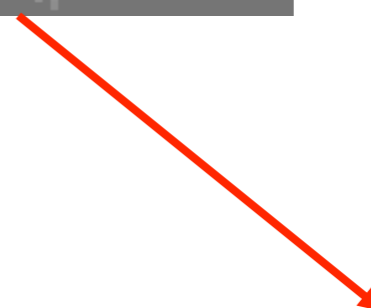
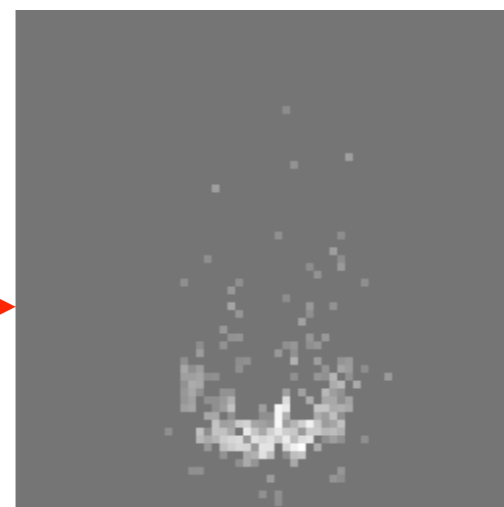
# Clustering cookbook

Instead of resel-based correction, we can do clustering

z stat image



Threshold at  
(arbitrary!) z level



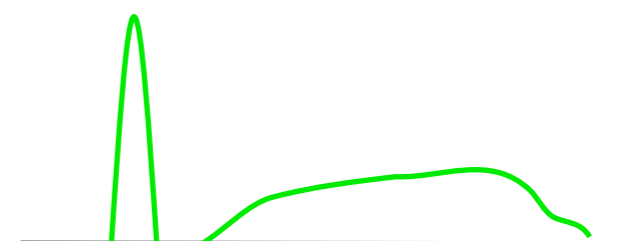
Form clusters from surviving voxels.  
Calculate the size threshold  $u(R,z)$ .  
Any cluster larger than  $u$  “survives” and we reject  
the null-hypothesis for that.





# How do we choose the (arbitrary!) z-threshold?

This is arbitrary and a trade-off

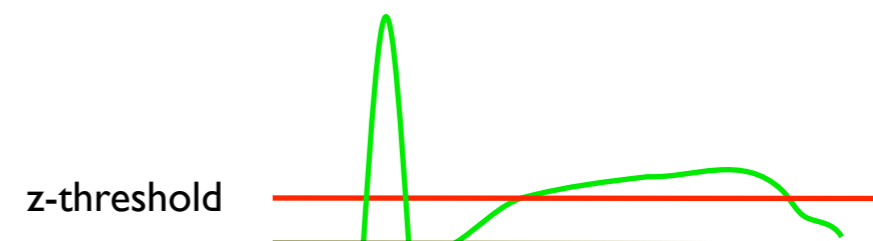




# How do we choose the (arbitrary!) z-threshold?

This is arbitrary and a trade-off

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



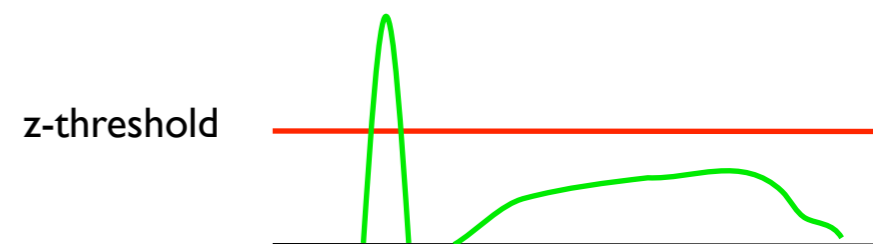
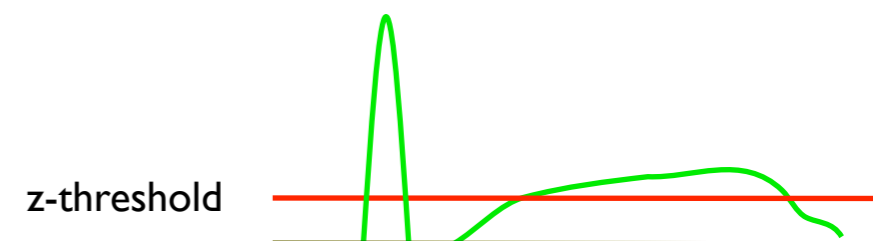


# How do we choose the (arbitrary!) z-threshold?

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

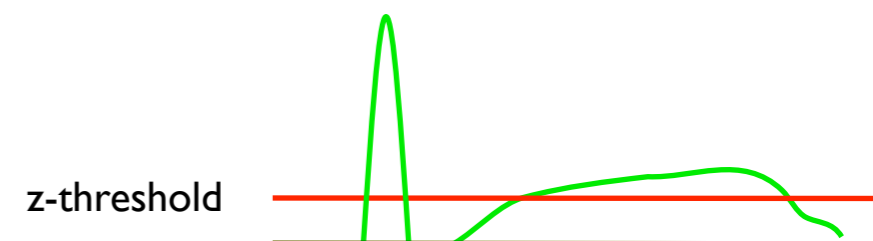




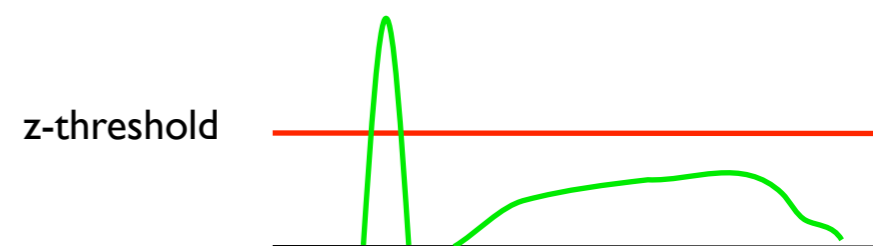
# How do we choose the (arbitrary!) z-threshold?

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



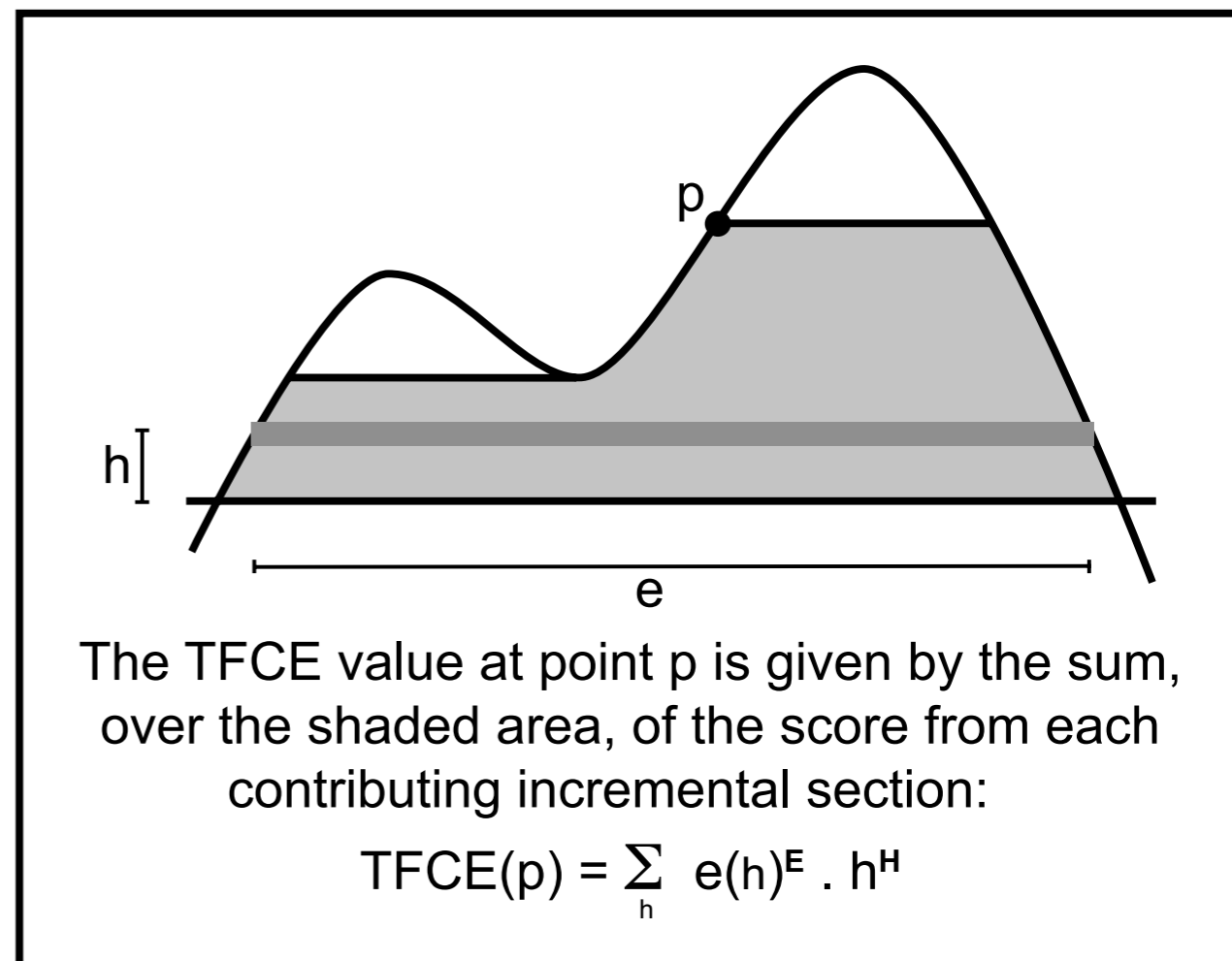


# TFCE

## Threshold-Free Cluster Enhancement

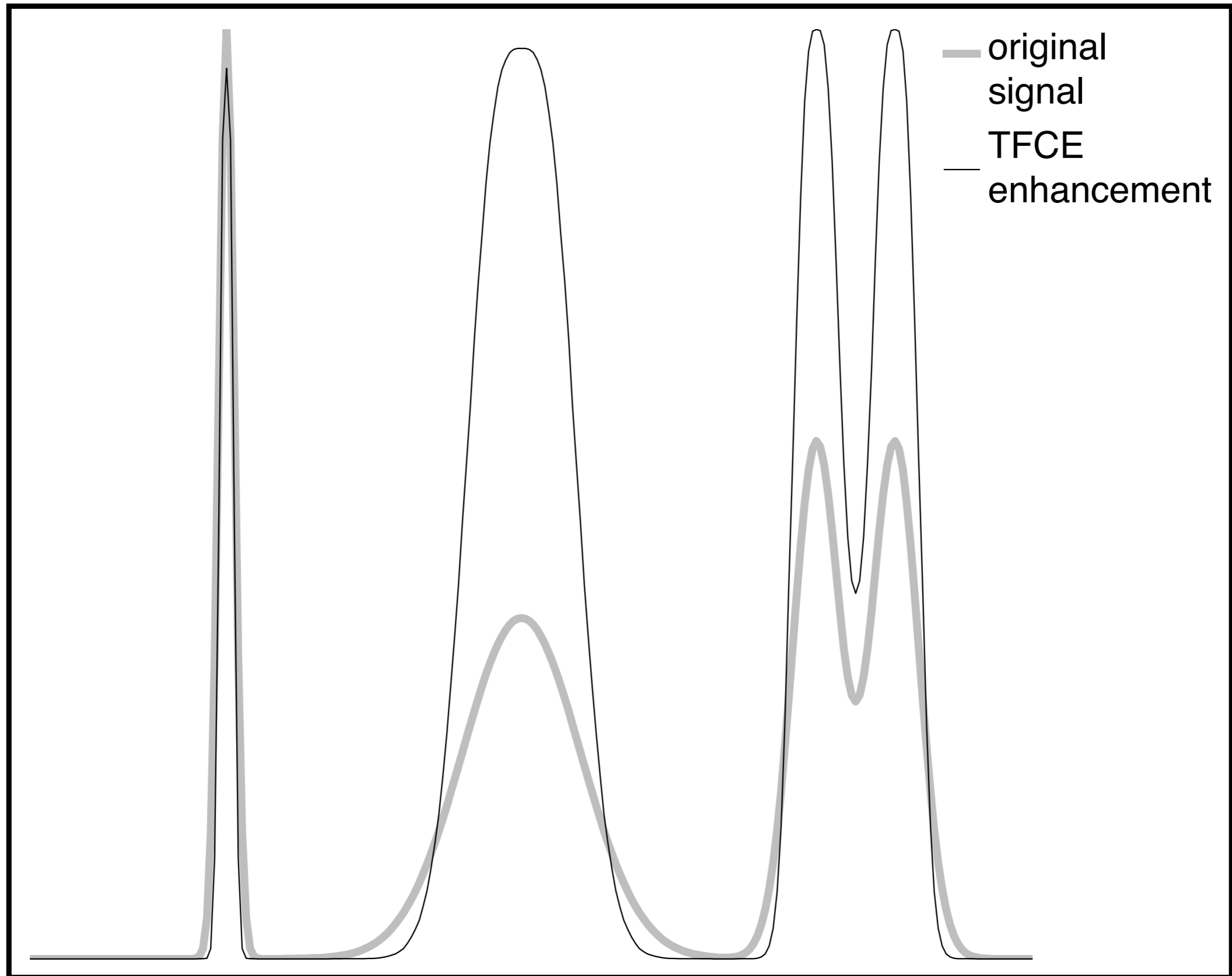
[Smith & Nichols, NeuroImage 2009]

- 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



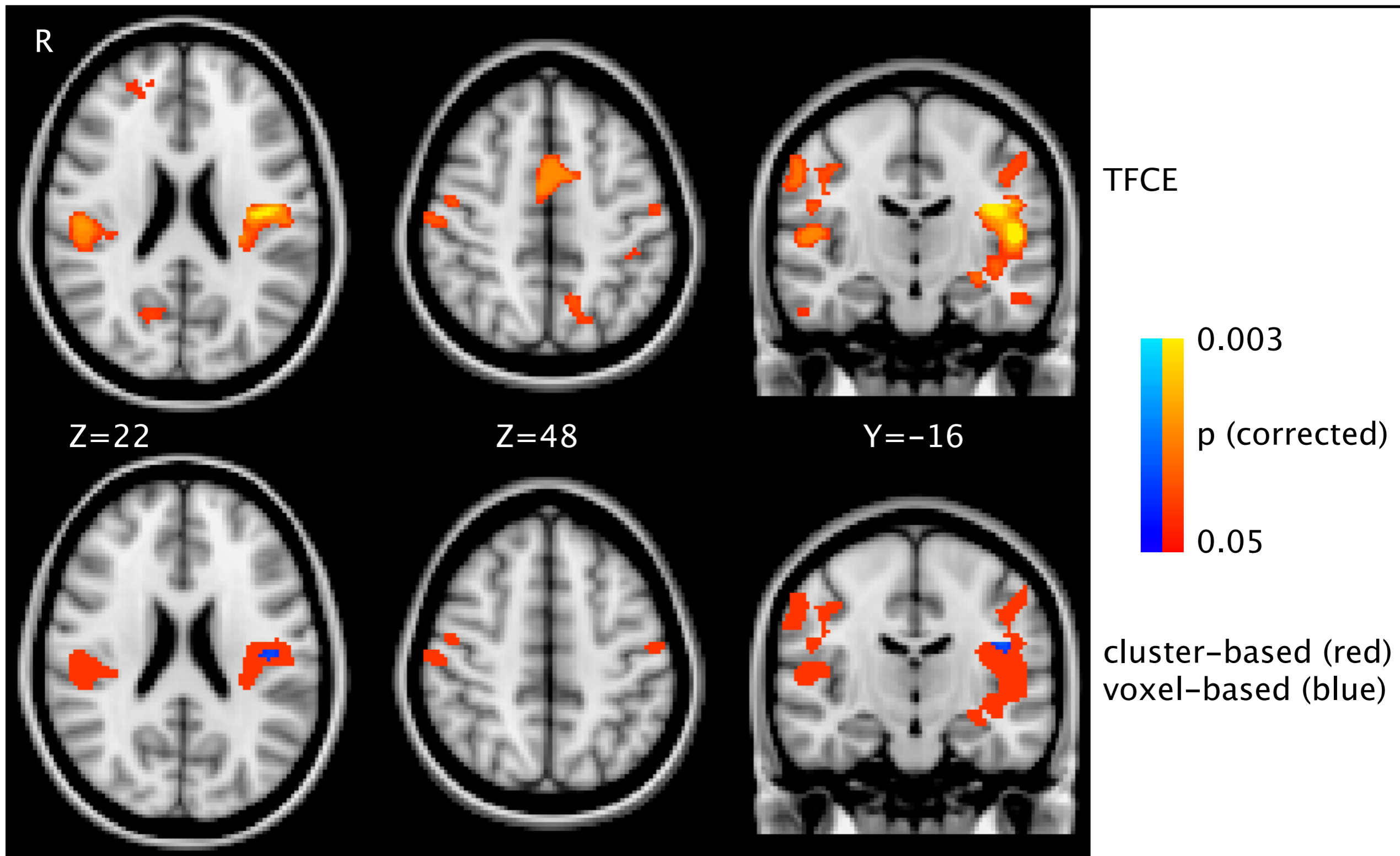


# Qualitative example





# TFCE for FSL-VBM

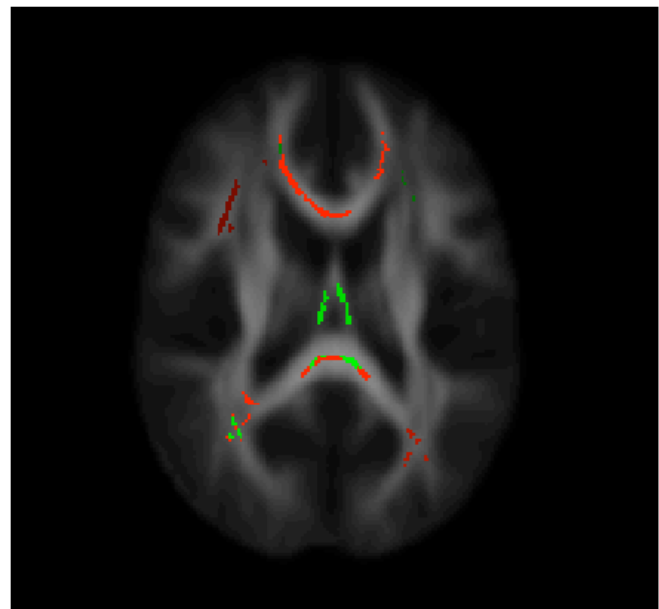
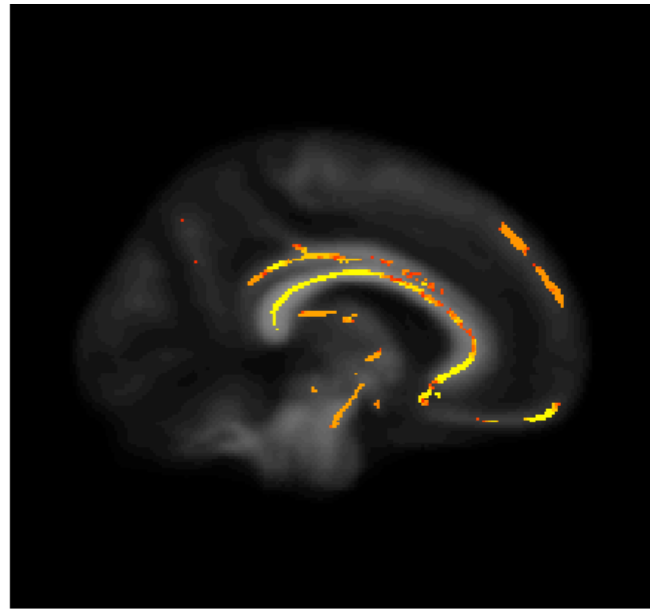
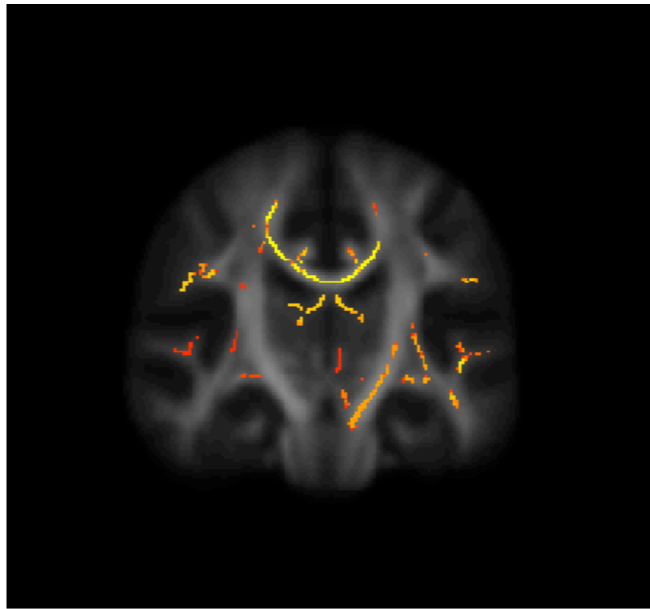
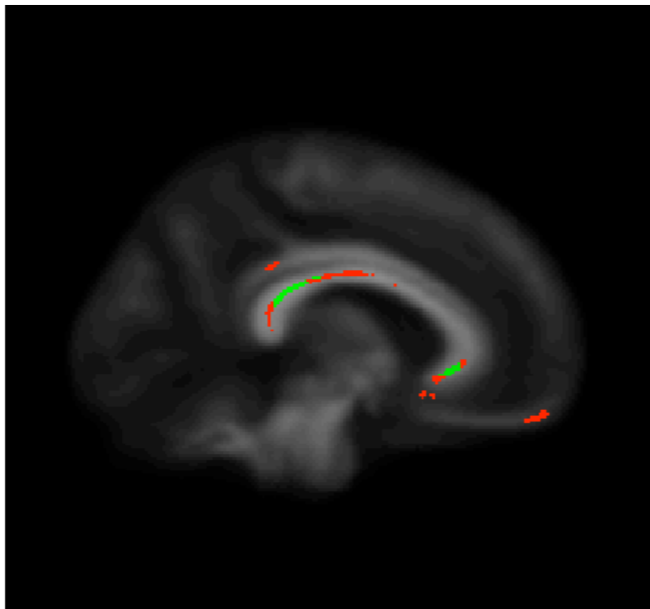
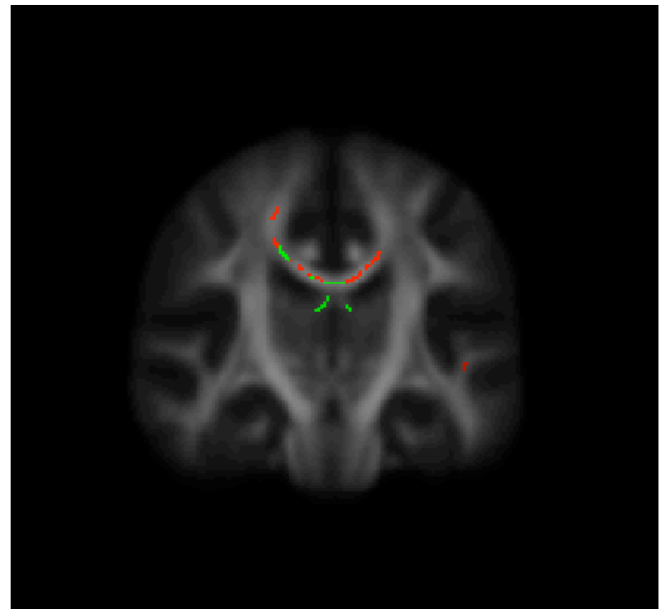




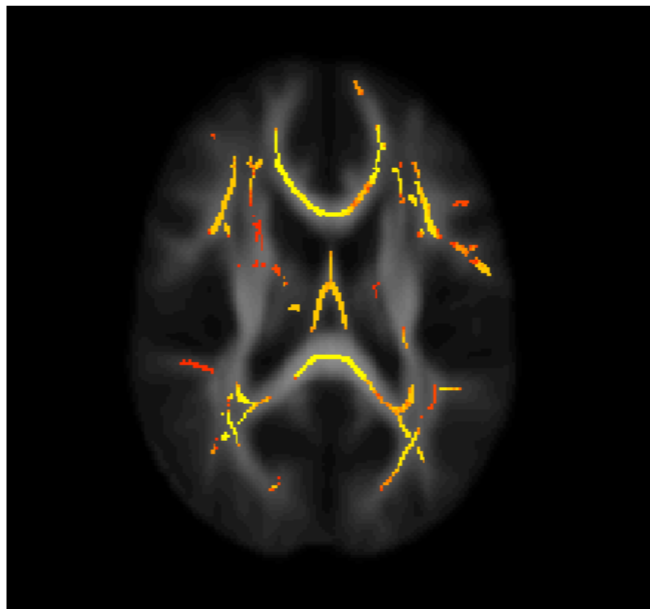
# TFCE for TBSS

controls > schizophrenics

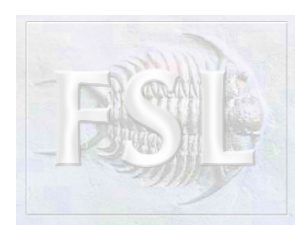
$p < 0.05$  corrected for multiple comparisons across space, using randomise



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

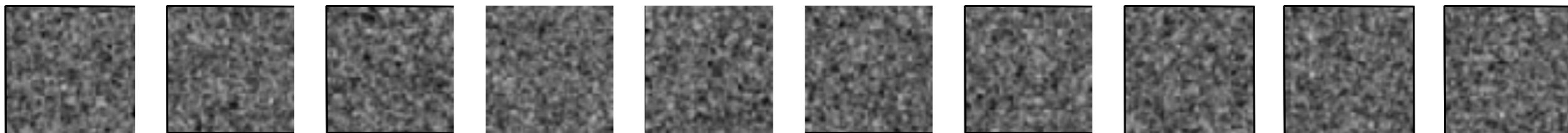


- **FDR: False Discovery Rate**  
A “new” way to look at inference.
- **Uncorrected (for multiple-comparisons):**
  - Is equivalent to saying: “I am happy to nearly always say something silly about my experiments”.
  - **On average, 5% of all voxels are false positives**
- **Family-Wise Error (FWE):**
  - 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**

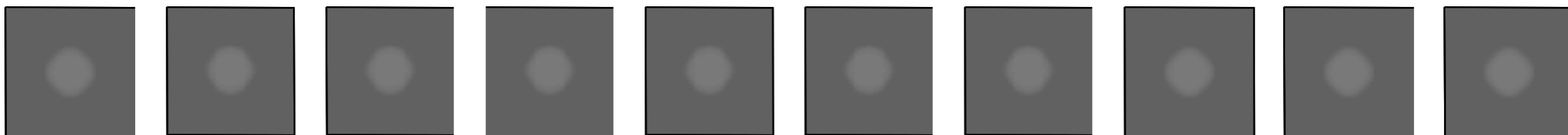


# Little imaging demonstration.

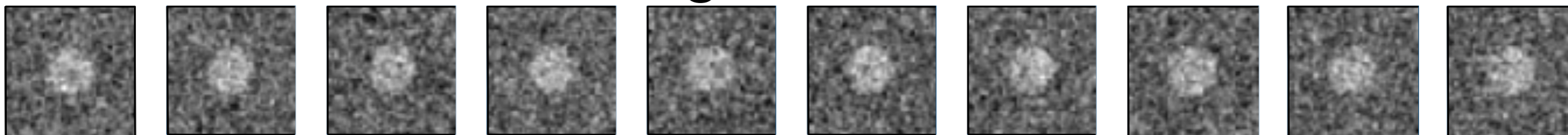
Noise



Signal



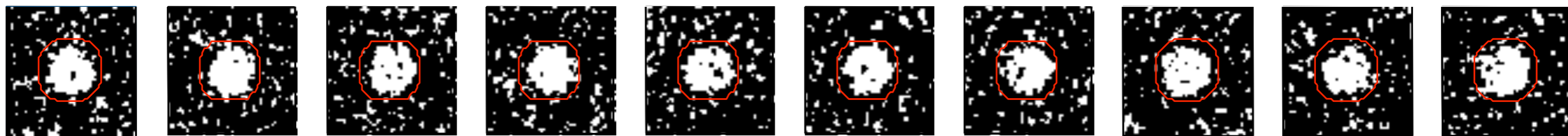
Signal+Noise





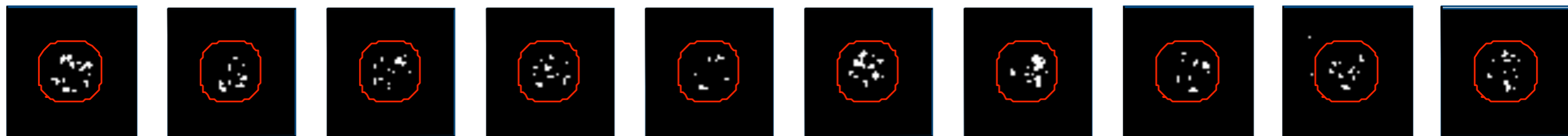


## uncorrected voxelwise control of FP rate at 10%



percentage of all null pixels that are False Positives

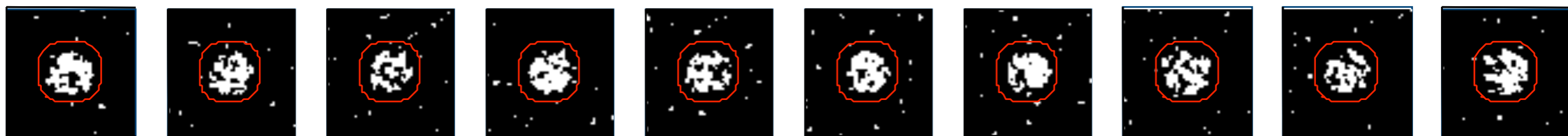
## control of FamilyWise Error rate at 10%



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.