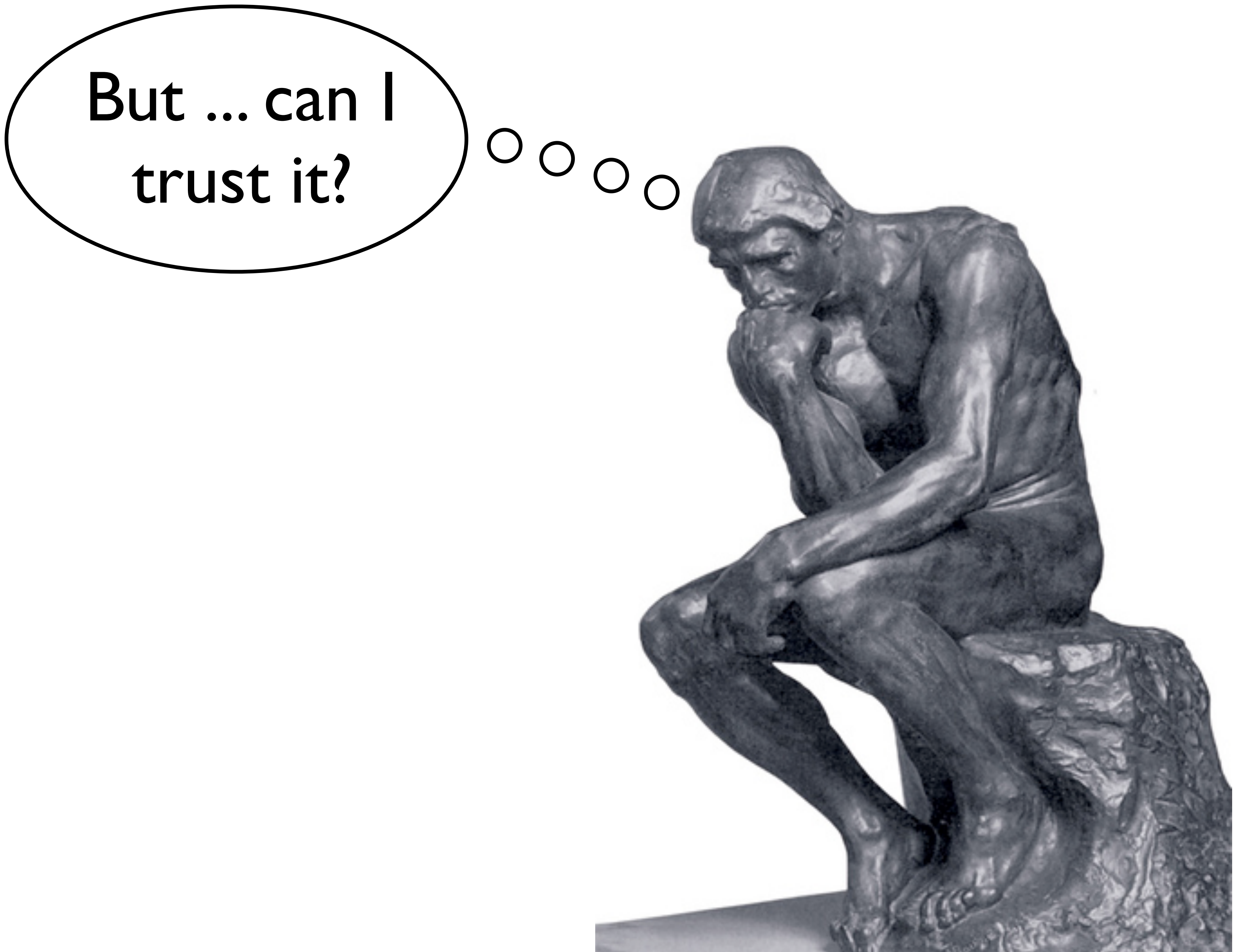
Inference

how surprising is your statistic? (thresholding)



Outline

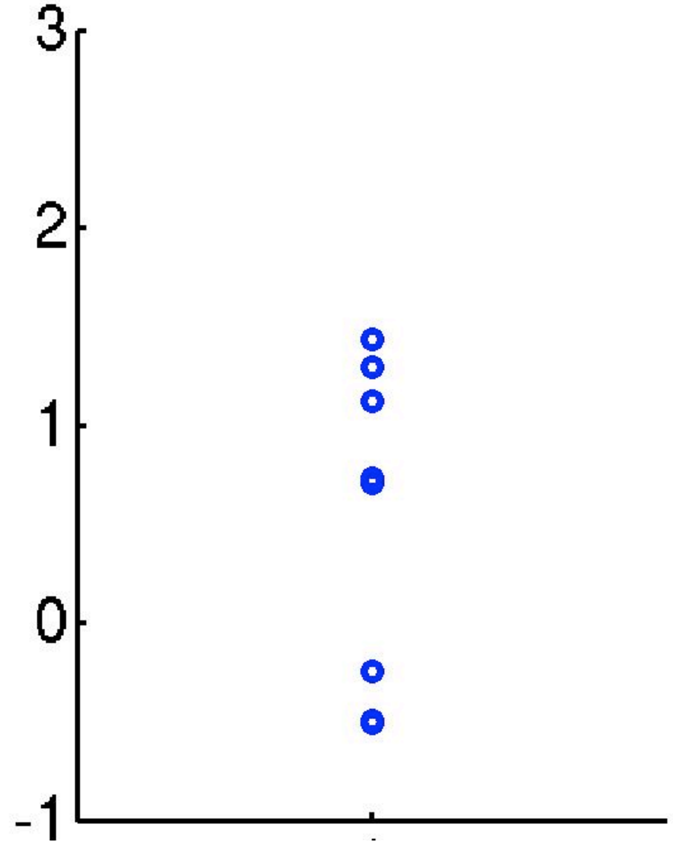
- Null-hypothesis and Null-distribution
- Multiple comparisons and Family-wise error
- Different ways of being surprised
  - Voxel-wise inference (Maximum z)
  - Cluster-wise inference (Maximum size)
- Parametric vs non-parametric tests
- Enhanced clusters
- FDR - False Discovery Rate

Outline

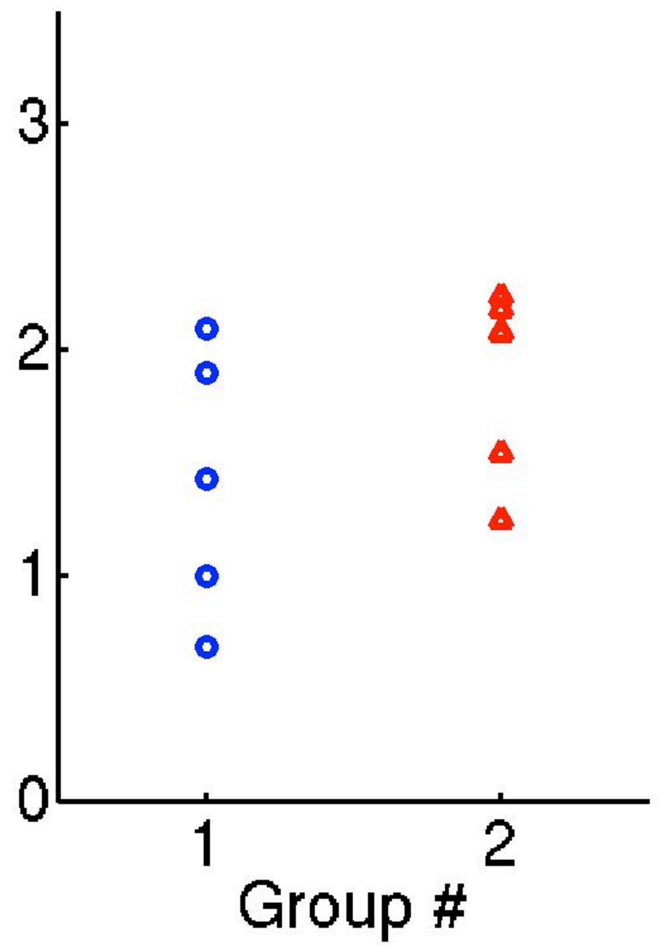
- Null-hypothesis and Null-distribution
- Multiple comparisons and Family-wise error
- Different ways of being surprised
  - Voxel-wise inference (Maximum z)
  - Cluster-wise inference (Maximum size)
- Parametric vs non-parametric tests
- Enhanced clusters
- FDR - False Discovery Rate

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 ?



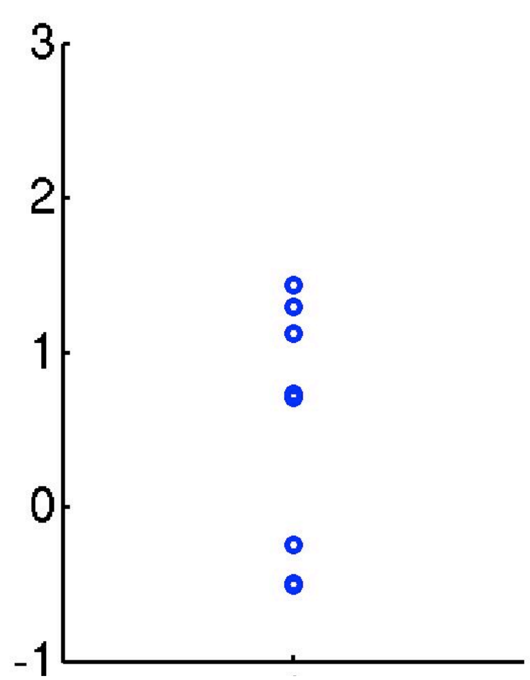
Different?

Tools of classical inference

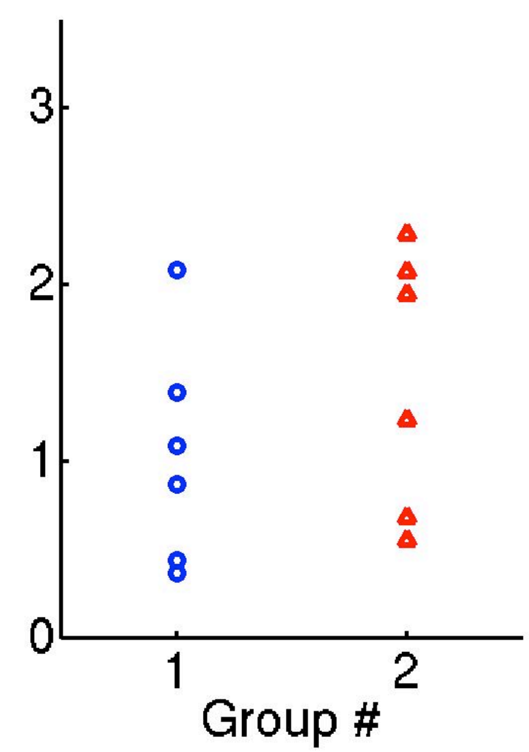
1. A null-hypothesis

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

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



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

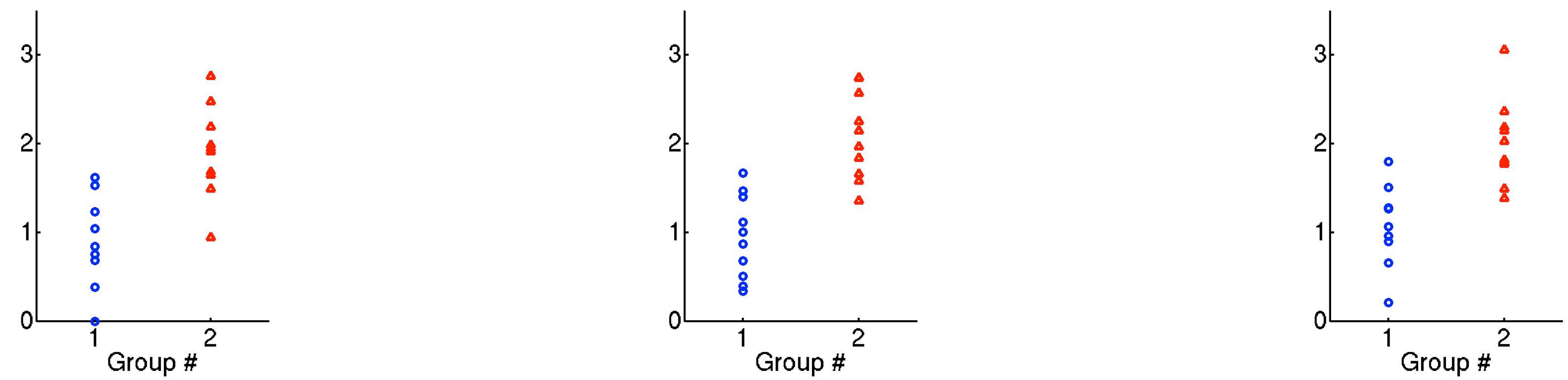


Tools of classical inference

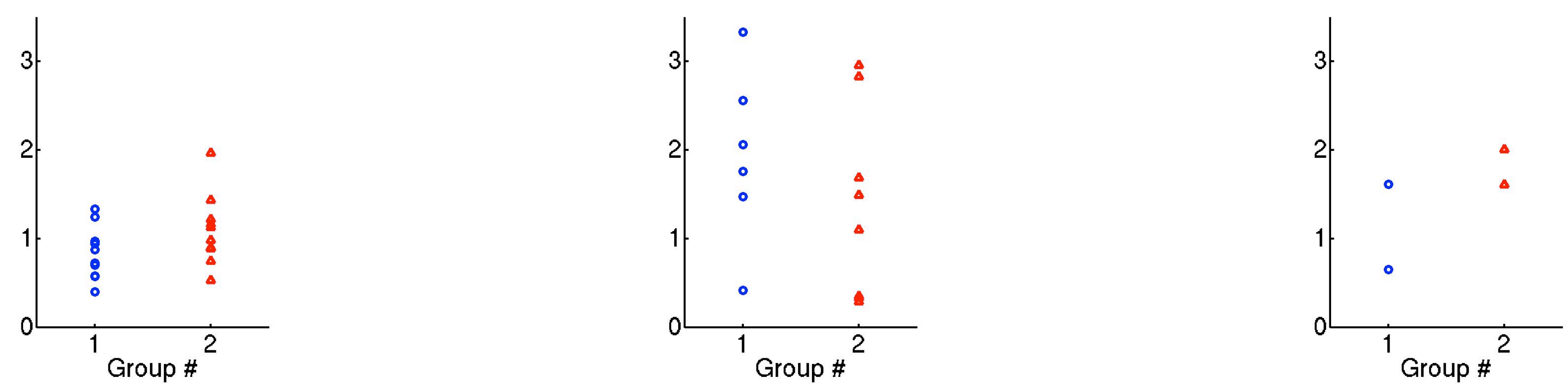
1. A null-hypothesis
2. A test-statistic

Assesses “trustworthiness”

Trustworthy



Dodgy



Tools of classical inference

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

Assesses “trustworthiness”

A *t*-statistic reflects precisely this

$$t = \sqrt{n}\frac{\overline{x_1} - \overline{x_2}}{\sqrt{\sigma^2}}$$

Large difference: Trustworthy

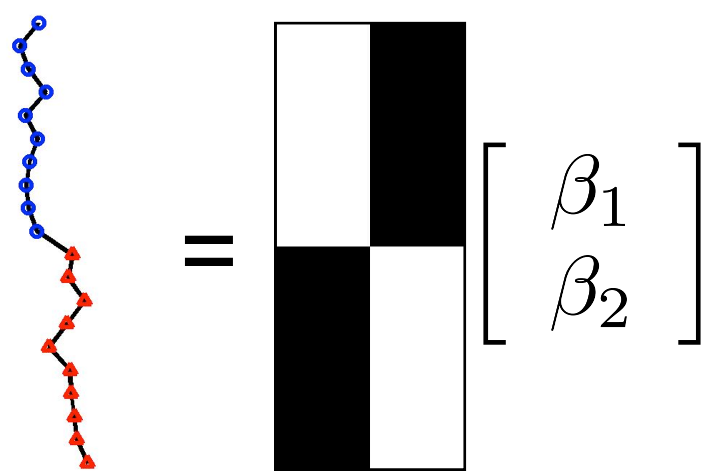
Many measurements: Trustworthy

Small variability: Trustworthy

Tools of classical inference

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

Or expressed in GLM lingo



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

Large difference: Trustworthy

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

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

Small variability: Trustworthy

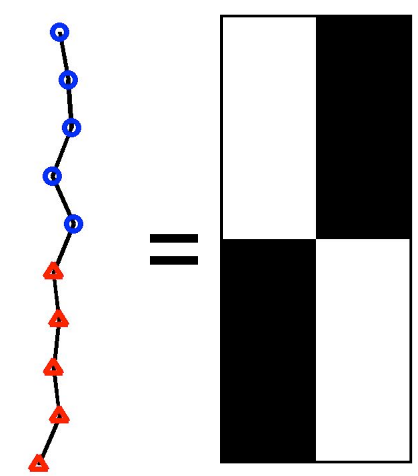
Many measurements: Trustworthy

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



$$= \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \mathbf{e}$$

$t = 2.19$

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

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

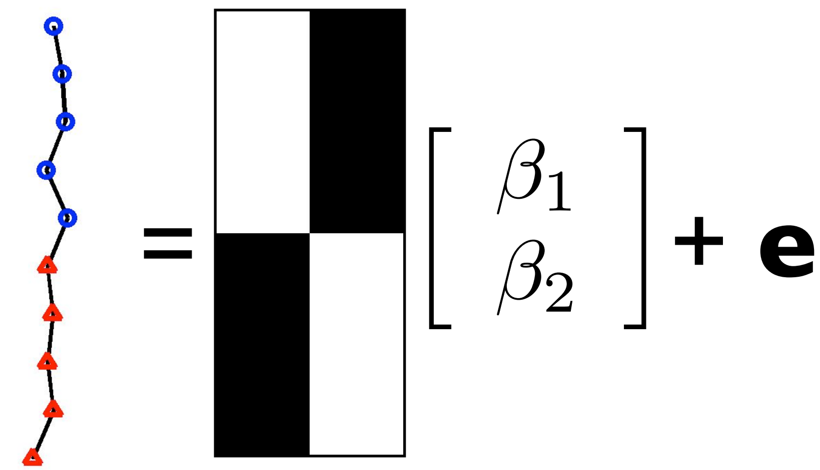
$\sigma^2 = 0.71$

Constant

Tools of classical inference

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

We might then get these data



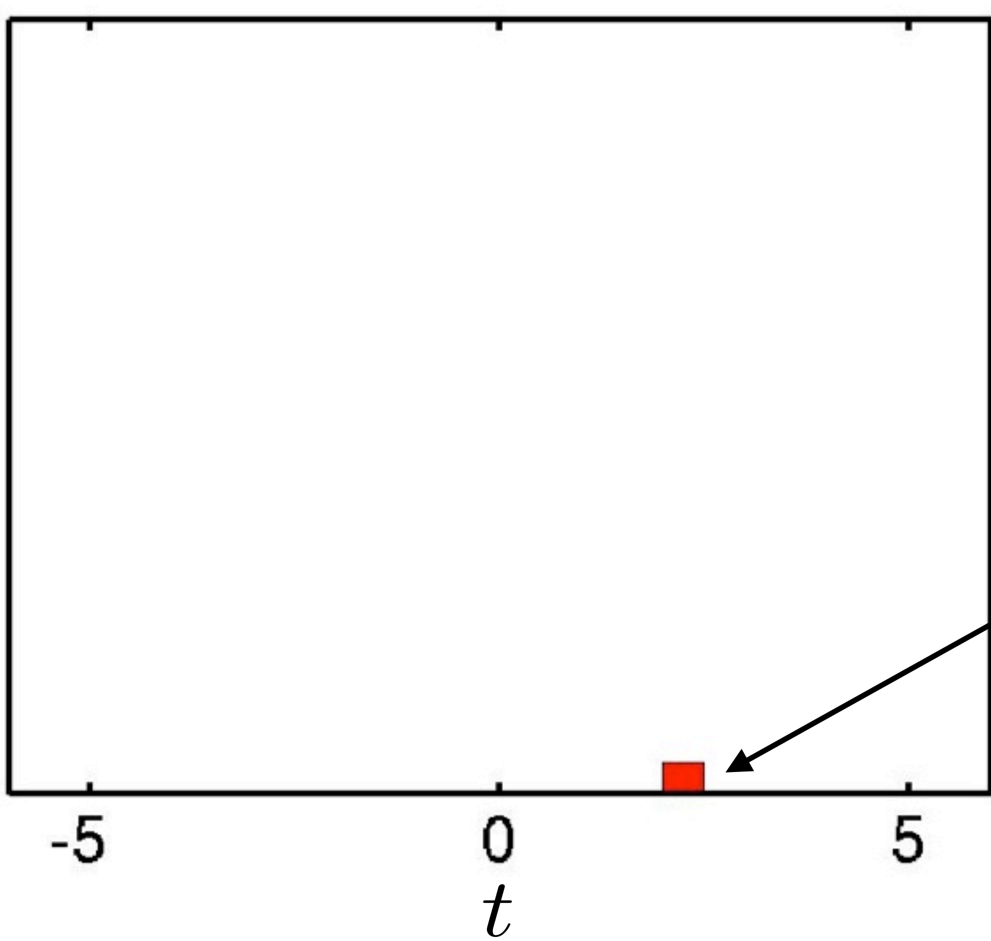
$\mathbf{c}^T\widehat{\boldsymbol{\beta}} = 1.17$

$t = 2.19$

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

$\sigma^2 = 0.71$

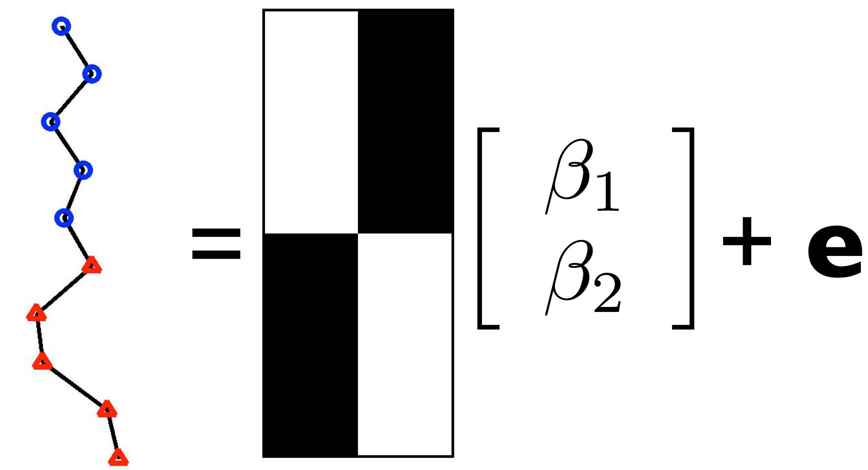
Constant

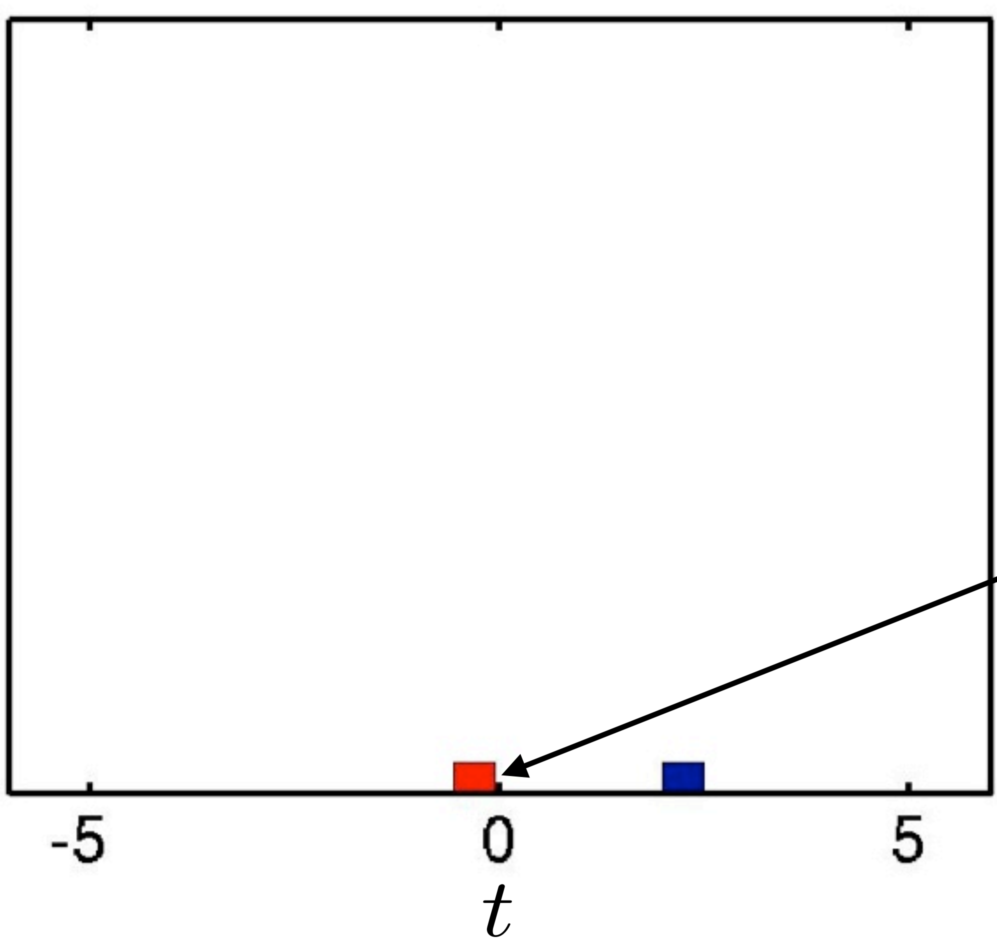


Tools of classical inference

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

or we could have gotten these





$t = -0.51$

$\mathbf{c}^T\widehat{\boldsymbol{\beta}} = -0.37$

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

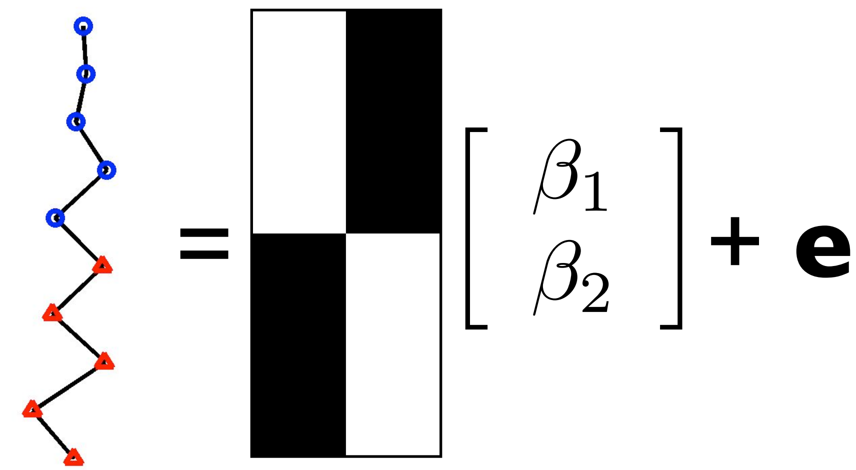
$\sigma^2 = 1.28$

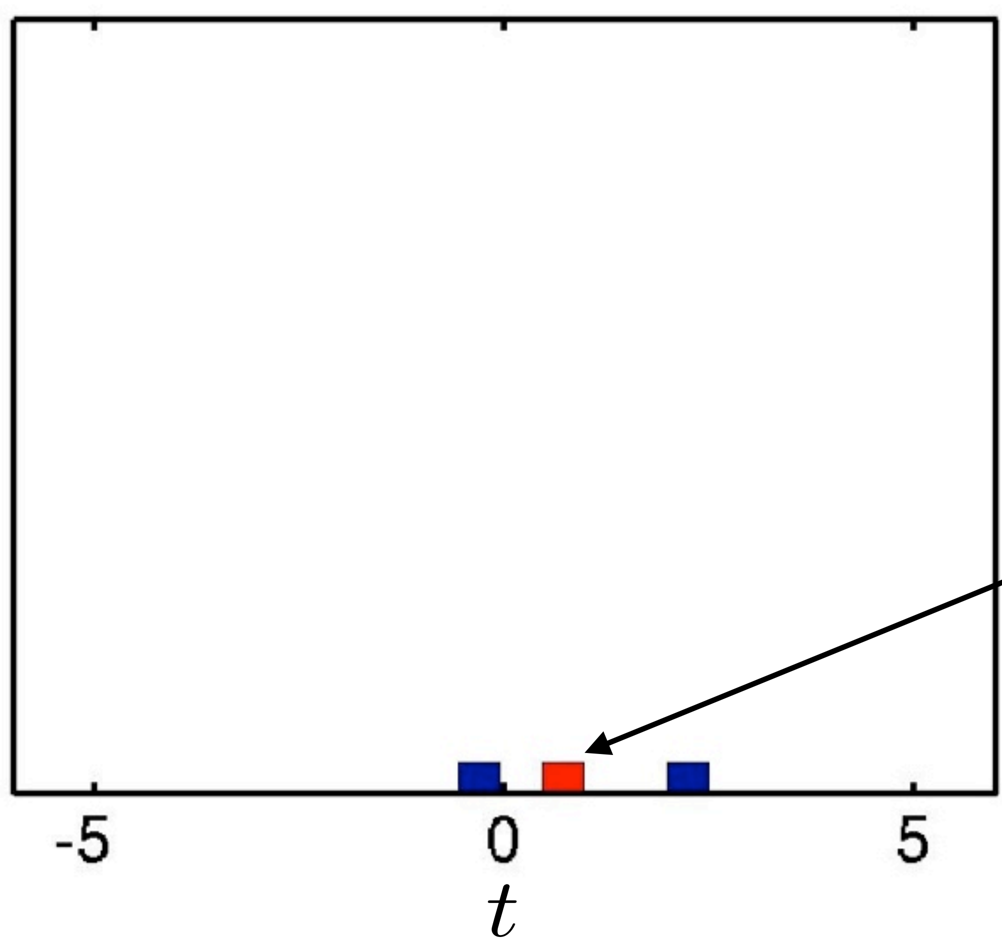
Constant

Tools of classical inference

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

maybe these





$t = 0.49$

$\mathbf{c}^T\widehat{\boldsymbol{\beta}} = 0.31$

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

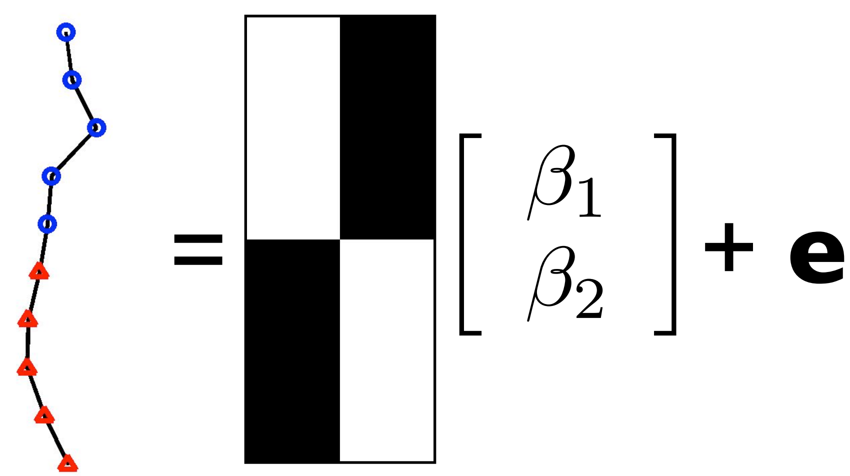
$\sigma^2 = 1.01$

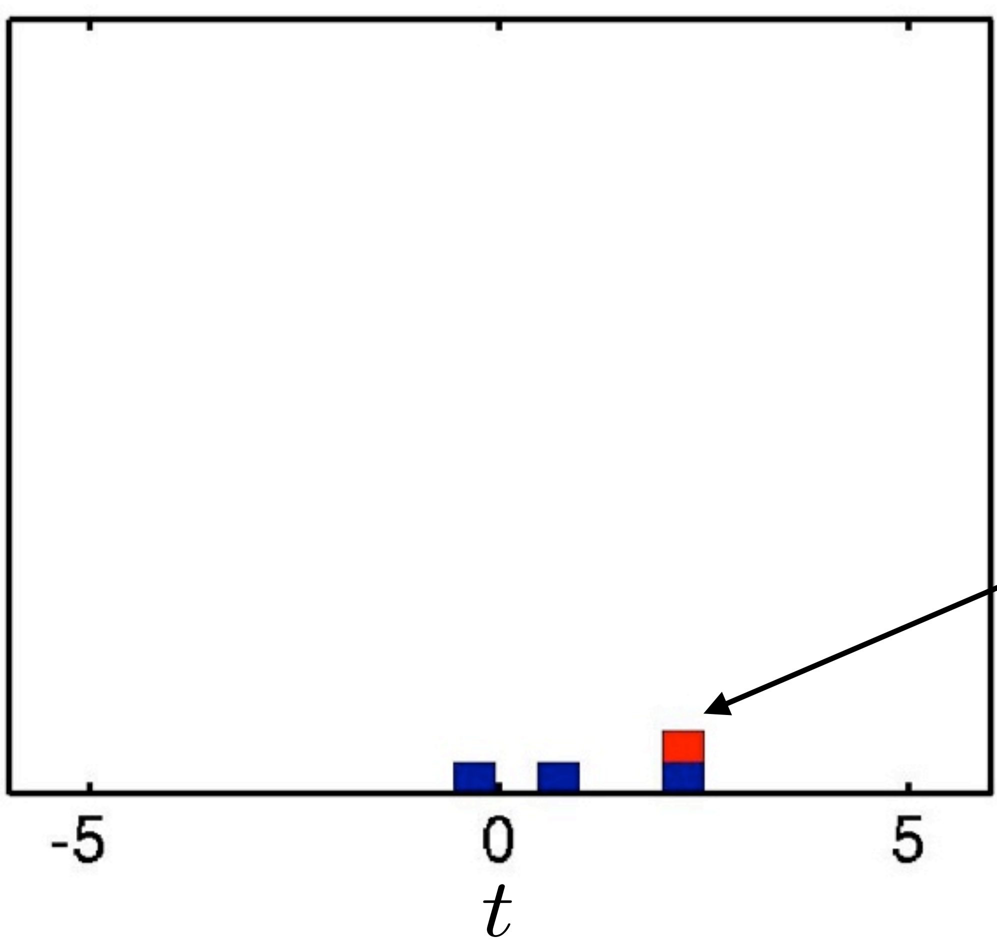
Constant

Tools of classical inference

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

or perhaps these





$$t = 2.19$$

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

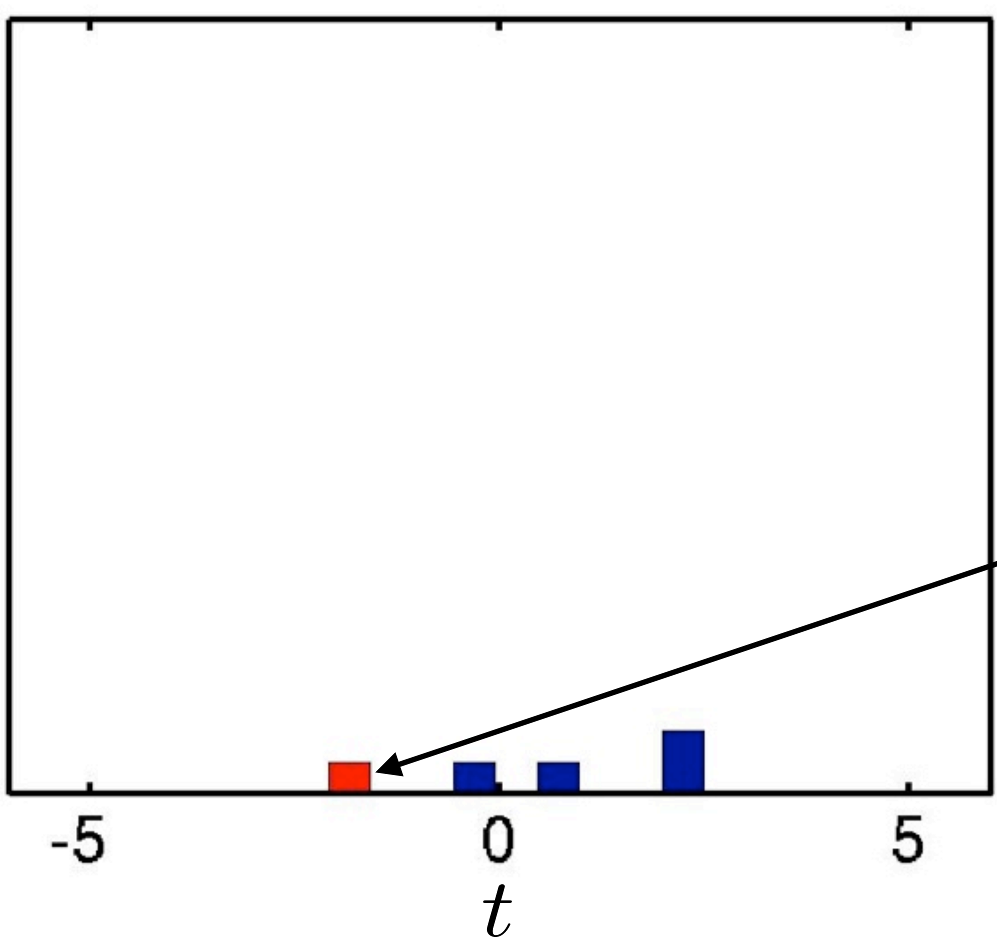
$$t = \frac{\mathbf{c}^T\widehat{\boldsymbol{\beta}}}{\sqrt{\sigma^2}\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}}$$

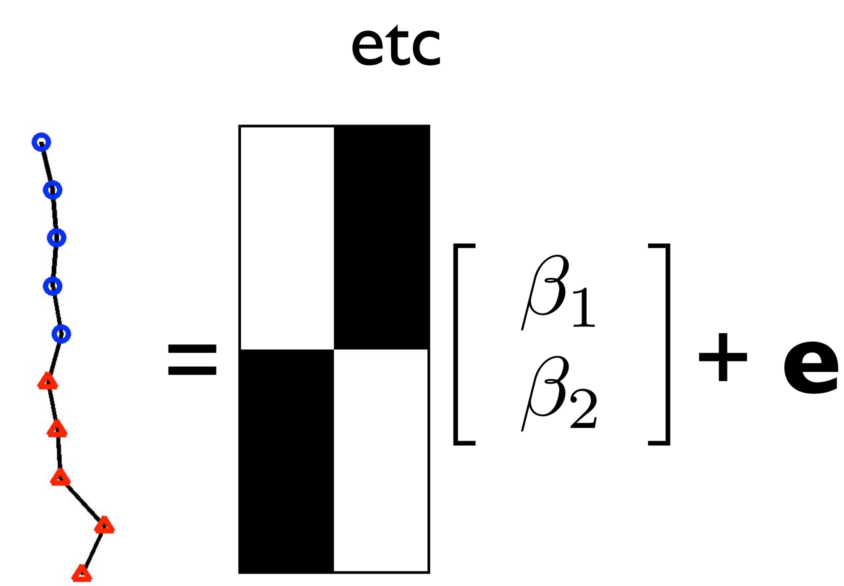
$$\sigma^2 = 0.78$$

Constant

Tools of classical inference

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





$$\mathbf{c}^T\hat{\boldsymbol{\beta}} = -0.69$$

$$t = -1.66$$

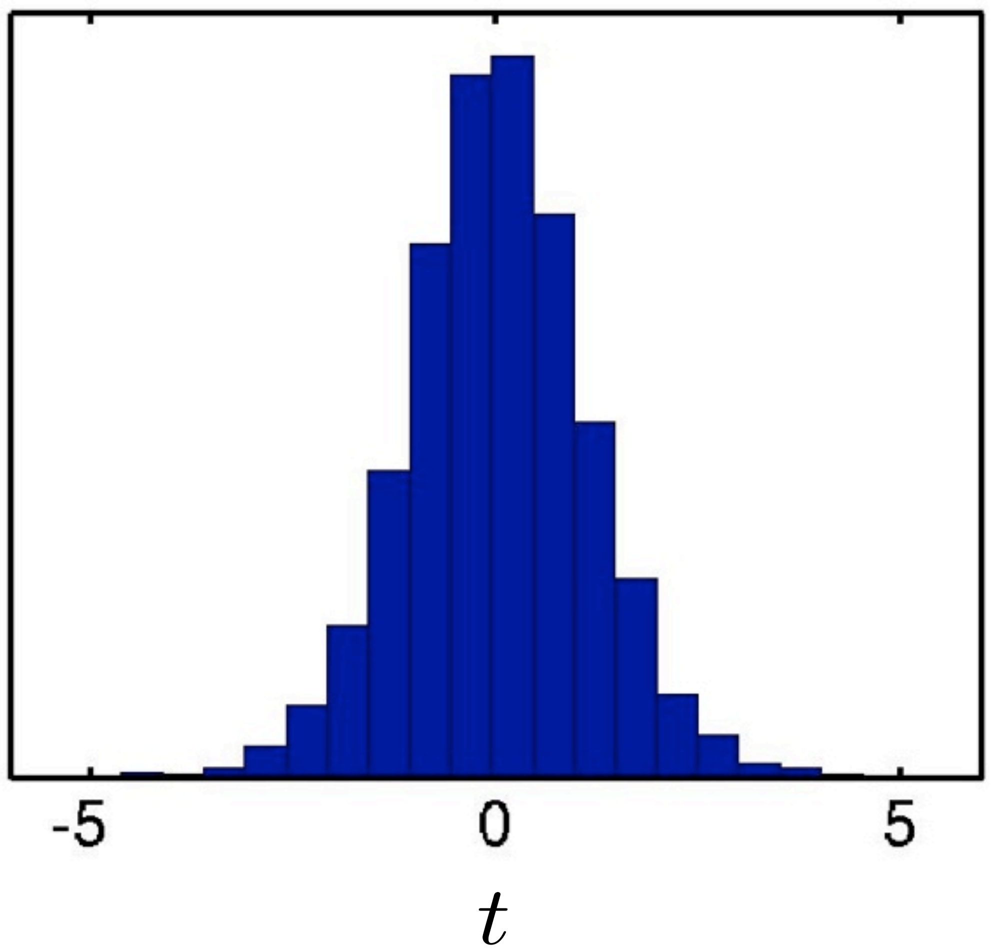
$$t = \frac{\mathbf{c}^T\hat{\boldsymbol{\beta}}}{\sqrt{\sigma^2}\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}}$$

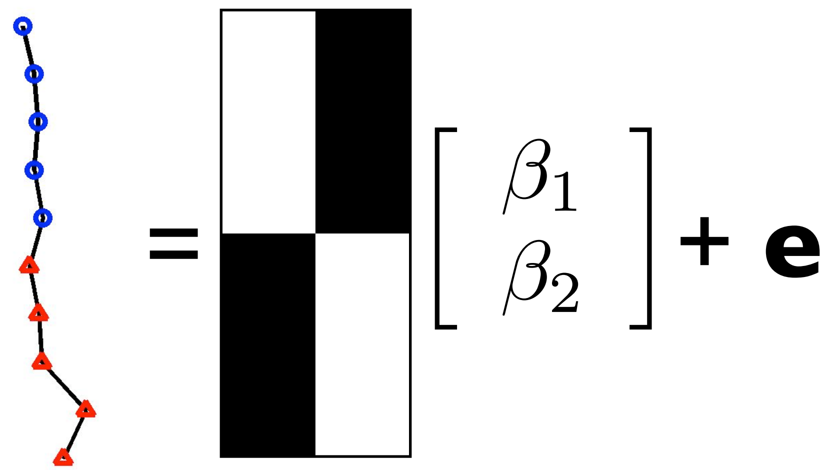
$$\sigma^2 = 0.44$$

Constant

Tools of classical inference

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

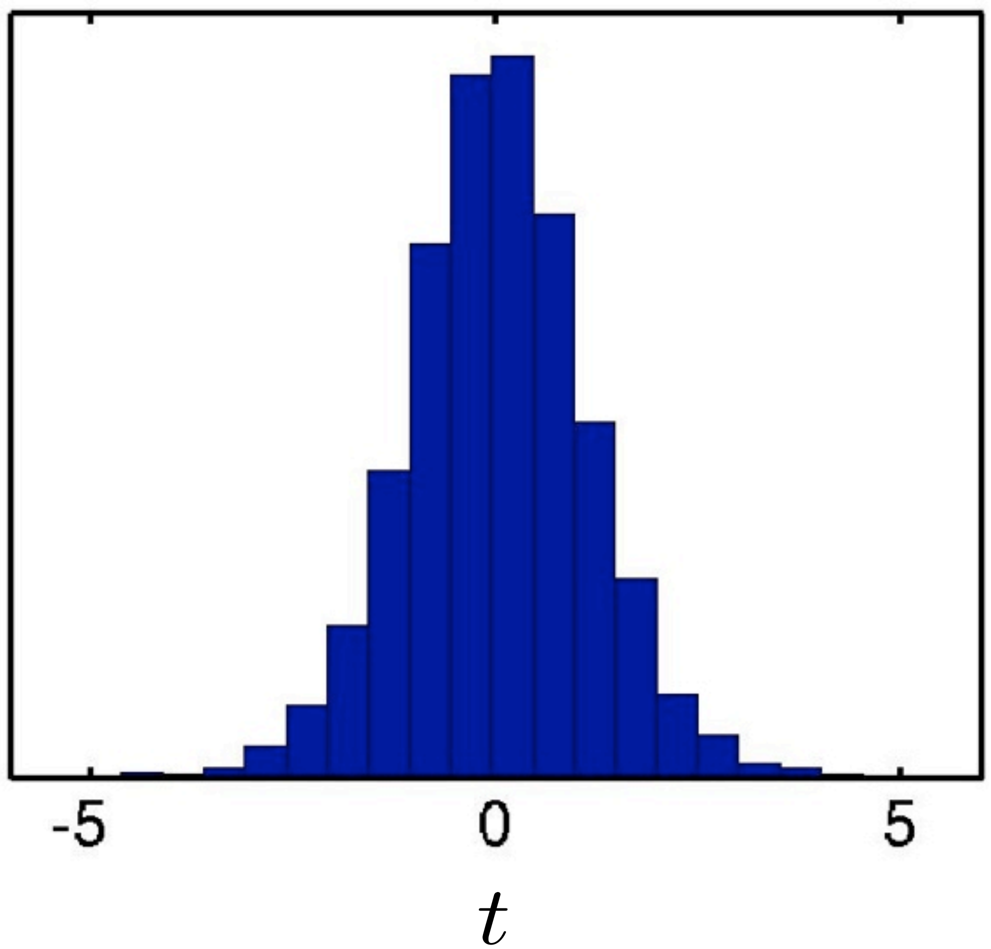




And if we do this til the cows come home

Tools of classical inference

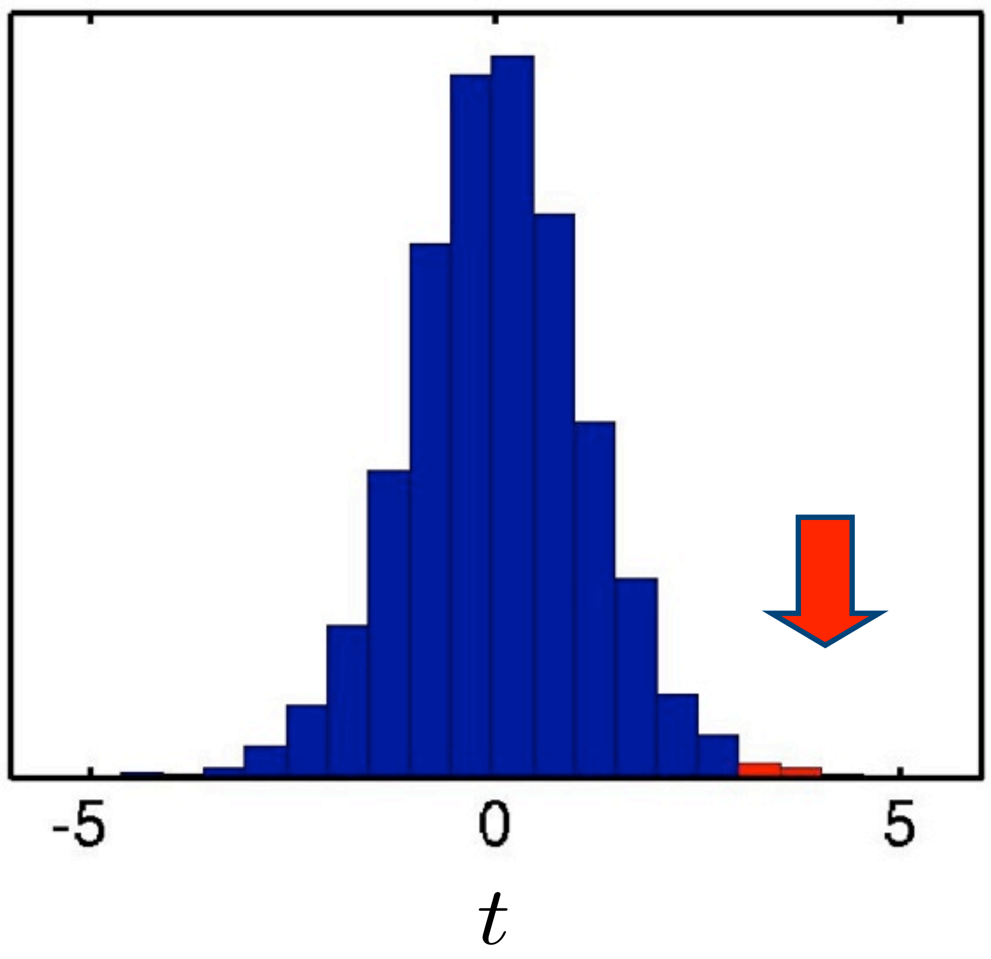
1. A null-hypothesis
2. A test-statistic
3. A null-distribution



So, why is this helpful?

Tools of classical inference

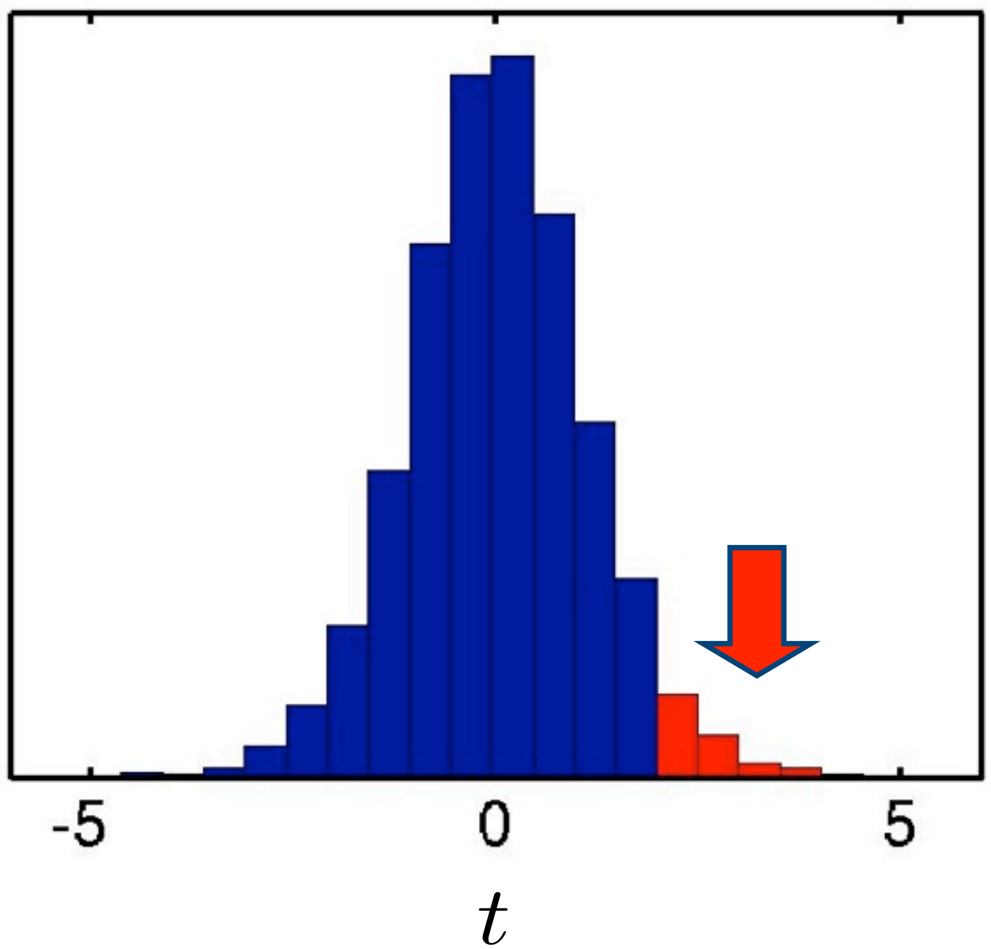
1. A null-hypothesis
2. A test-statistic
3. A null-distribution



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

Tools of classical inference

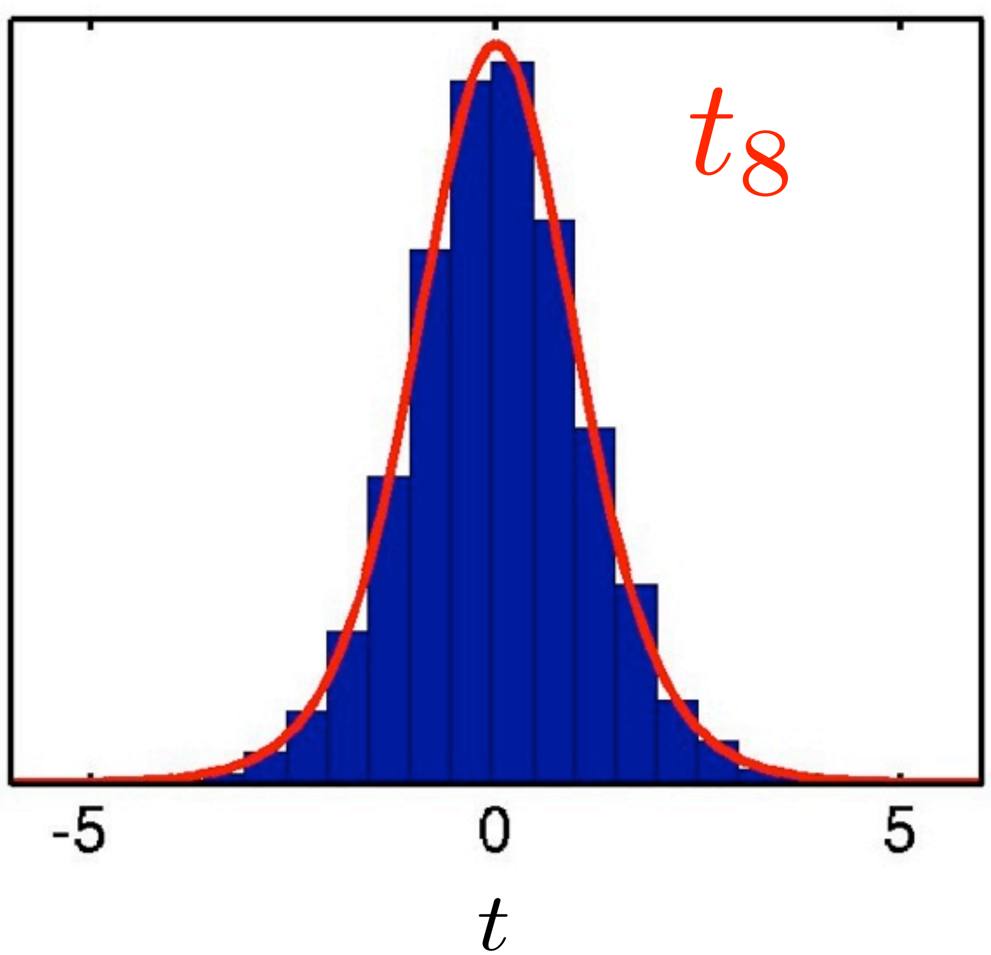
1. A null-hypothesis
2. A test-statistic
3. A null-distribution



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

Tools of classical inference

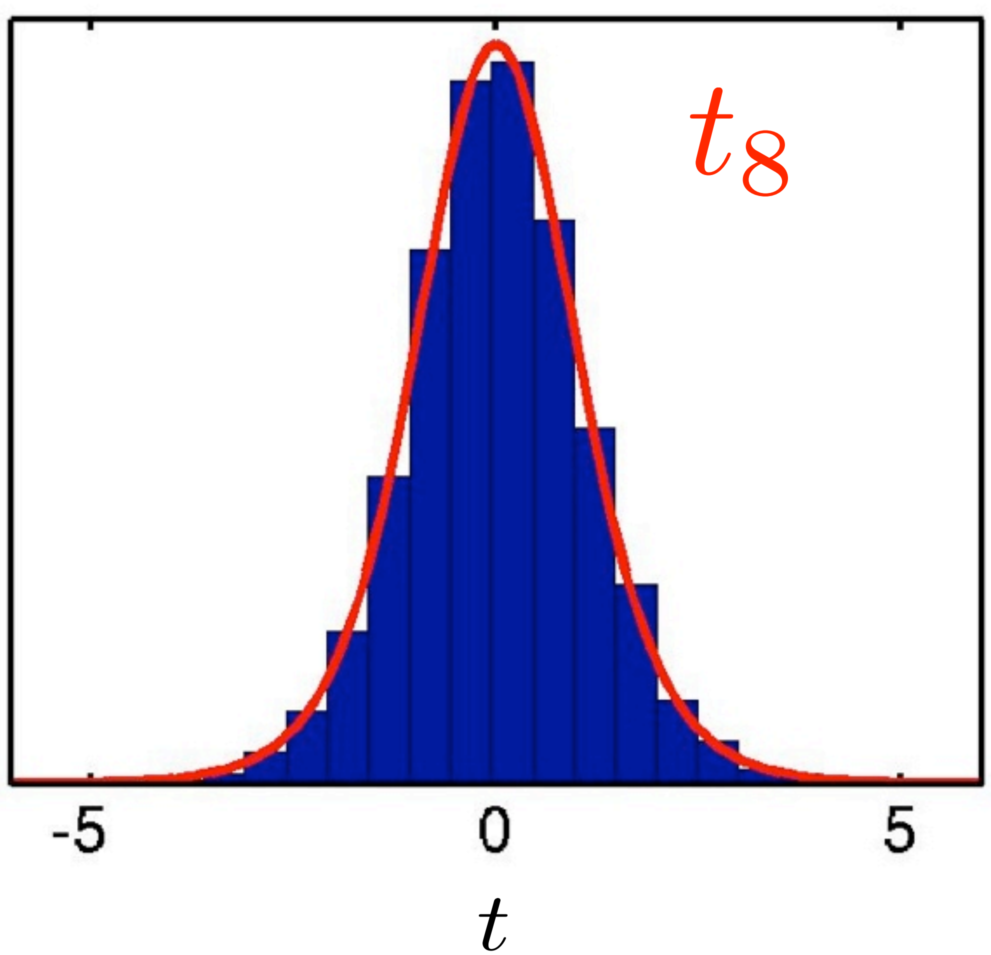
1. A null-hypothesis
2. A test-statistic
3. A null-distribution



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

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

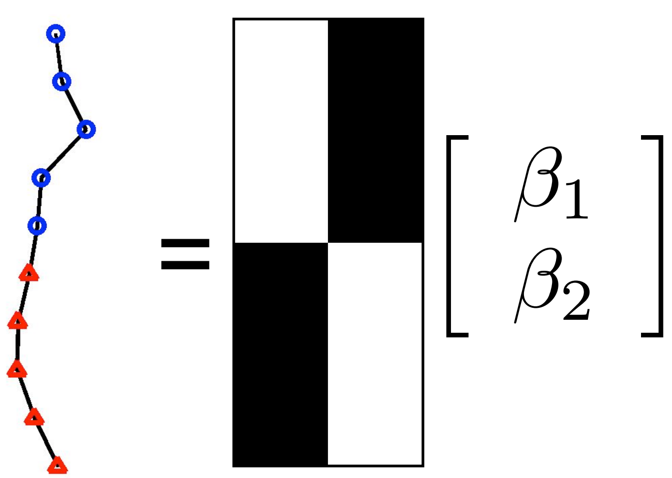
An example experiment

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

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

$t_8 = 2.64$

So, with these tools let us do an experiment



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

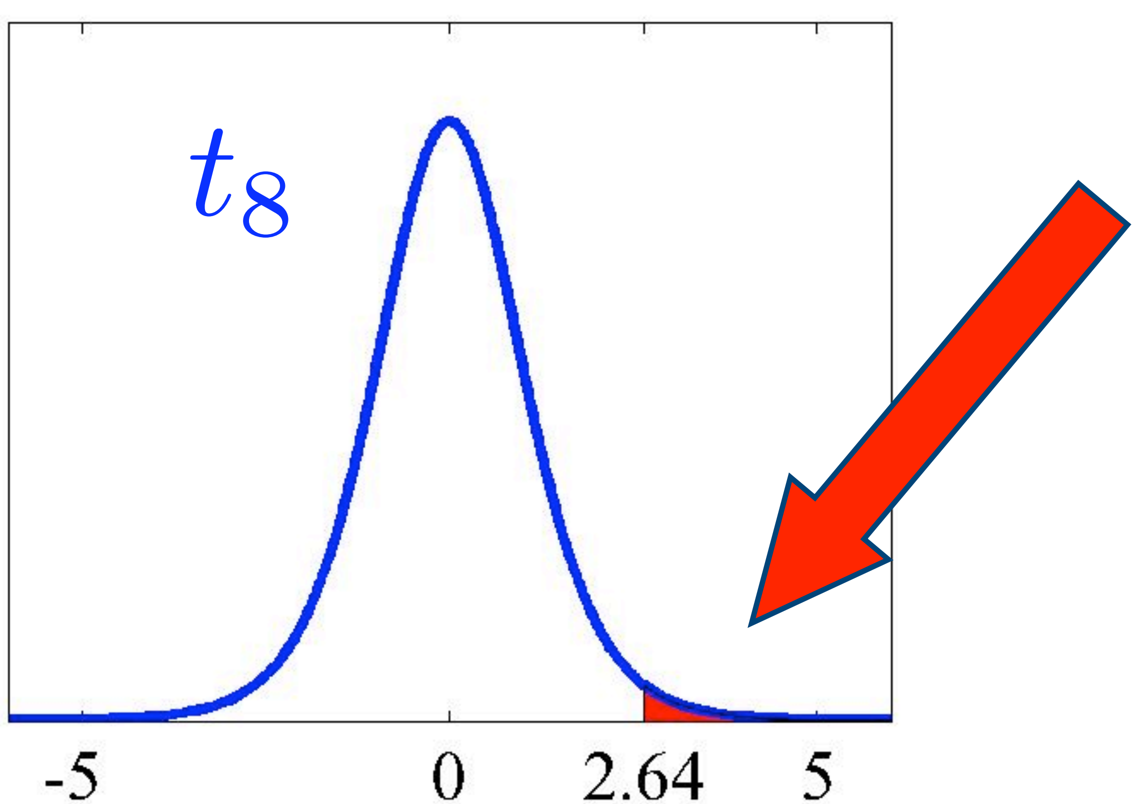
An example experiment

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

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

$t_8 = 2.64$

So, with these tools let us do an experiment

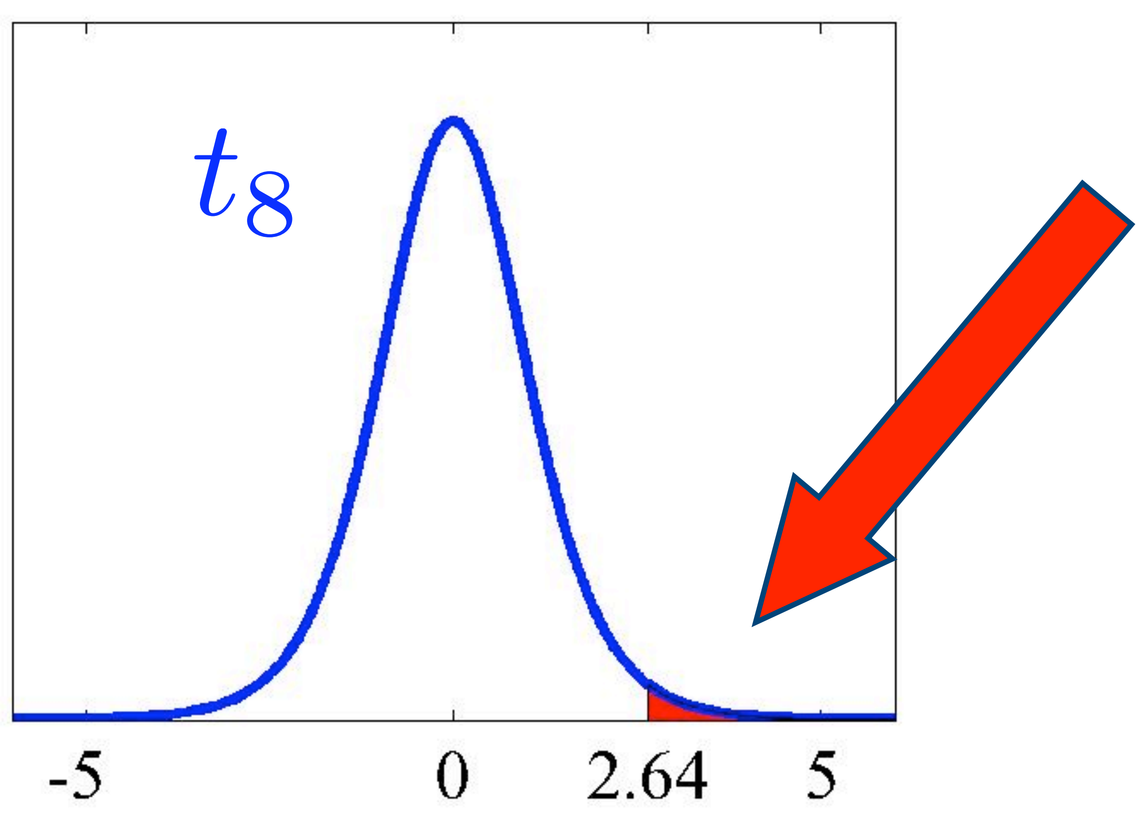


If the null-hypothesis is true, we would expect to have a ~1.46% chance of finding a t-value this large or larger

An example experiment

1. A null-hypothesis — H_0: $\overline{x}_1 = \overline{x}_2$, H_1: $\overline{x}_1 > \overline{x}_2$
2. A test-statistic — $t_8 = 2.64$
3. A null-distribution — $t_8 = 2.64^*$

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 (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?

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

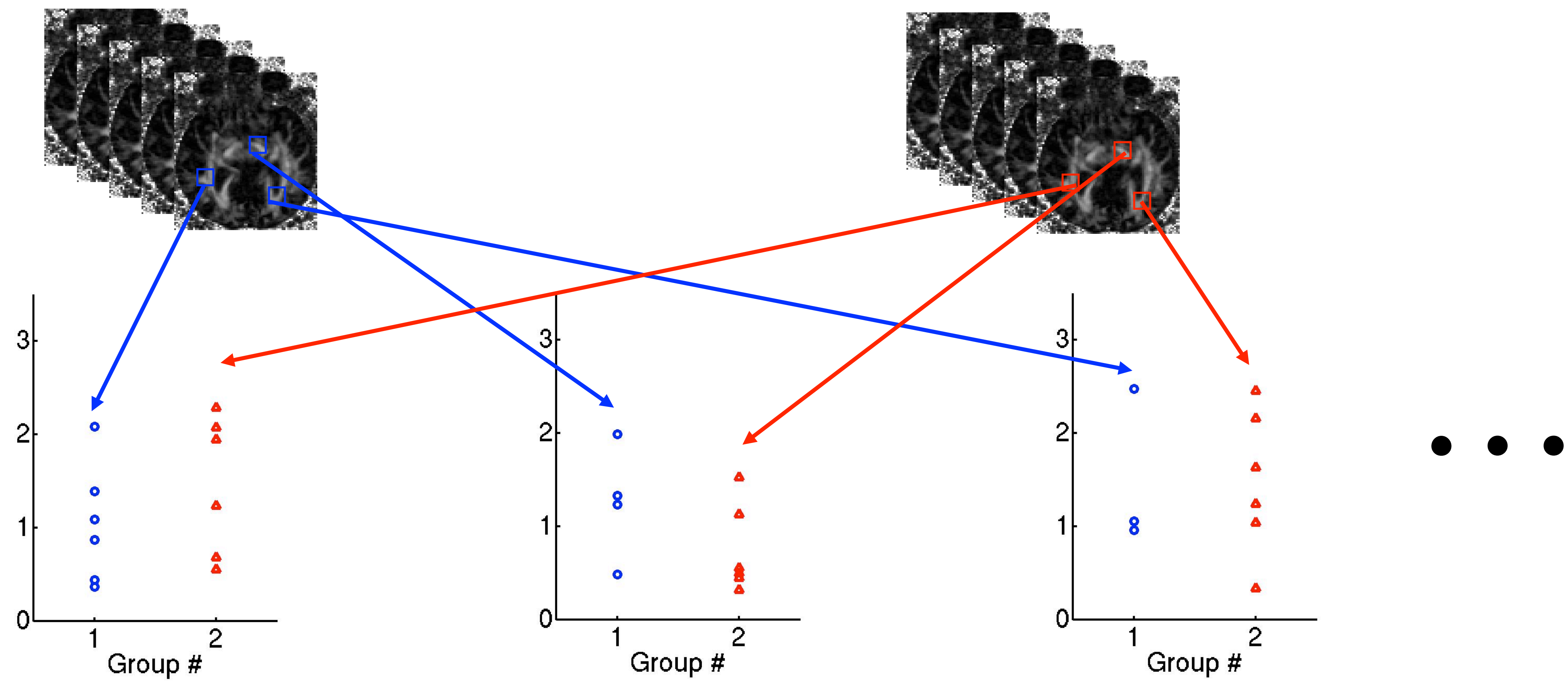
	We don't reject H_0	We reject H_0
H_0 is true	☺️	False positive Type I error
H_0 is false	False negative Type II error	☺️

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

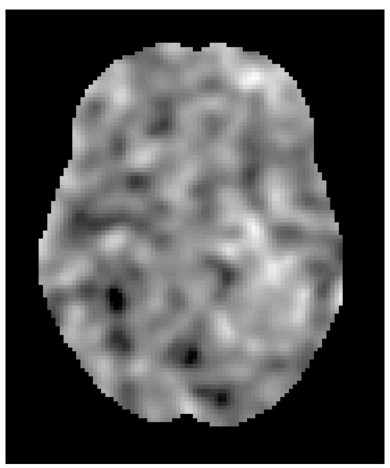
Multiple Comparisons

- In neuroimaging we typically perform <u>many</u> tests as part of a study



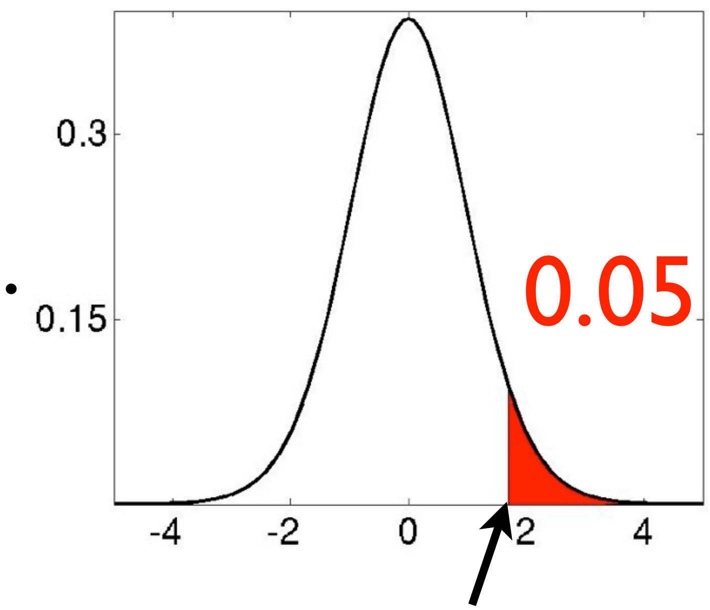
Different here? Maybe here? Or here? . . .

What happens when we apply this to imaging data?



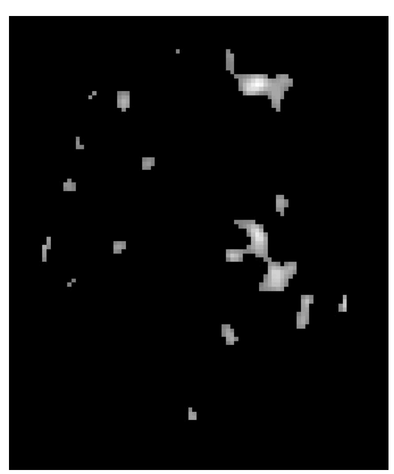
z-map where each voxel ~*N*.
Null-hypothesis true everywhere, i.e.
NO ACTIVATIONS

z



1.64

z-map
thresholded at
1.64

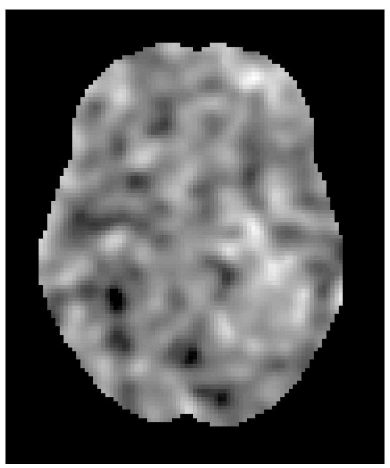


16 clusters
288 voxels
~5.5% of the voxels

That's a LOT of false positives

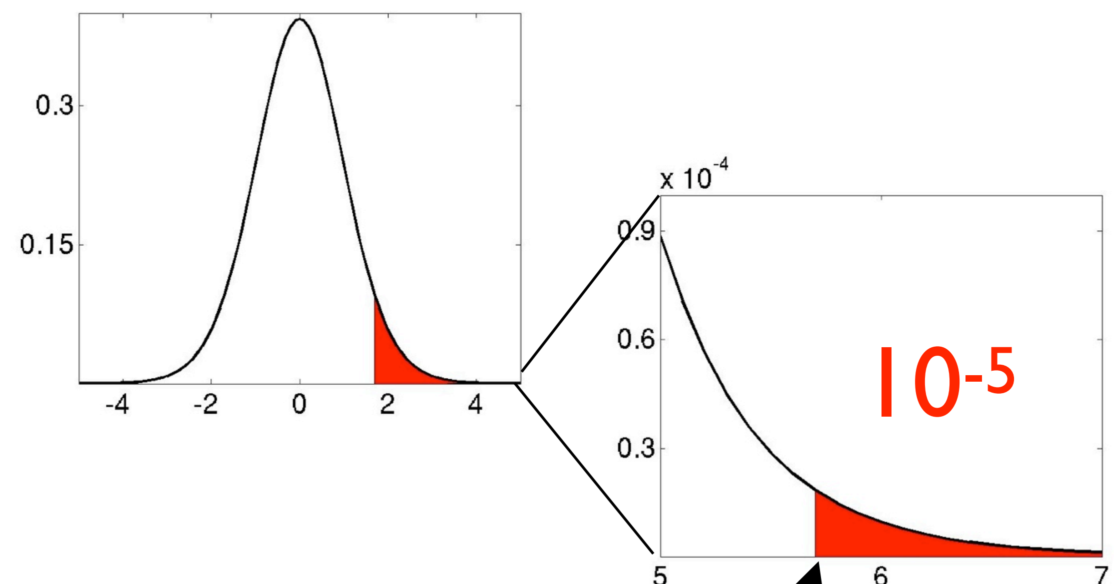
Italians doing maths:
The Bonferroni correction

Bonferroni says threshold at α divided by # of tests



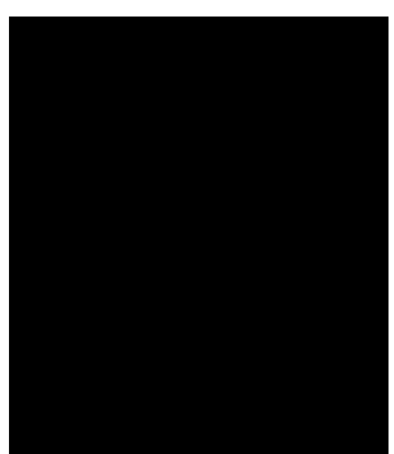
5255 voxels

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



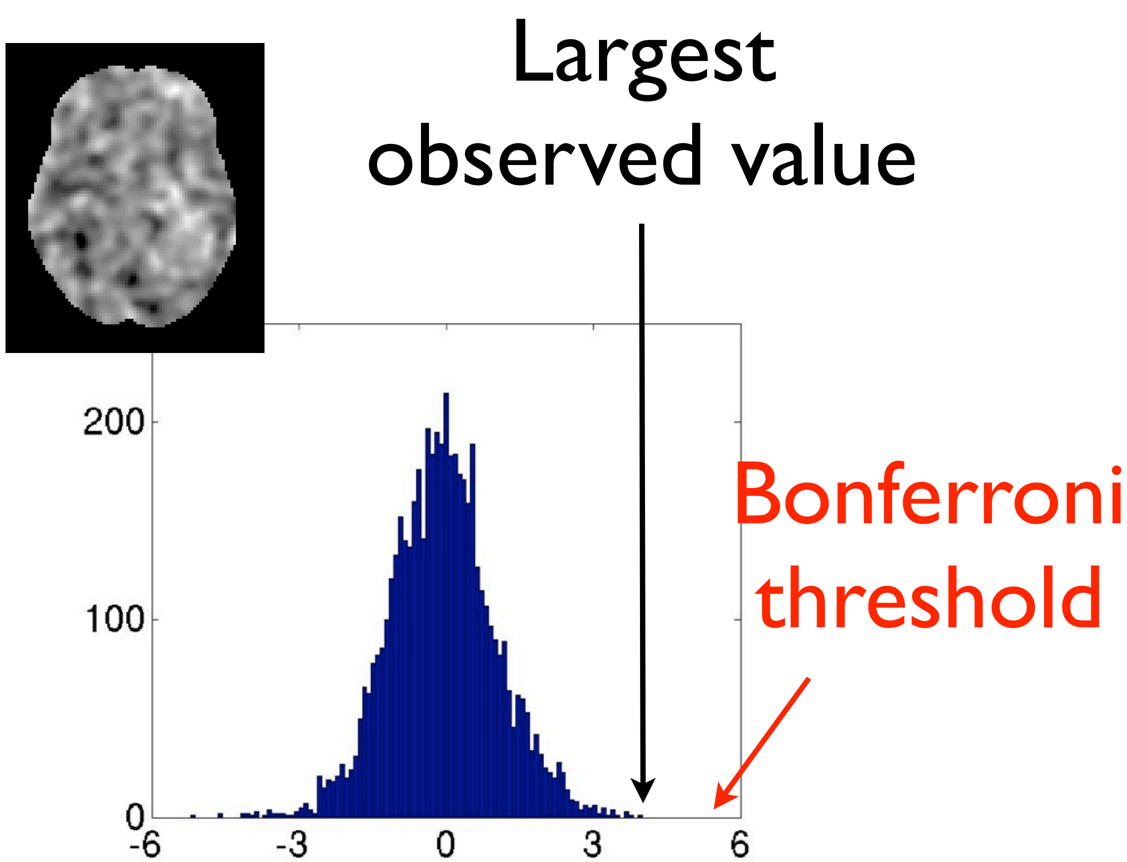
5.65

z-map thresholded at 5.65

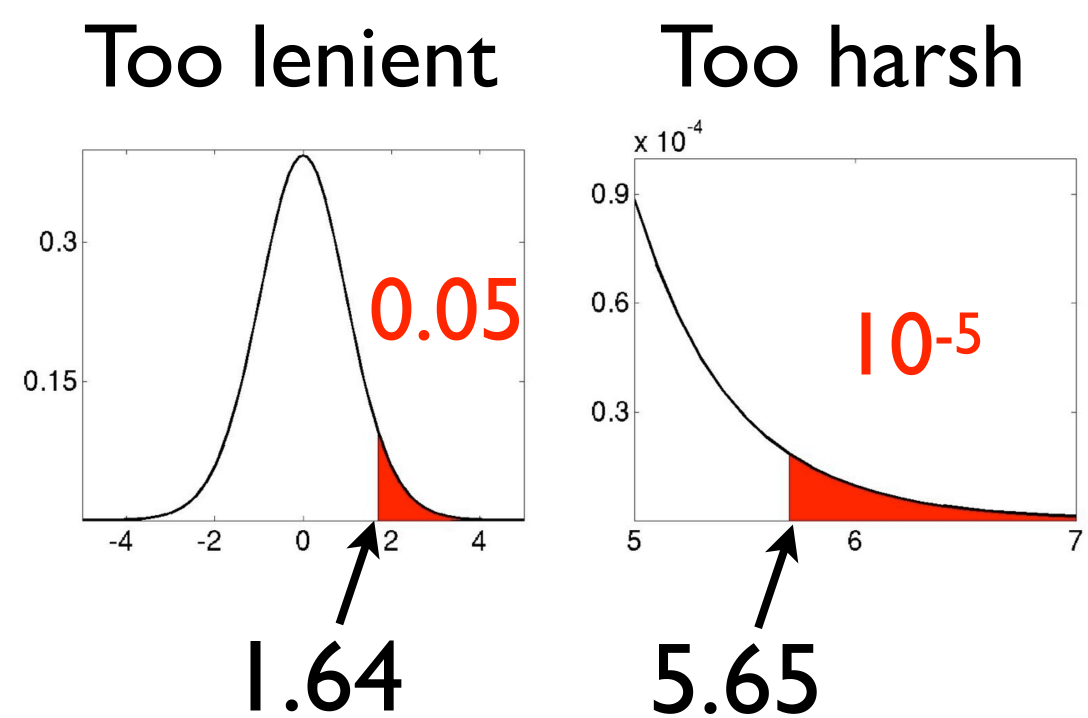


No false positives.
Hurrah for Italy!

But ... doesn’t 5.65 sound very high?



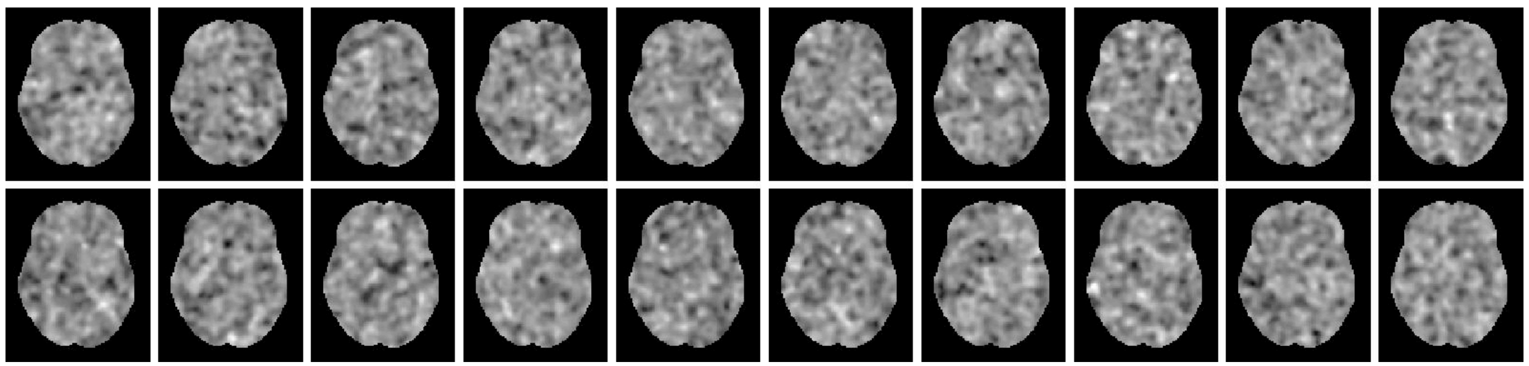
Observed values
in the z-map



So what do we want then?

Family-wise error

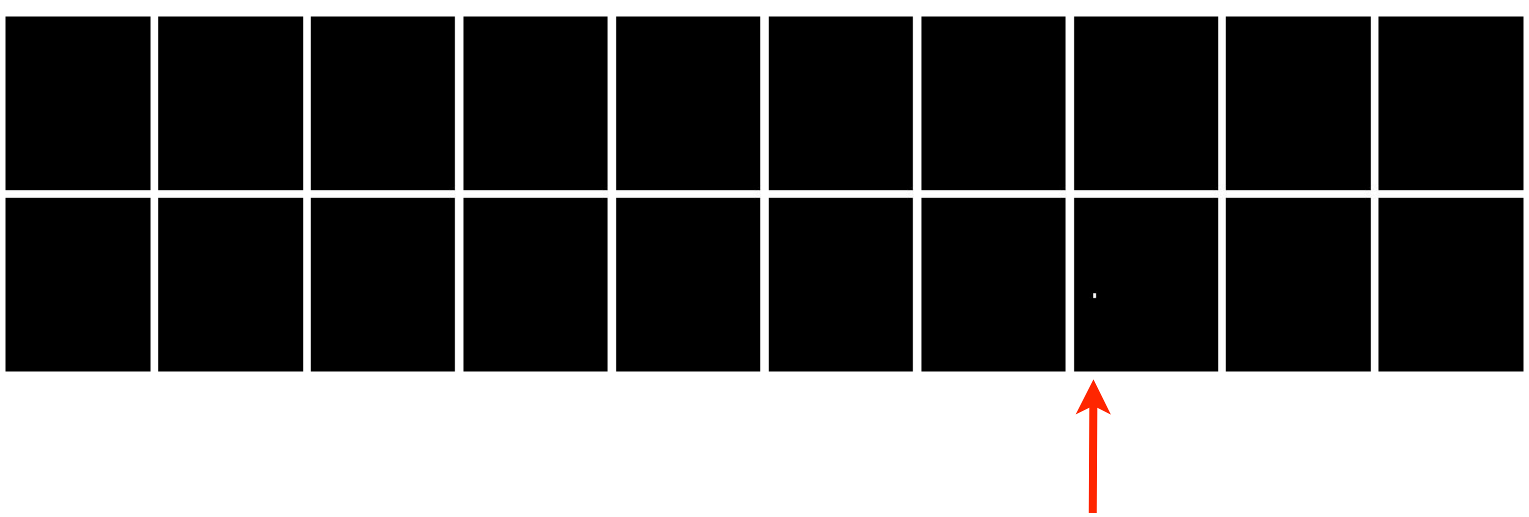
Let's say we perform a series of identical studies



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



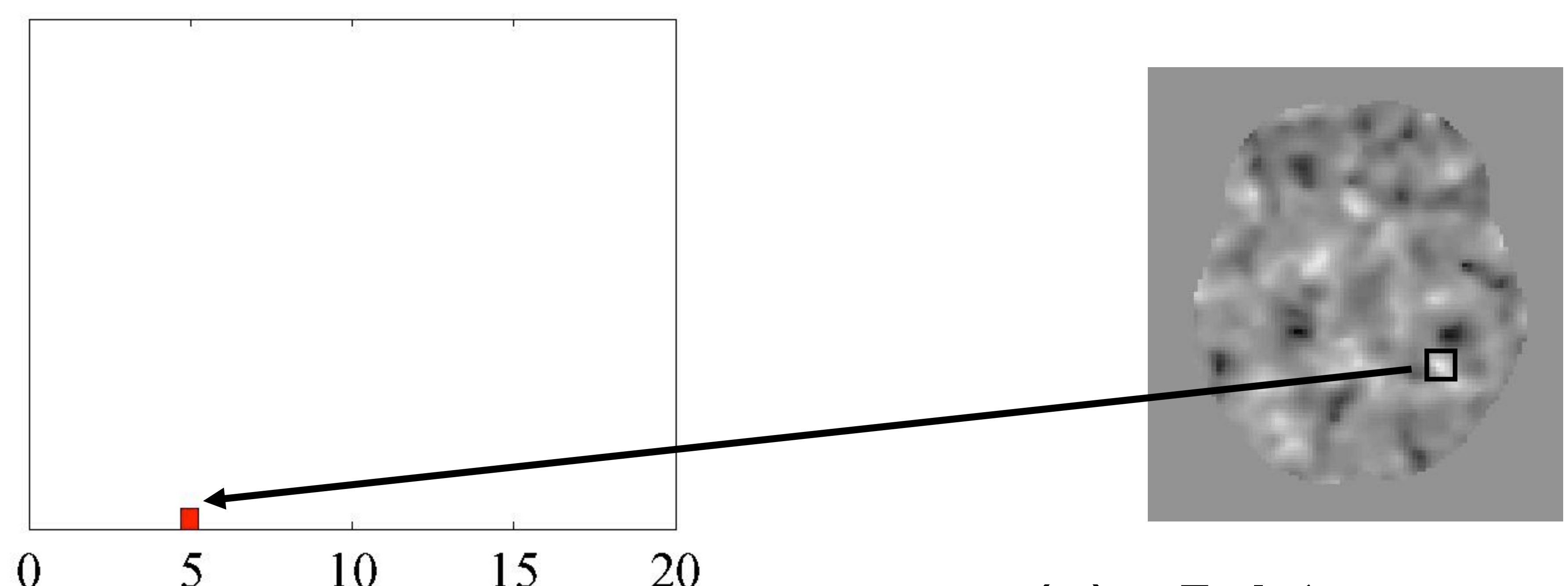
But how do we find such a threshold?

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

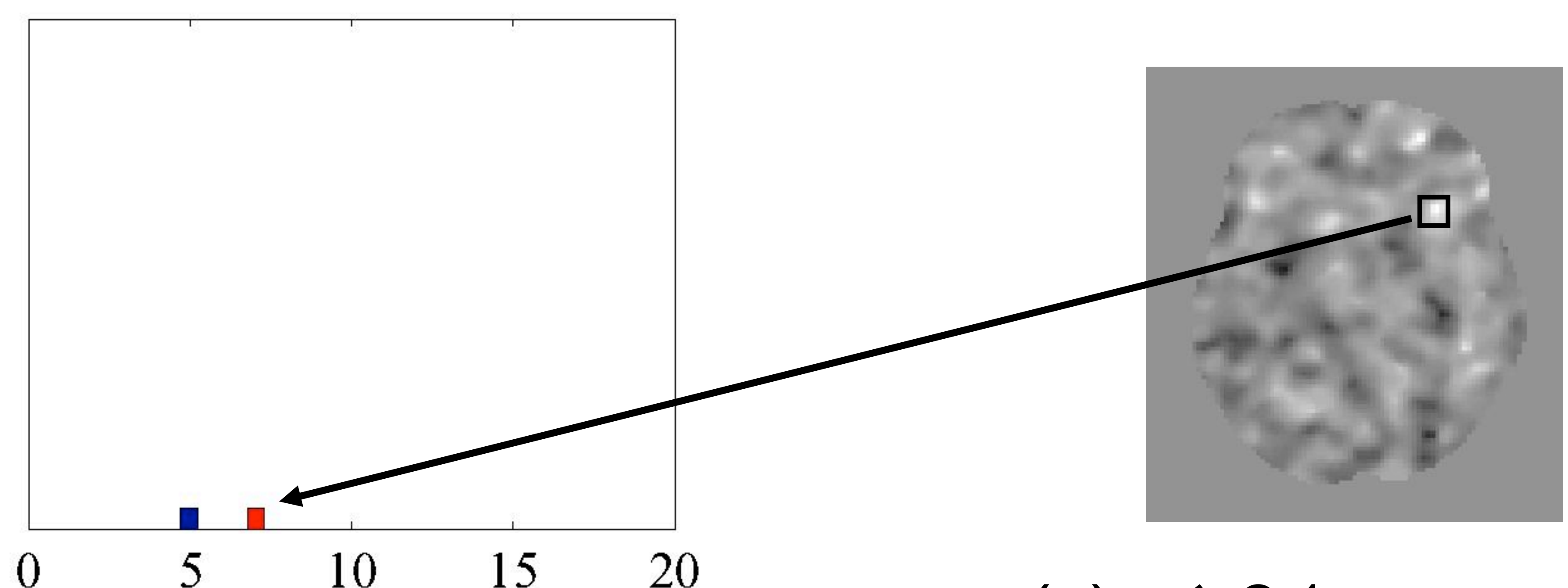
- When we want to control "family-wise error", what do we in practice want?
- If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
- And if we reject anything, we will definitely reject the most "extreme" value (max(z)) in the brain.



max(z)=5.16

Maximum z

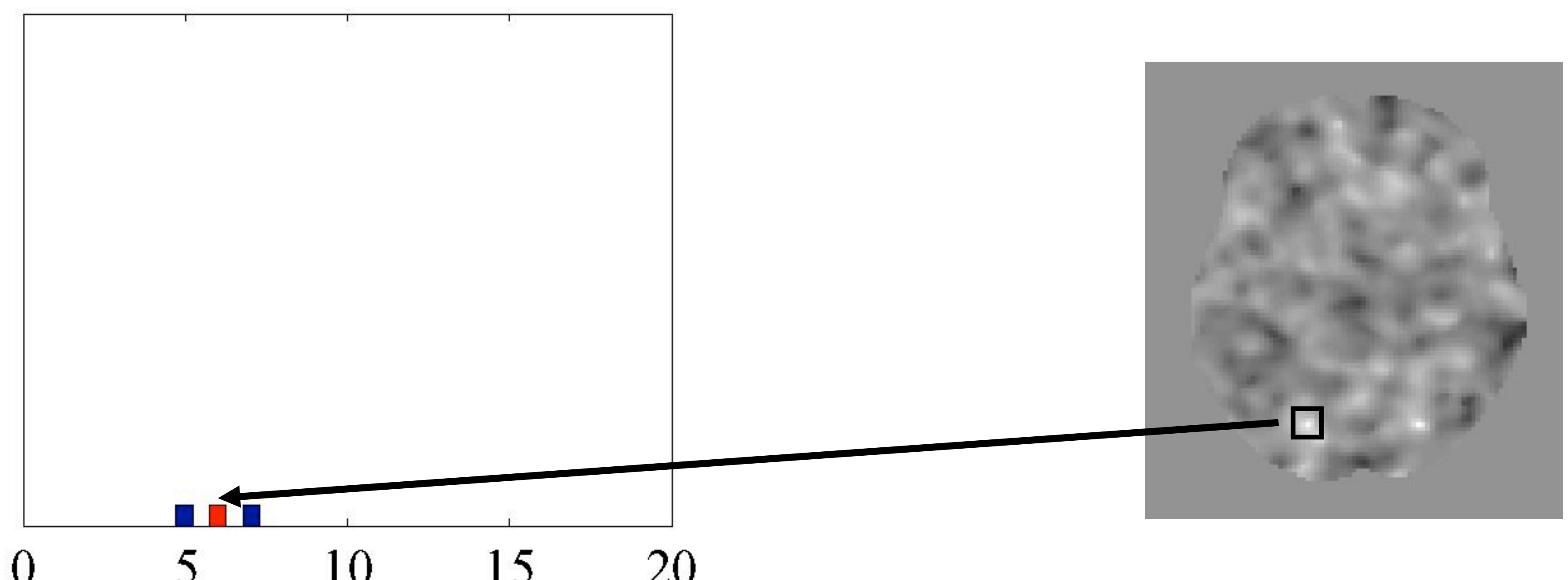
- When we want to control "family-wise error", what do we in practice want?
- If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
- And if we reject anything, we will definitely reject the most "extreme" value in the brain.



max(z)=6.84

Maximum z

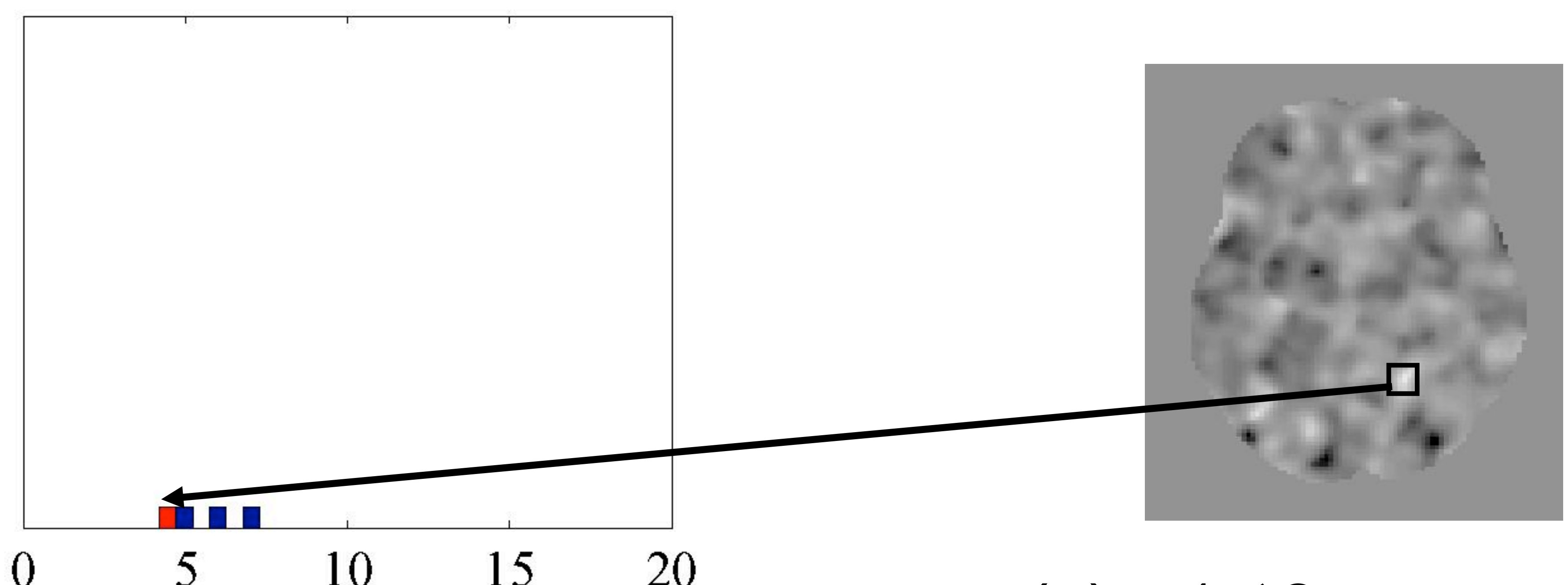
- When we want to control "family-wise error", what do we in practice want?
- If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
- And if we reject anything, we will definitely reject the most "extreme" value in the brain.



max(z)=5.93

Maximum z

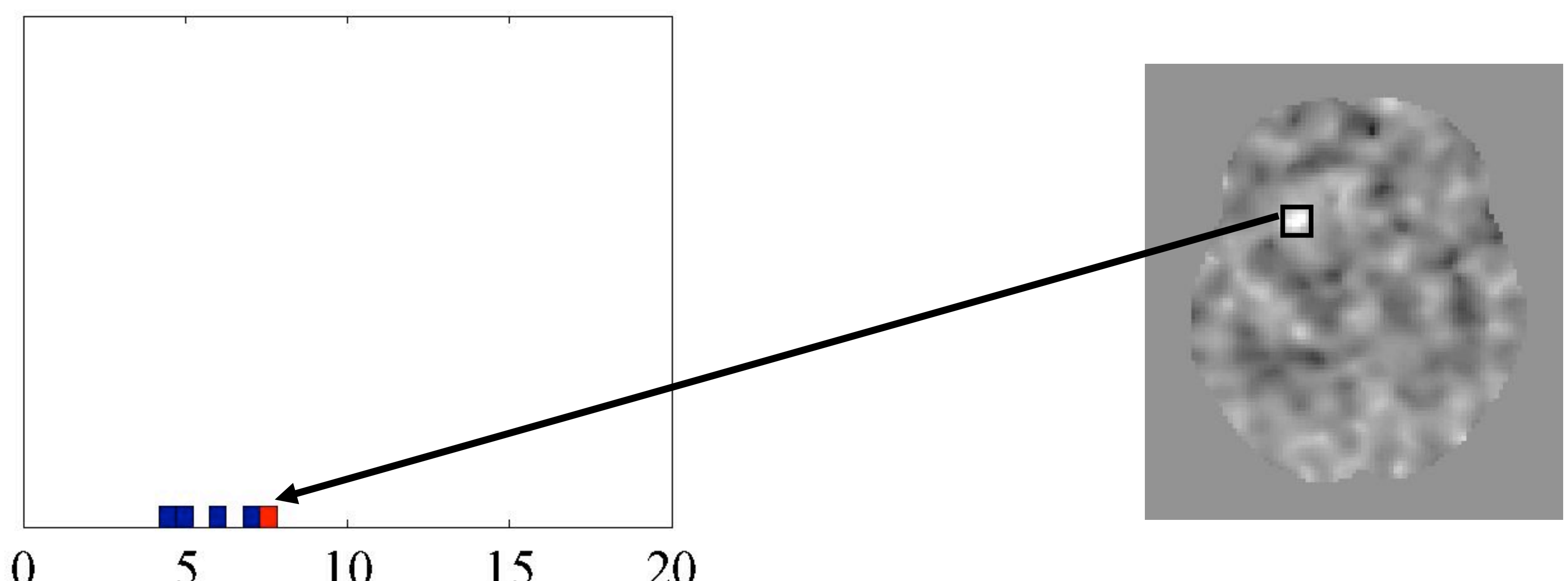
- When we want to control "family-wise error", what do we in practice want?
- If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
- And if we reject anything, we will definitely reject the most "extreme" value in the brain.



max(z)=4.62

Maximum z

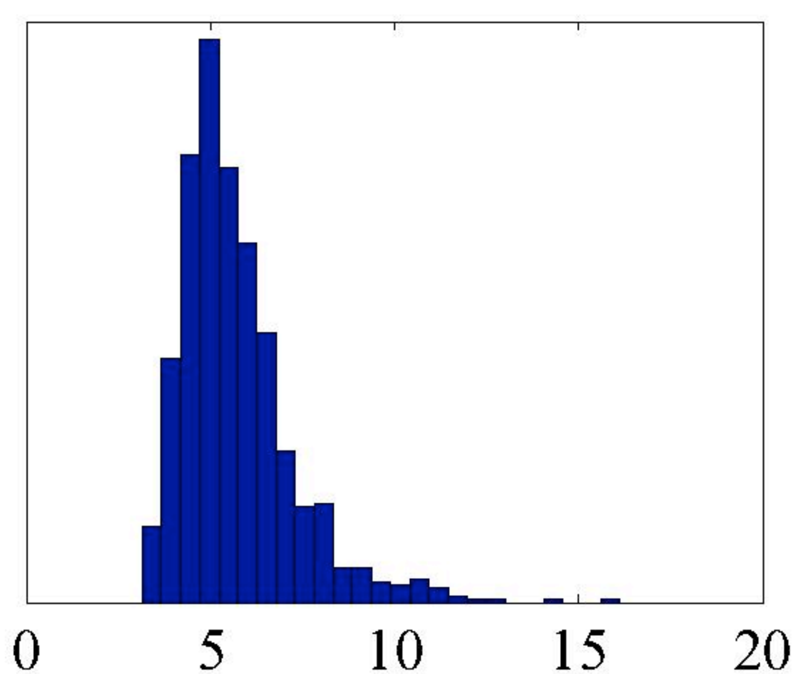
- When we want to control "family-wise error", what do we in practice want?
- If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
- And if we reject anything, we will definitely reject the most "extreme" value in the brain.



max(z)=7.36

Maximum *z*

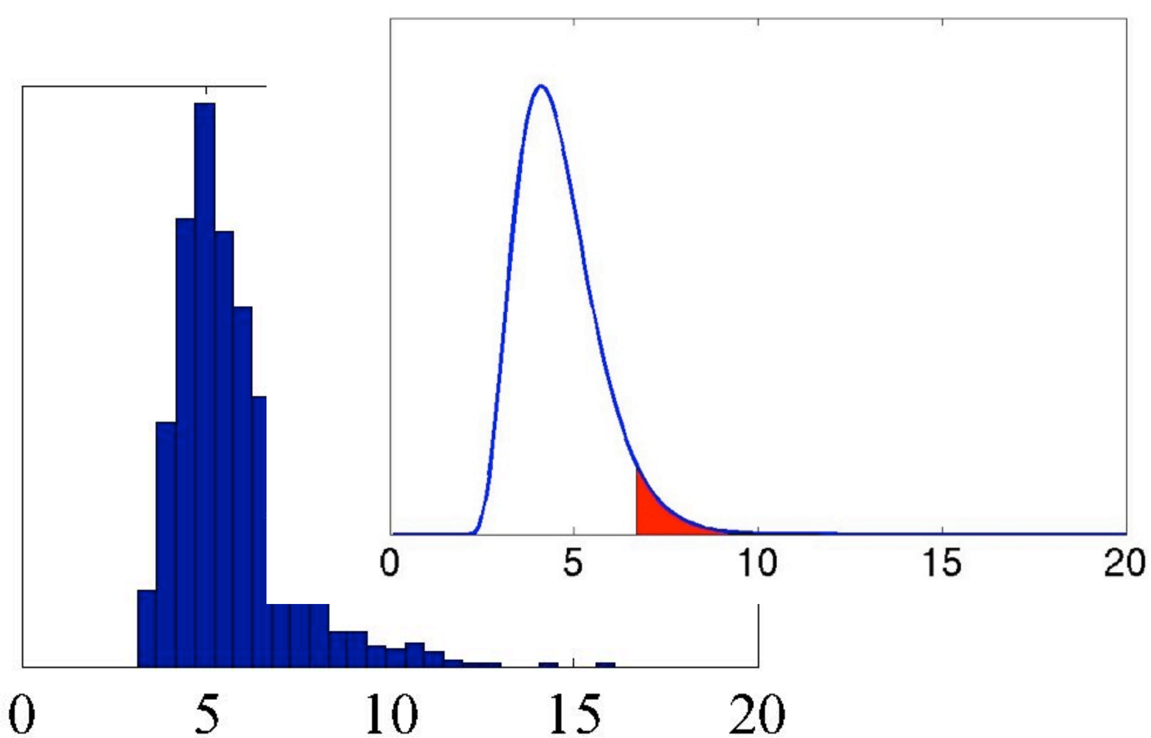
- When we want to control "family-wise error", what do we in practice want?
- If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
- And if we reject anything, we will definitely reject the most "extreme" value in the brain.



Etc…

Maximum z

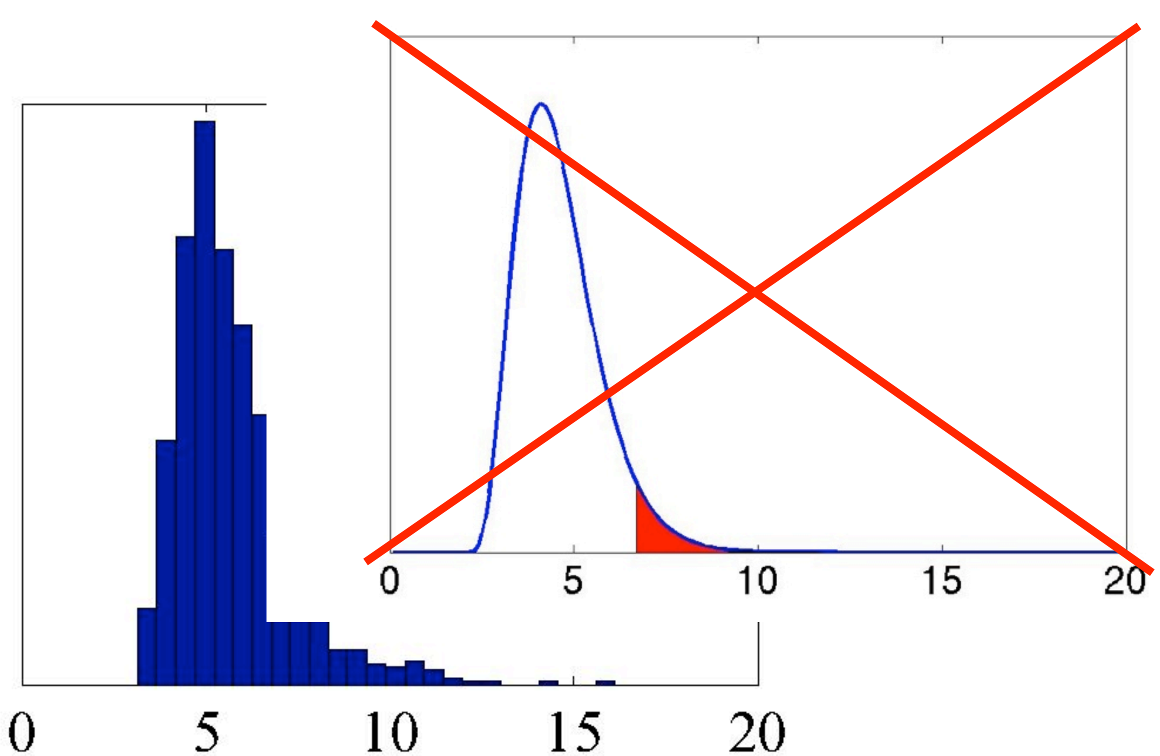
- When we want to control "family-wise error", what do we in practice want?
- If the null-hypothesis is true (no activation) we want to reject it no more than 5% of the time.
- And if we reject anything, we will definitely reject the most "extreme" value in the brain.



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.

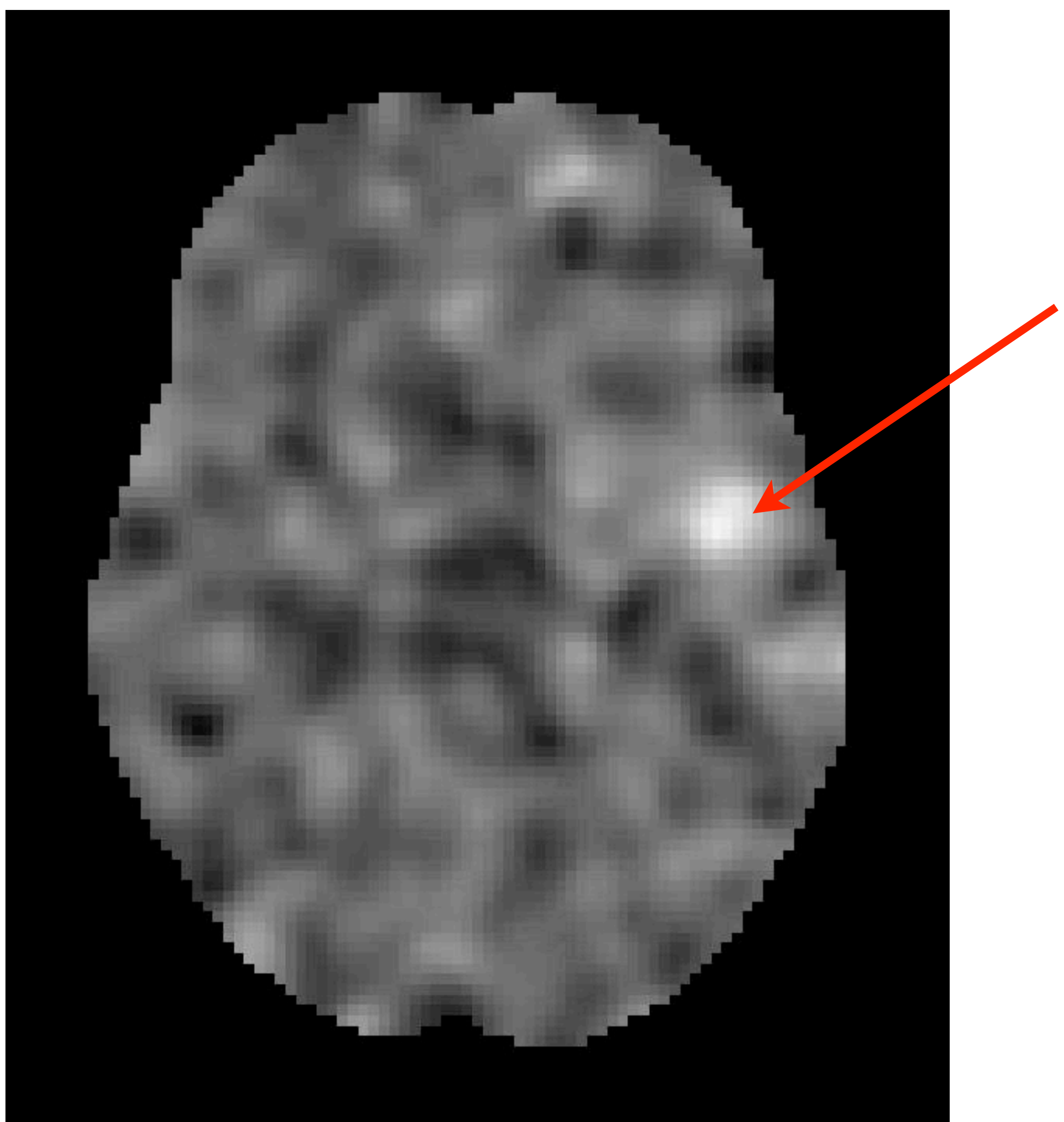
But there is no known expression for it! ☹

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

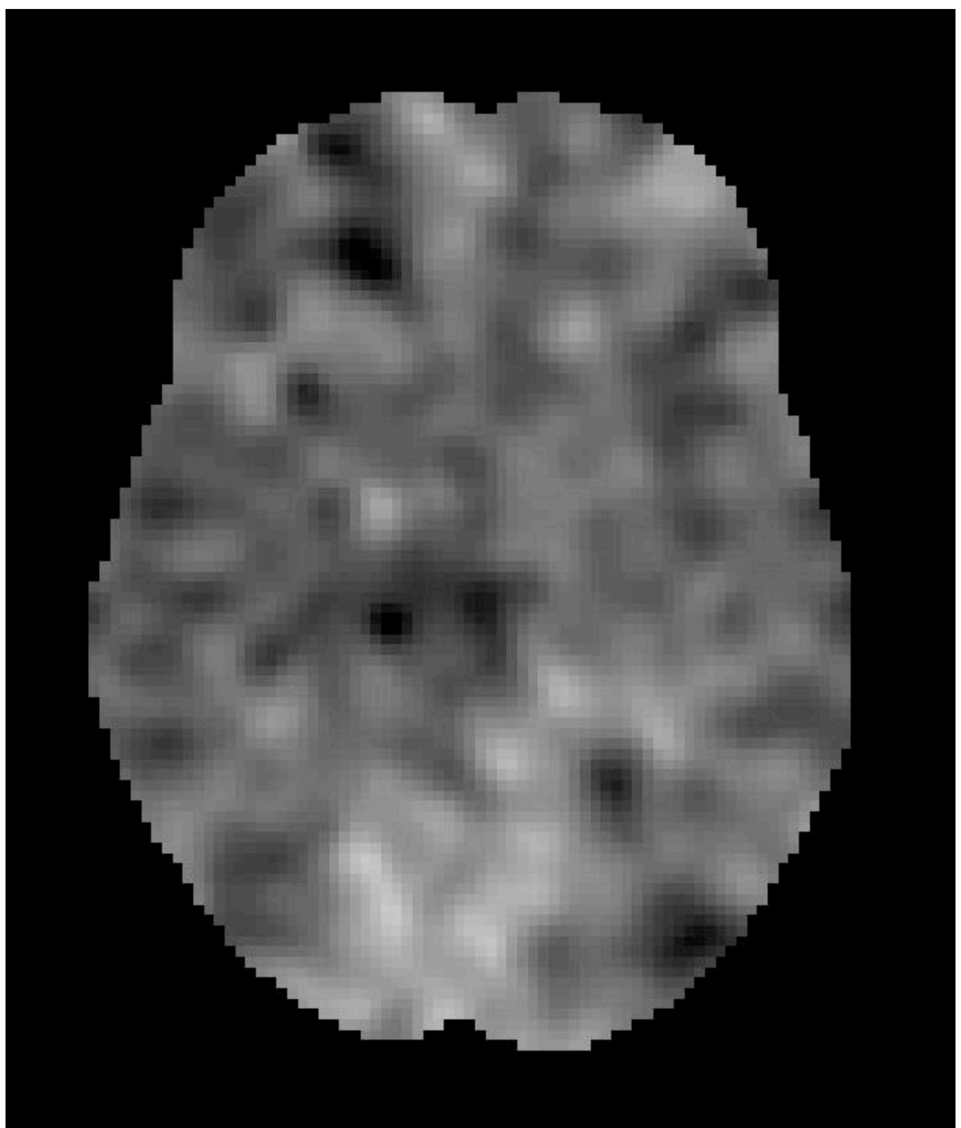
This far we have talked about voxel-based tests



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

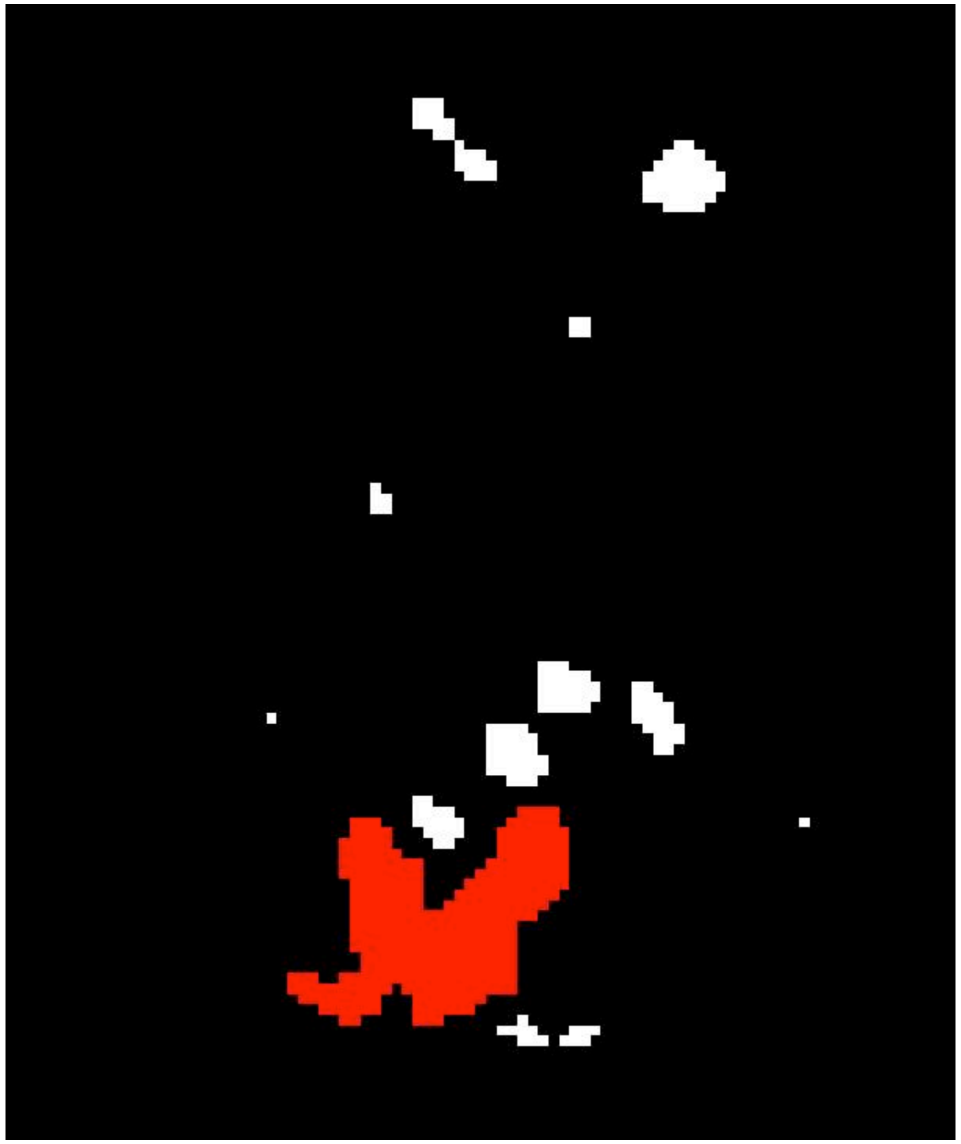
But sometimes our data just aren't that surprising.



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

Spatial extent: another way to be surprised

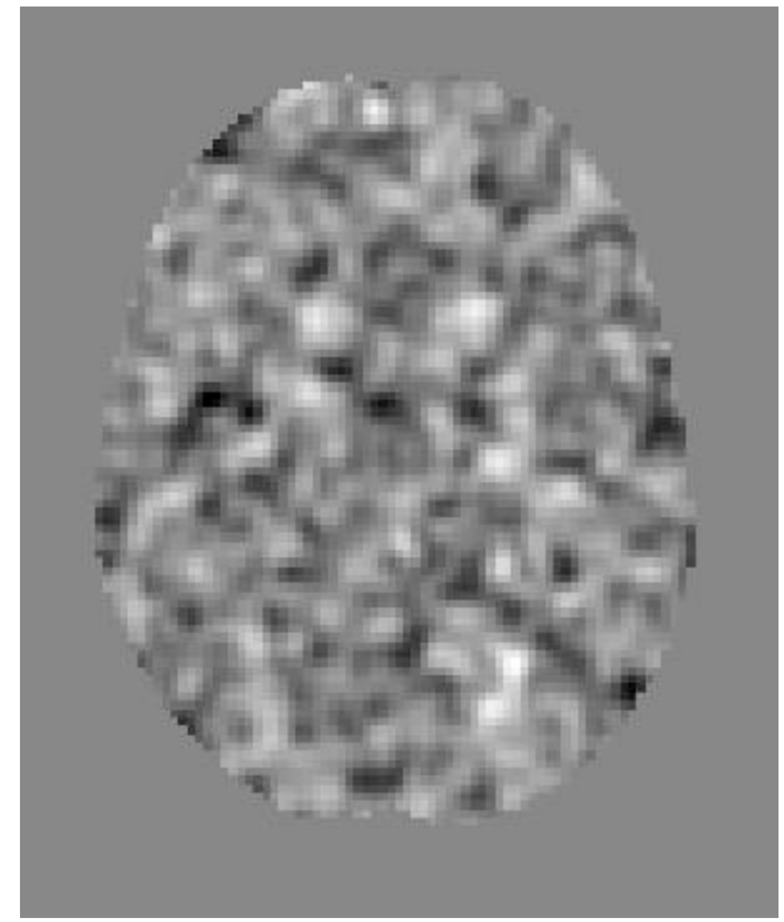
So we threshold the z-map at 2.3 (arbitrary threshold) and look at the spatial extent of clusters



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

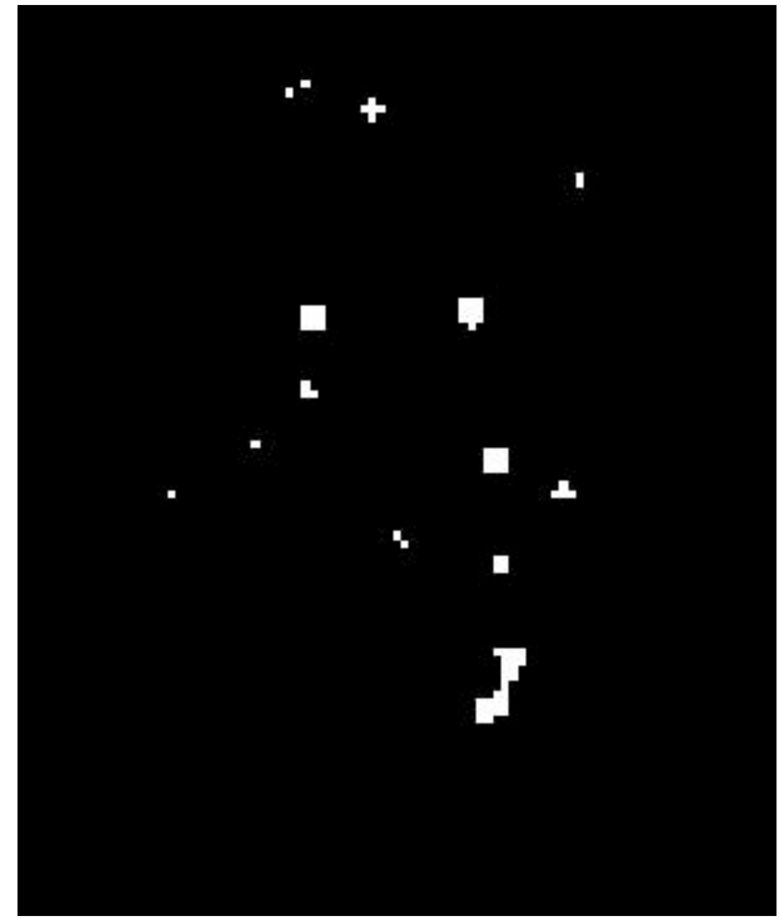
As with the *z*-values we need a "null-distribution". What would that look like in this case?



Let's say we have acquired some data

Distribution of Max Cluster Size

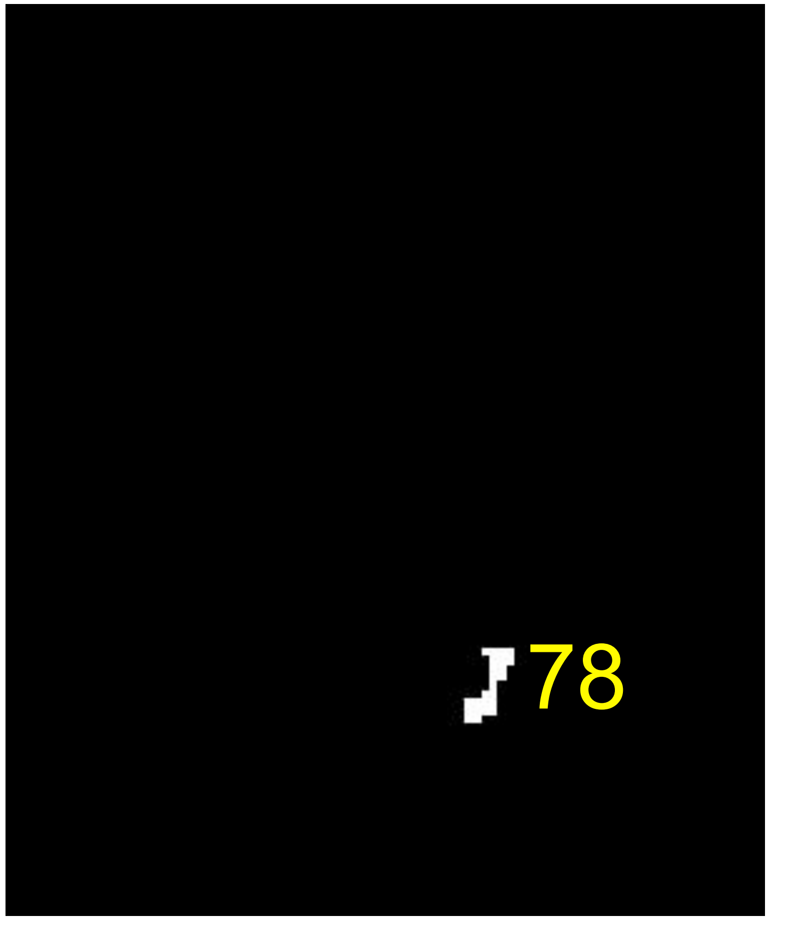
If we reject any cluster we will reject the largest. So what we want is the distribution of the largest cluster, under the null-hypothesis.



Threshold the z-map at 2.3 (arbitrary)

Distribution of Max Cluster Size

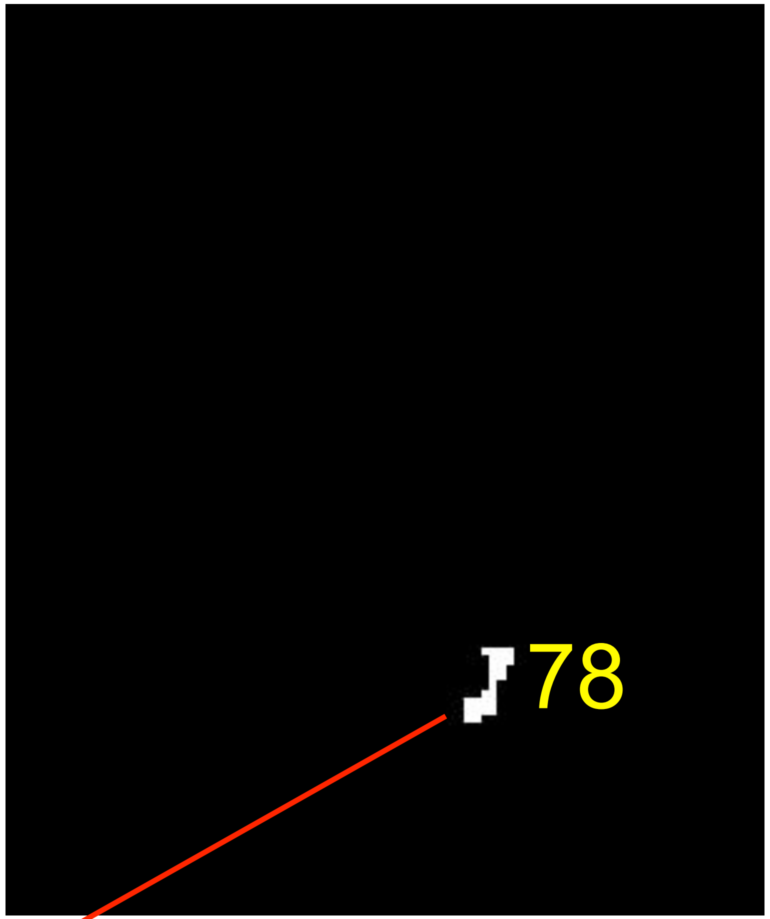
If we reject any cluster we will reject the largest. So what we want is the distribution of the largest cluster, under the null-hypothesis.



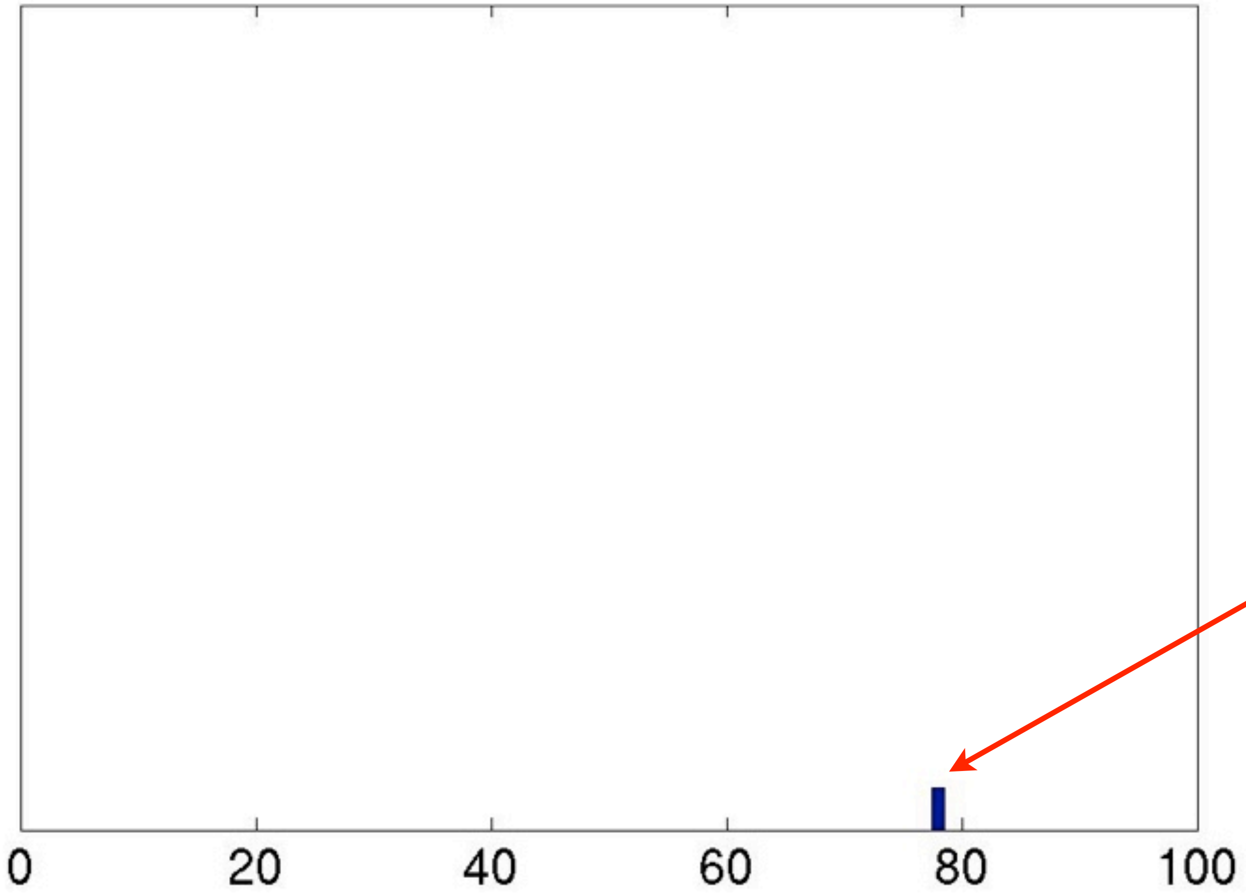
Locate the largest cluster anywhere in the brain.

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.

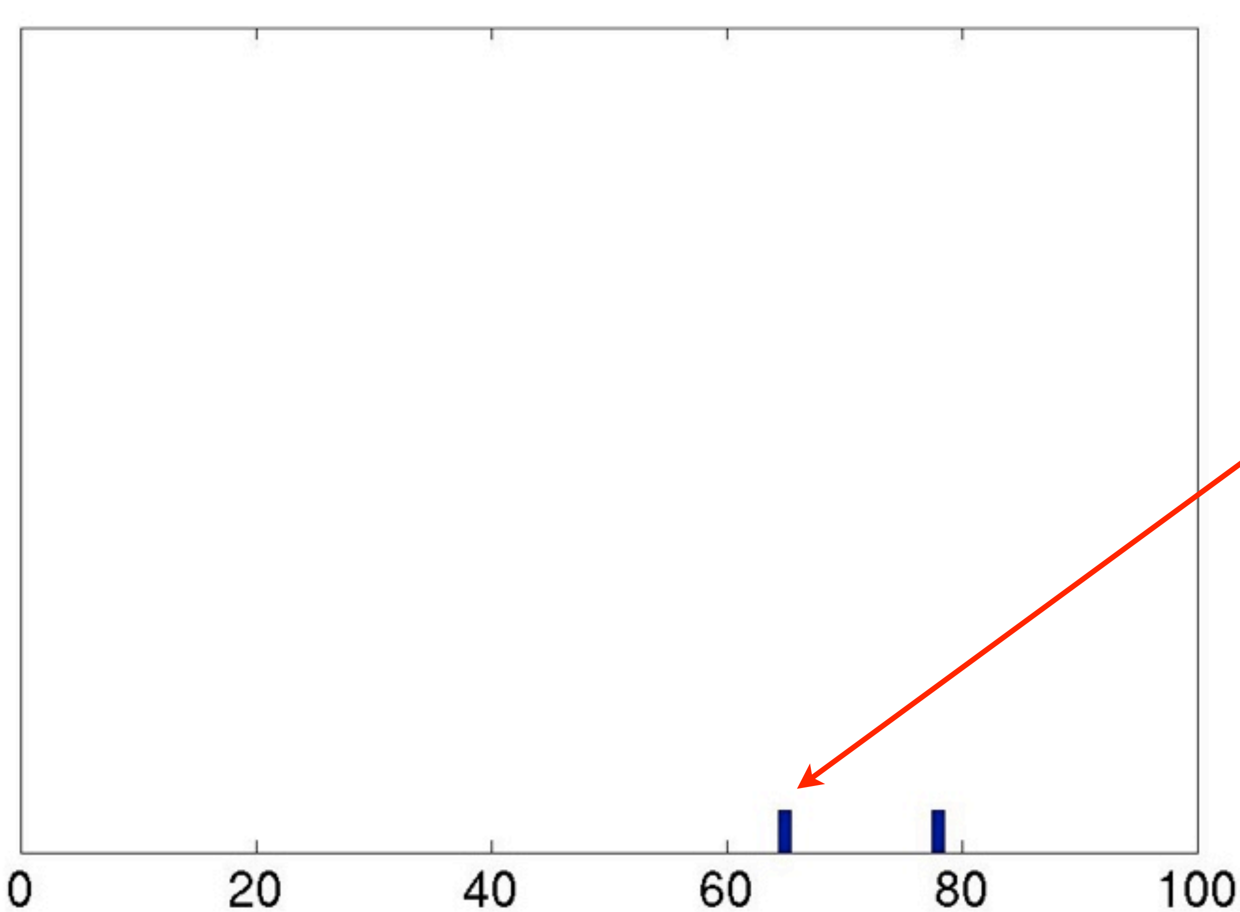


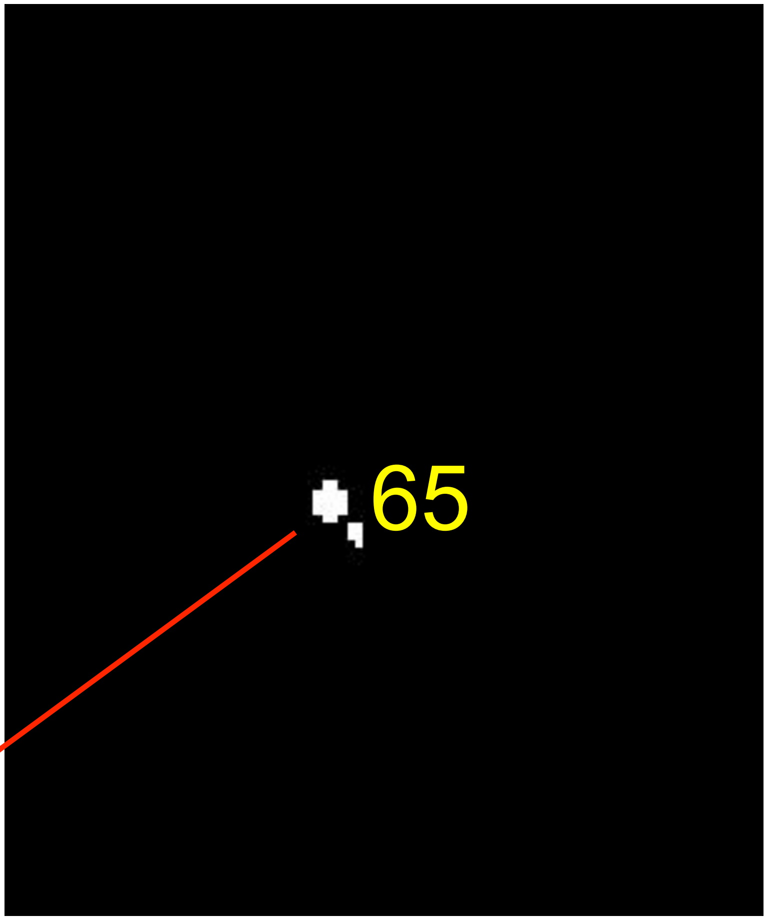
And record how large it is.



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.

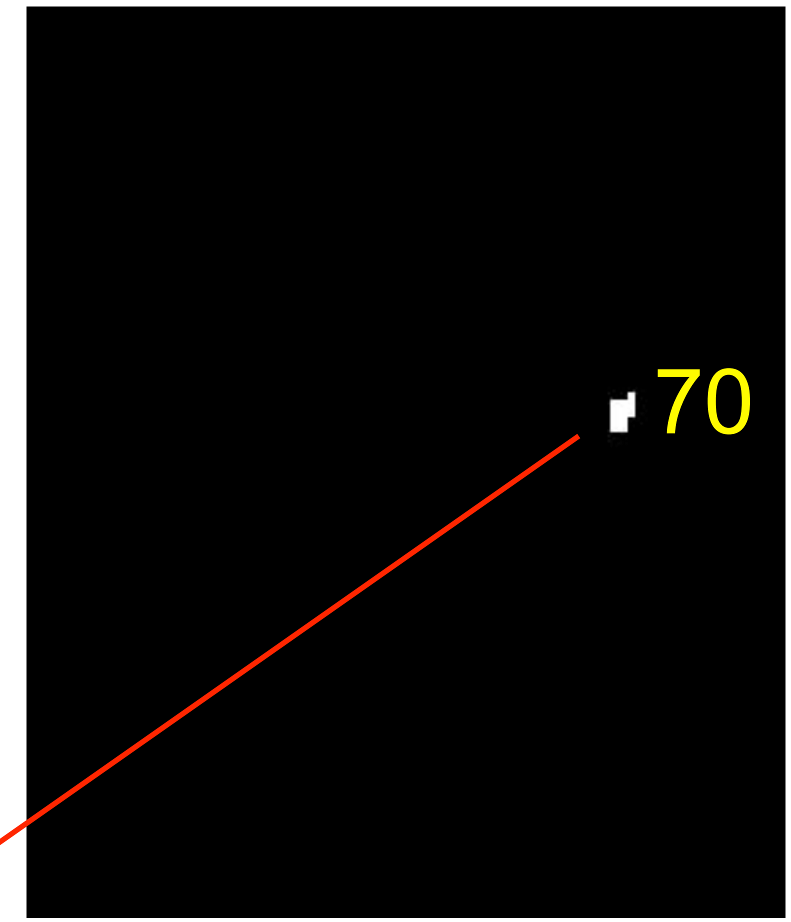




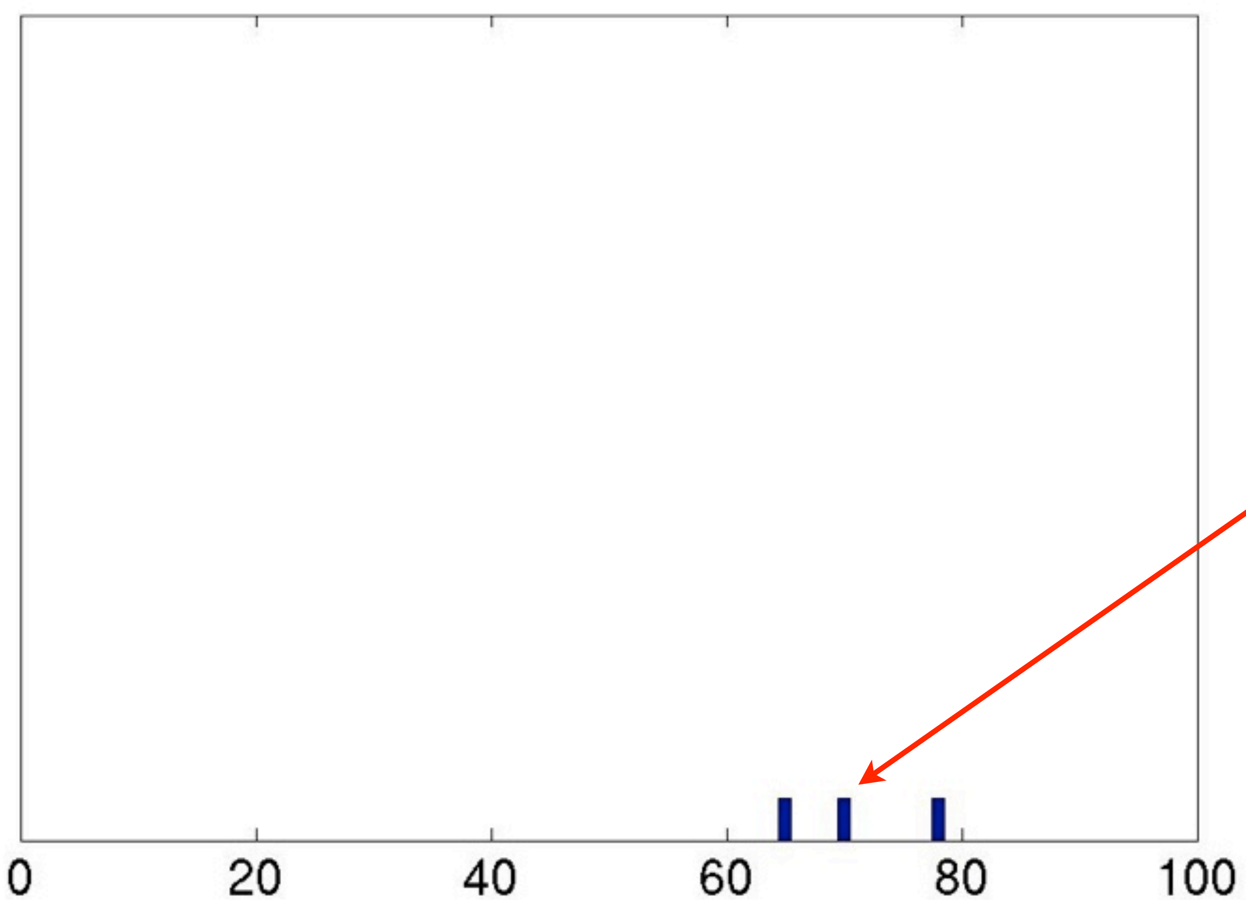
And do the same for another experiment...

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.

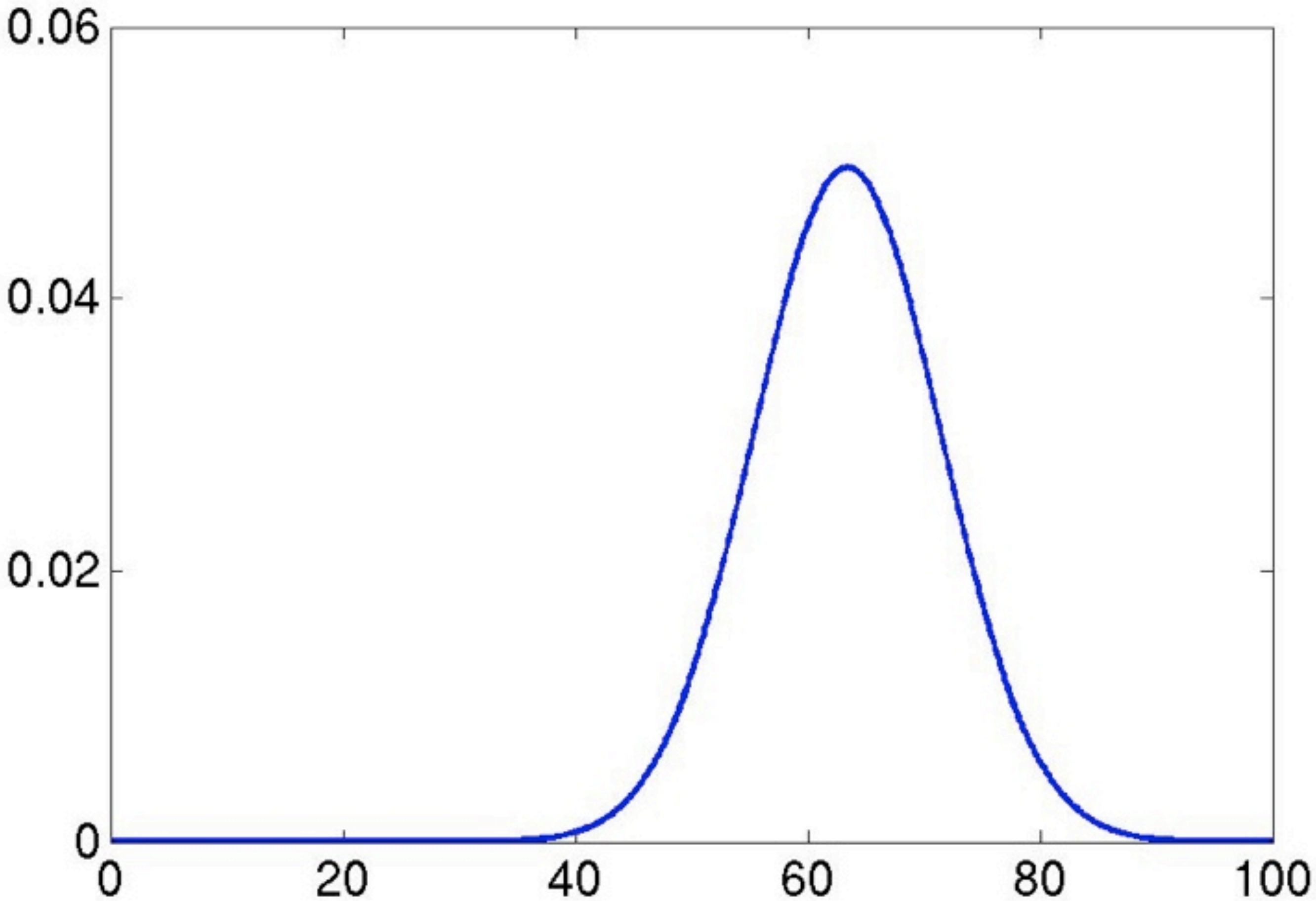


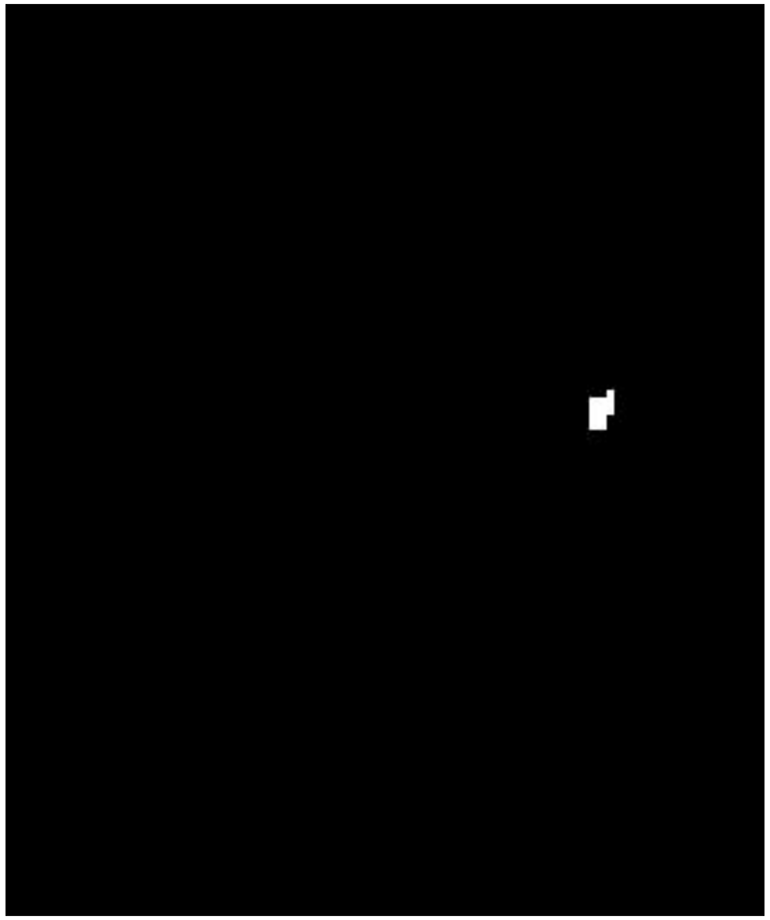
Etc ...



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.



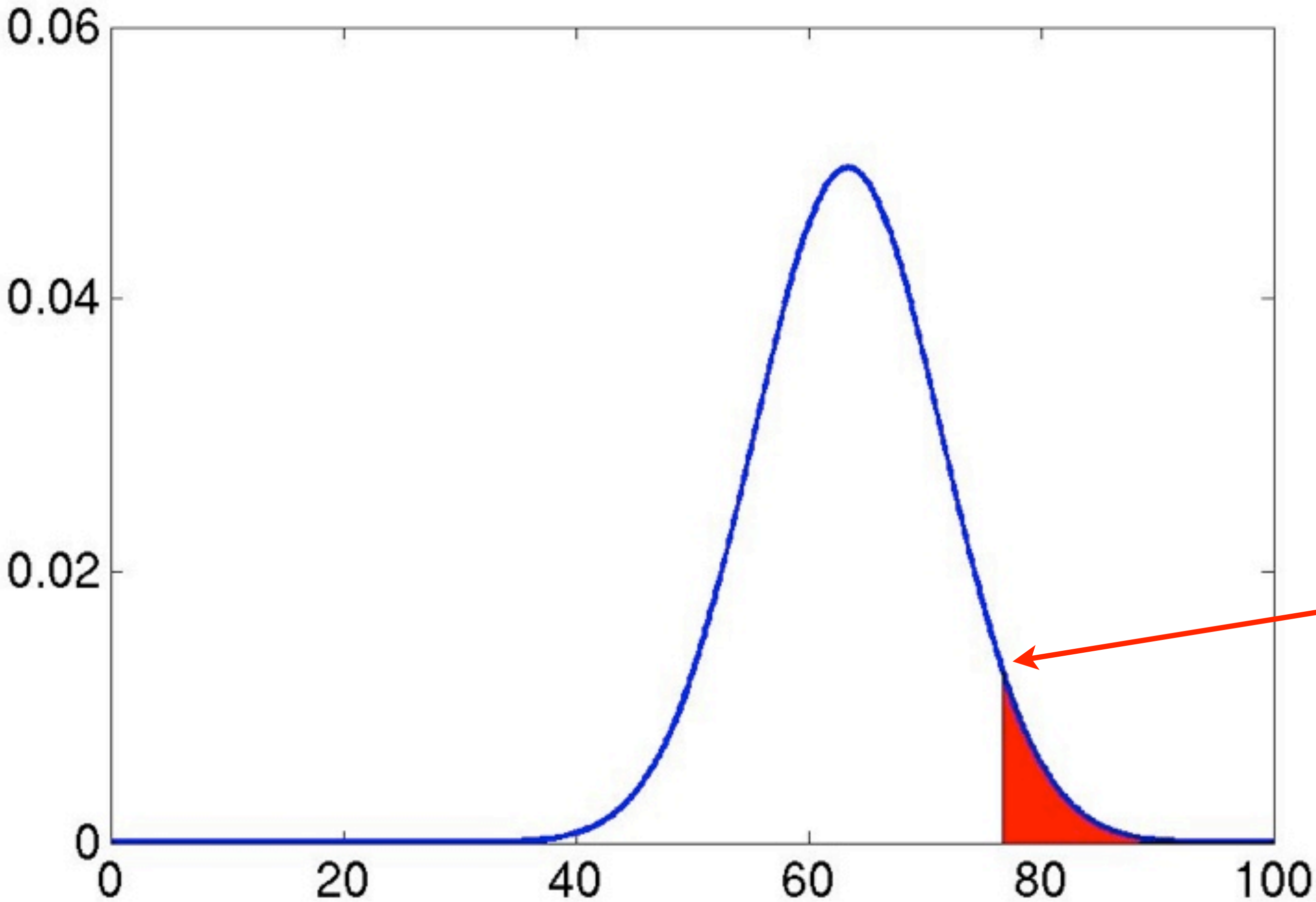


Until we have ...

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.

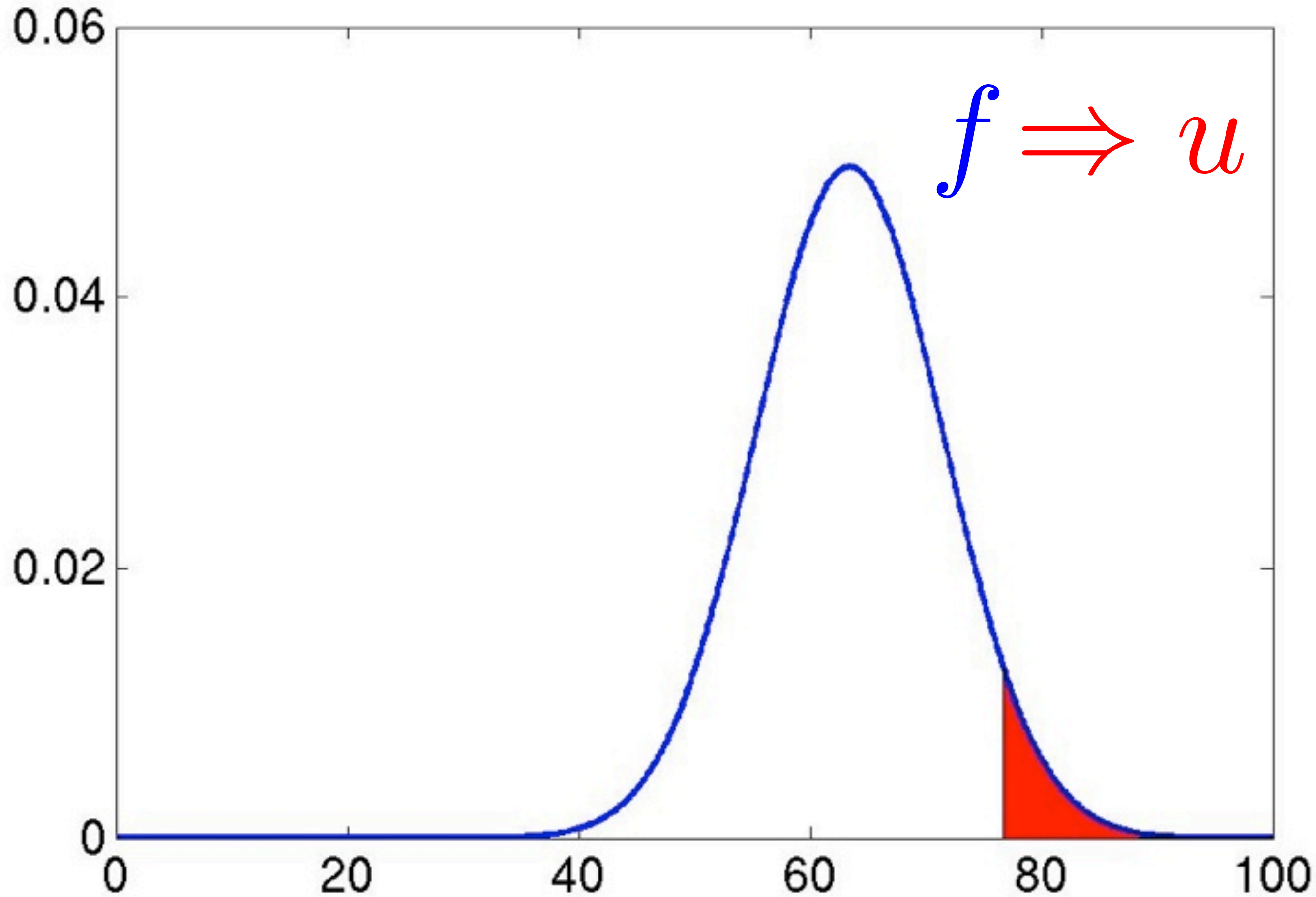
If we find a cluster larger than 76 voxels we reject the null-hypothesis.



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.

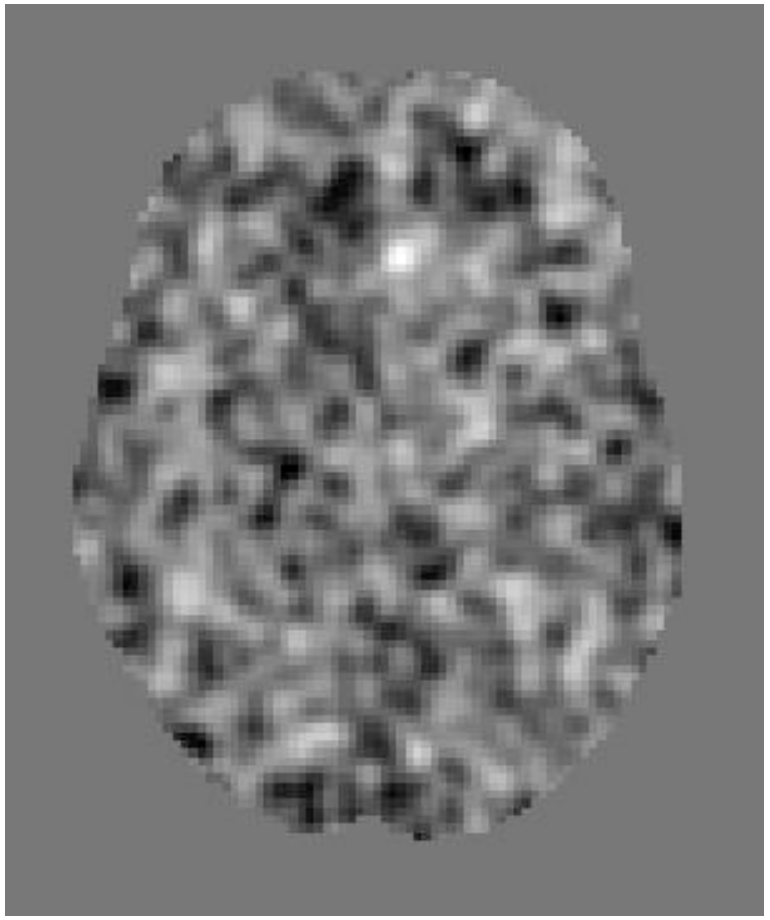


But what does f and u crucially depend on?

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.

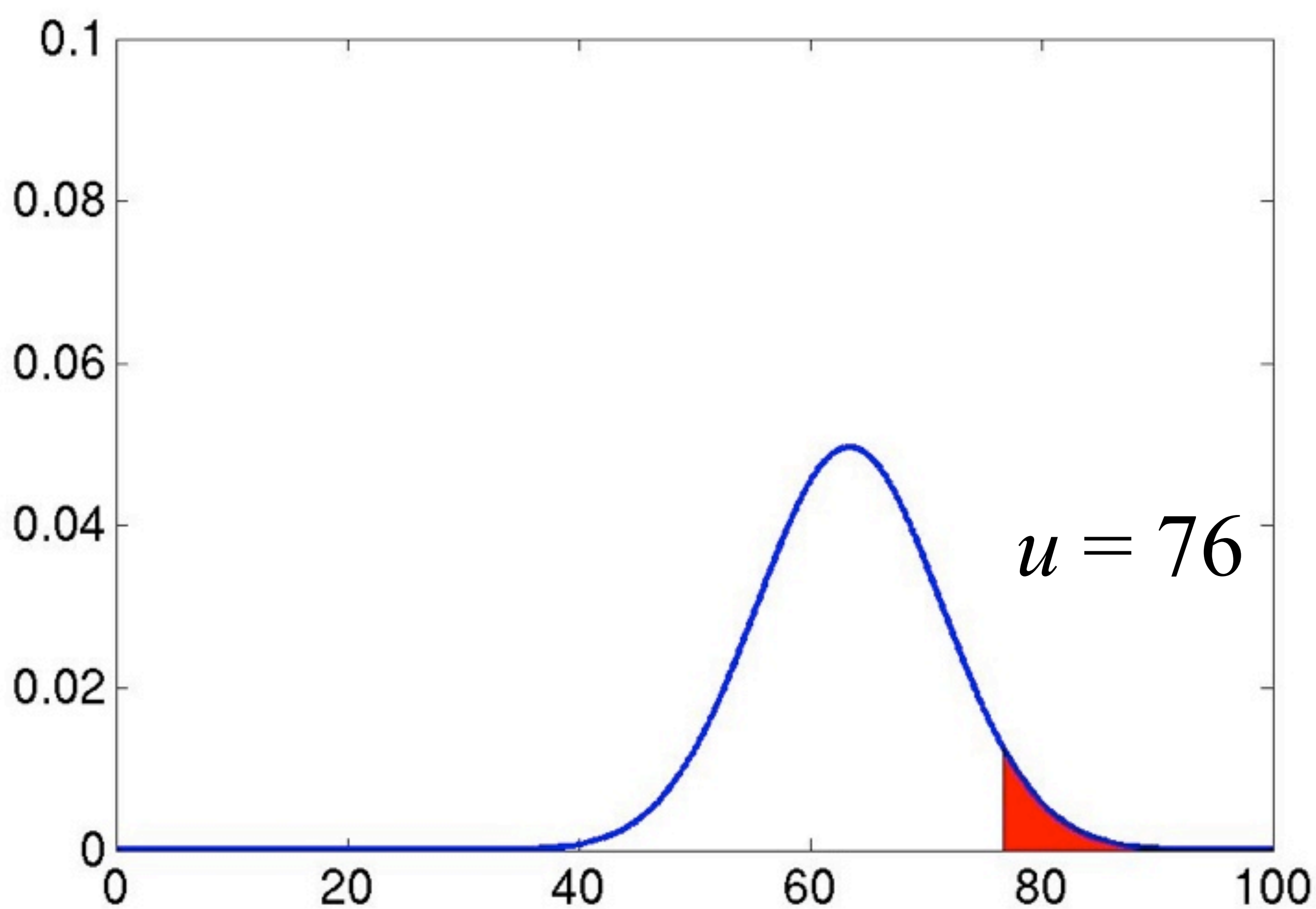
f depends crucially on the initial “cluster-forming” threshold?



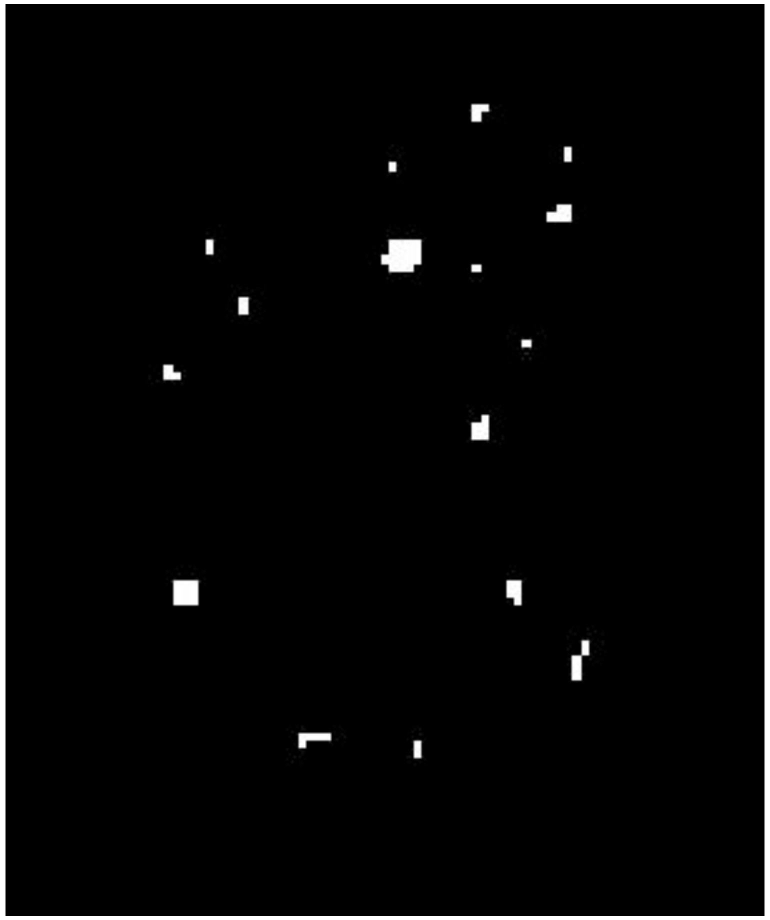
$z = 2.3$

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.



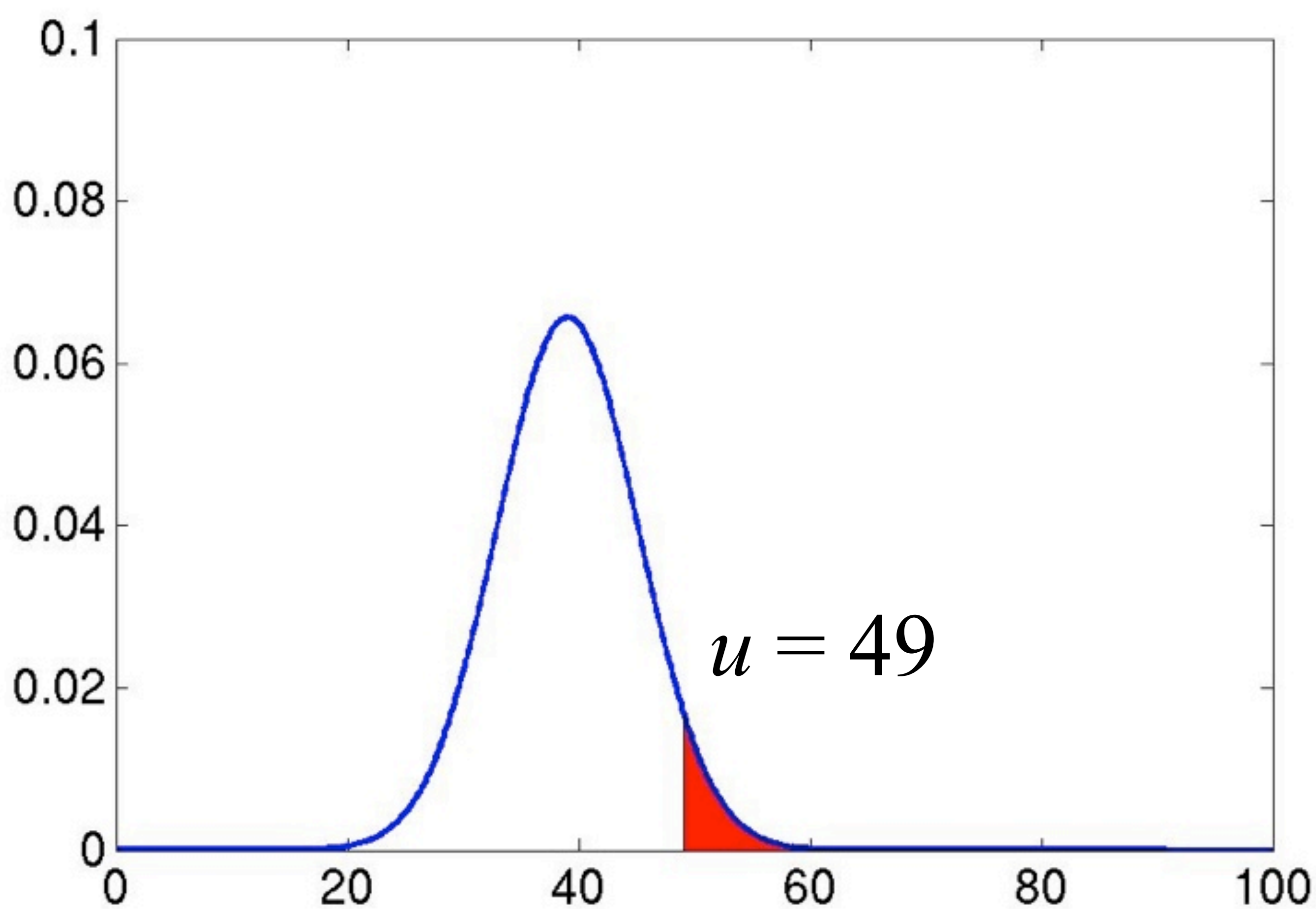
f depends crucially on the initial “cluster-forming” threshold?



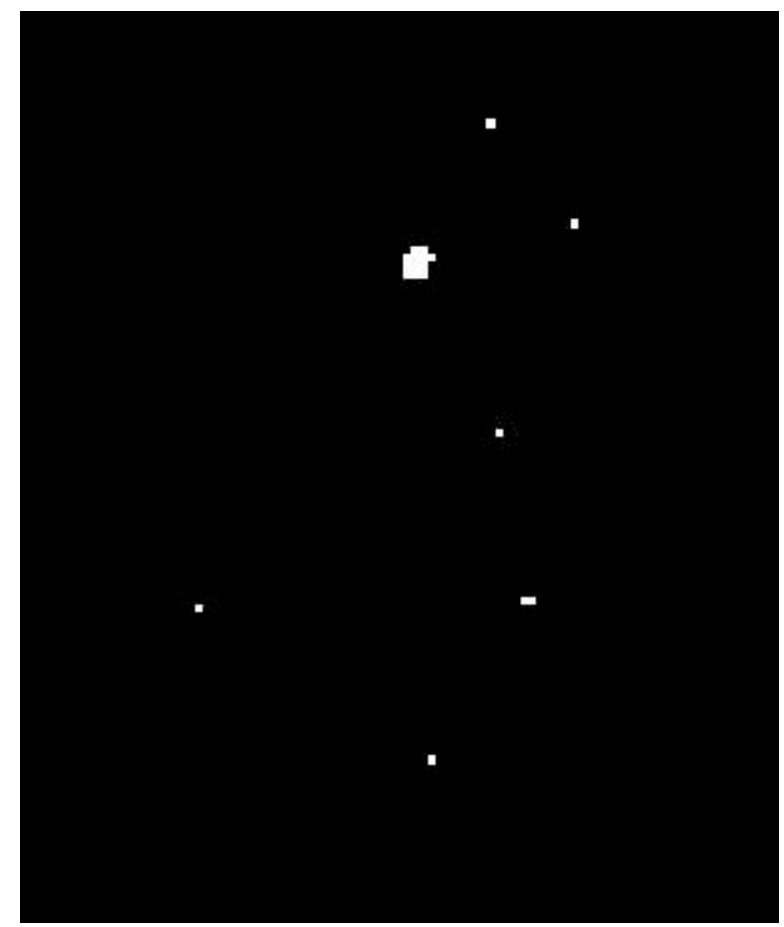
$z = 2.3$

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.



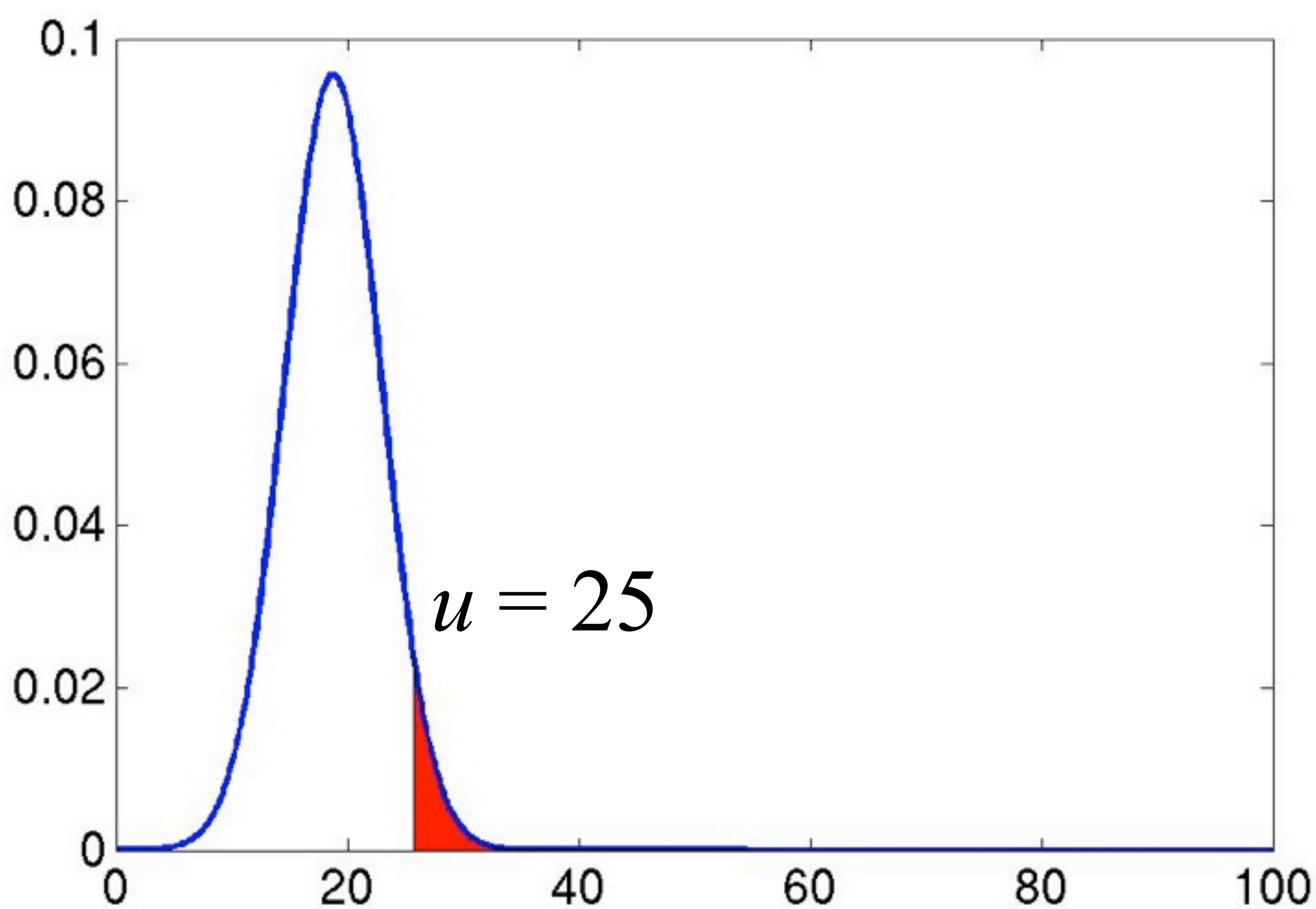
f depends crucially on the initial "cluster-forming" threshold?



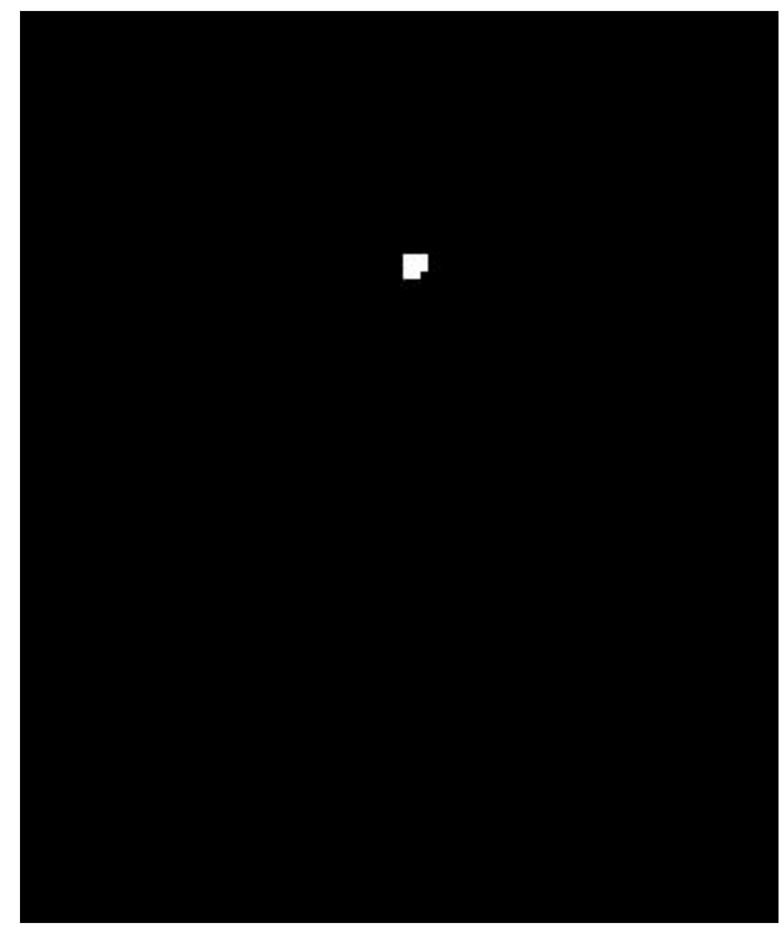
$z = 2.7$

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.



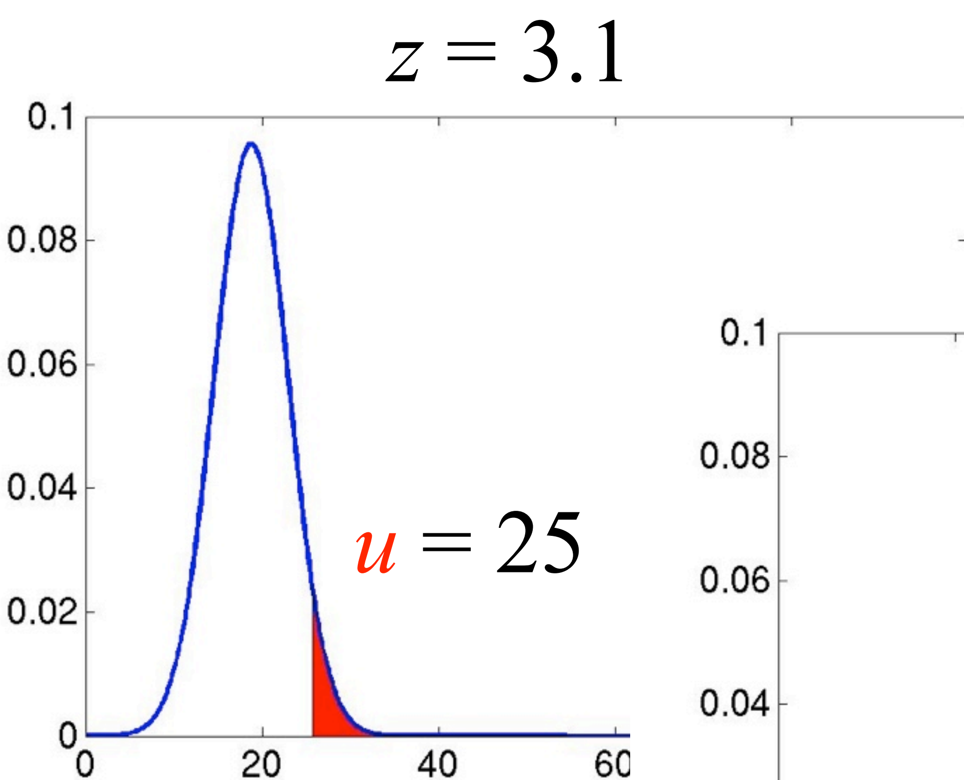
f depends crucially on the initial "cluster-forming" threshold?

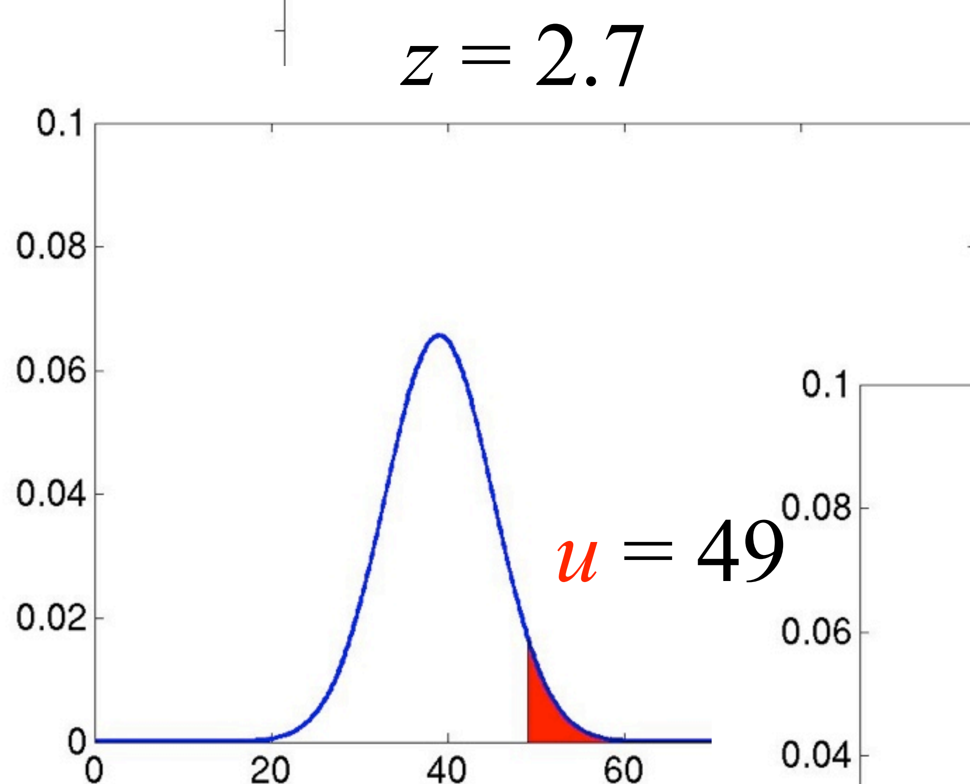


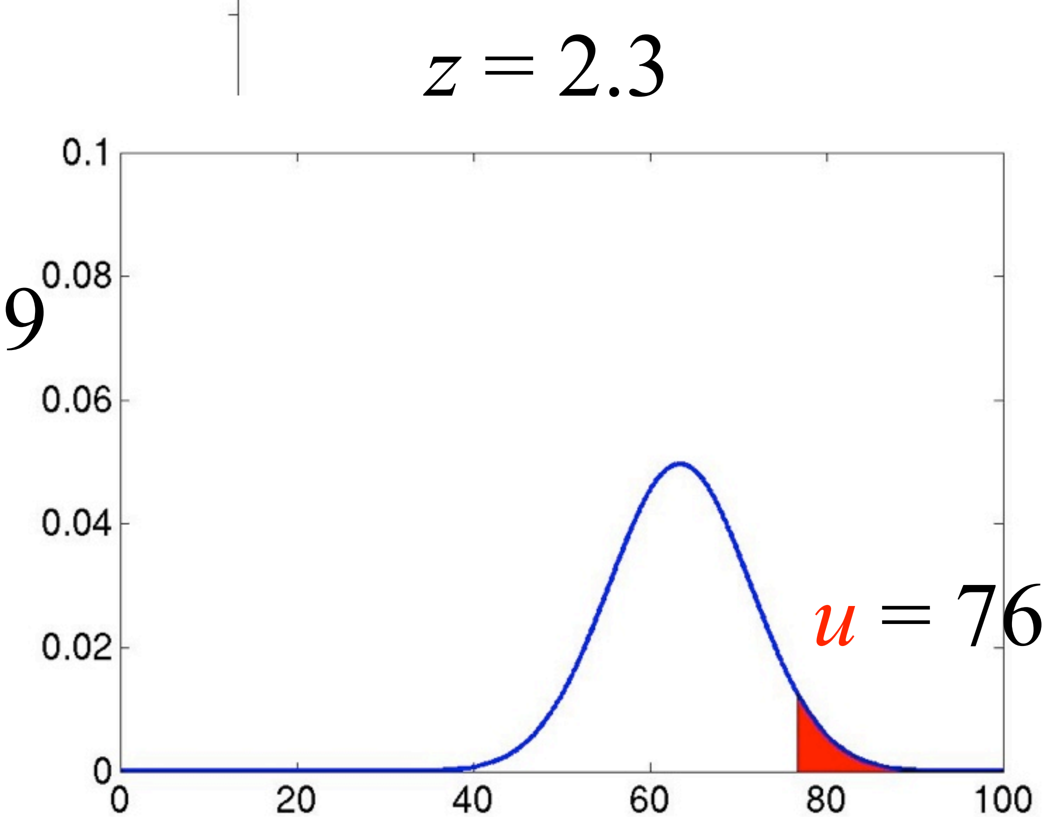
$z = 3.1$

Distribution of Max Cluster Size

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







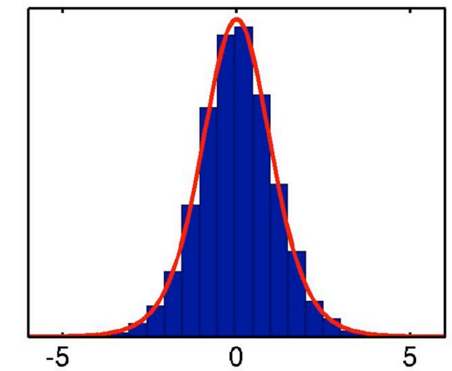
But as before we don't have an expression for these distributions.

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

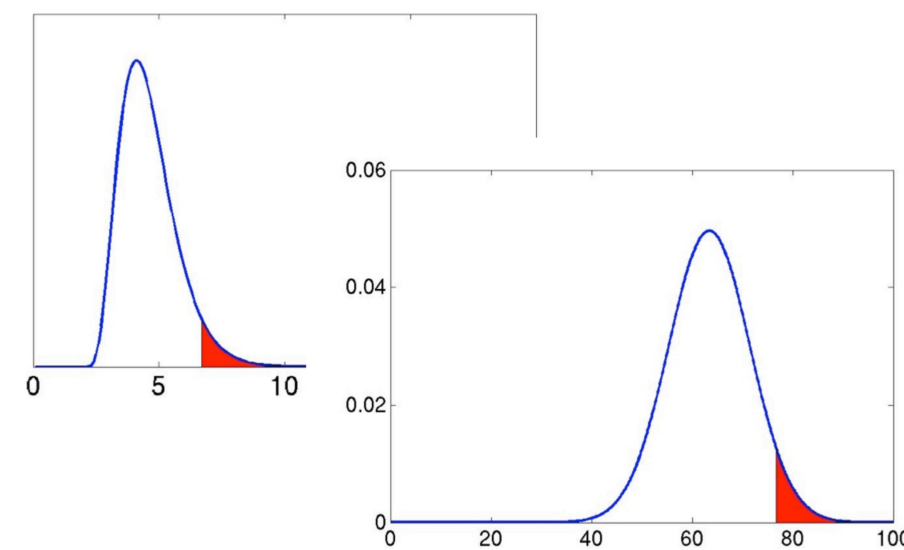
Parametric vs non-parametric

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

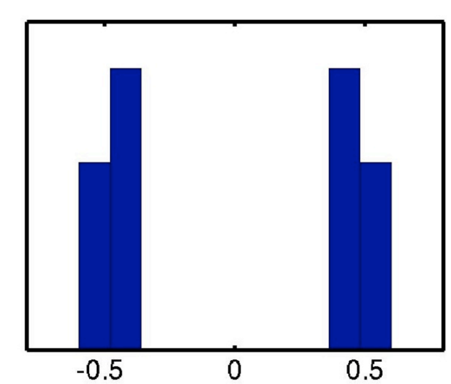


Provided that **e** ~ $N(0,\sigma^2)$

- But most distributions are not that simple



- And errors are not always normal-distributed

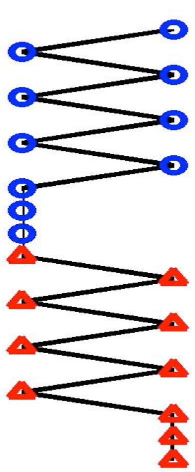


Example: VBM-style analysis

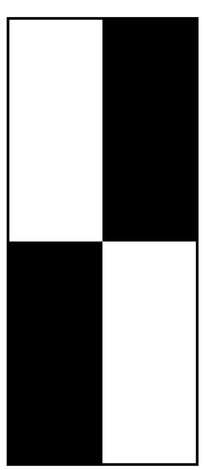
- Our data is segmented grey matter maps
- A voxel is either grey matter, or not.

Group #1
(Oxford students)

Group #2
(Train spotters)



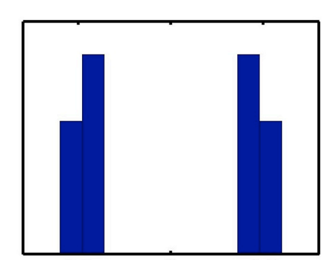
=



$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

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

hist($\mathbf{e}$)

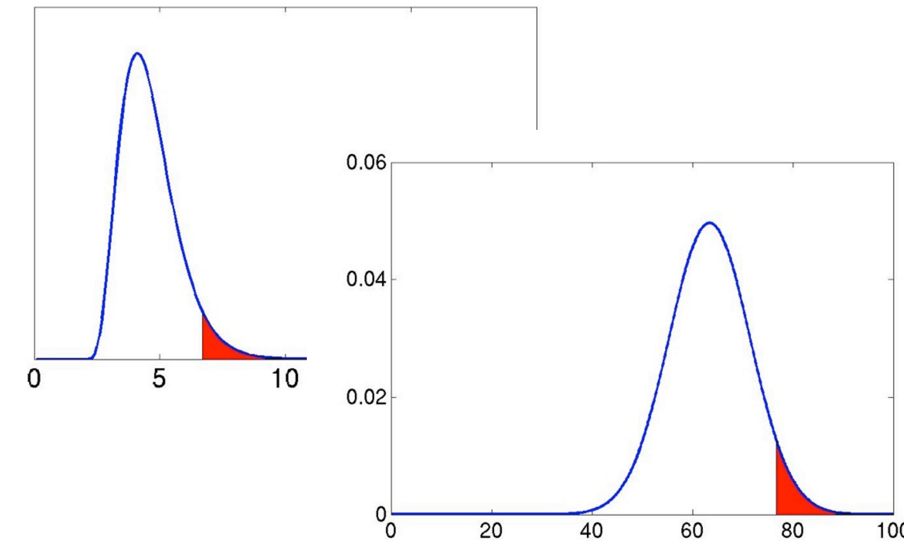


$\sim N$?

☹

Parametric vs non-parametric

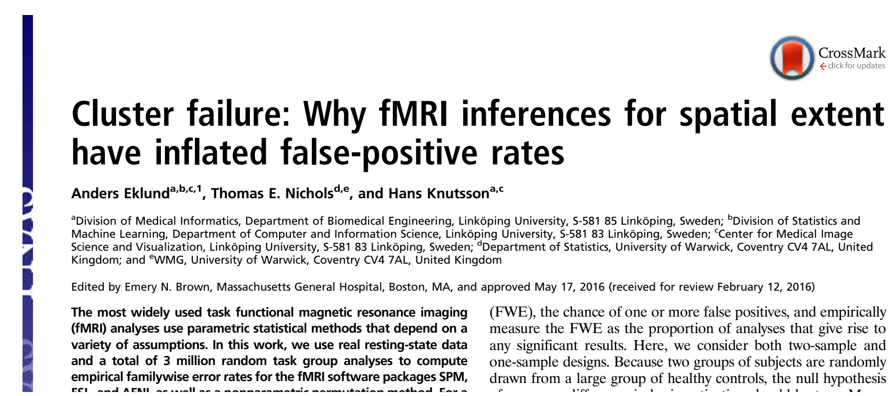
- There are <u>approximations</u> to the Max-z and Max-size statistics



- These are valid under certain sets of assumptions

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

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



CrossMark click for updates

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

Anders Eklund[a,b,c,1], Thomas E. Nichols[d,e], and Hans Knutsson[a,c]

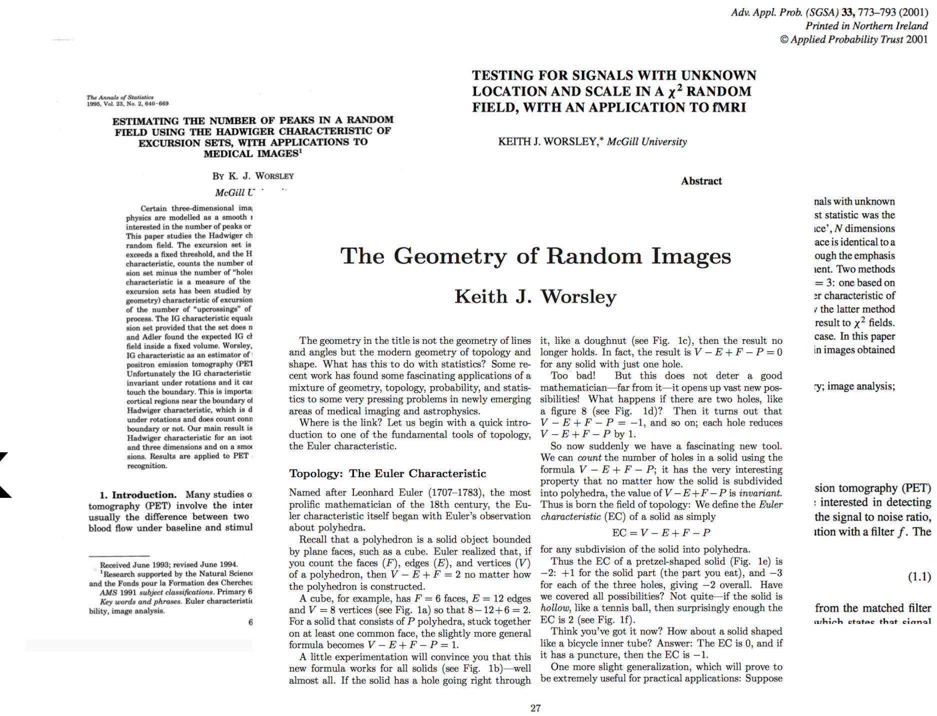
[a]Division of Medical Informatics, Department of Biomedical Engineering, Linköping University, S-581 85 Linköping, Sweden; [b]Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University, S-581 83 Linköping, Sweden; [c]Center for Medical Image Science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

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

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,

(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

TESTING FOR SIGNALS WITH UNKNOWN LOCATION AND SCALE IN A χ^2 RANDOM FIELD, WITH AN APPLICATION TO fMRI

KEITH J. WORSLEY,* McGill University

Abstract

ESTIMATING THE NUMBER OF PEAKS IN A RANDOM FIELD USING THE HADWIGER CHARACTERISTIC OF EXCURSION SETS, WITH APPLICATIONS TO MEDICAL IMAGES[1]

By K. J. Worsley

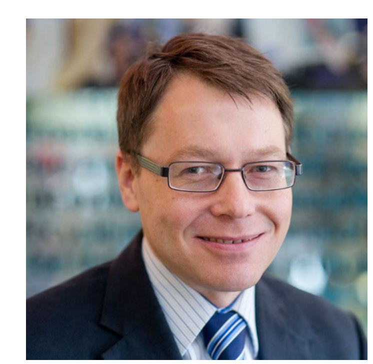
The Geometry of Random Images

Keith J. Worsley

Topology: The Euler Characteristic

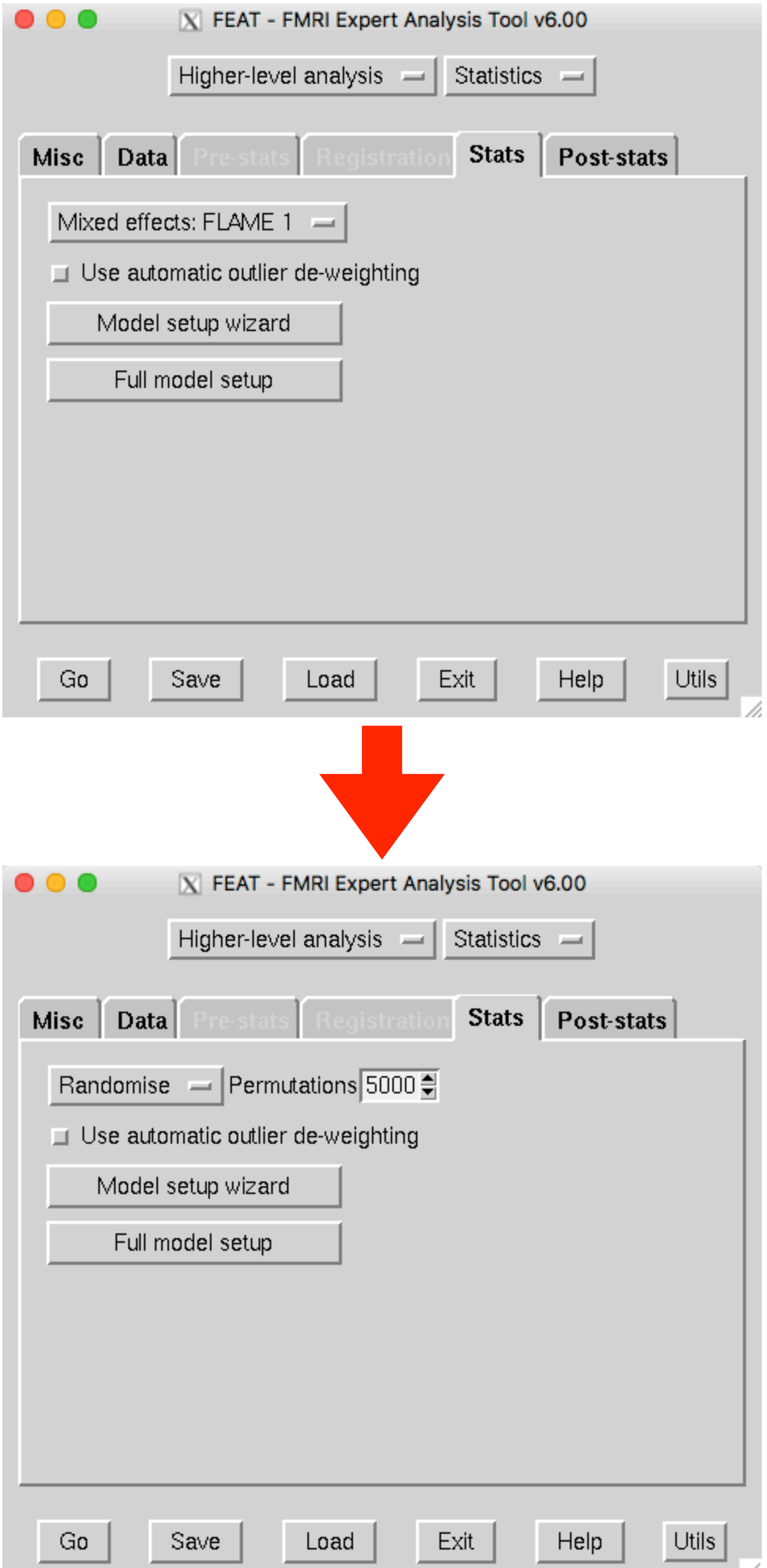
- Those approximations were based on Gaussian Random Field Theory, and was an impressive body of work

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



- But the future is non-parametric

Parametric vs non-parametric



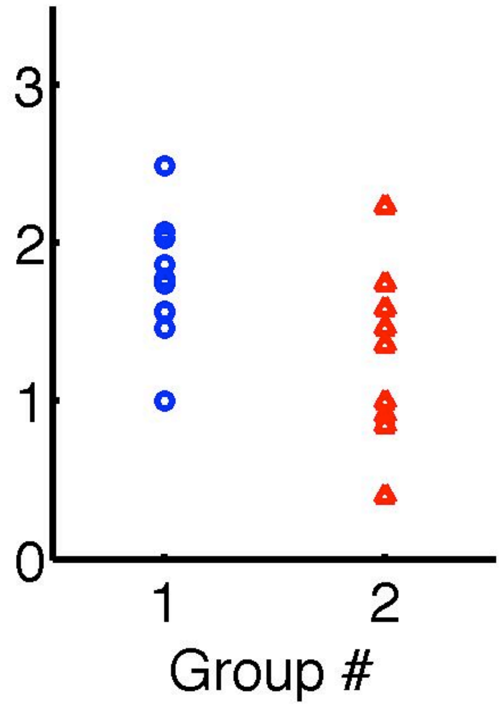
The Red
(randomise) Baron

FLAME going
down in flames

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

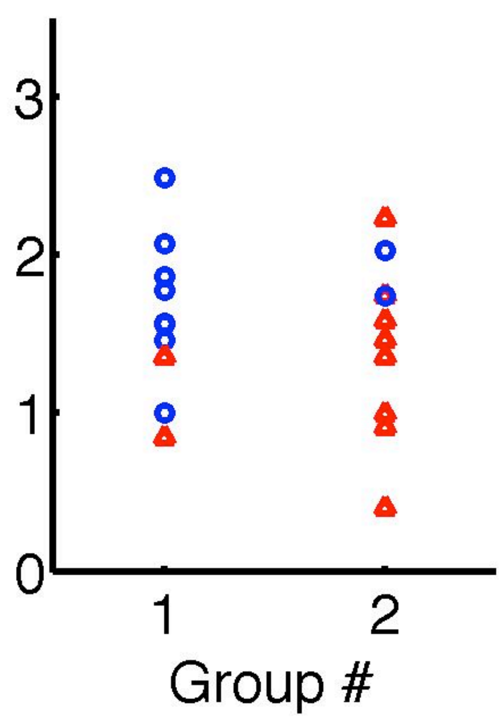
$$t = 2.27$$

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

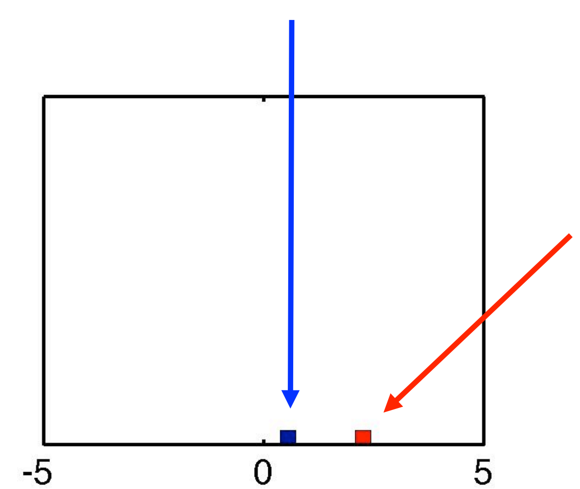
- We can permute the data itself to create a distribution that we can use to test our statistic.
  - + Makes very few assumptions about the data
  - + Works for any test statistic

One re-labelling



t-value after re-labelling

$$t = 0.67$$



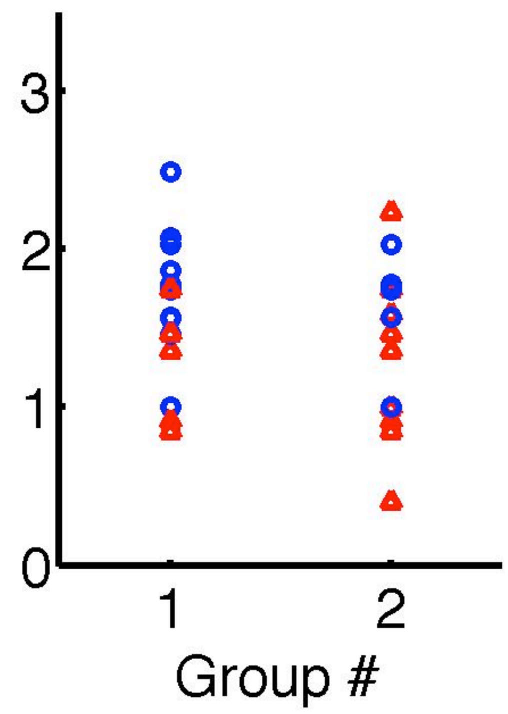
Original labelling

Let's start collecting them

A simple permutation test

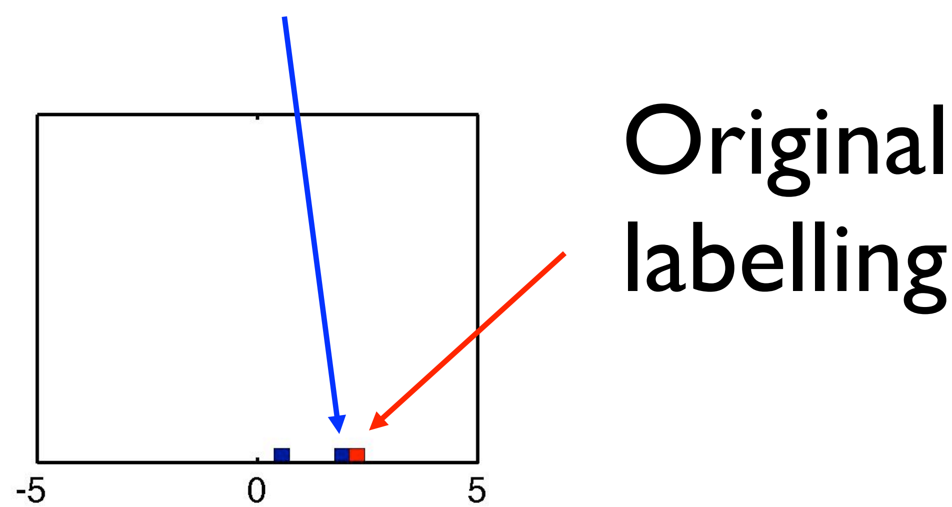
- We can permute the data itself to create a distribution that we can use to test our statistic.
  - \+ Makes very few assumptions about the data
  - \+ Works for any test statistic

Second re-labelling



t-value after re-labelling

$t = 1.97$



And another one

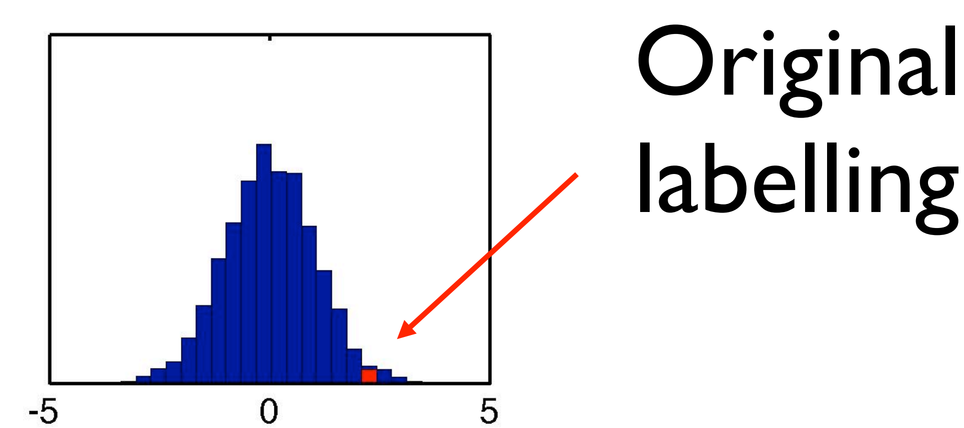
A simple permutation test

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

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

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

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

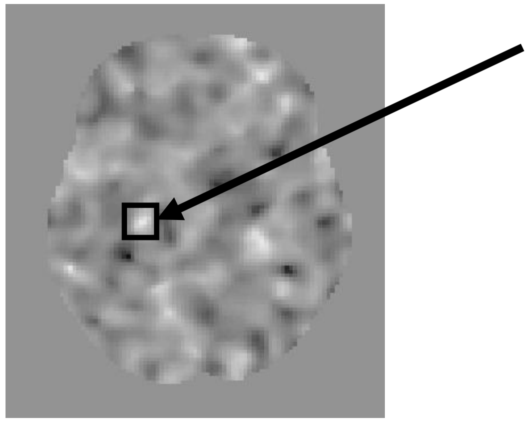


5000 re-labellings. Phew!

And we can use this for any statistic

This is what we got

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



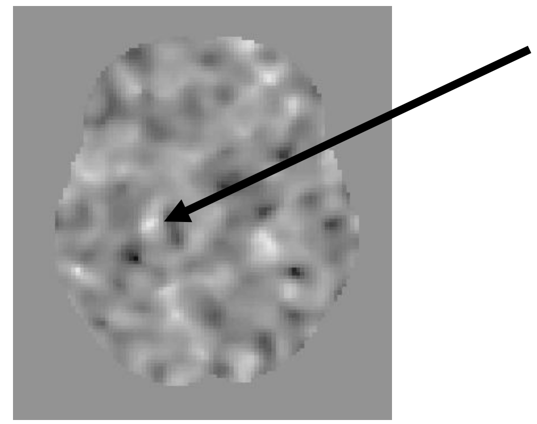
Very intriguing activation. $t_8 = 4.65$

Prof. ran to write to Science. **But**, did she 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

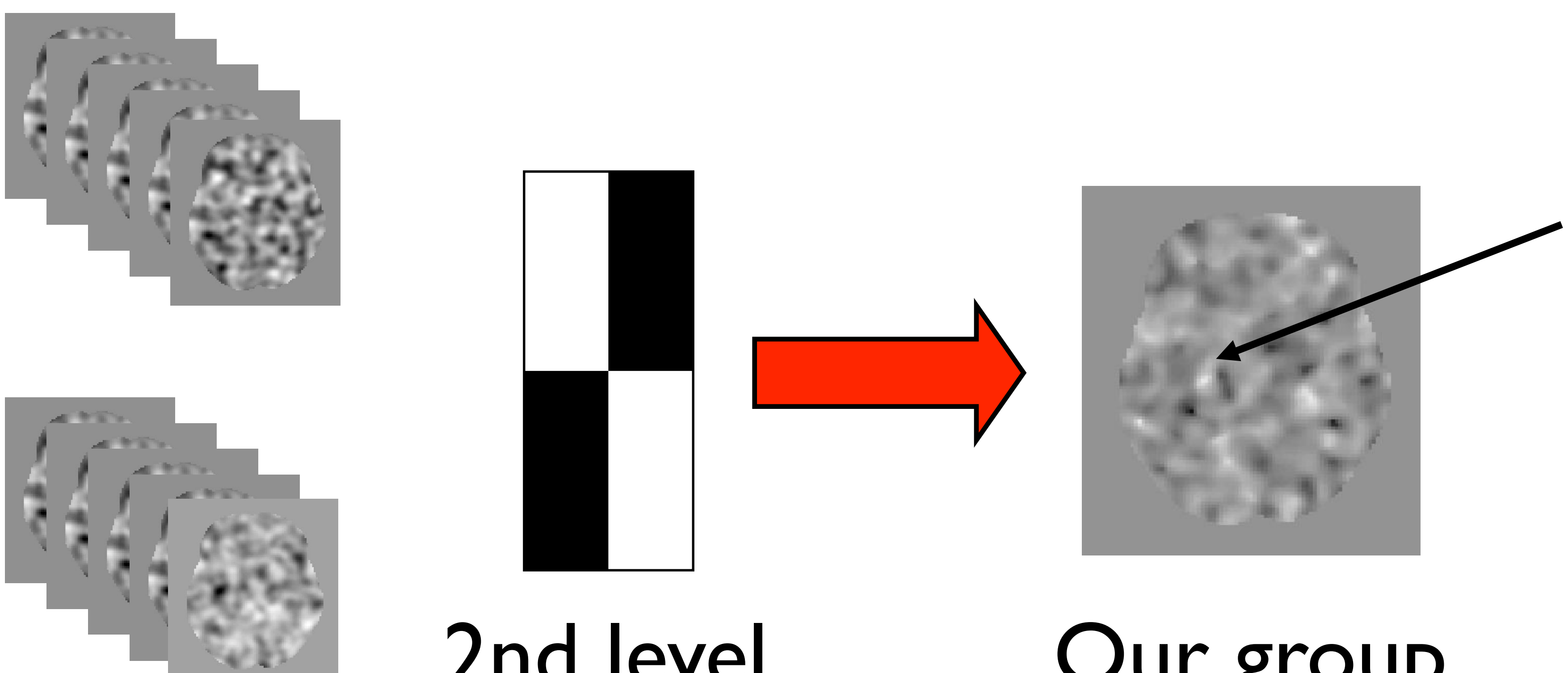


<u>Very</u> intriguing activation. $t_8 = 4.65$

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

Group 1

Group 2



max(t)=4.65

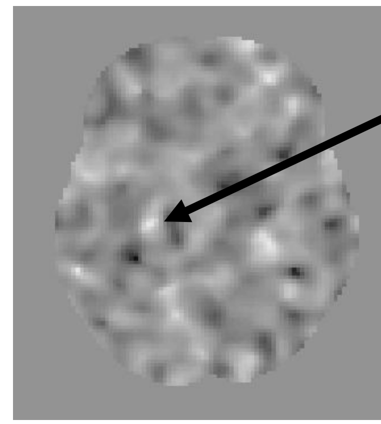
2nd level model

Our group difference map

And we can use this for any statistic

This is what we got

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

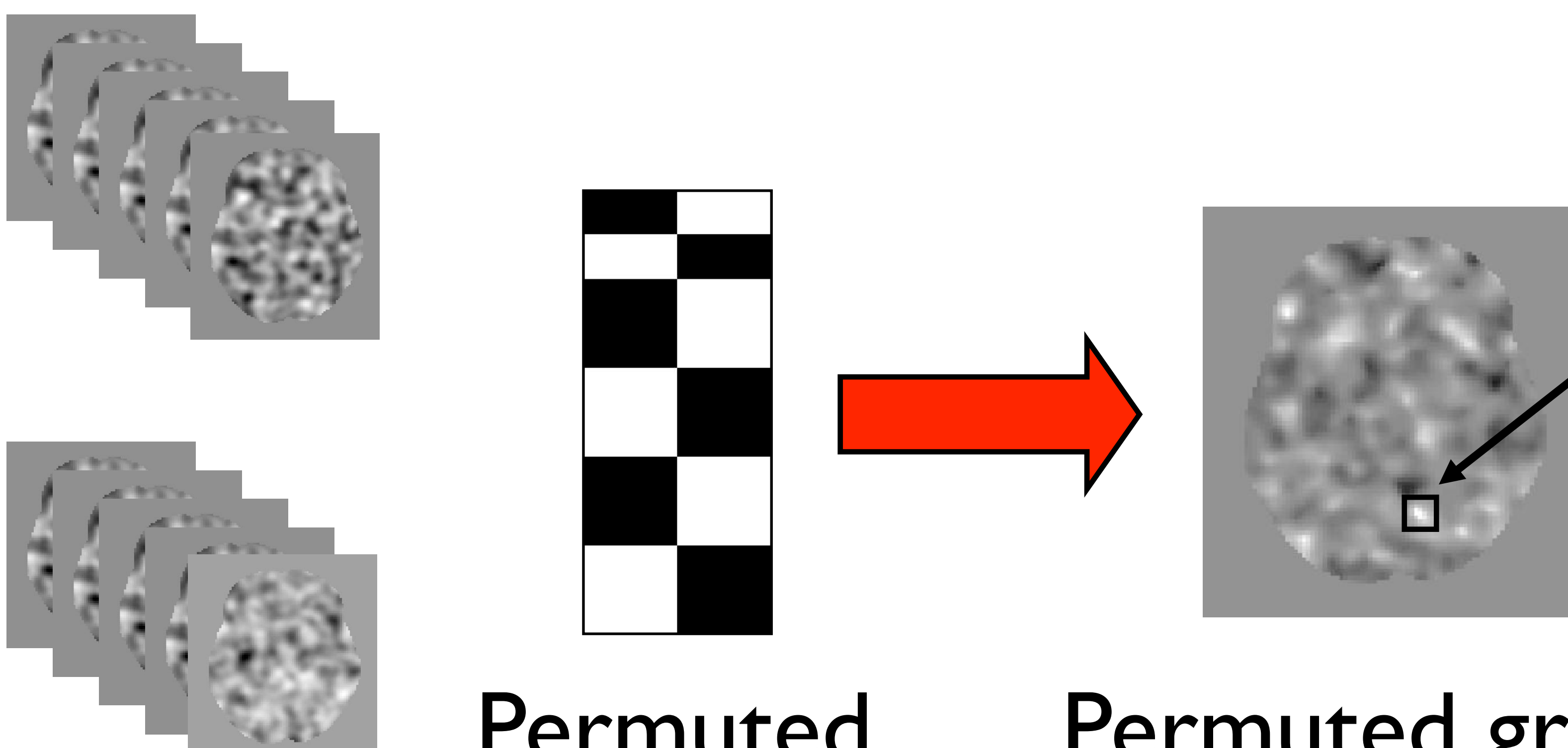


Very intriguing activation. $t_8 = 4.65$

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

Group 1

Group 2



$\max(t)$=8.23

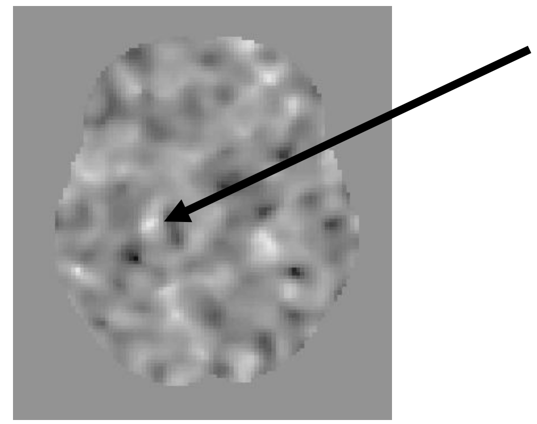
Permuted model

Permuted group difference map

And we can use this for any statistic

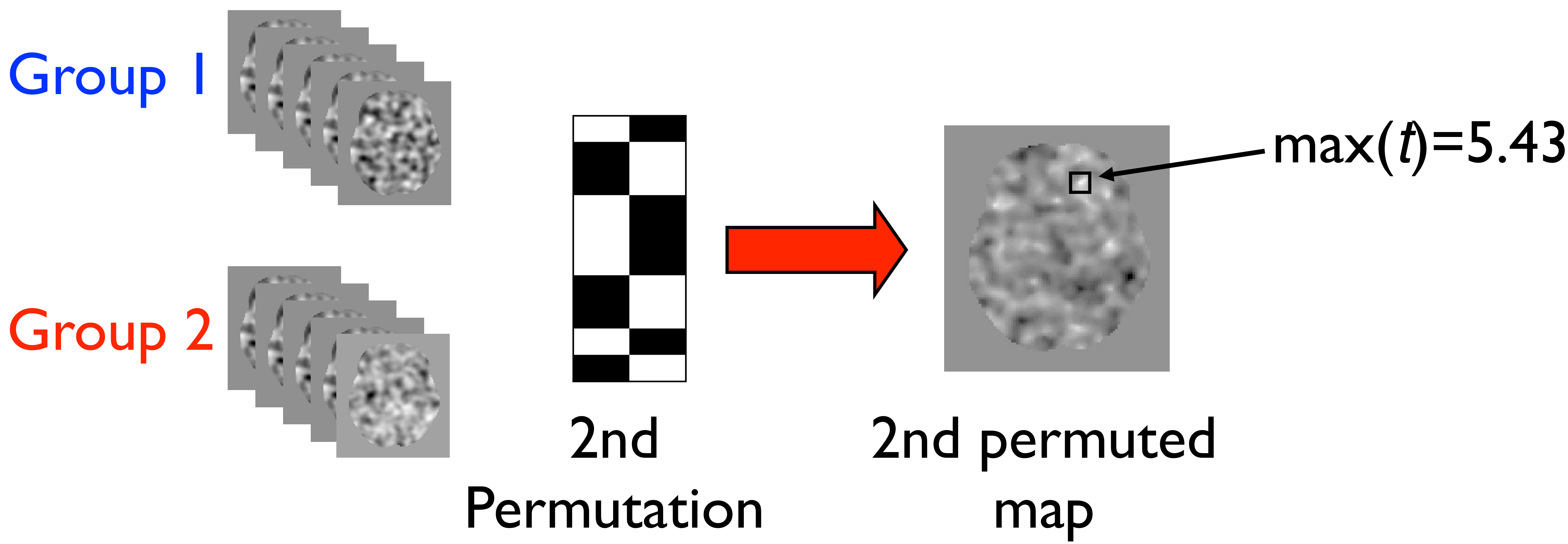
We compared activation by painful stimuli in two groups of 5 subjects each.

This is what we got



Very intriguing activation. $t_8 = 4.65$

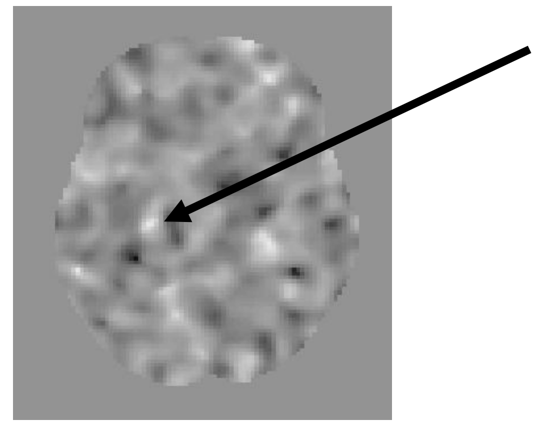
Prof. ran to write to Science. **But**, did she jump the gun?



And we can use this for any statistic

This is what we got

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

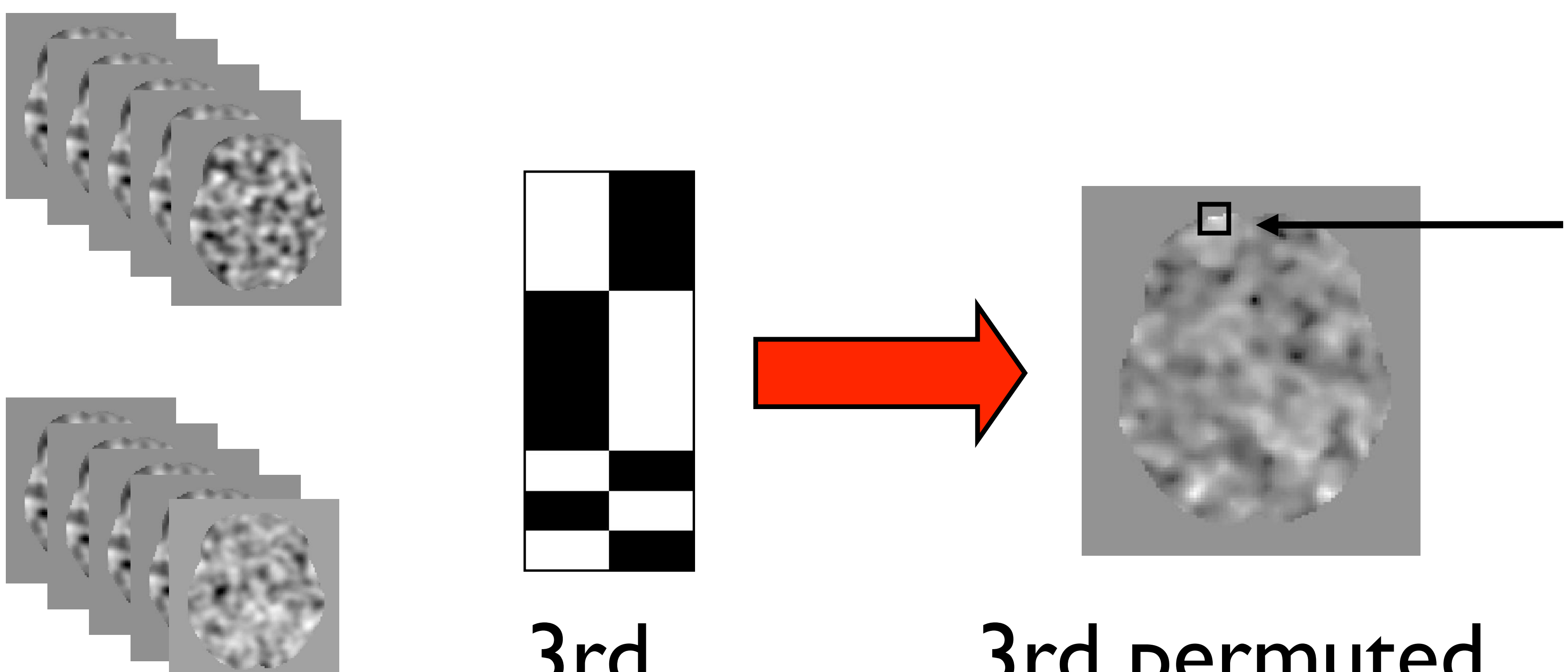


<u>Very</u> intriguing activation. $t_8 = 4.65$

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

Group 1

Group 2



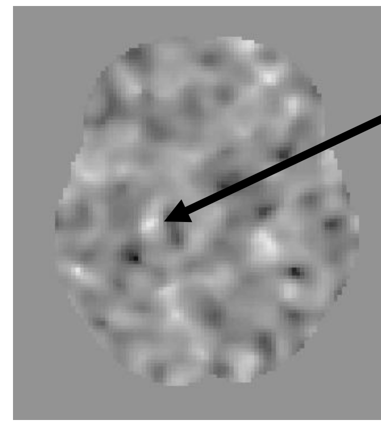
3rd Permutation

3rd permuted map

And we can use this for any statistic

This is what we got

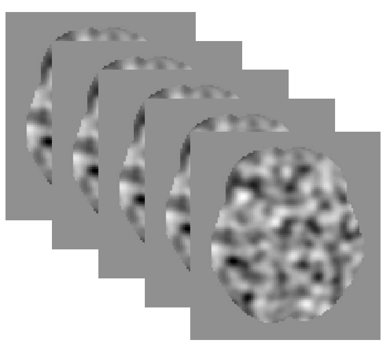
We compared activation by painful stimuli in two groups of 5 subjects each.



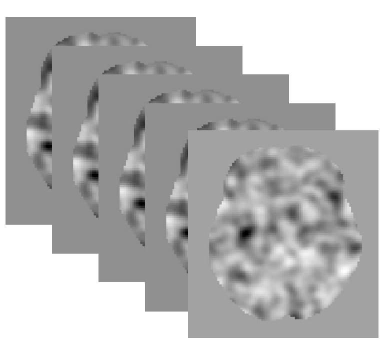
Very intriguing activation. $t_8 = 4.65$

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

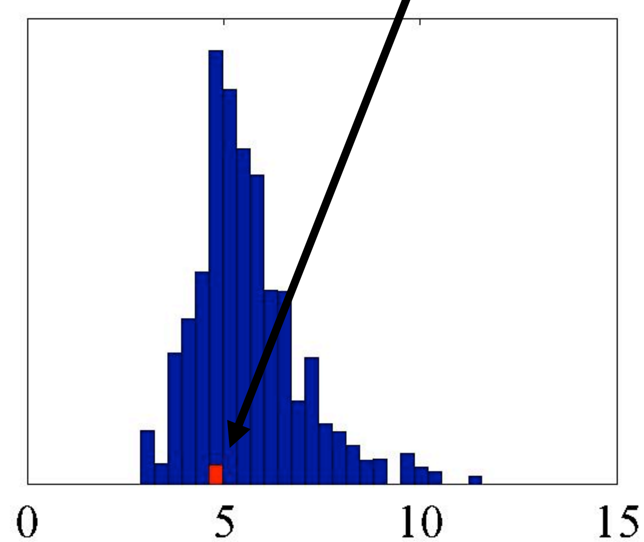
Group 1



Group 2



Original labelling

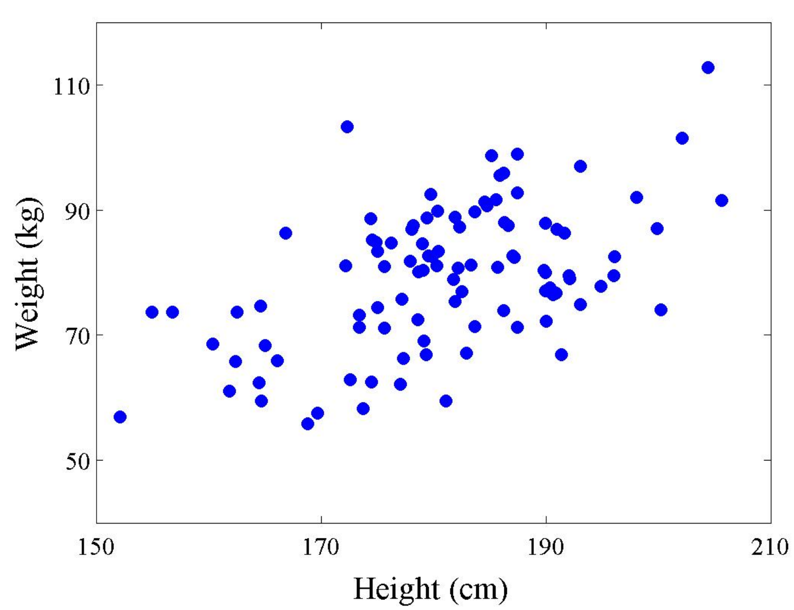


5000 permutations

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”
- I will start to explain “exchangeability” through a case that is not
- But first we need to learn about covariance matrices

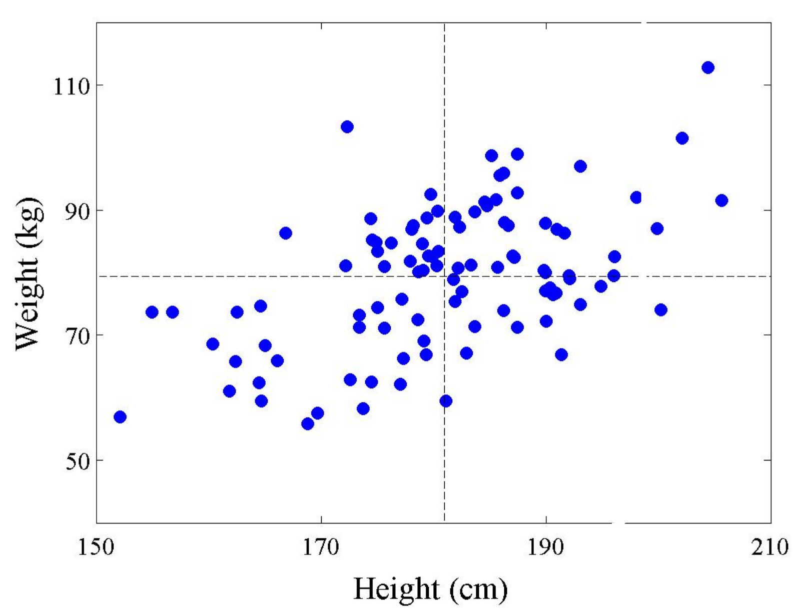


Height and weight of a random sample of Swedish men

Covariance matrices

- When we swap the labels of two data-points we need to make sure that they are "exchangeable"
- I will start to explain "exchangeability" through a case that is not
- But first we need to learn about covariance matrices

Mean height ≈181 cm

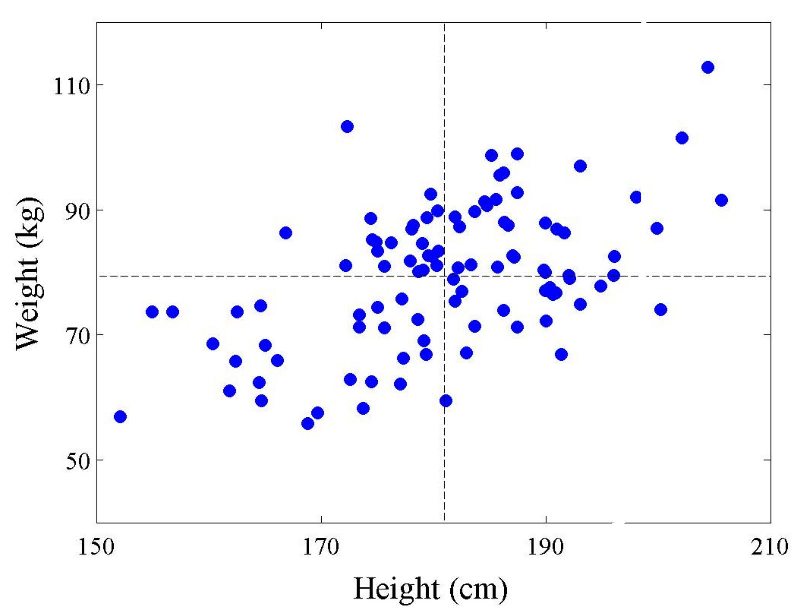


Mean weight ≈79.4 kg

Characterised by two means

Covariance matrices

- When we swap the labels of two data-points we need to make sure that they are "exchangeable"
- I will start to explain "exchangeability" through a case that is not
- But first we need to learn about covariance matrices

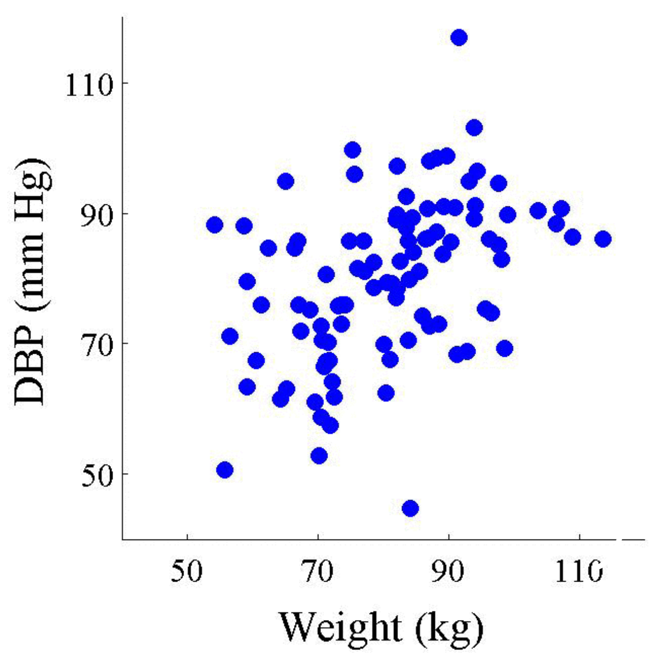


$$\Sigma = \begin{bmatrix} 130 & 52 \\ 52 & 165 \end{bmatrix}$$

And a covariance - matrix

Covariance matrices

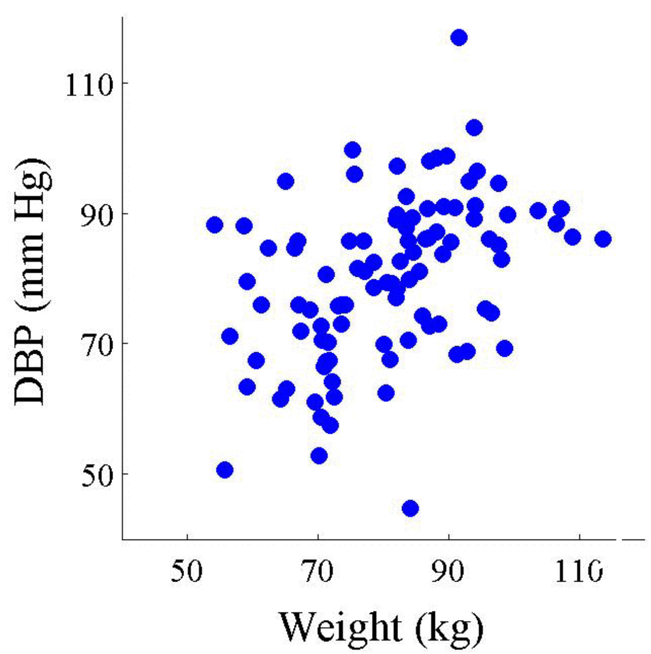
- When we swap the labels of two data-points we need to make sure that they are "exchangeable"
- I will start to explain "exchangeability" through a case that is not
- But first we need to learn about covariance matrices



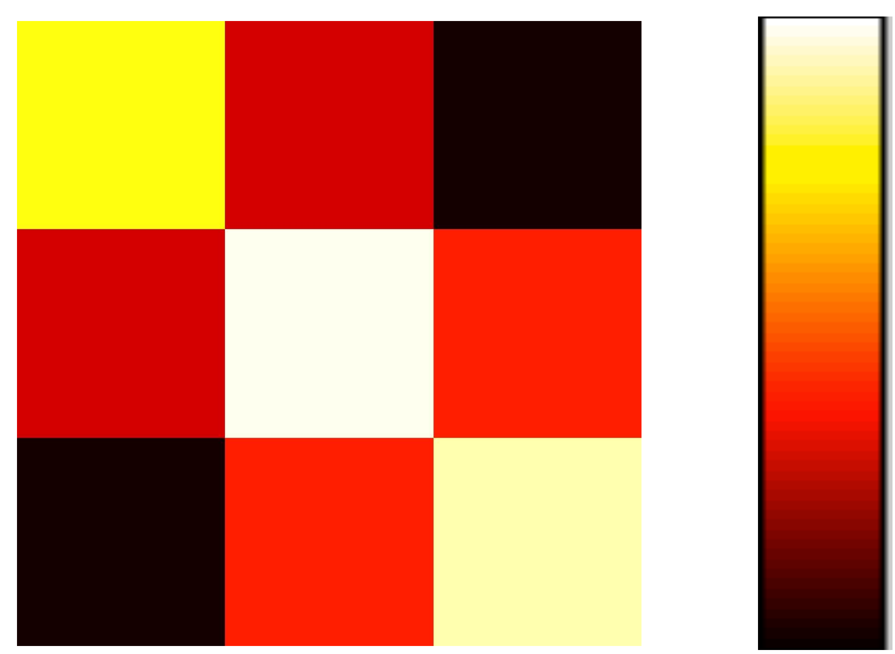
$$\Sigma = \begin{bmatrix} 130 & 52 & 4.8 \\ 52 & 165 & 69 \\ 4.8 & 69 & 156 \end{bmatrix}$$

Covariance matrices

- When we swap the labels of two data-points we need to make sure that they are "exchangeable"
- I will start to explain "exchangeability" through a case that is not
- But first we need to learn about covariance matrices

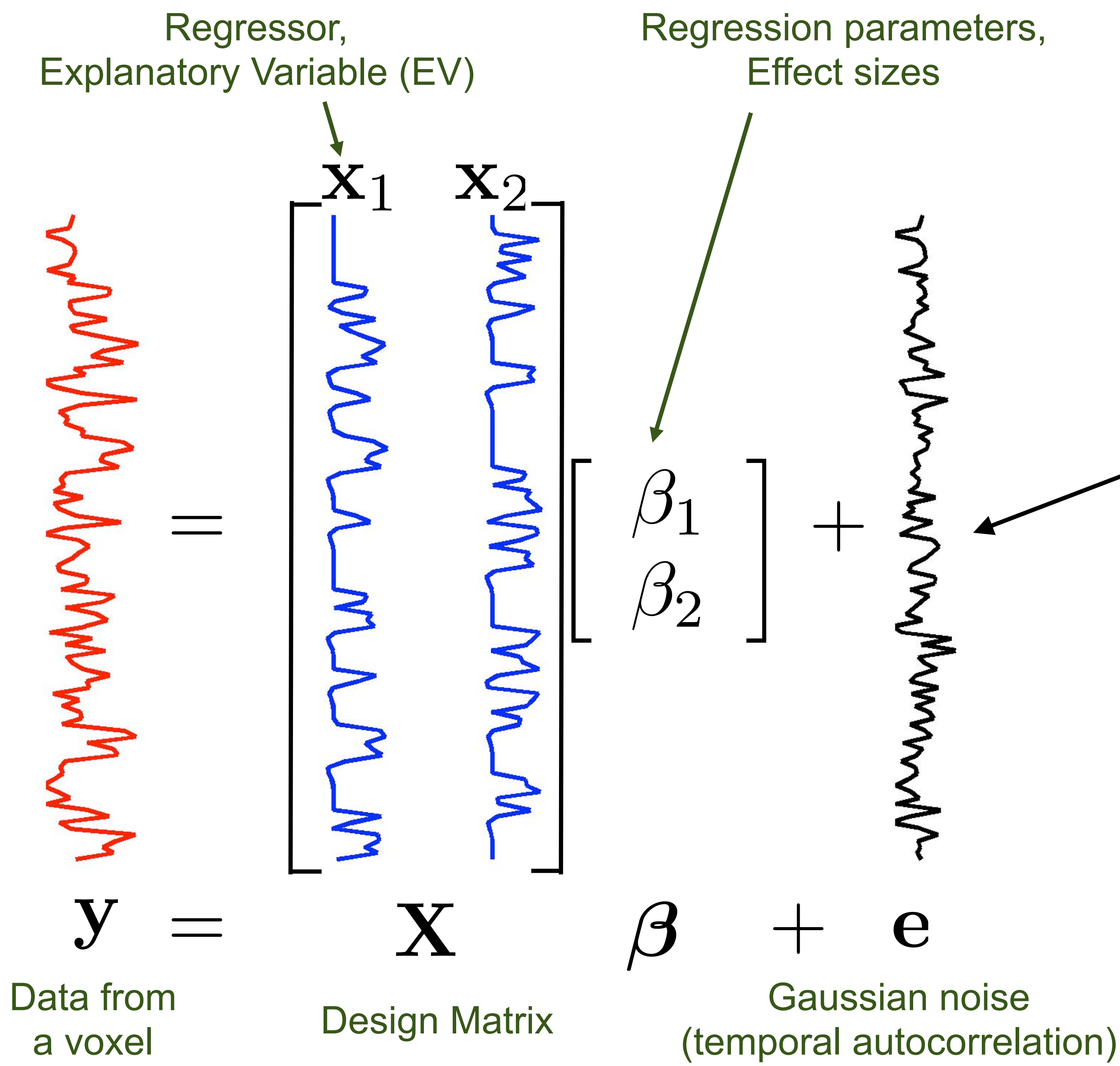


$\Sigma =$



1st level fMRI data is not exchangeable

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



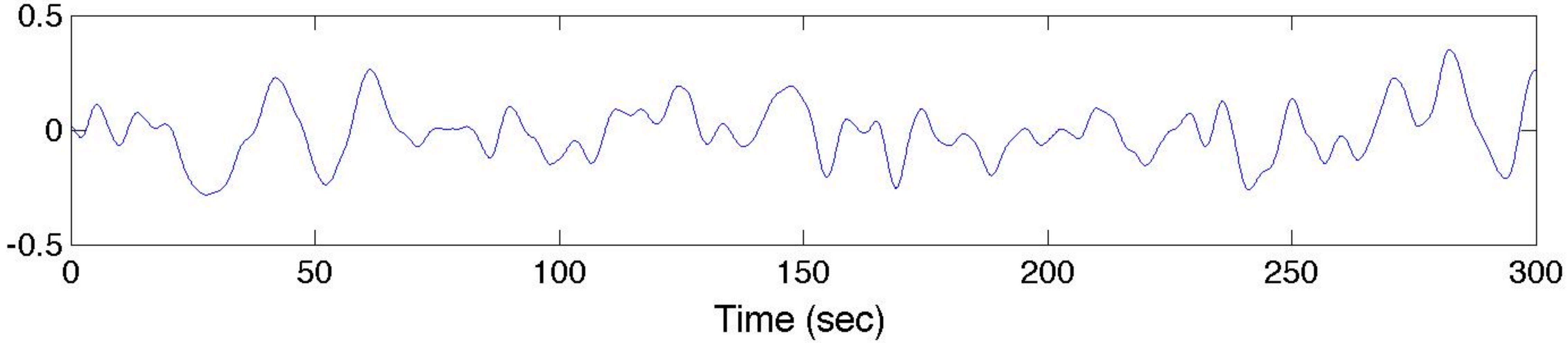
This time we will look more closely at this part

$$\mathbf{e} \sim N(\mathbf{0}, \Sigma)$$

Our old friend "the covariance matrix"

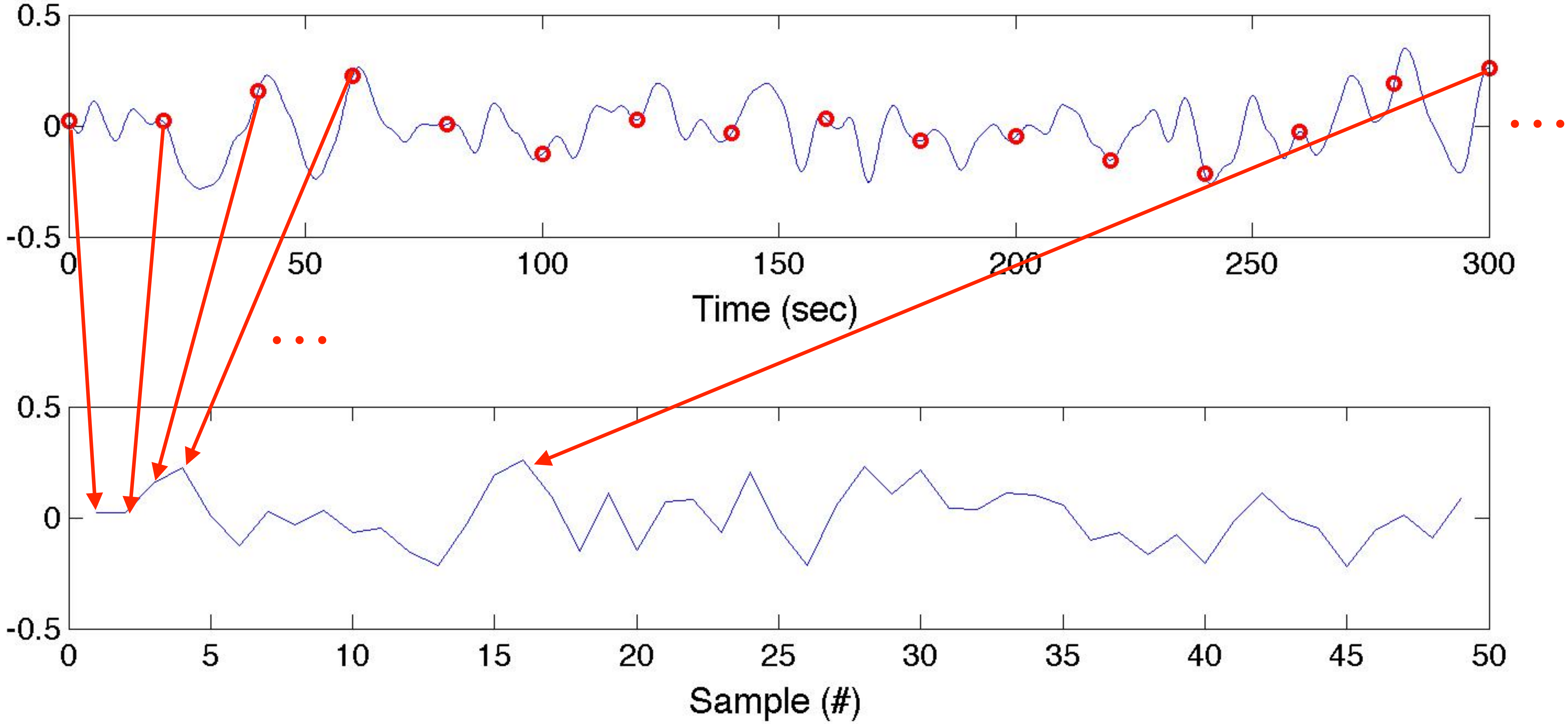
1st level fMRI data is not exchangeable

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



1st level fMRI data is not exchangeable

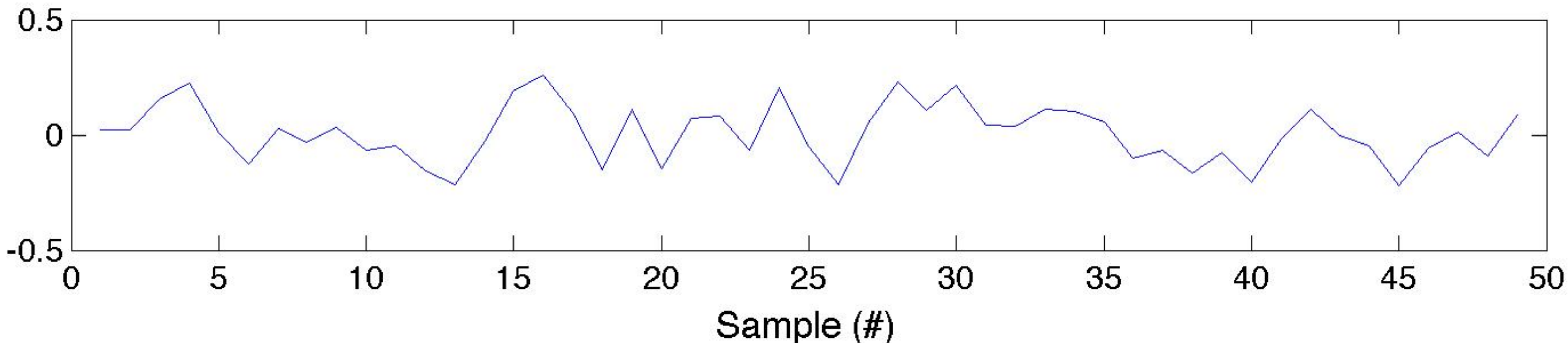
- One important component of noise in fMRI consists of physiological/neuronal events convolved by the HRF

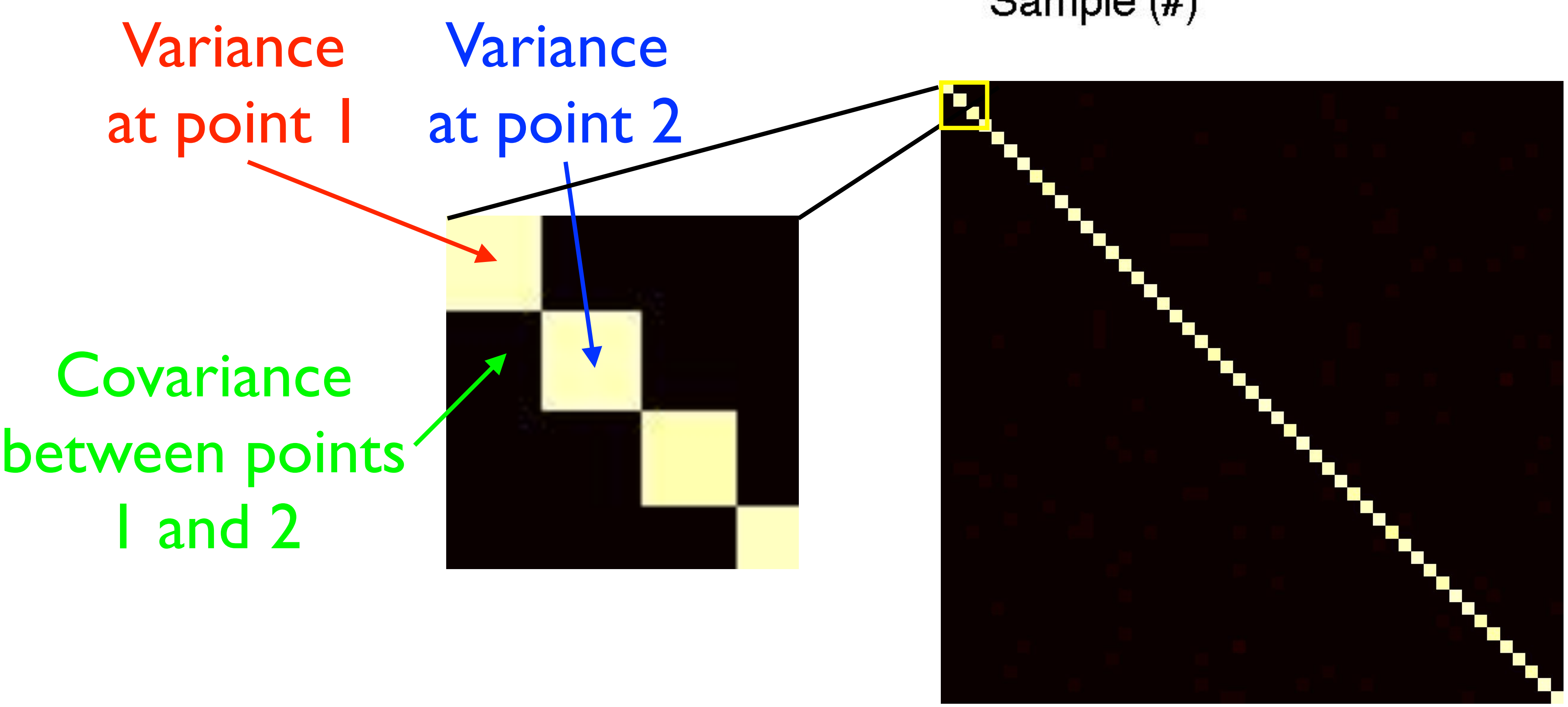


If we sample this every 20 seconds it no longer looks "smooth"

1st level fMRI data is not exchangeable

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

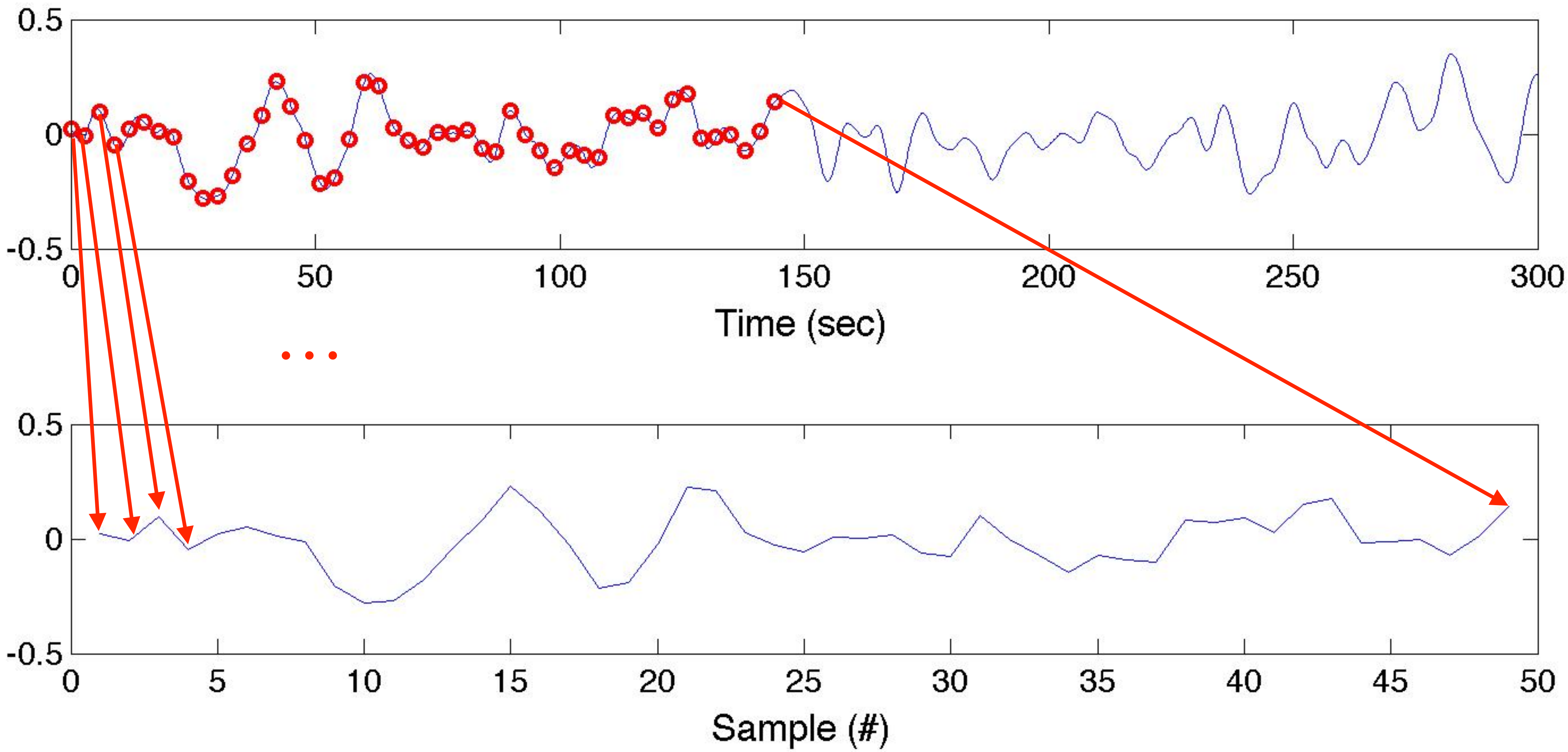




$$\mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

1st level fMRI data is not exchangeable

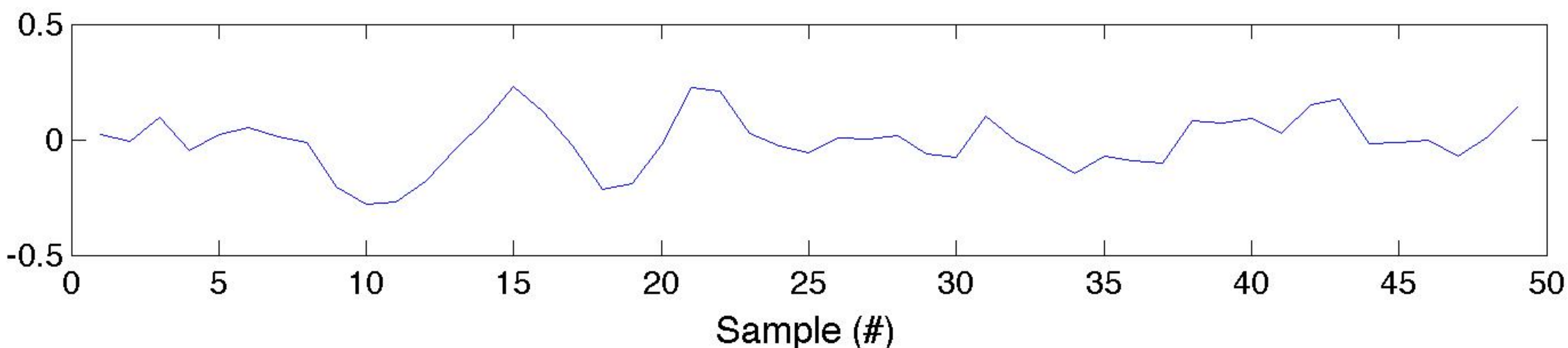
- One important component of noise in fMRI consists of physiological/neuronal events convolved by the HRF

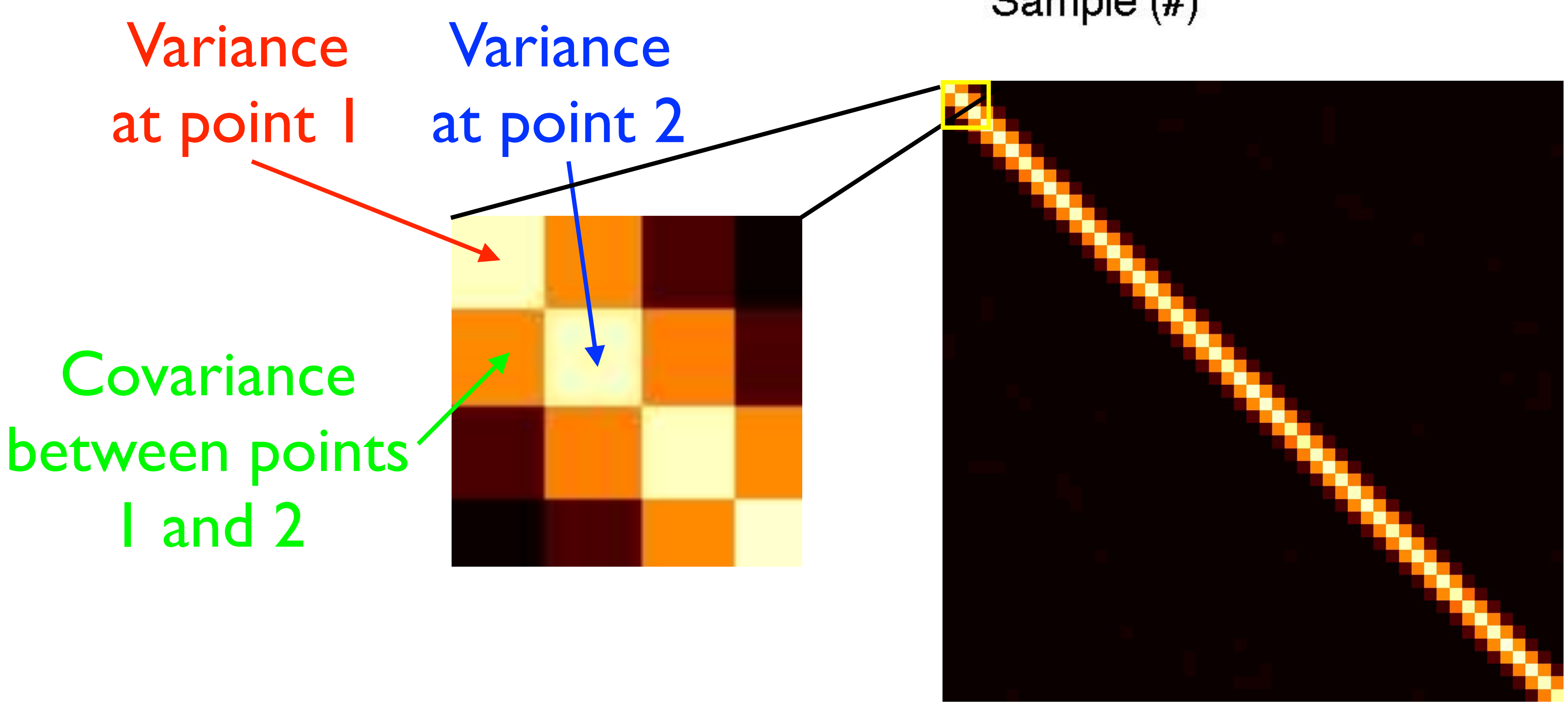


But that is not a realistic TR. What about every 3 seconds?

1st level fMRI data is not exchangeable

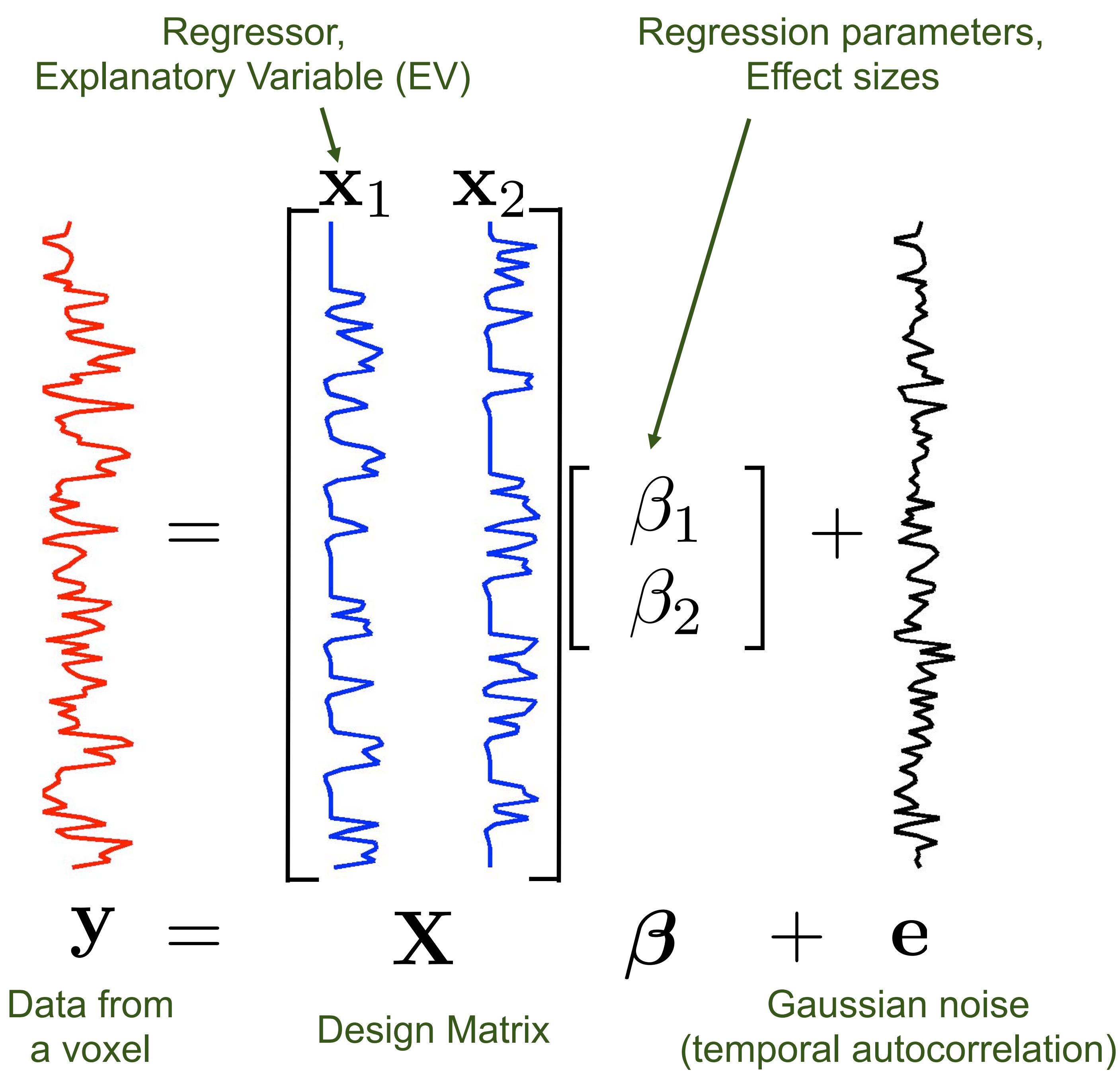
- One important component of noise in fMRI consists of physiological/neuronal events convolved by the HRF





1st level fMRI data is not exchangeable

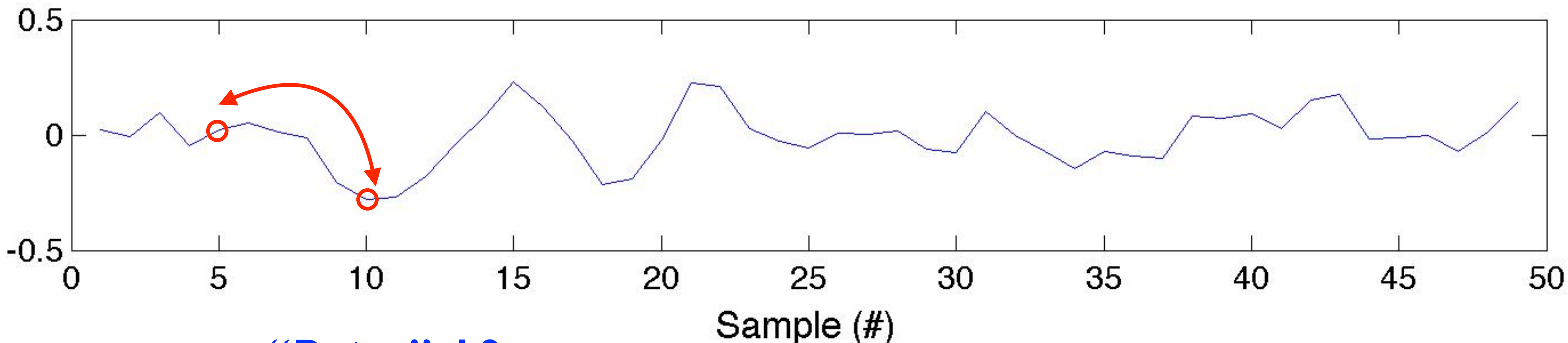
- Let us now return to our model again

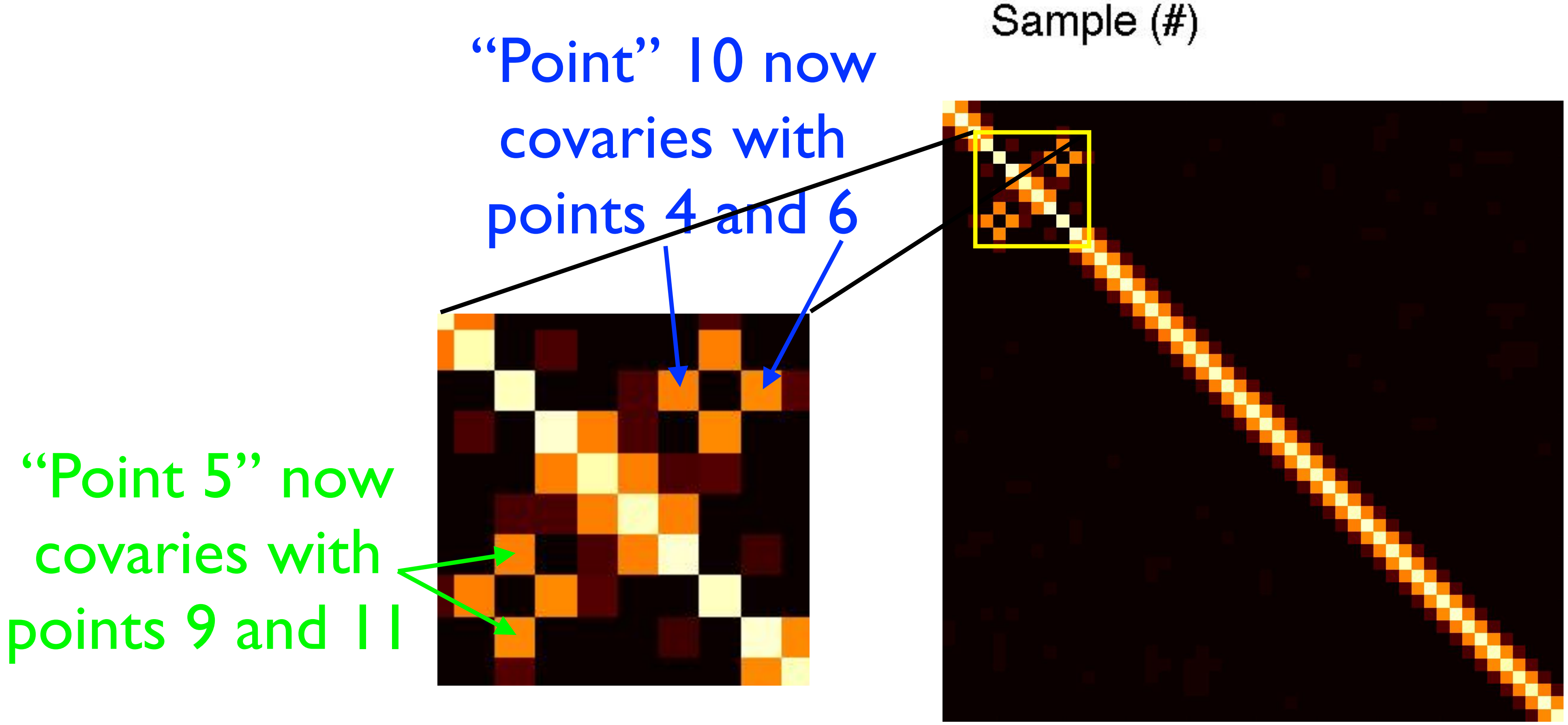


- The model consists of our regressors $\mathbf{X}$ and the noise model
- All permutations must result in "equivalent models"
- Let us now see what happens if we swap two data-points (points 5 and 10)

1st level fMRI data is not exchangeable

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

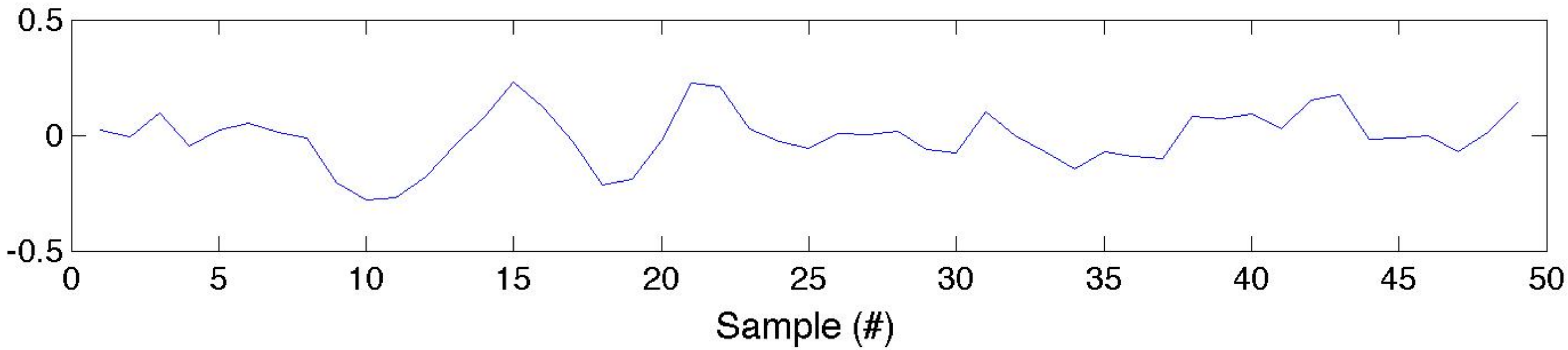




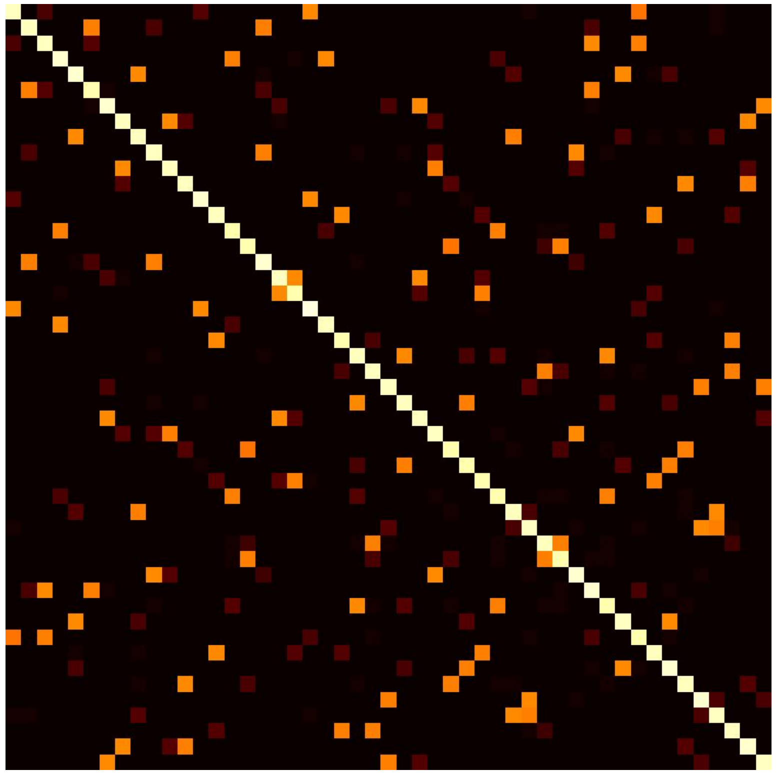
And the models are no longer equivalent

1st level fMRI data is not exchangeable

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



And for a random permutation …

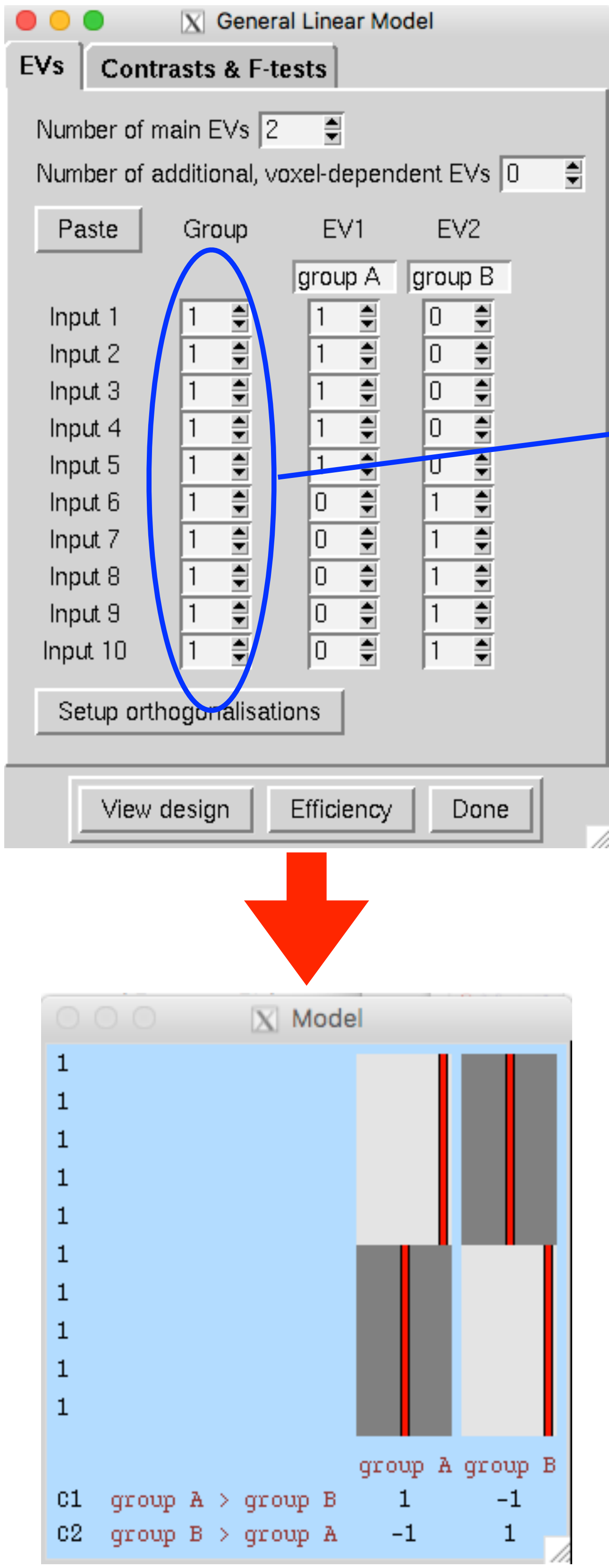


And the models are no longer equivalent

Back to exchangeability

- Data-points are not "exchangeable" if swapping them means that the noise covariance-matrix ends up looking different.
- Formally "The joint distribution of the data must be unchanged by the permutations under the null-hypothesis".
- If the noise covariance-matrix has non-zero off-diagonal elements (covariances) you need to beware.
- You typically never estimate or see the covariance-matrix. You need to "imagine it" and determine from that if there is a problem.

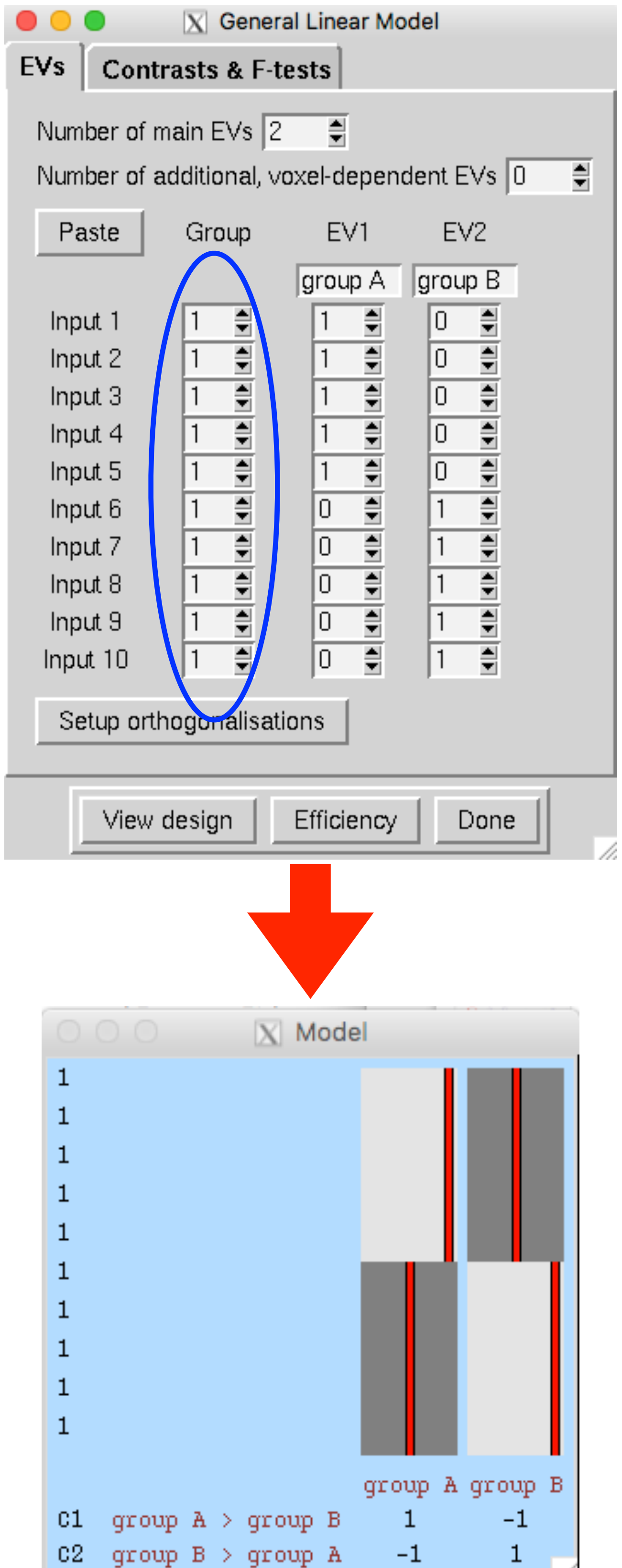
Examples of exchangeability: Two groups unpaired



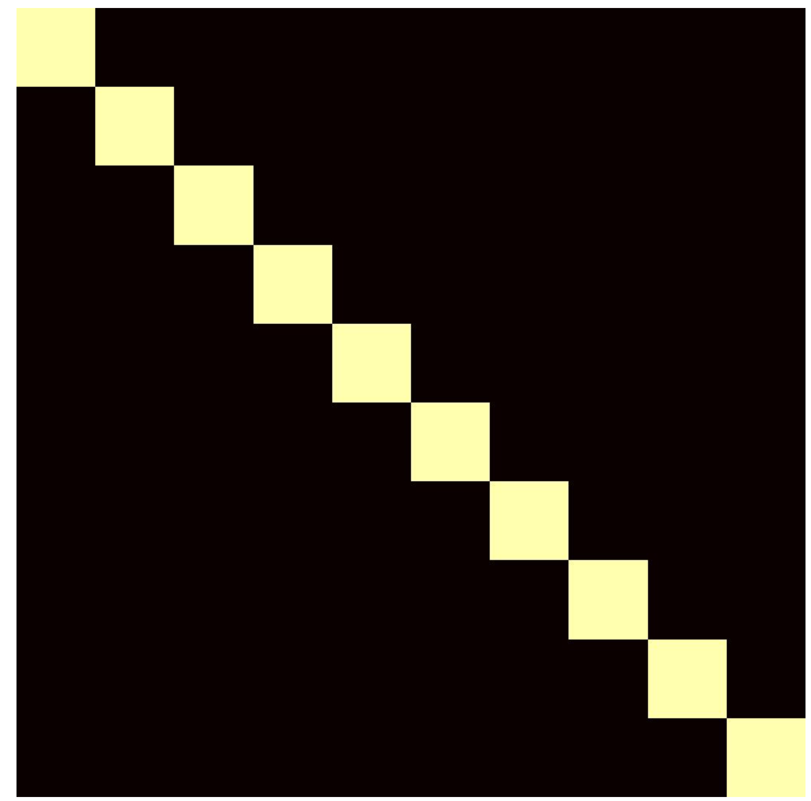
This is the "exchangeability group". Here all scans are in the same group, which means any scan can be exchanged for any other.

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

Examples of exchangeability: Two groups unpaired

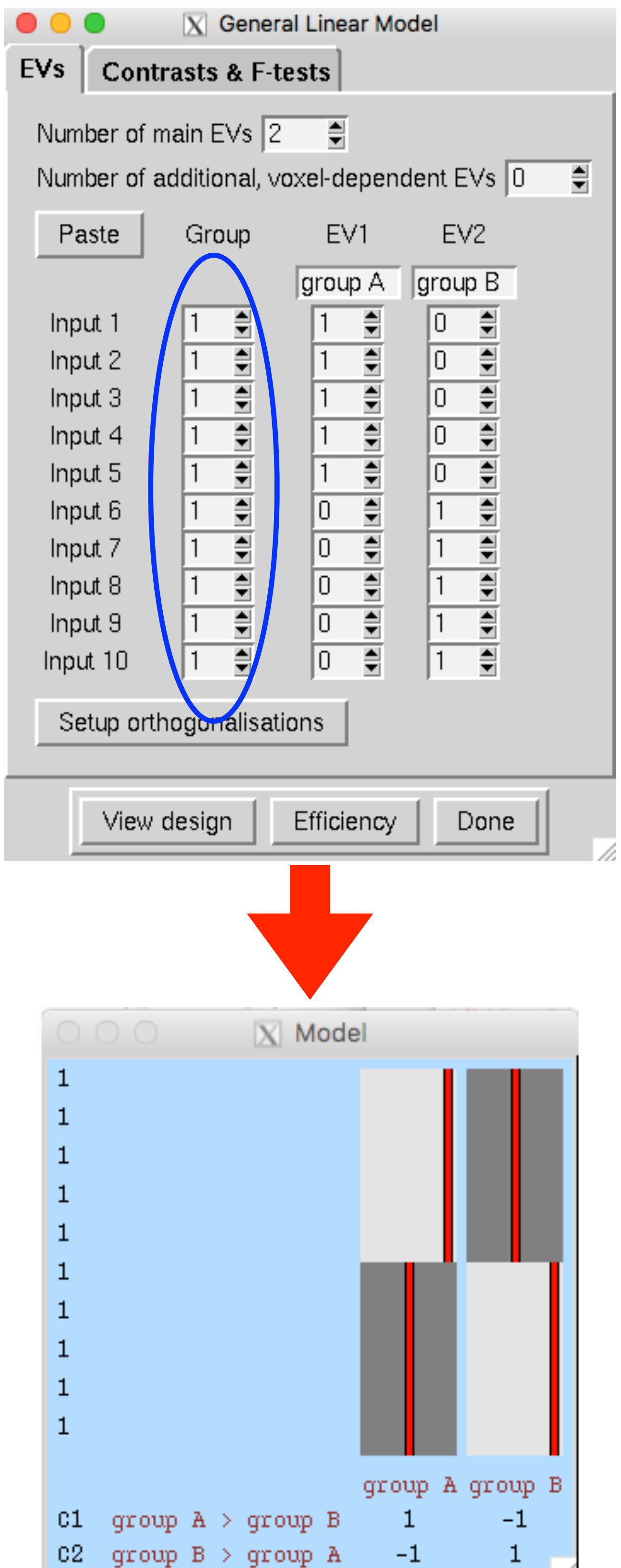


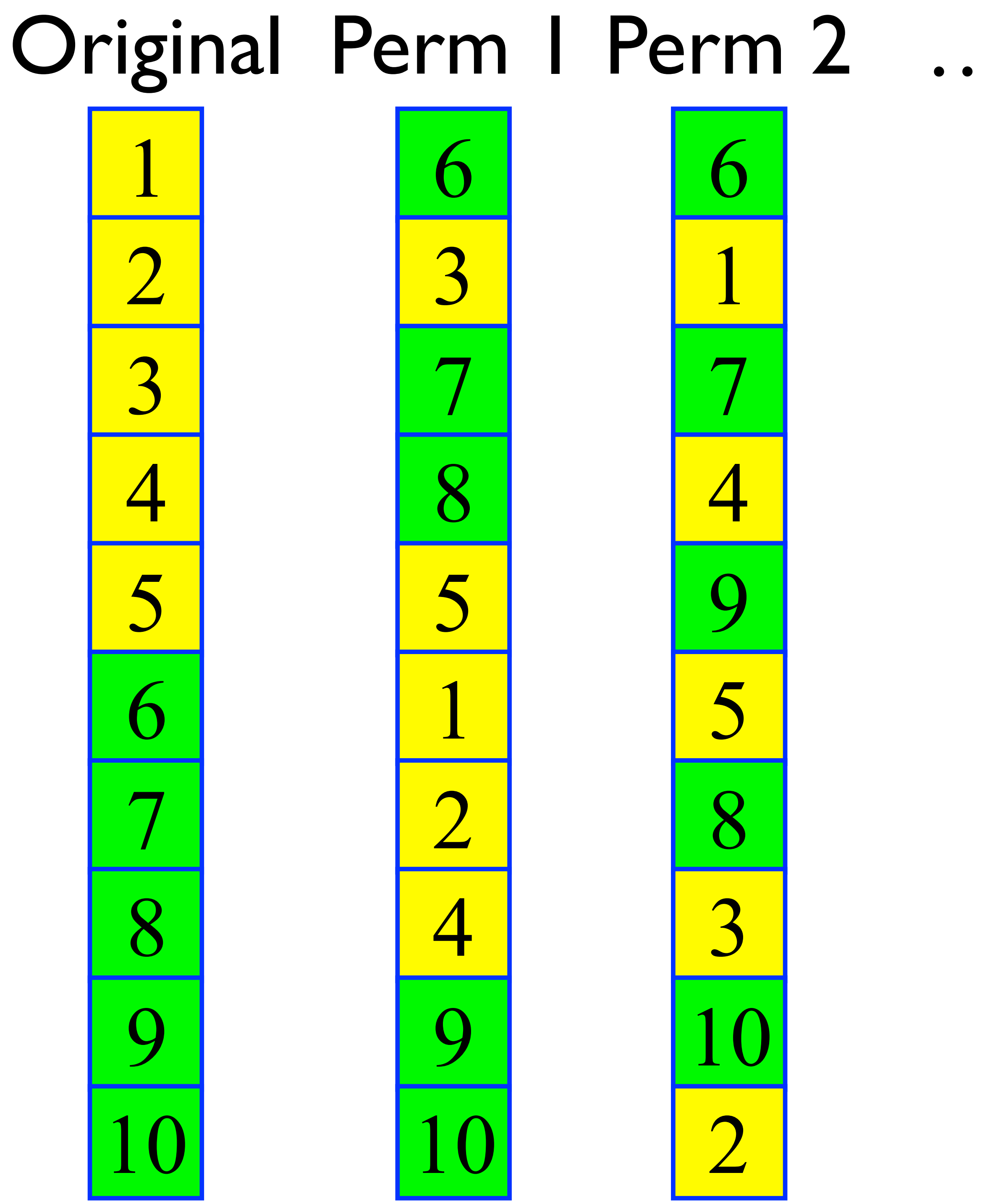
Assumed covariance matrix



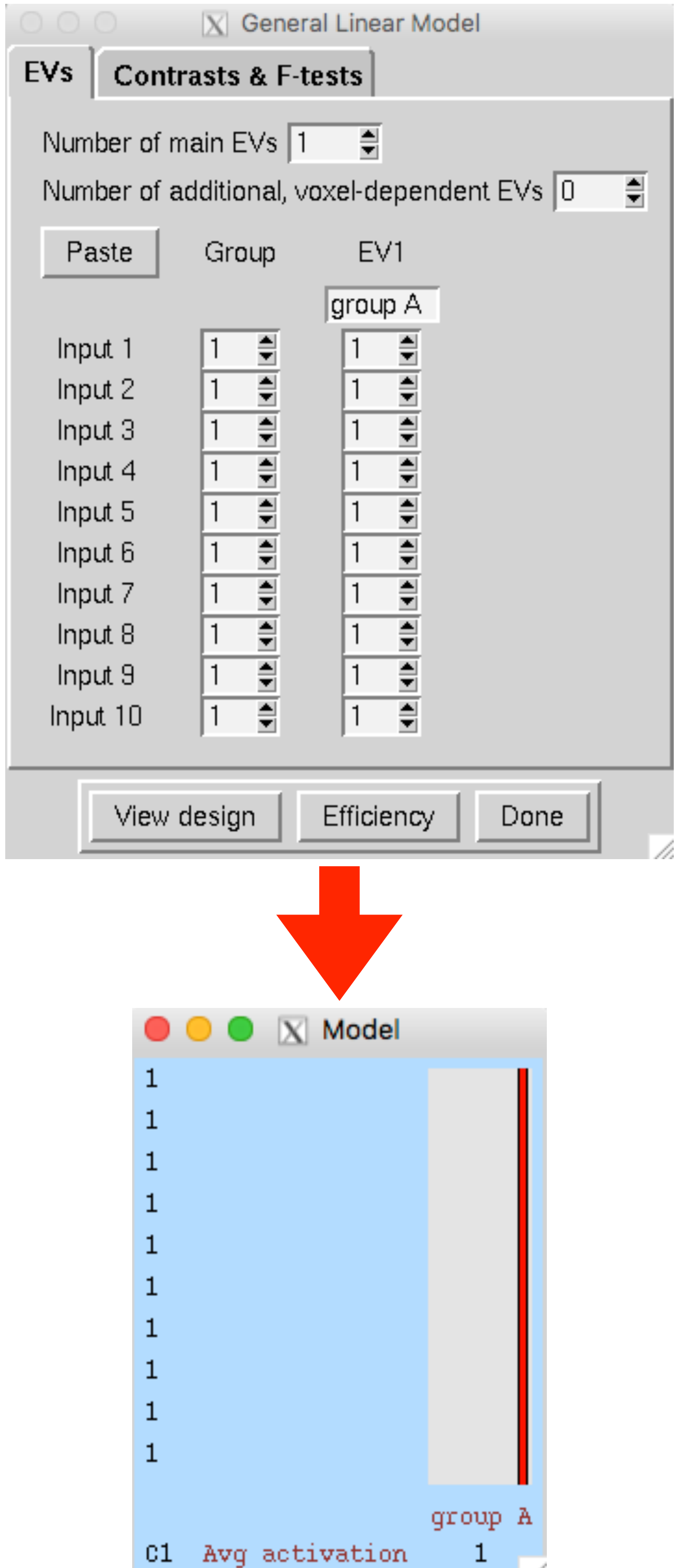
The implicit assumption here is that data from all subjects have the same uncertainty and are all independent

Examples of exchangeability: Two groups unpaired





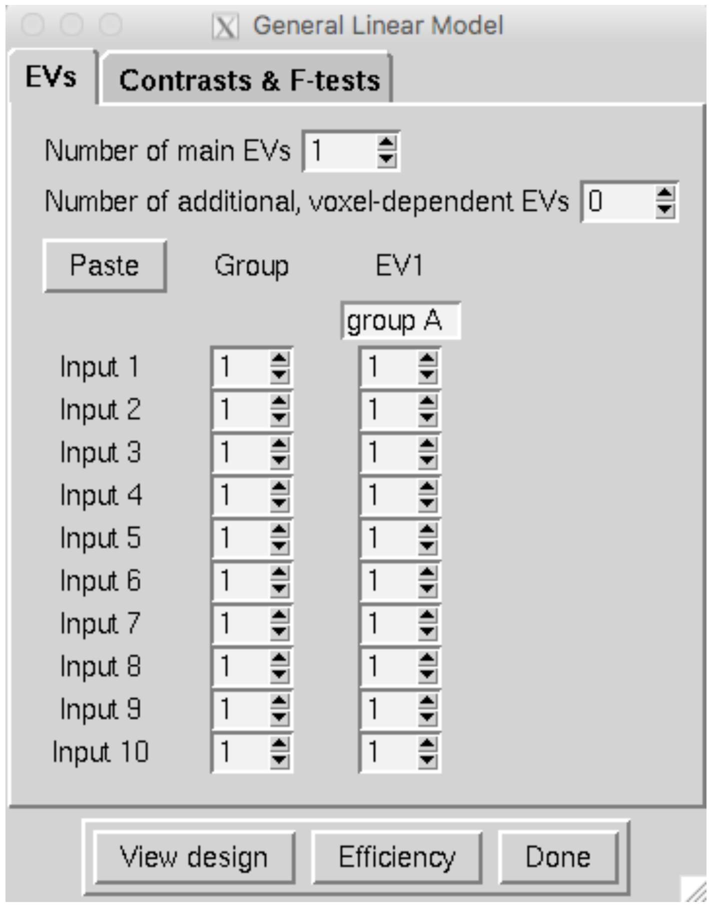
Examples of exchangeability: Single group average

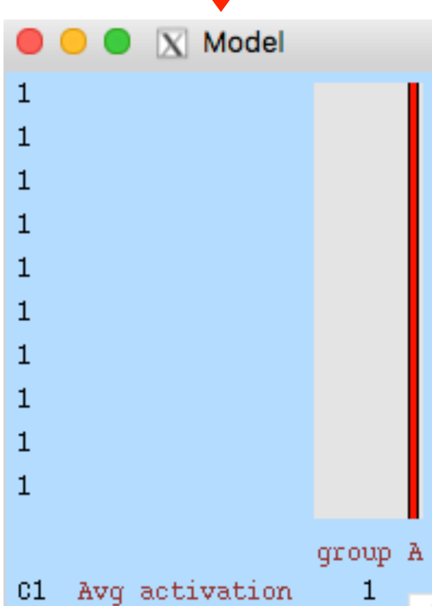


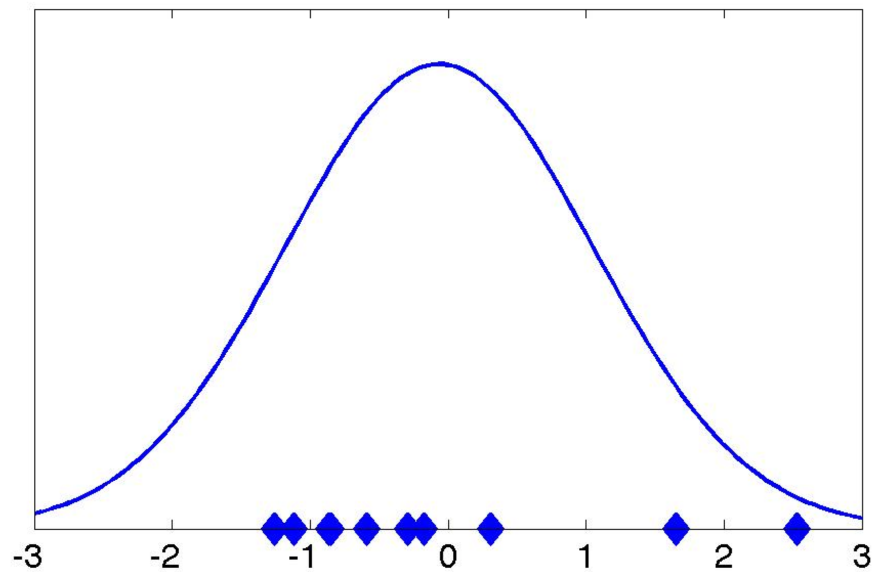
Here we model a single mean and want to know if that is different from zero

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

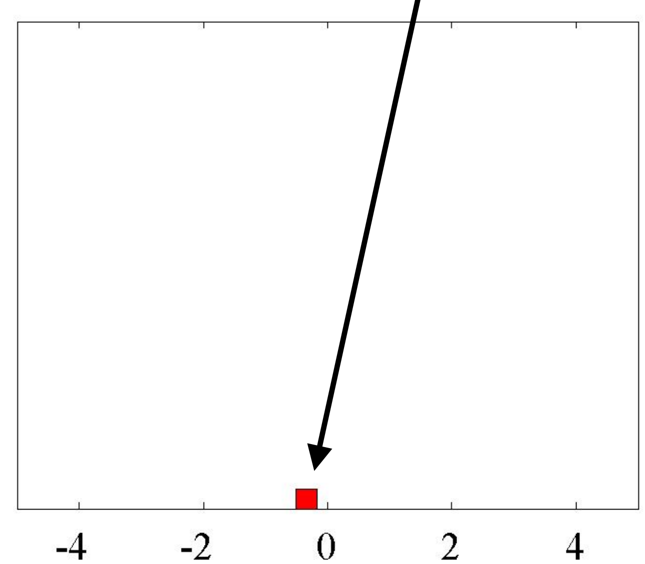
Examples of exchangeability: Single group average



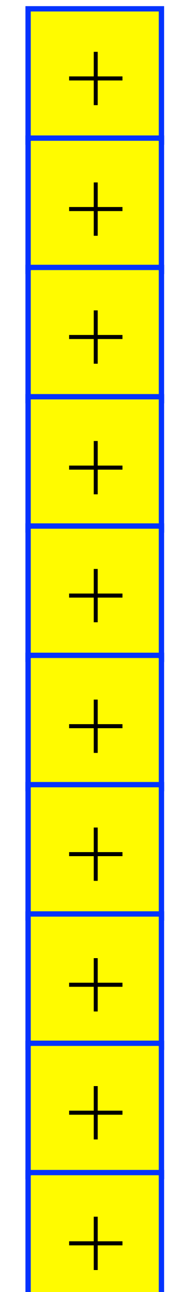




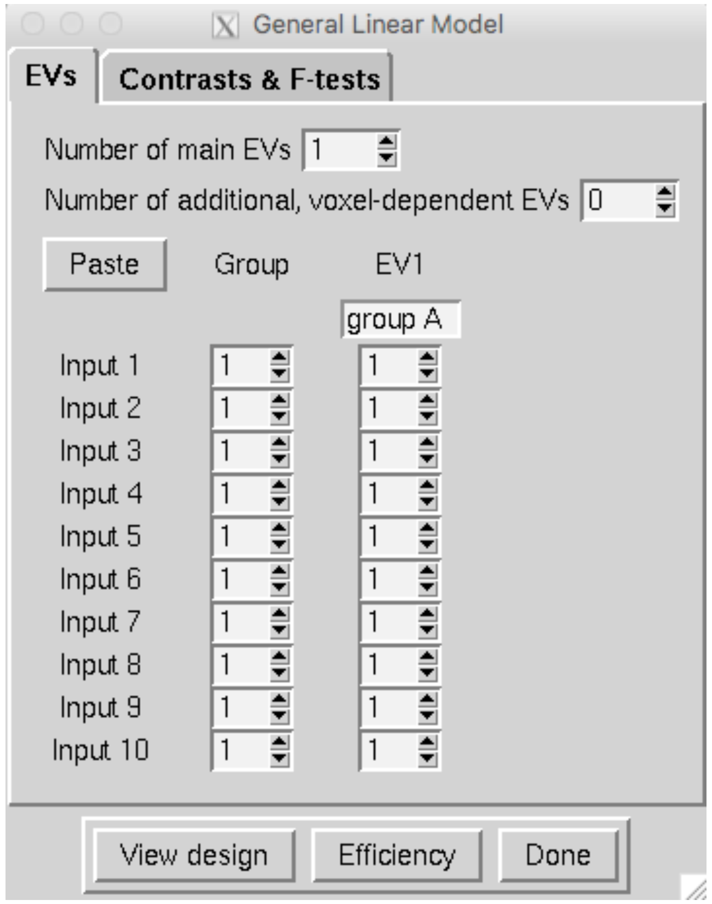
$$t = -0.17$$

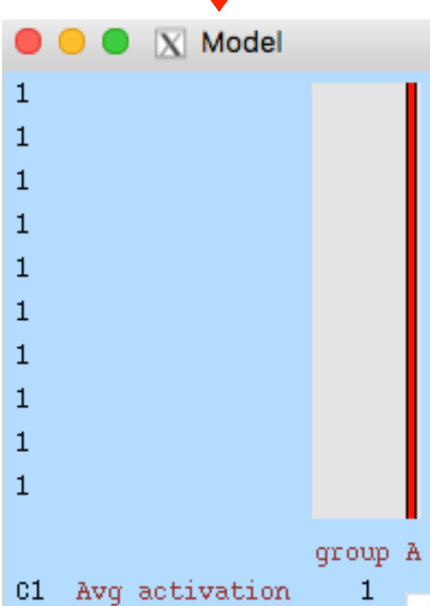


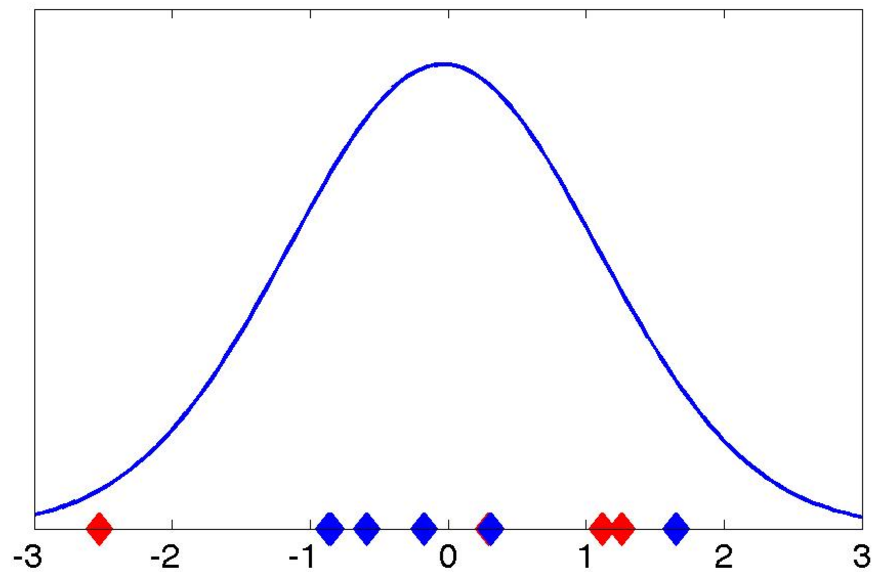
Original



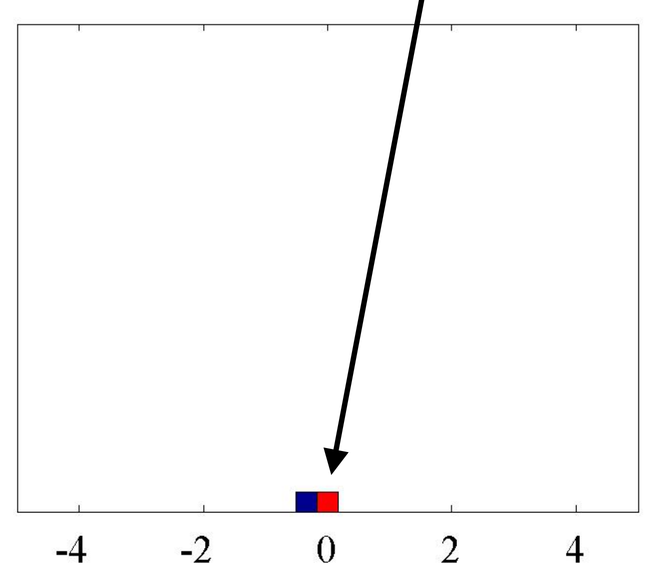
Examples of exchangeability: Single group average



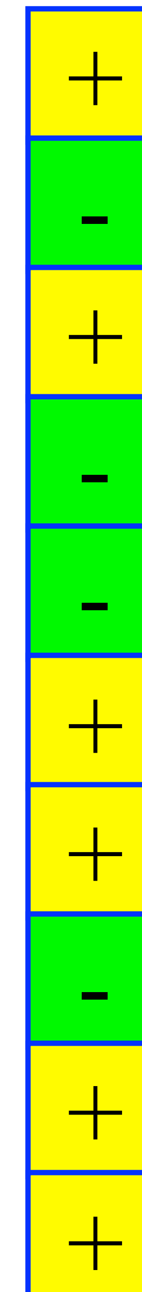




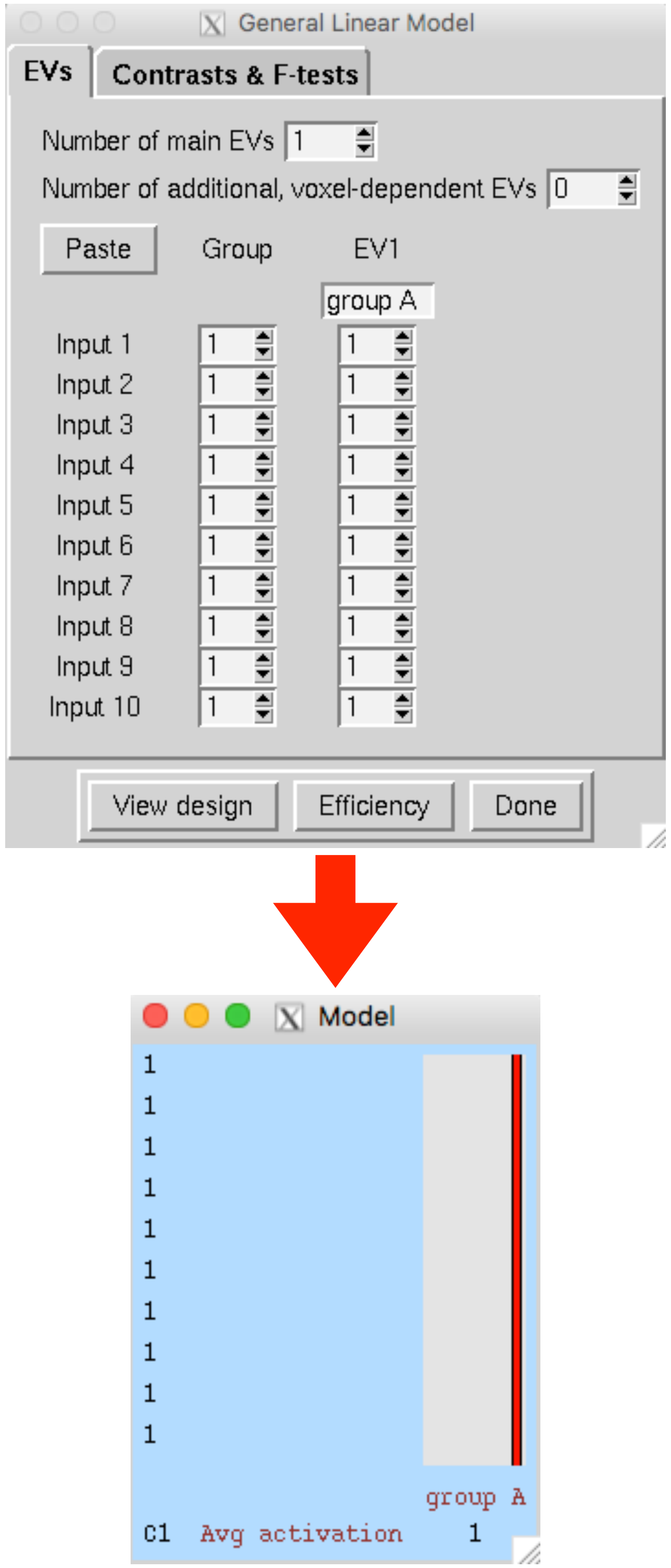
$$t = -0.09$$

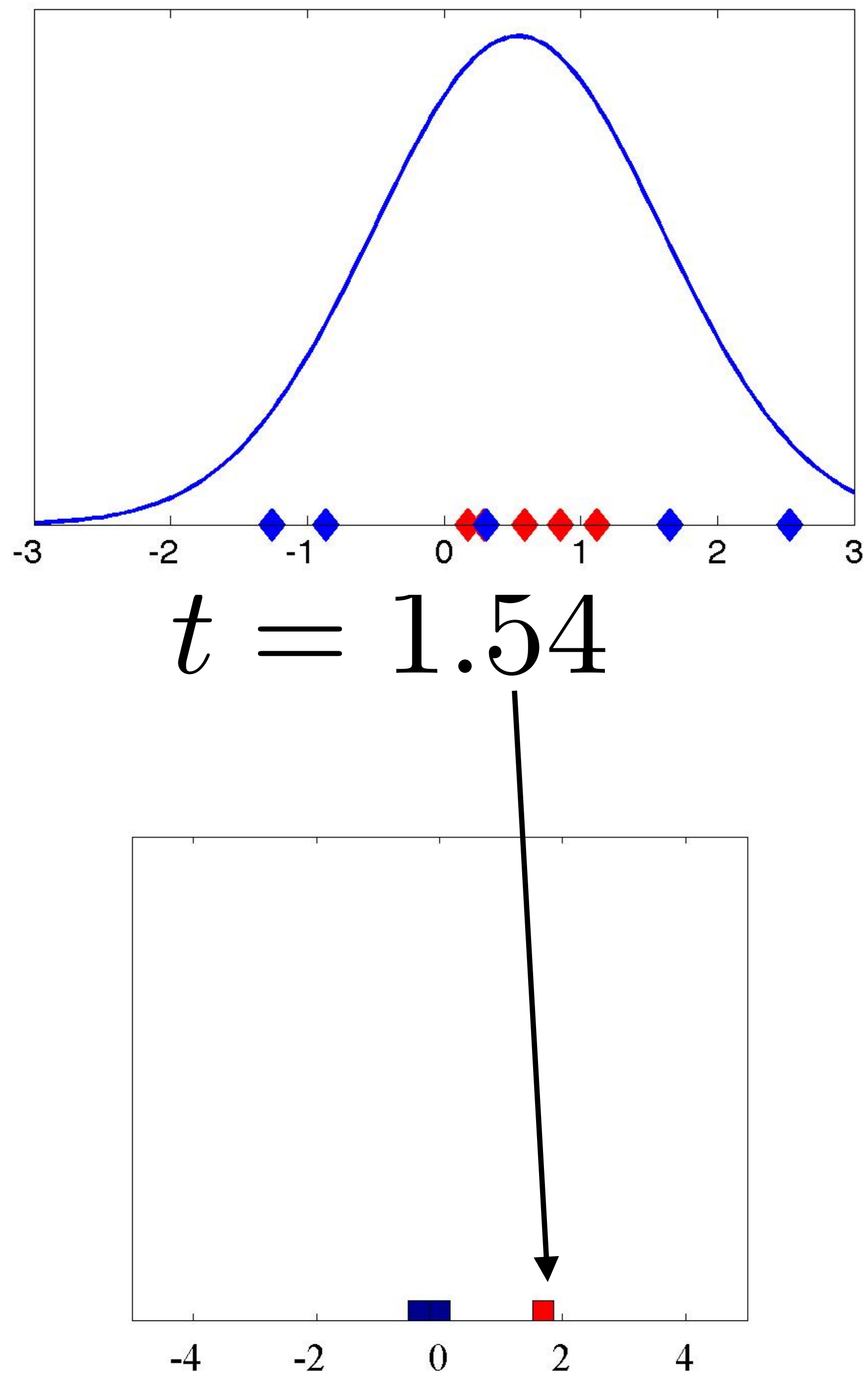


First flip

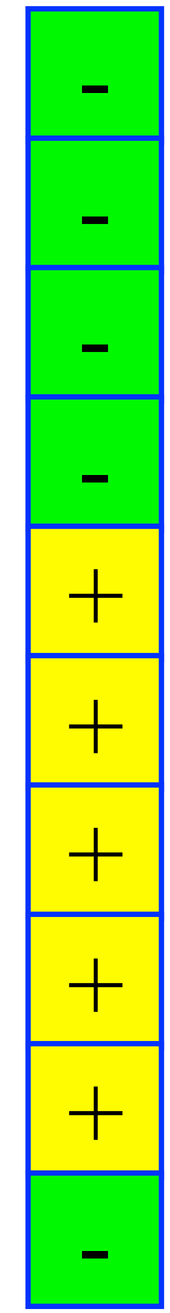


Examples of exchangeability: Single group average

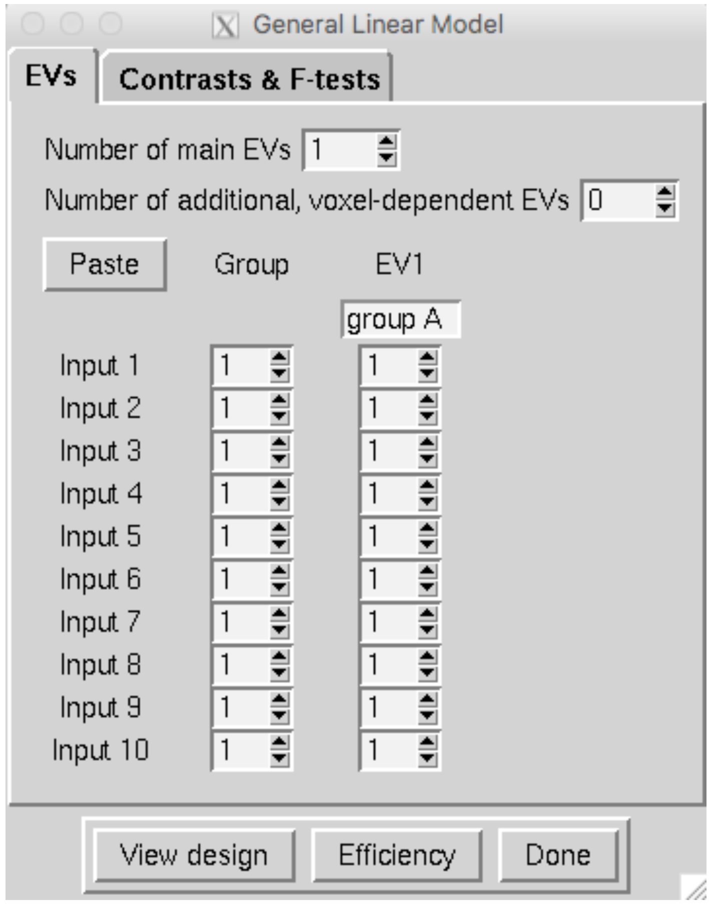


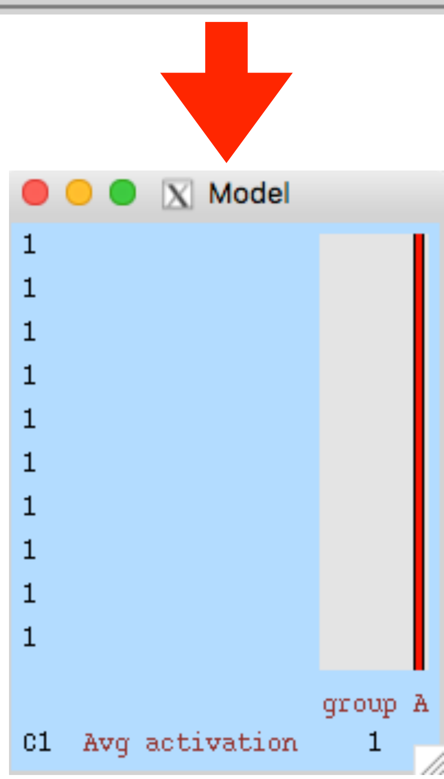


Second flip

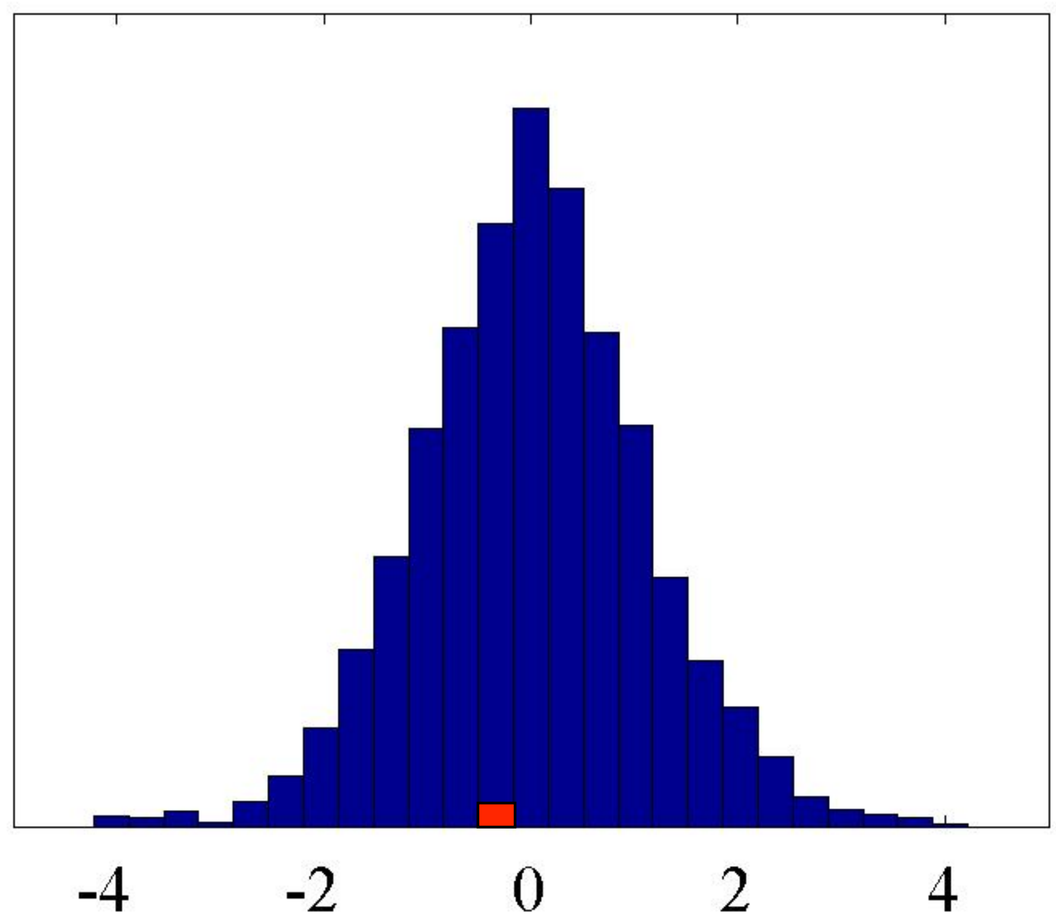


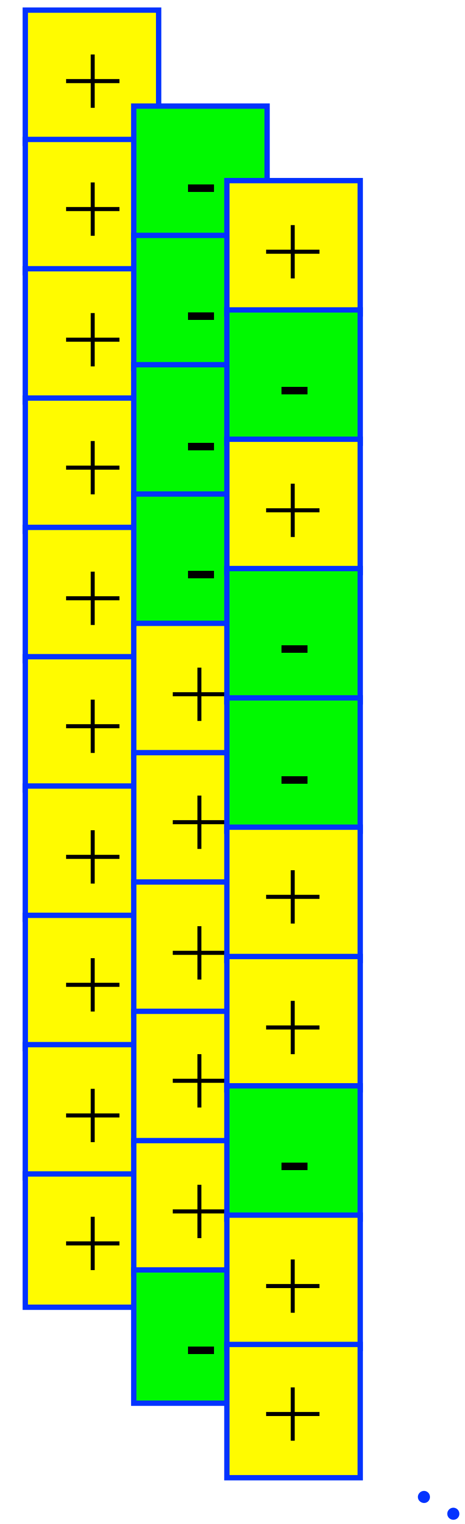
Examples of exchangeability: Single group average



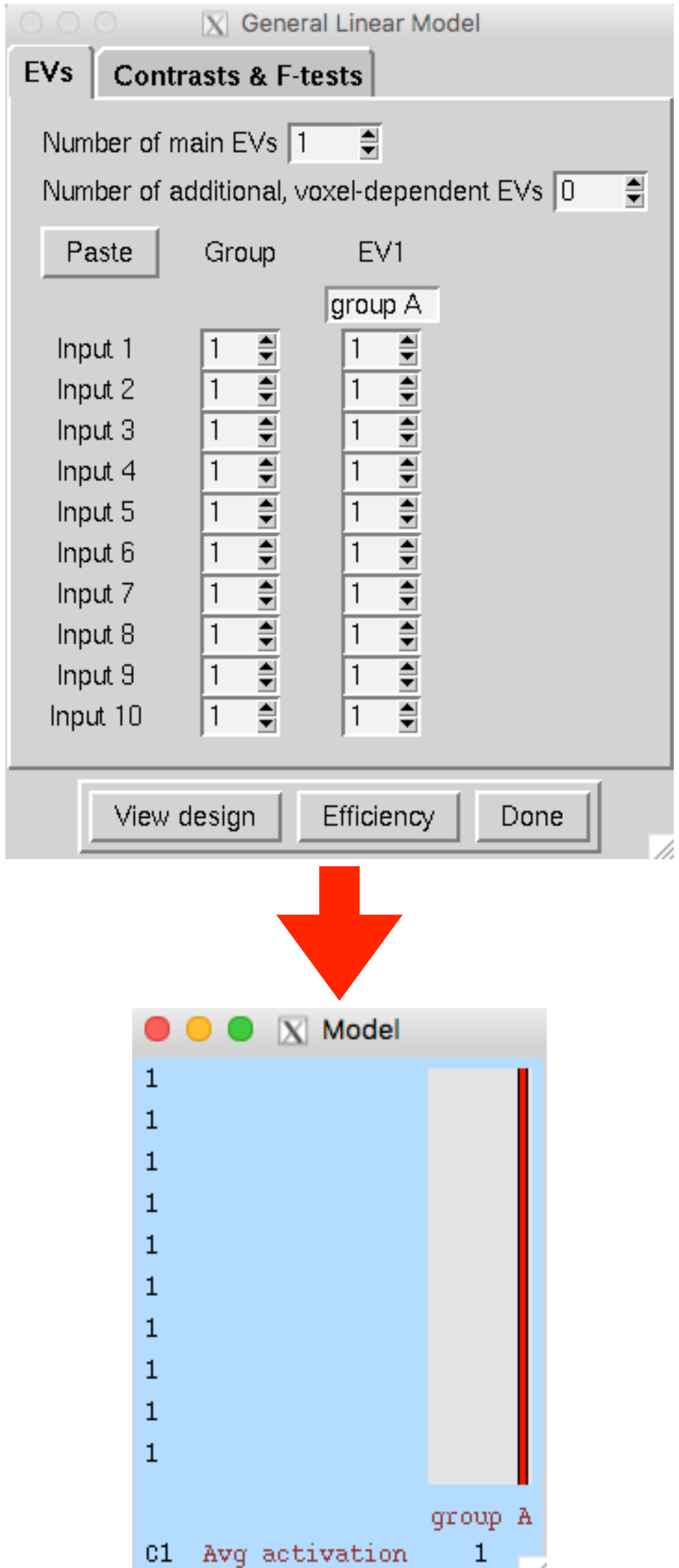


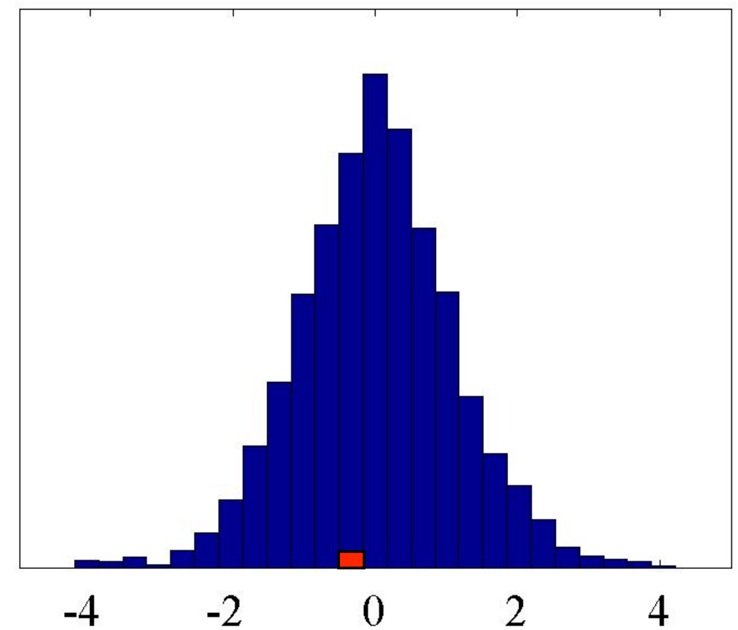
Etc ...





Examples of exchangeability: Single group average

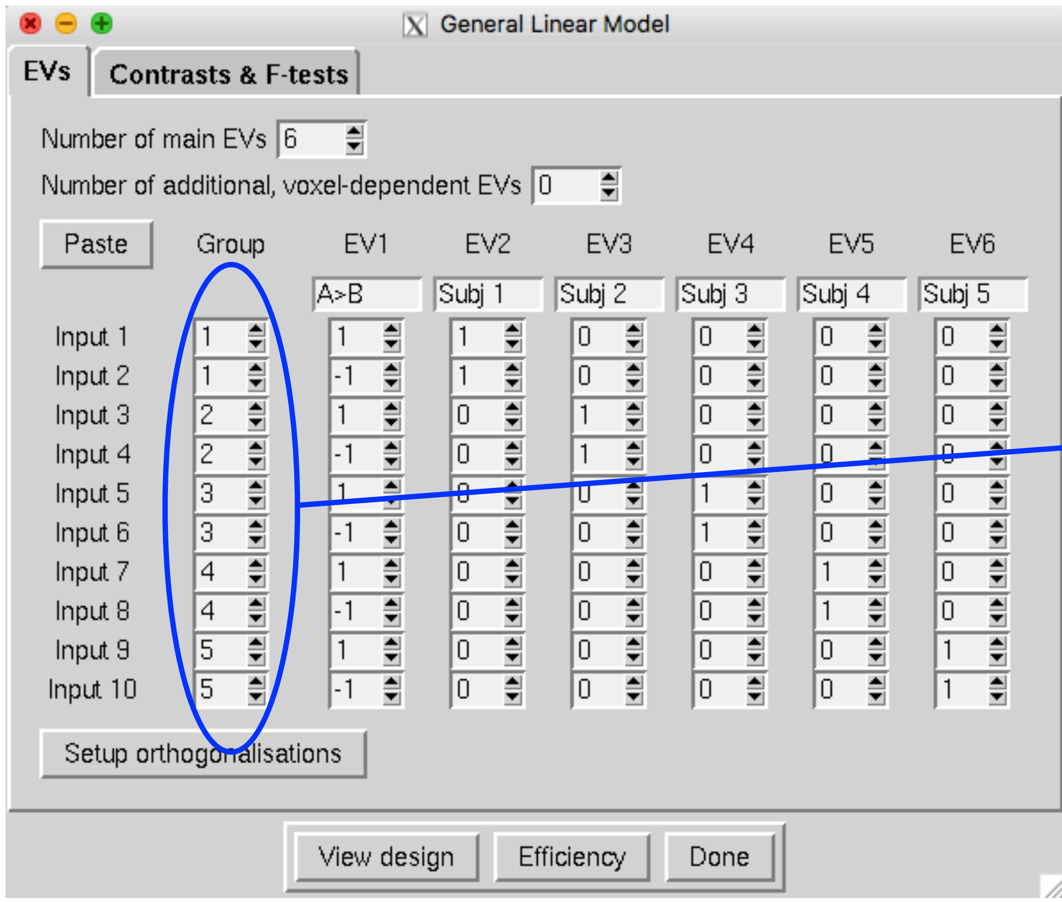




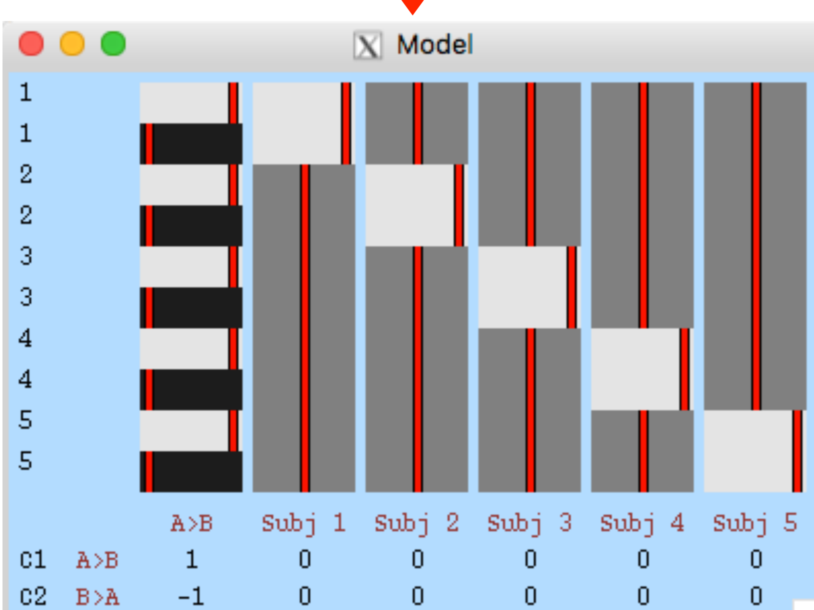
And the assumptions are:

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

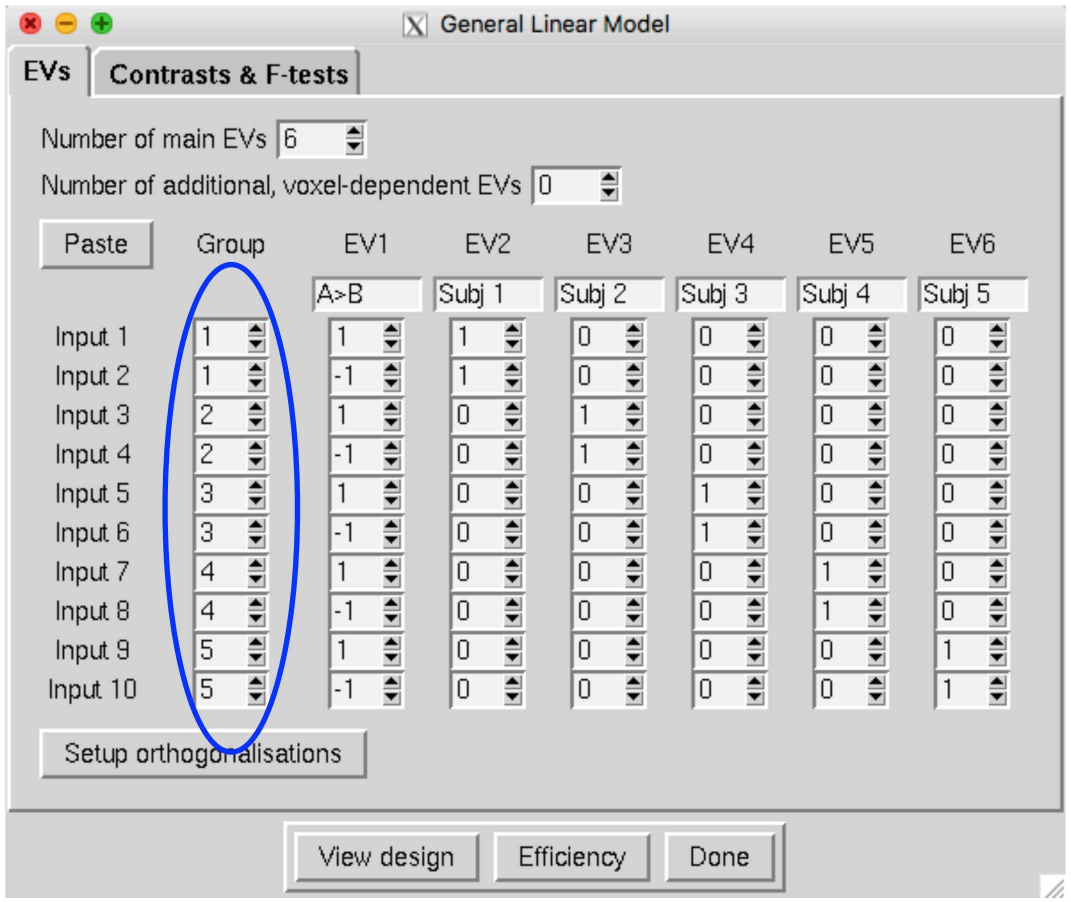
Examples of exchangeability: Two groups paired

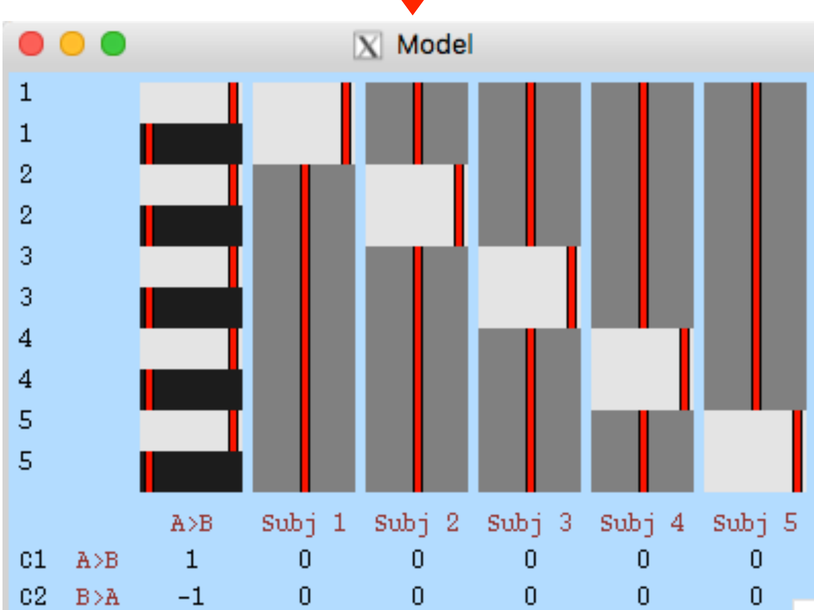


Here we can only exchange scans within each subject. I.e. Input 1 for Input 2, Input 3 for Input 4 etc



Examples of exchangeability: Two groups paired

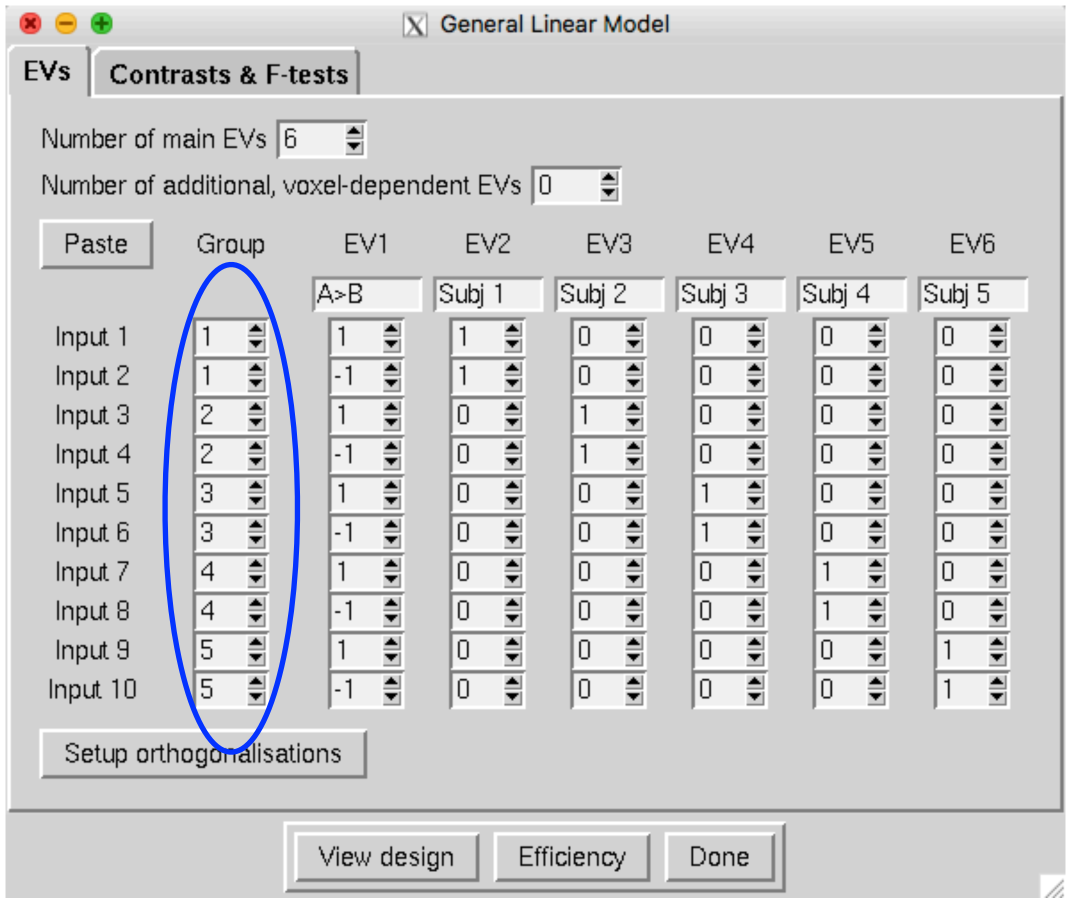


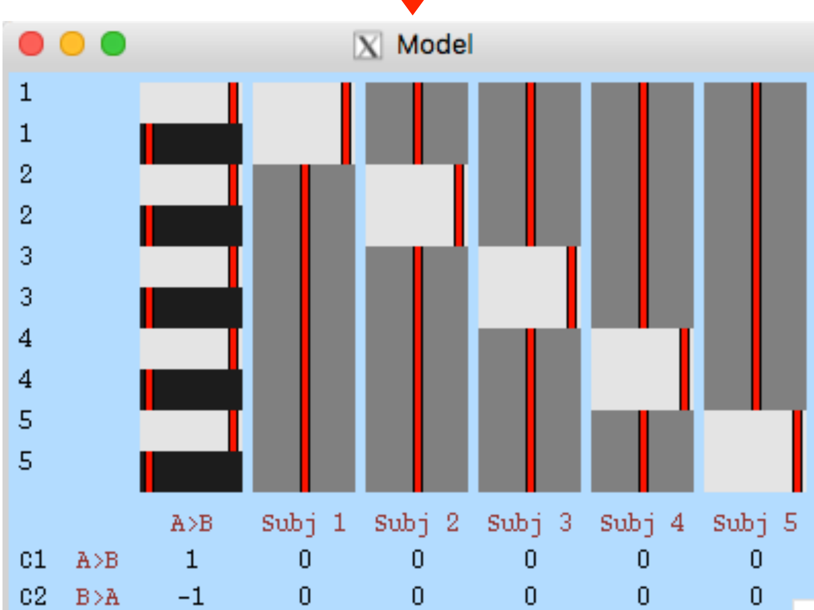


Assumed covariance matrix

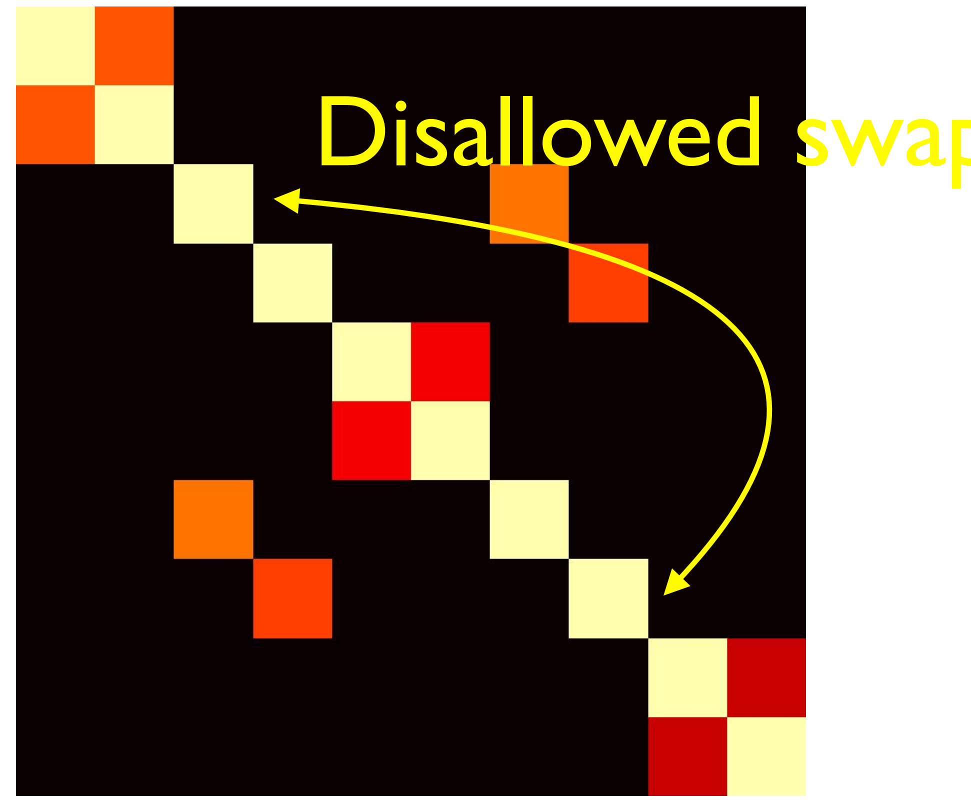
The implicit assumption here is that data from all subjects have the same uncertainty and that there is no dependence between subjects

Examples of exchangeability: Two groups paired



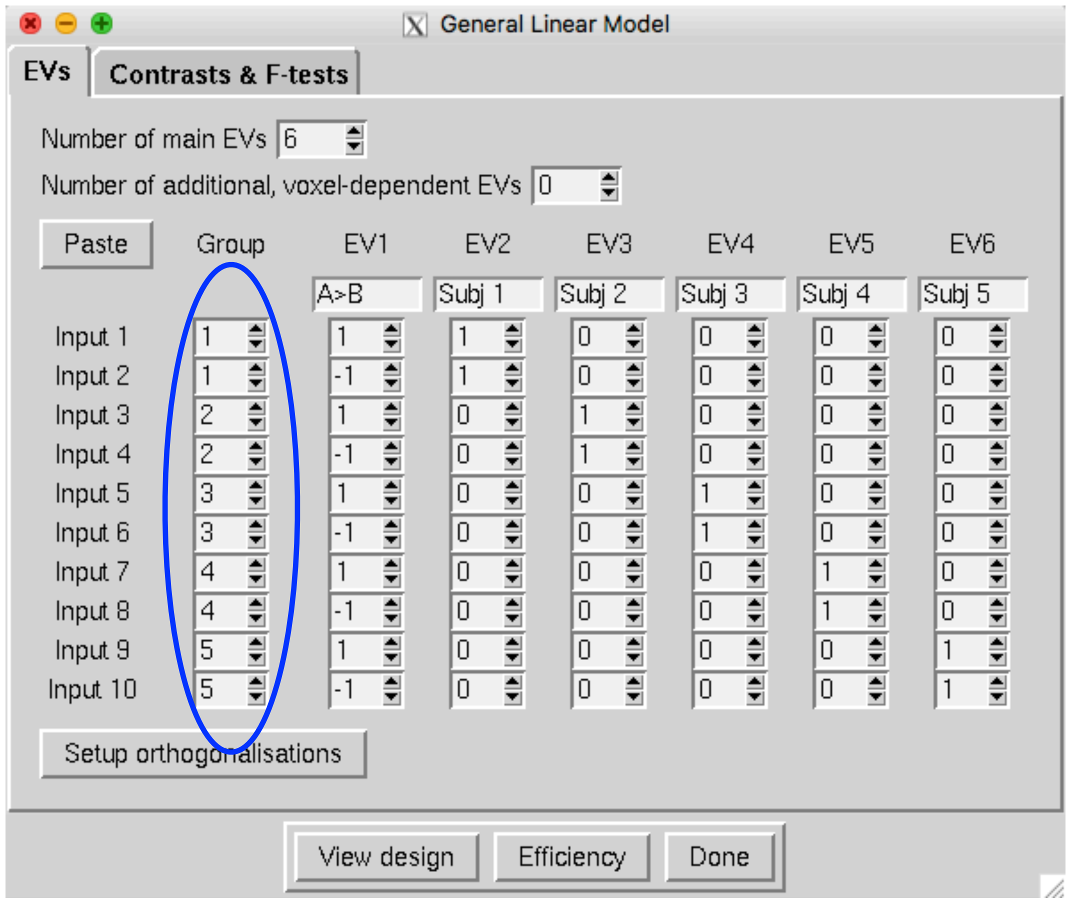


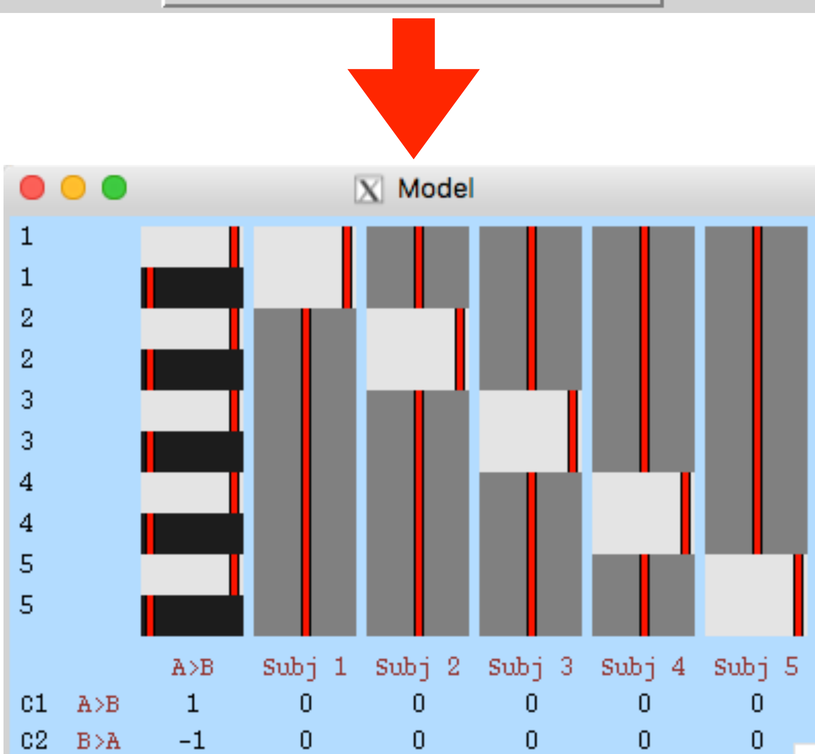
Assumed covariance matrix

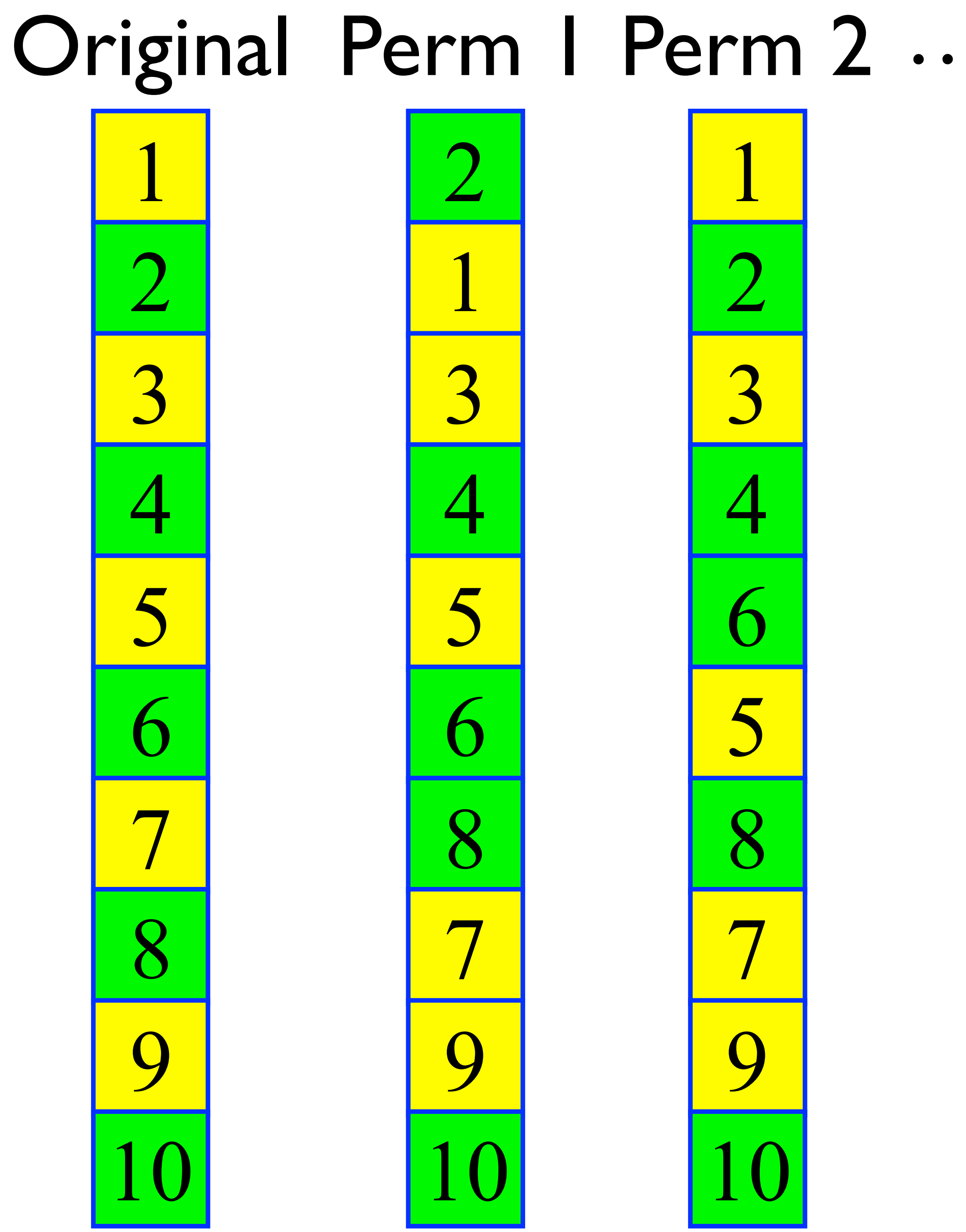


The implicit assumption here is that data from all subjects have the same uncertainty and that there is no dependence between subjects

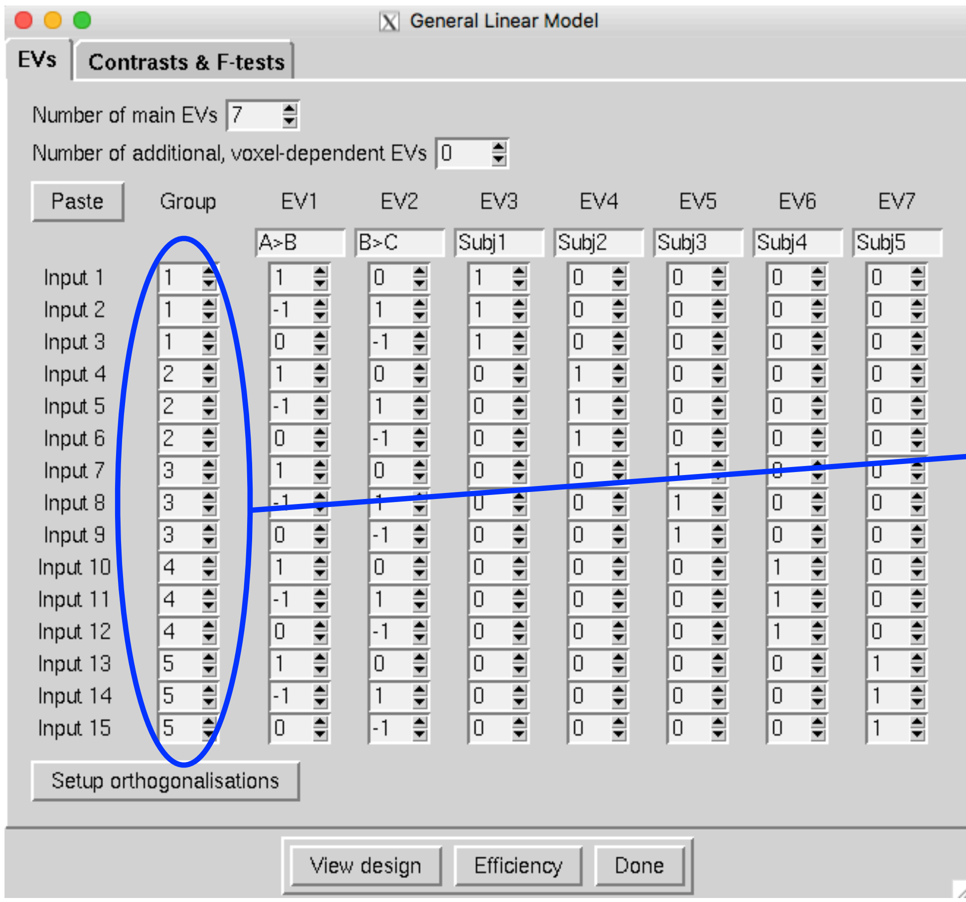
Examples of exchangeability: Two groups paired



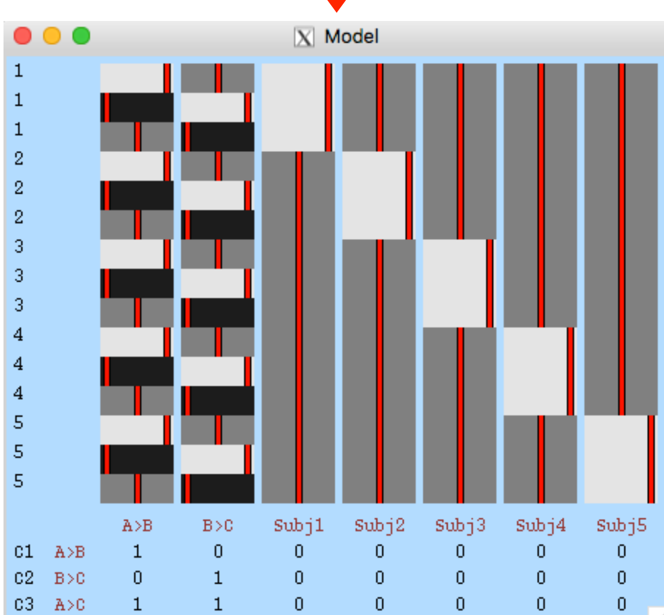




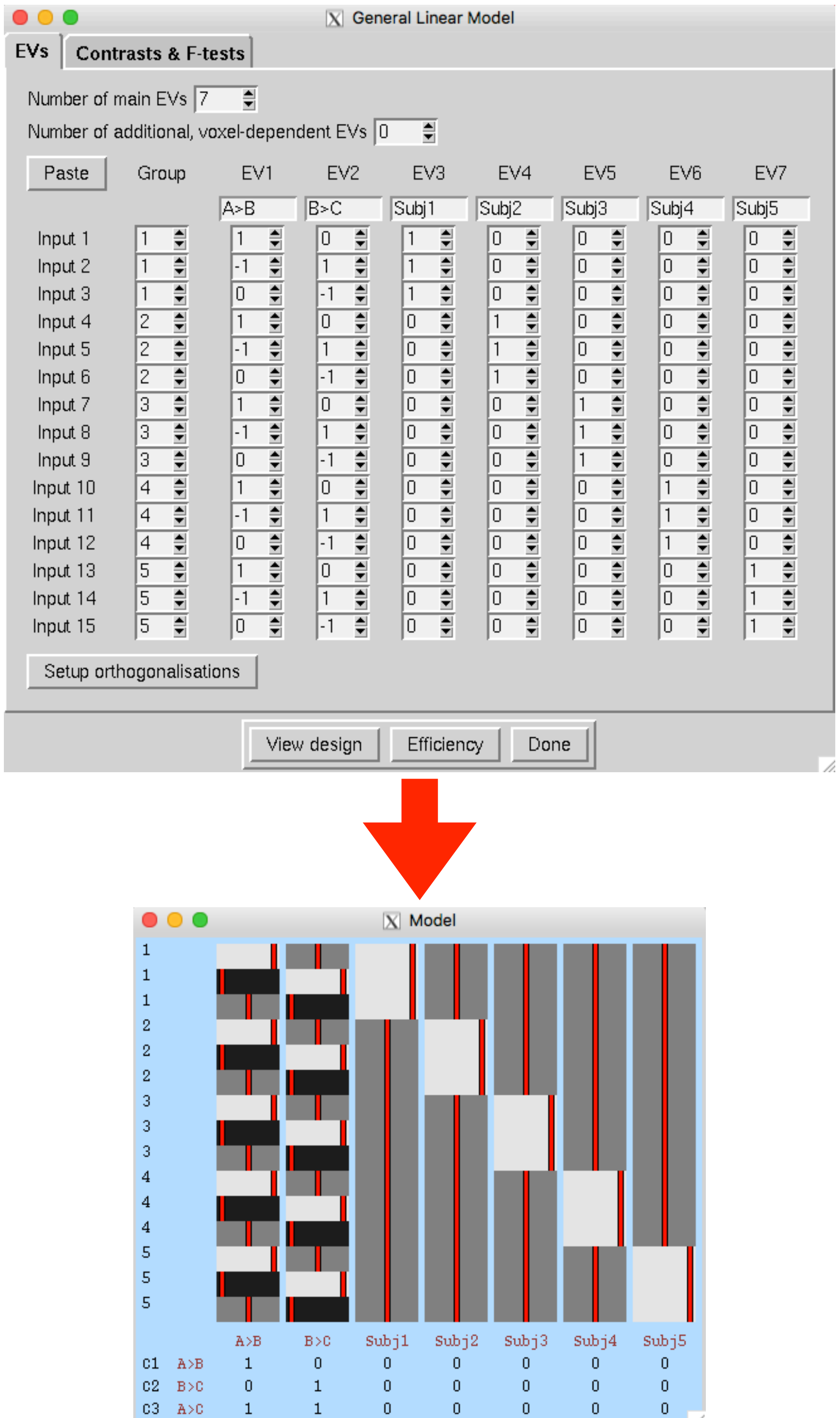
Examples of exchangeability: blocked ANOVA



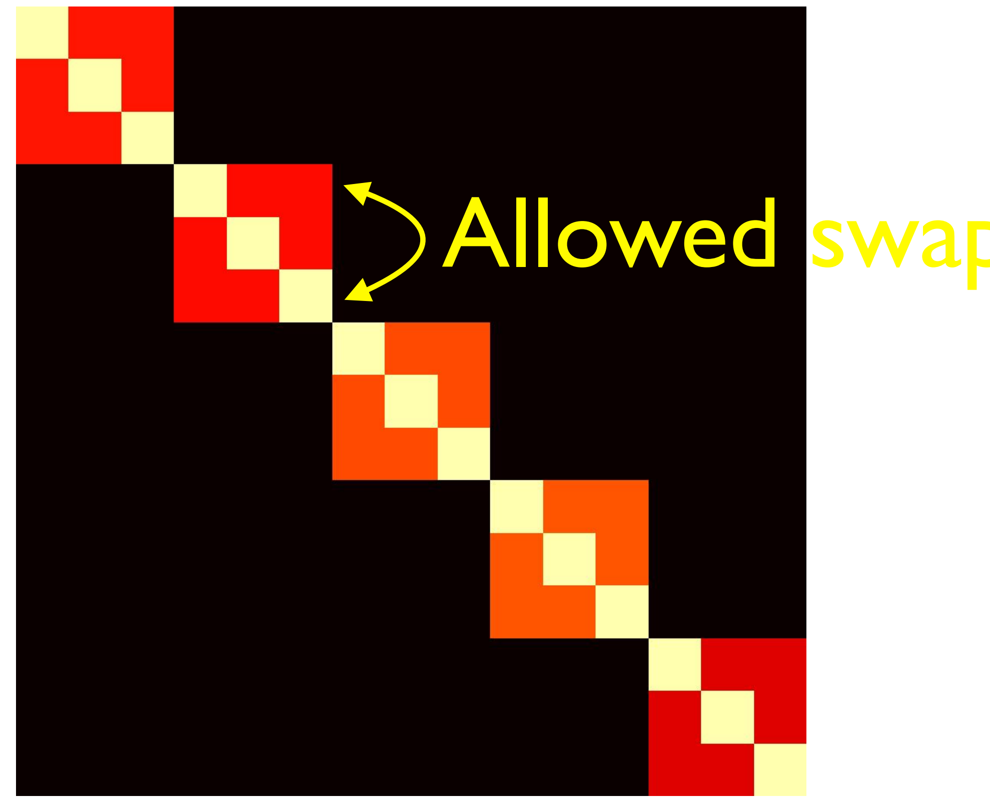
Same as previous: We can only swap labels within each subject



Examples of exchangeability: blocked ANOVA

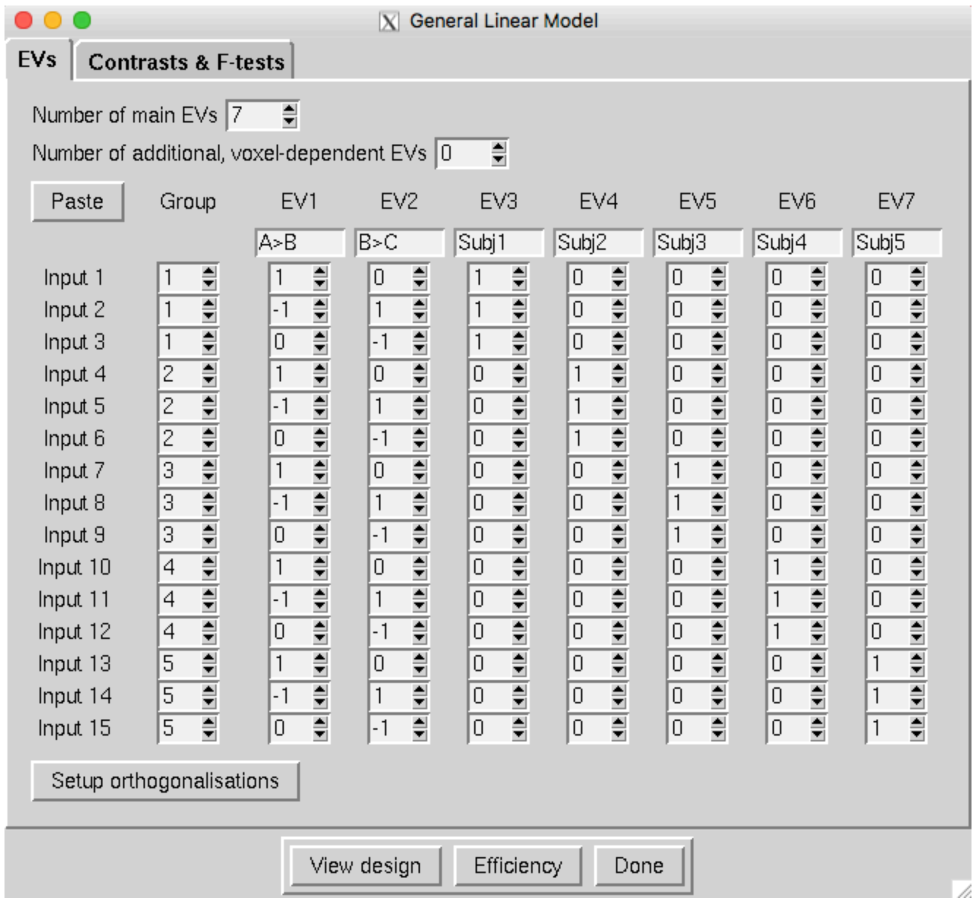


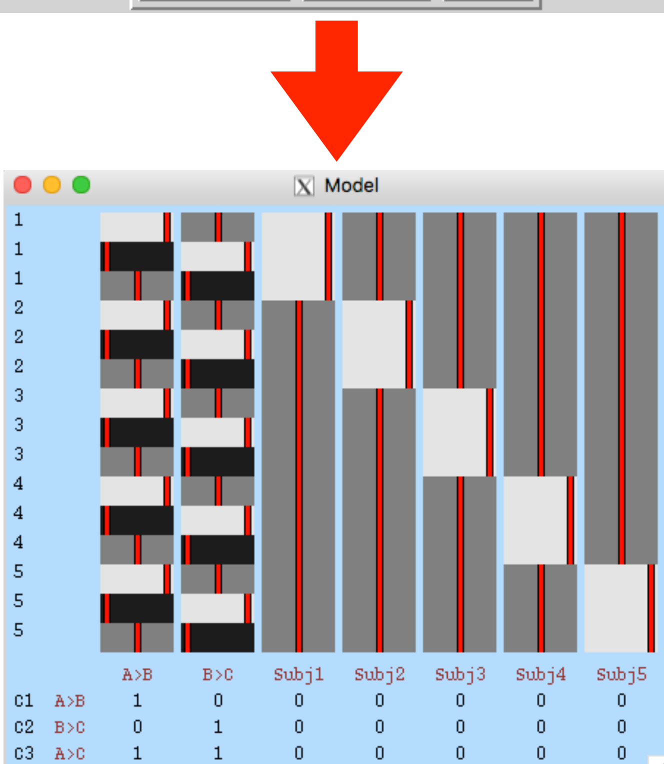
Assumed covariance matrix



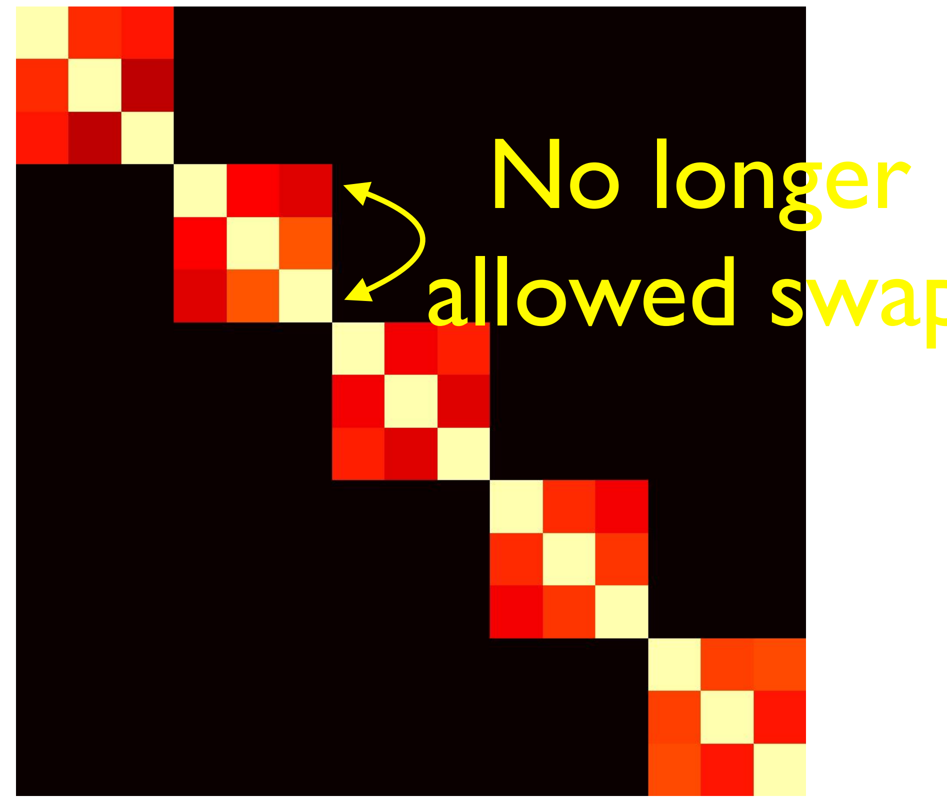
Assumptions: All subjects from the same "population", no dependence between subjects and "compound symmetry" within subjects

Examples of exchangeability: blocked ANOVA



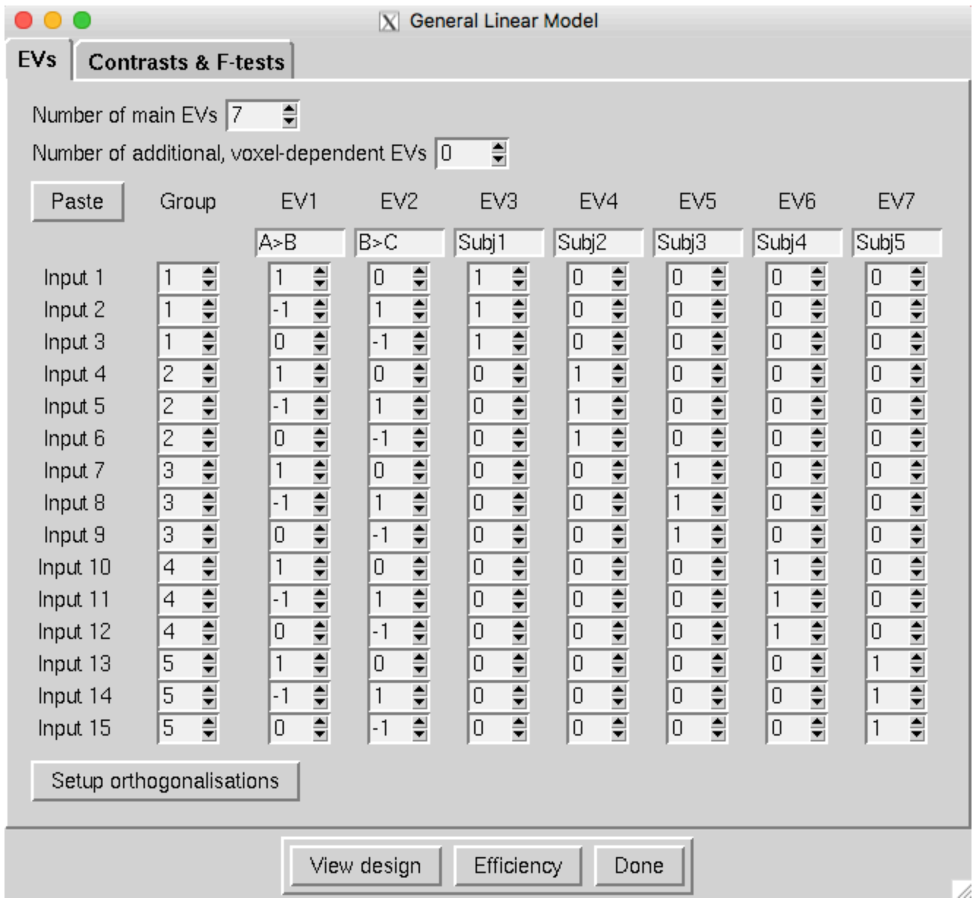


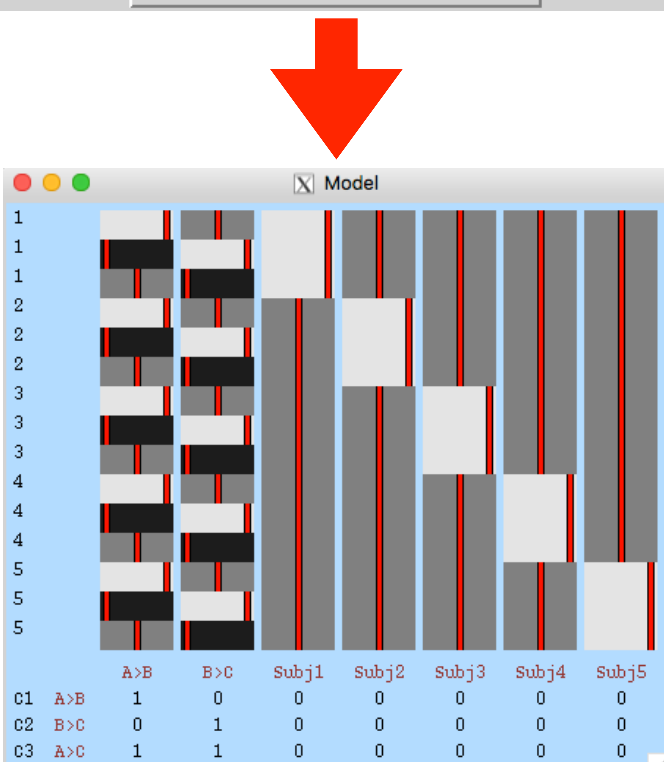
Assumed covariance matrix



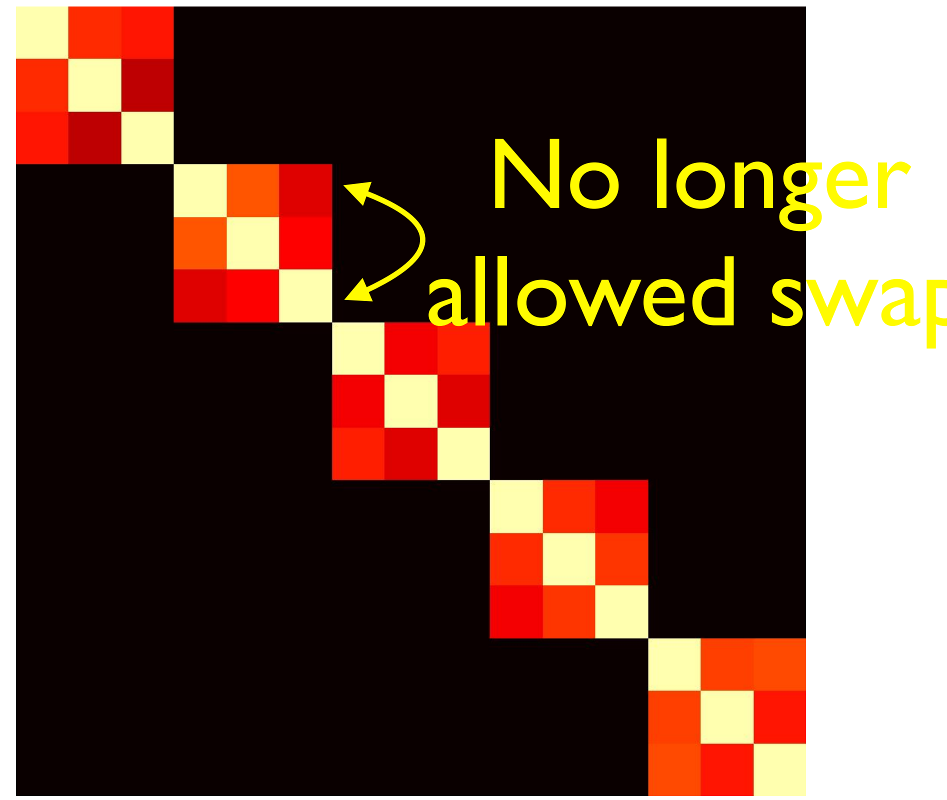
Assumptions: All subjects from the same "population", no dependence between subjects and "compound symmetry" within subjects

Examples of exchangeability: blocked ANOVA





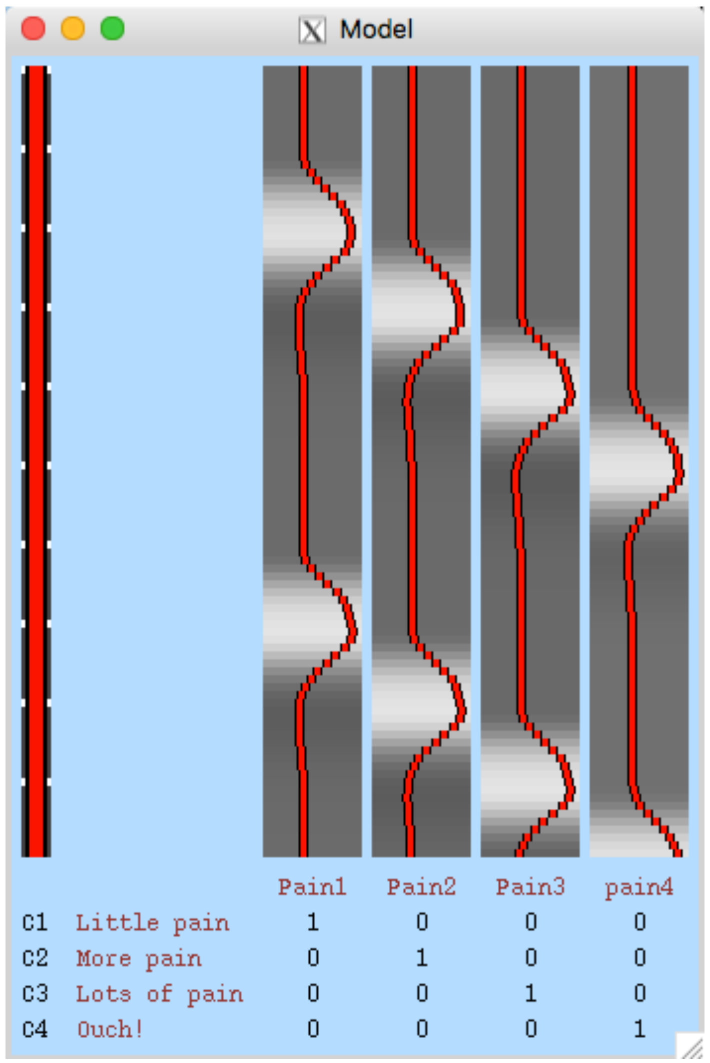
Assumed covariance matrix



Assumptions: All subjects from the same "population", no dependence between subjects and "compound symmetry" within subjects

My advice: Keep it simple!

Each subject scanned like this

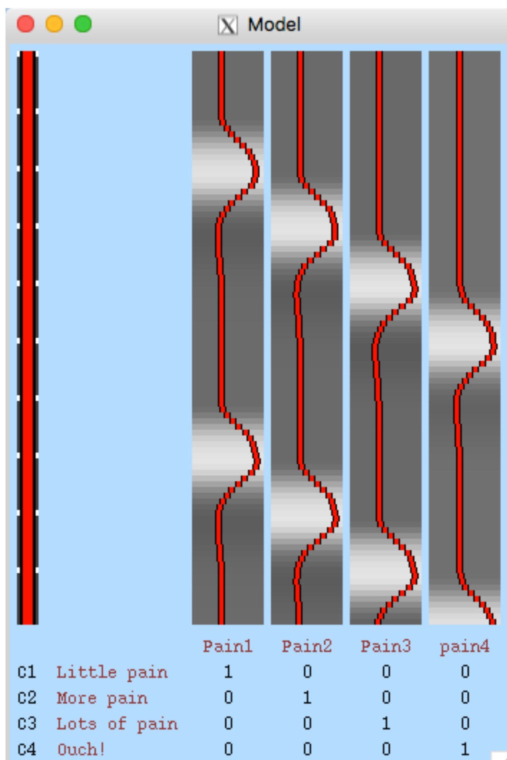


Taking 4 contrasts to 2nd level

We want to find areas that respond "linearly" to pain.

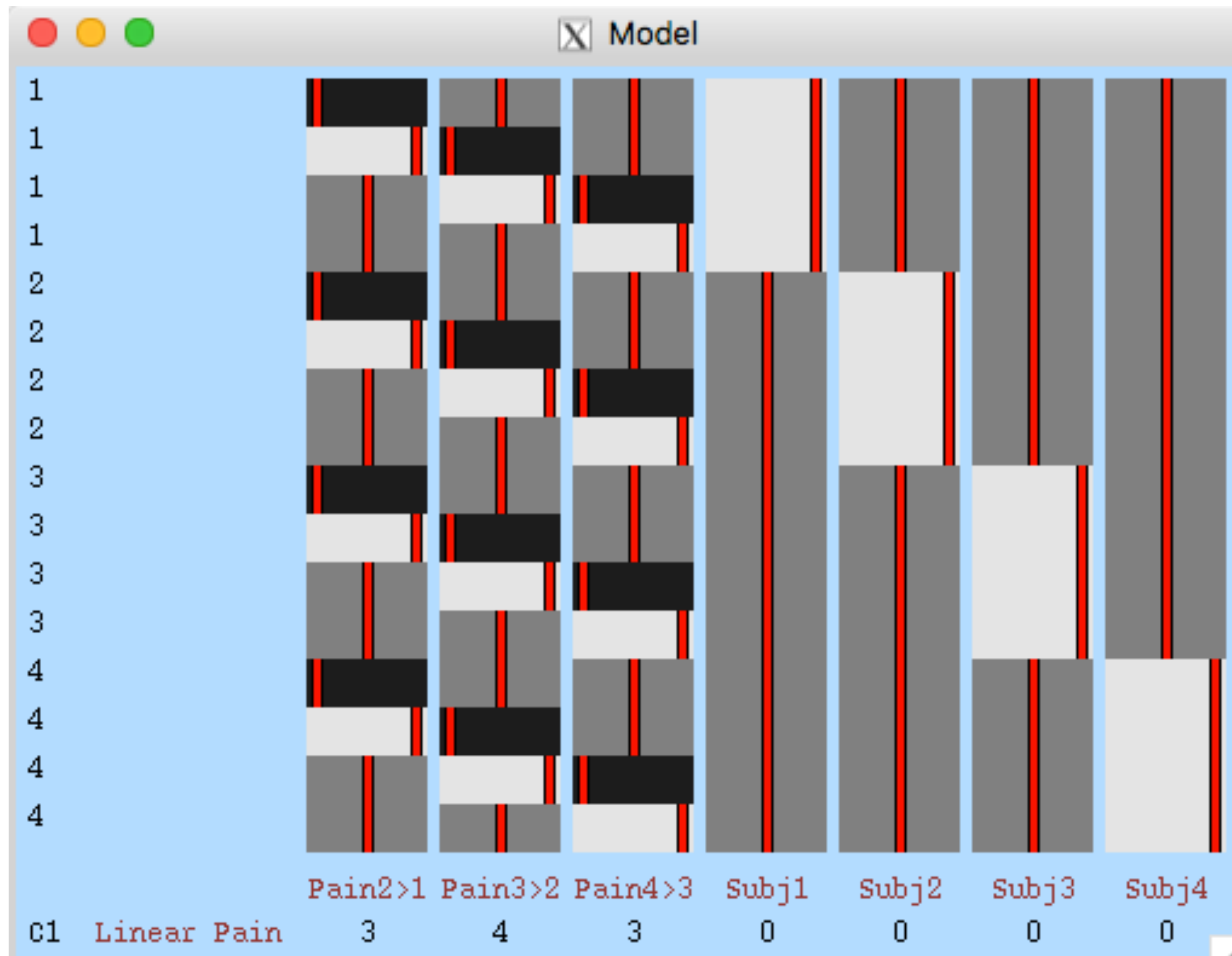
My advice: Keep it simple!

Each subject scanned like this

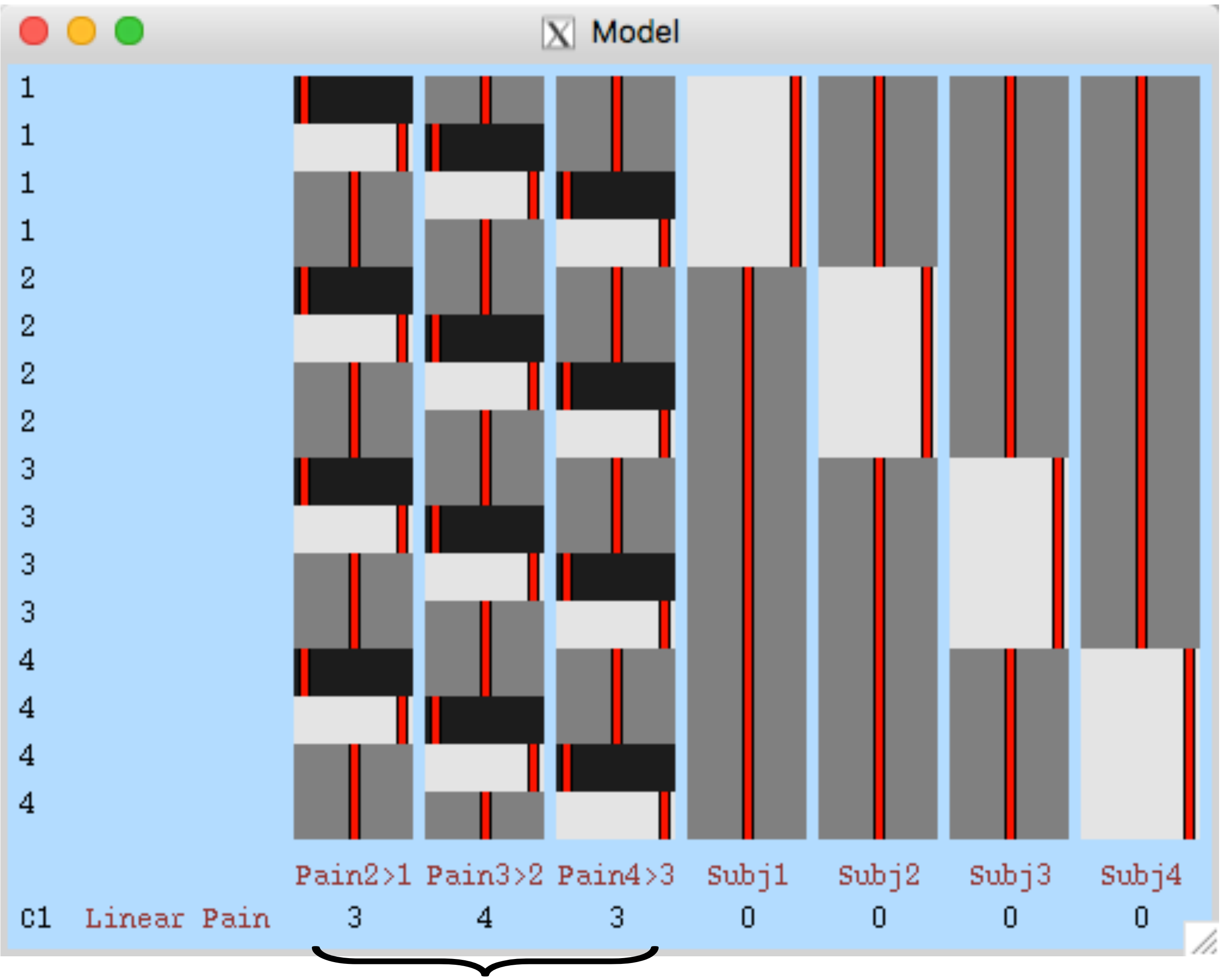


Taking 4 contrasts to 2nd level

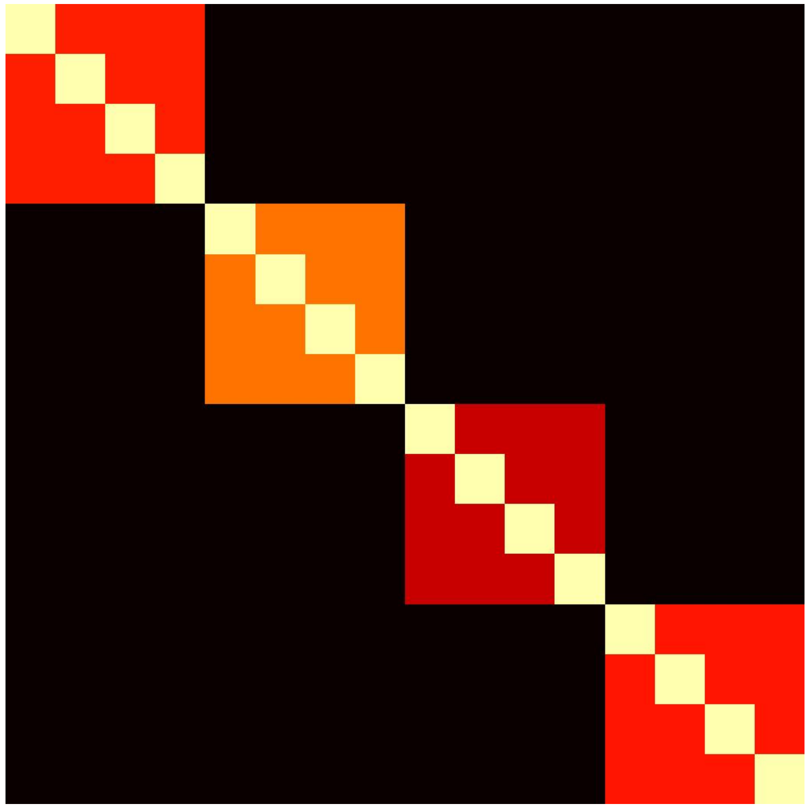
Repeating this for four subjects



My advice: Keep it simple!



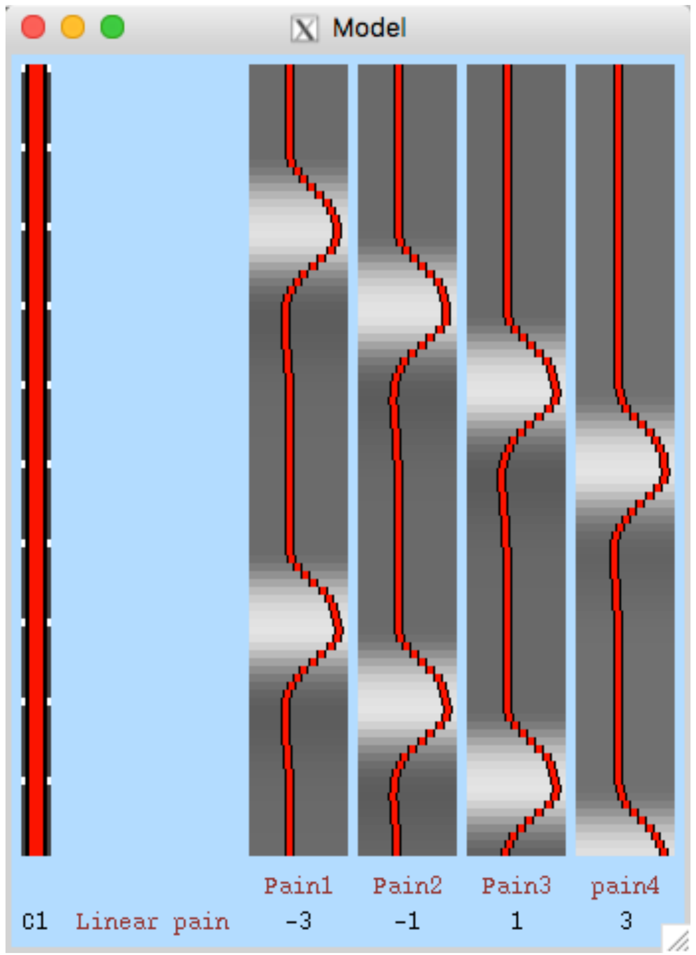
And figure out this contrast



You have to assume this covariance matrix

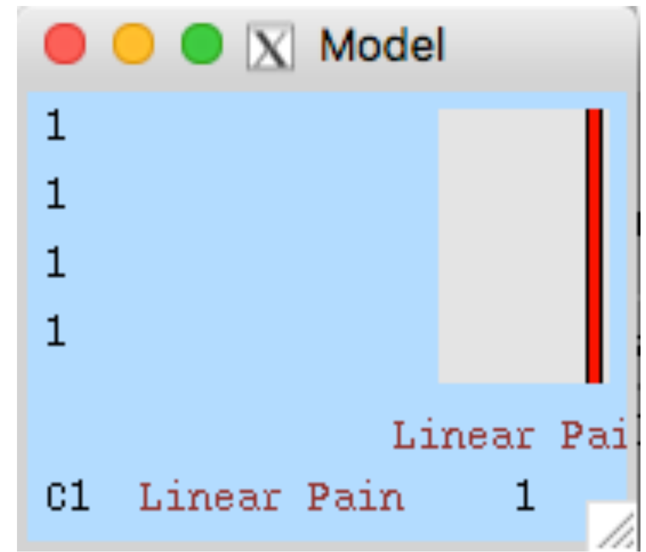
Why put yourself through all that pain?

My advice: Keep it simple!



When you can take a single contrast from the first level

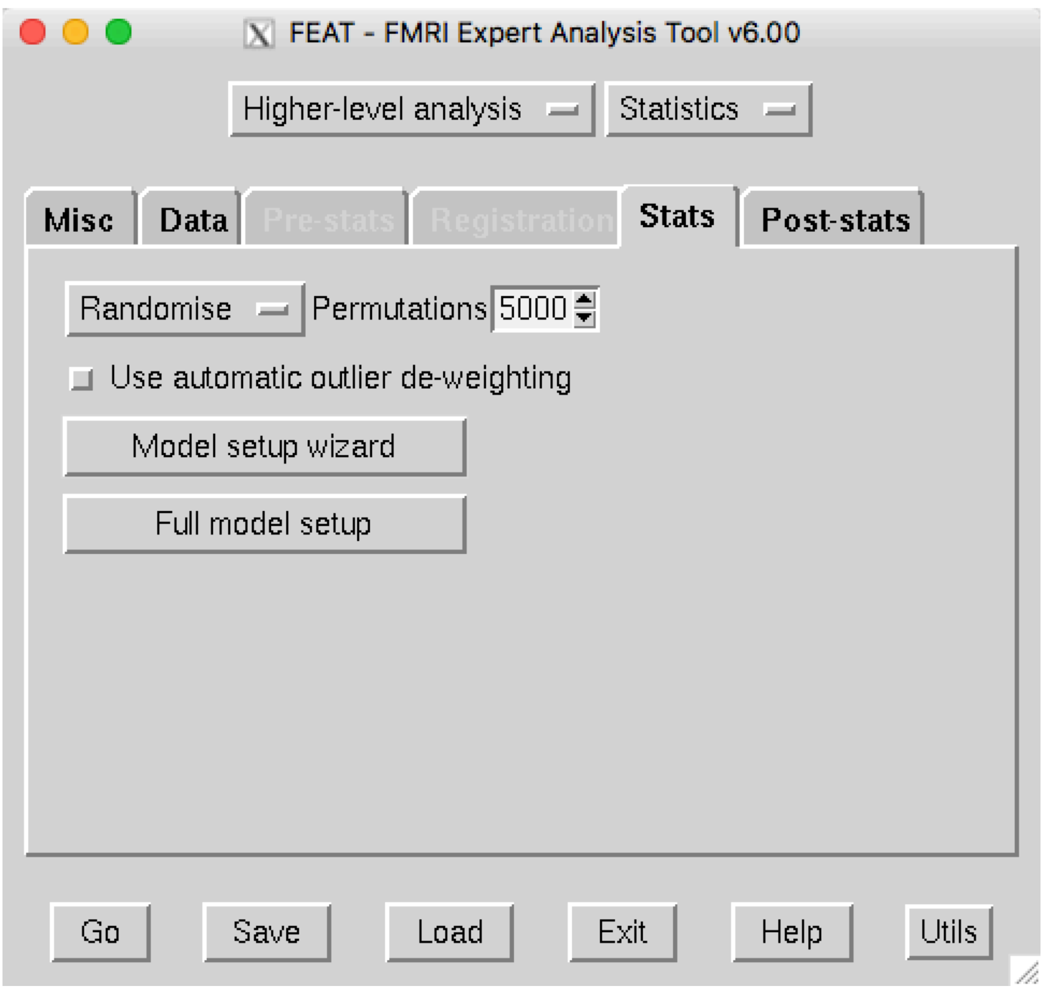
And get this at the second level

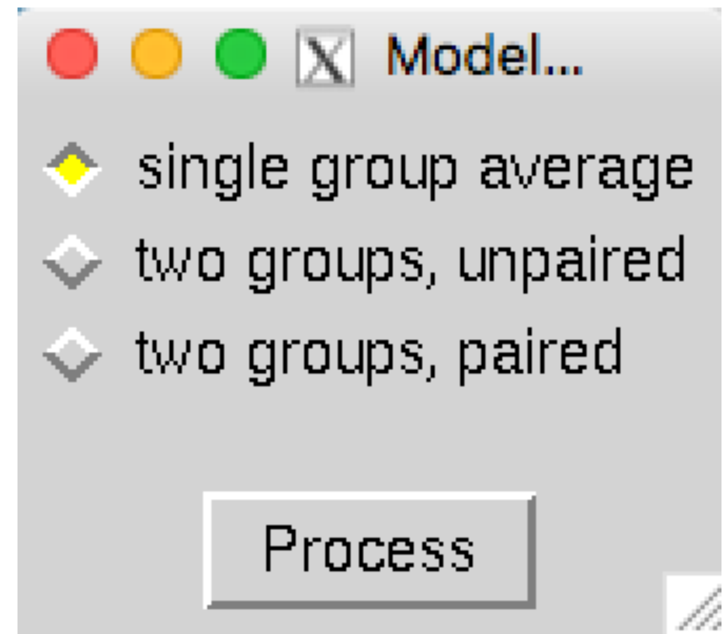


Assuming only symmetric errors

Much nicer, no?

Warning pertaining to FSL 6.0.1





Do not use the Model setup wizard together with Randomise in FSL 6.0.1

Outline

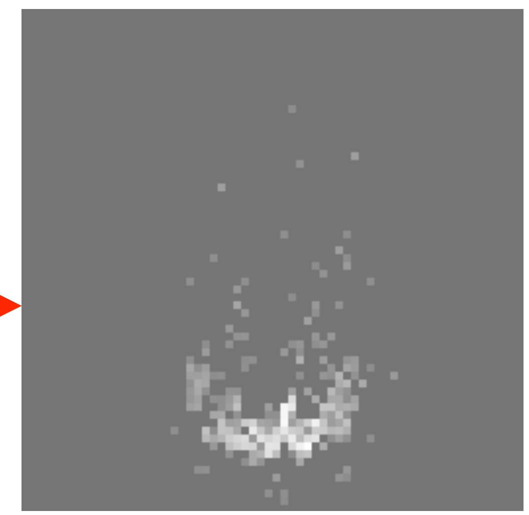
- Null-hypothesis and Null-distribution
- Multiple comparisons and Family-wise error
- Different ways of being surprised
  - Voxel-wise inference (Maximum z)
  - Cluster-wise inference (Maximum size)
- Parametric vs non-parametric tests
- Enhanced clusters
- FDR - False Discovery Rate

Clustering cookbook

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

z stat image

Threshold at
(arbitrary!) z level

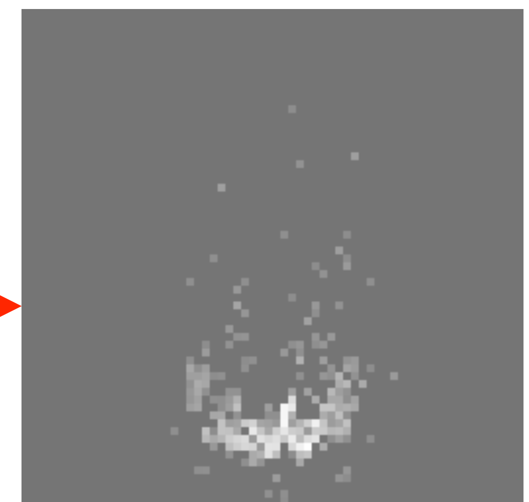


Clustering cookbook

Instead of resel-based correction, we can do clustering

z stat image

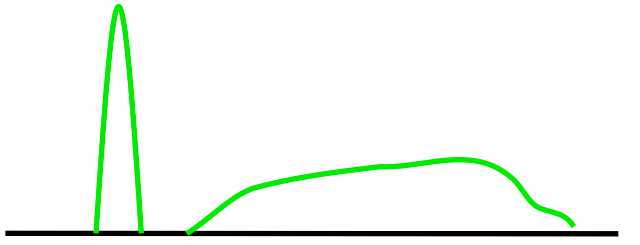
Threshold at
(arbitrary!) z level



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?

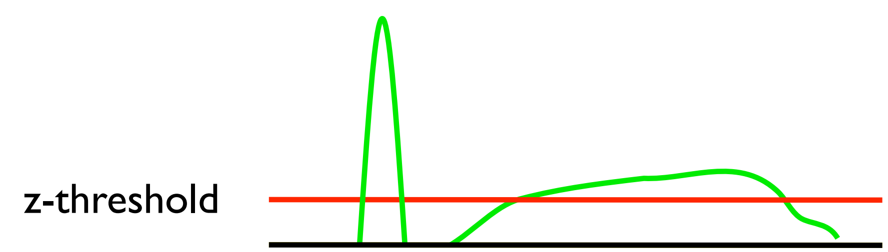
This is arbitrary and a trade-off



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

This is arbitrary and a trade-off

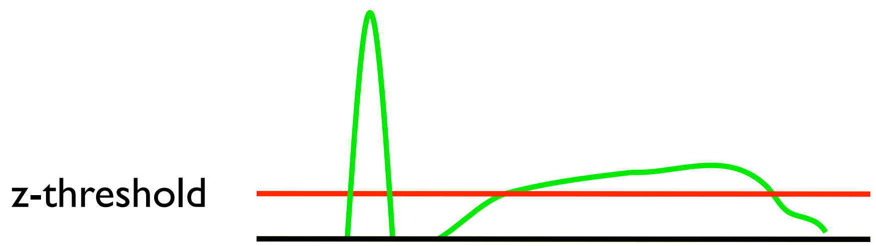
1. **Low threshold** - can violate RFT assumptions, but can detect clusters with large spatial extent and low z



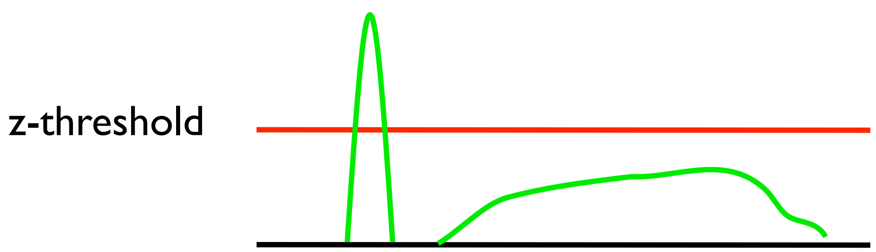
How do we choose the (arbitrary!) z-threshold?

This is arbitrary and a trade-off

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



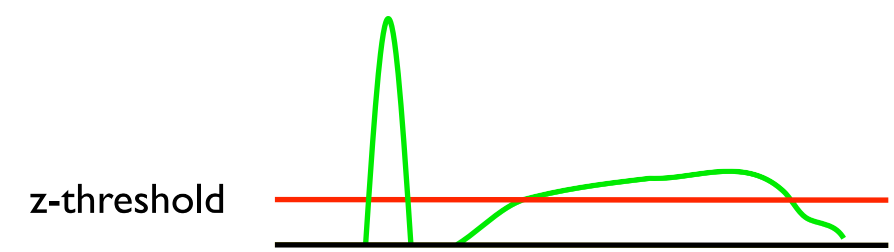
2. **High threshold** - gives more power to clusters with small spatial extent and high z



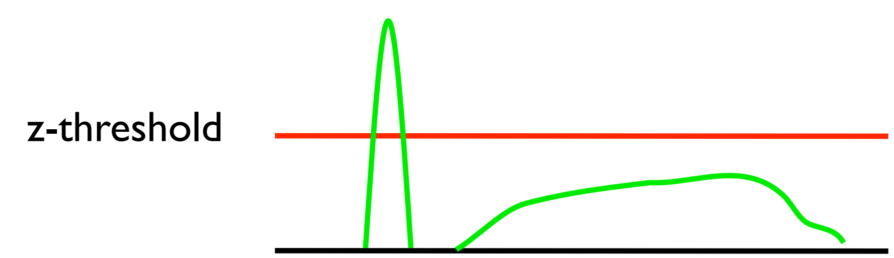
How do we choose the (arbitrary!) z-threshold?

This is arbitrary and a trade-off

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



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



Tends to be more sensitive than voxel-wise corrected testing

Results depend on extent of spatial smoothing in pre-processing

TFCE

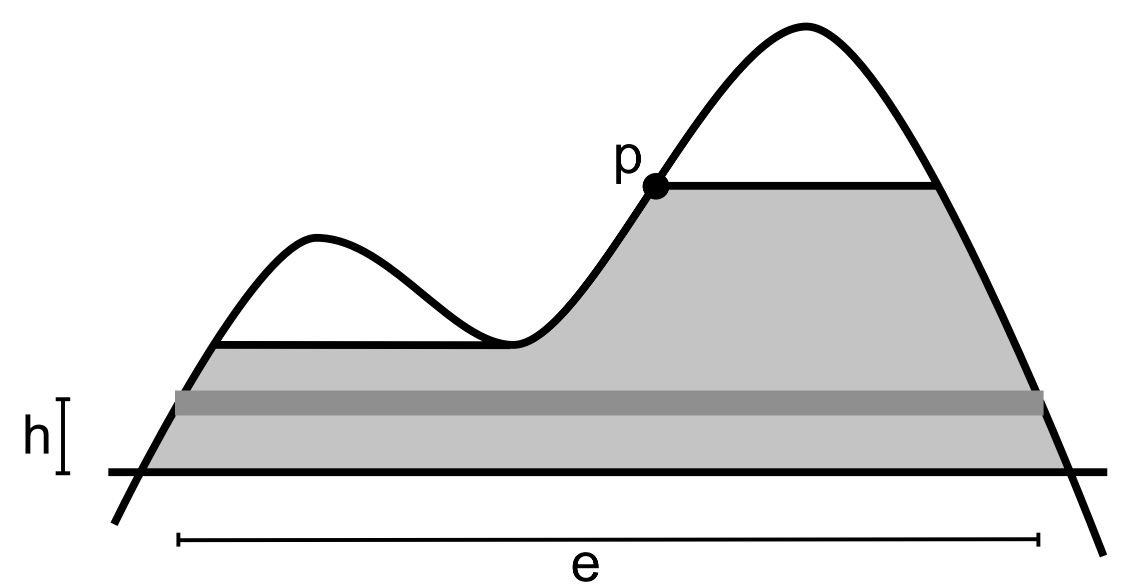
Threshold-Free Cluster Enhancement

[Smith & Nichols, NeuroImage 2009]

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

  ➡ unstable and arbitrary

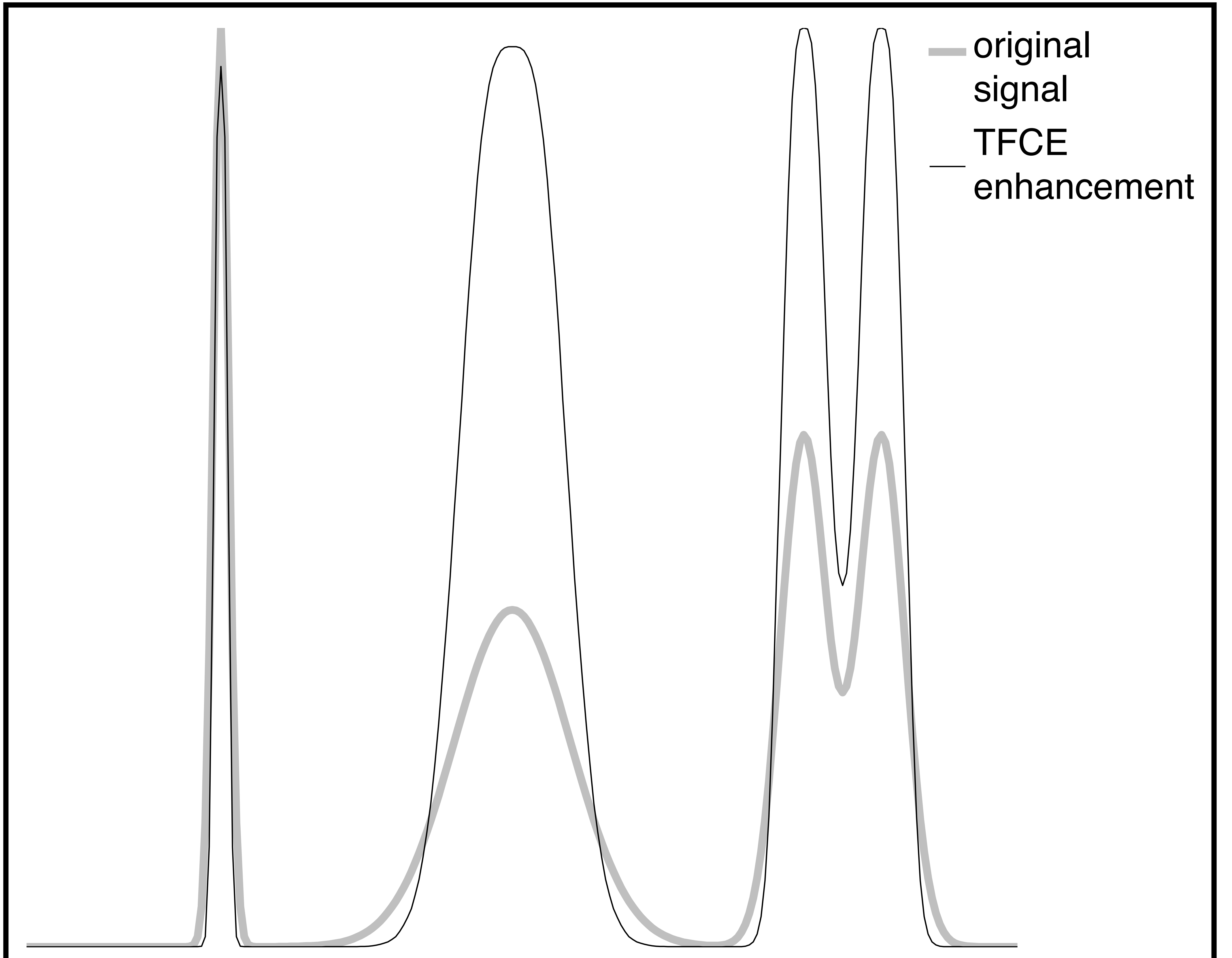
- TFCE
  - integrates cluster "scores" over all possible thresholds
  - output at each voxel is measure of local cluster-like support
  - similar sensitivity to optimal cluster-thresholding, but stable and non-arbitrary



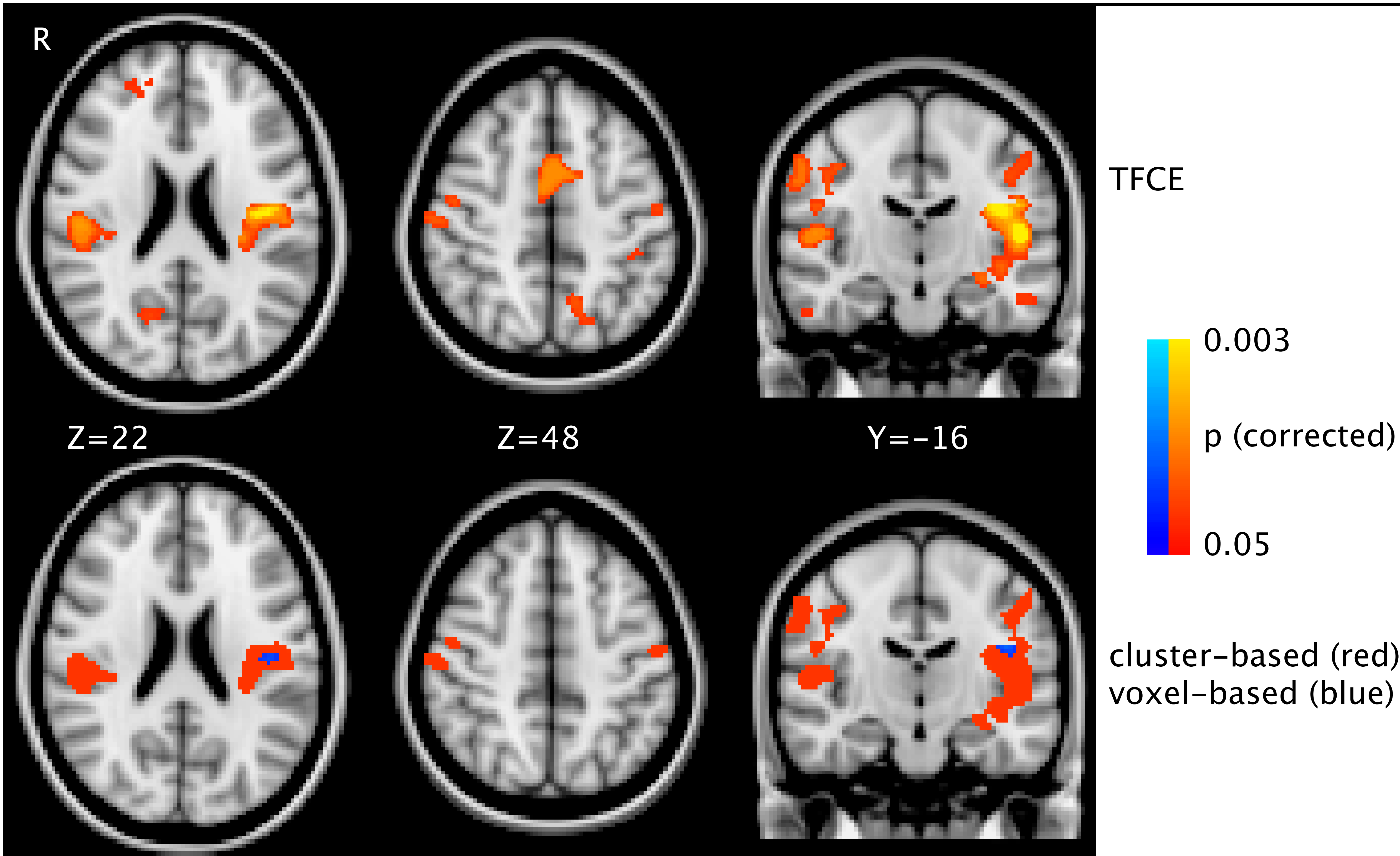
The TFCE value at point p is given by the sum, over the shaded area, of the score from each contributing incremental section:

$$TFCE(p) = \sum_h e(h)^E \cdot h^H$$

Qualitative example

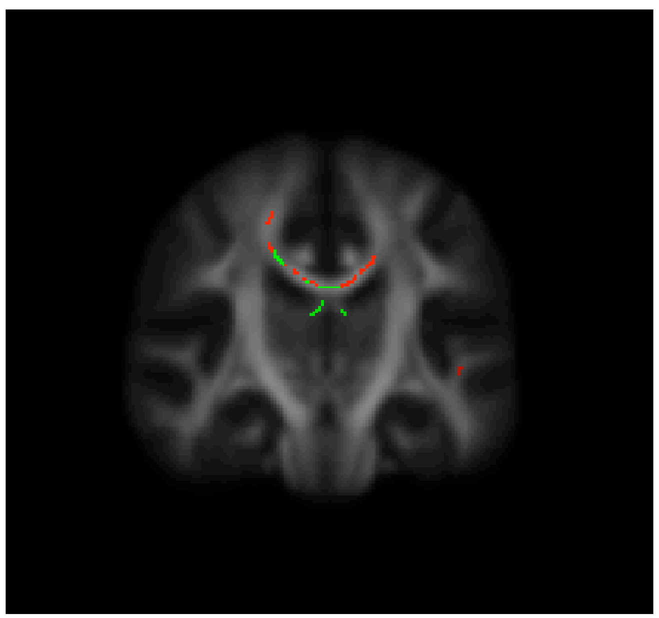


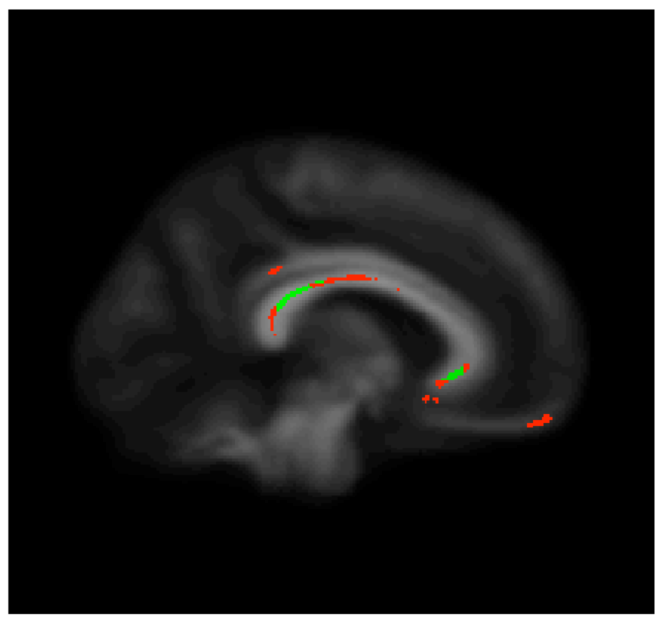
TFCE for FSL-VBM

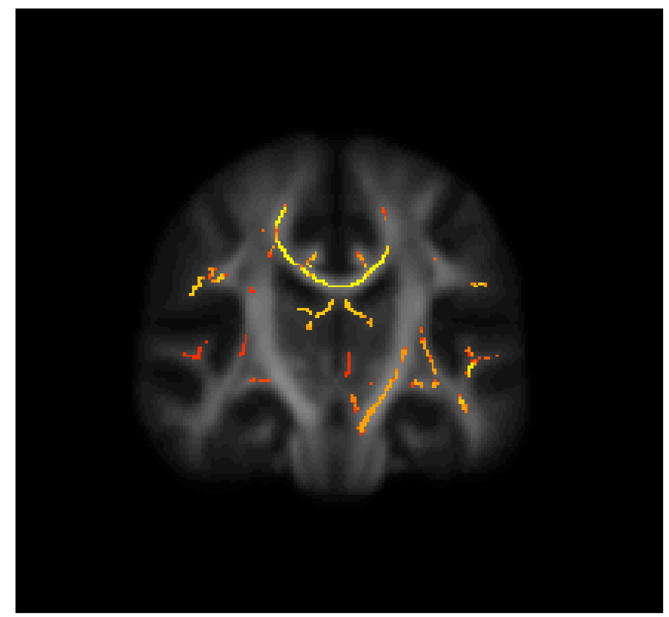


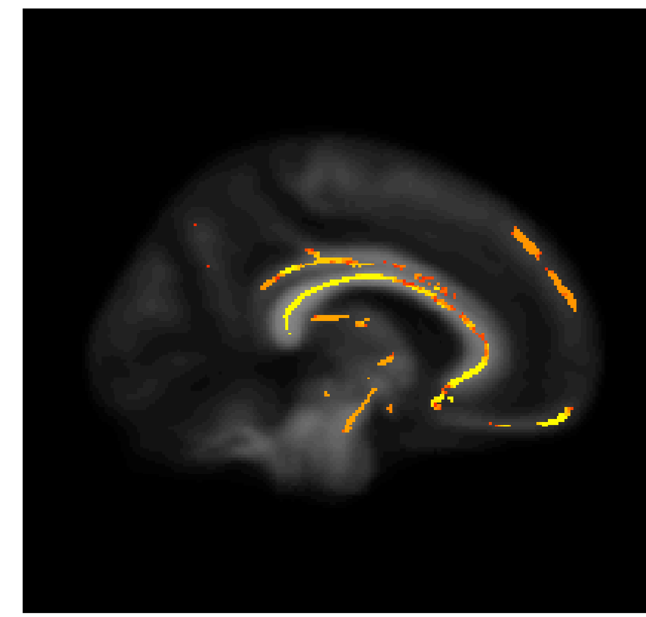
TFCE for TBSS

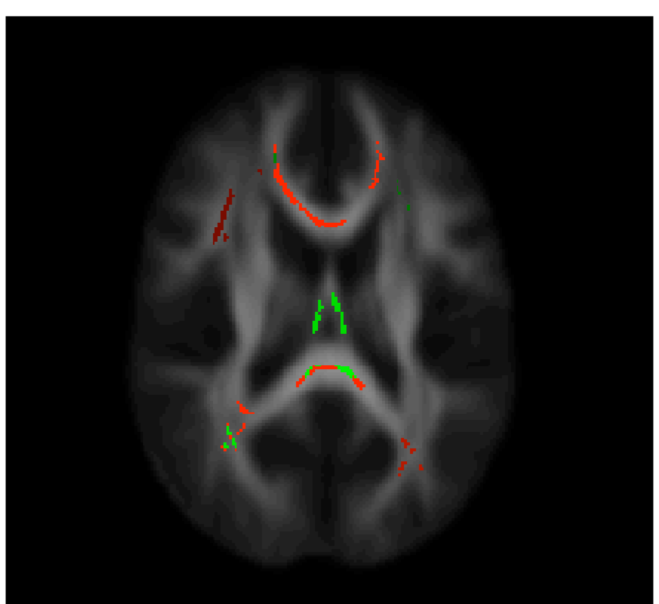
controls > schizophrenics
$p<0.05$ corrected for multiple comparisons across space, using randomise



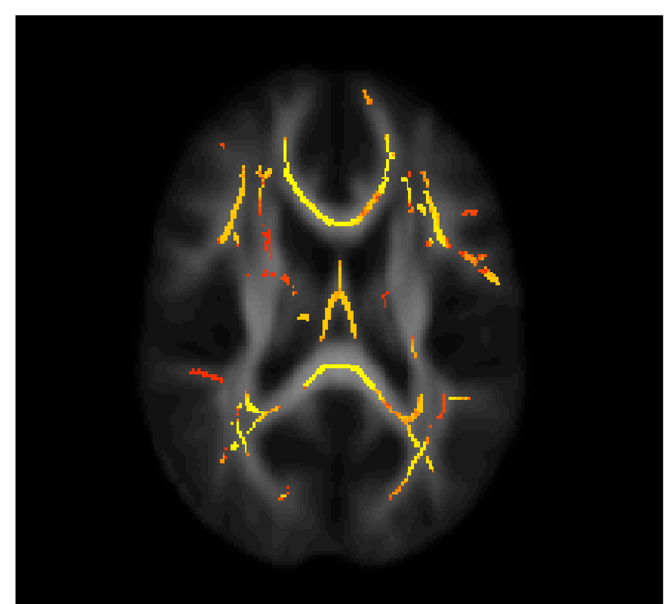








cluster-based:
cluster-forming threshold =
2 or 3



TFCE

Outline

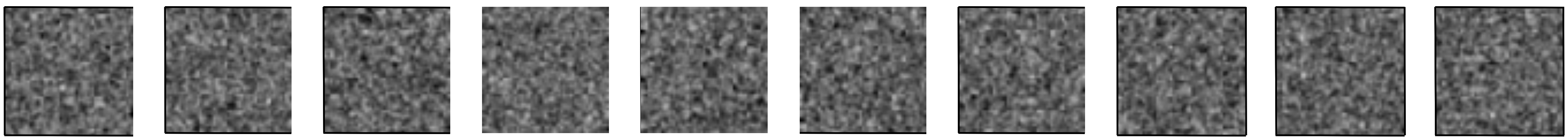
- Null-hypothesis and Null-distribution
- Multiple comparisons and Family-wise error
- Different ways of being surprised
  - Voxel-wise inference (Maximum z)
  - Cluster-wise inference (Maximum size)
- Parametric vs non-parametric tests
- Enhanced clusters
- FDR - False Discovery Rate

False Discovery Rate

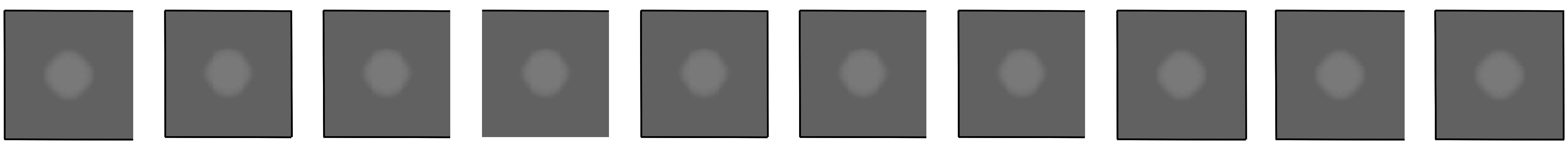
- FDR: False Discovery Rate
  A “new” way to look at inference.
- Uncorrected (for multiple-comparisons):
  - Is equivalent to saying: “I am happy to nearly always say something silly about my experiments”.
  - On average, **5% of all voxels** are false positives
- Family-Wise Error (FWE):
  - Is equivalent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

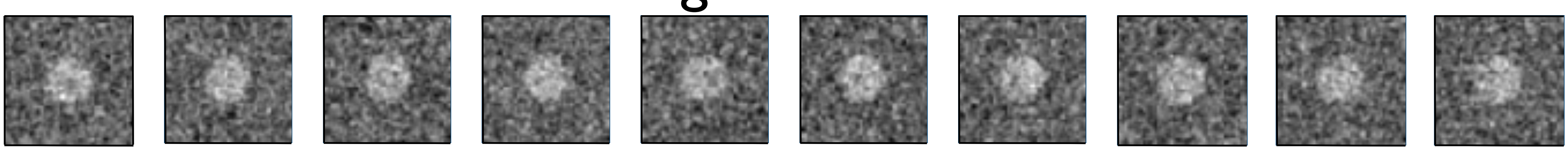
Noise



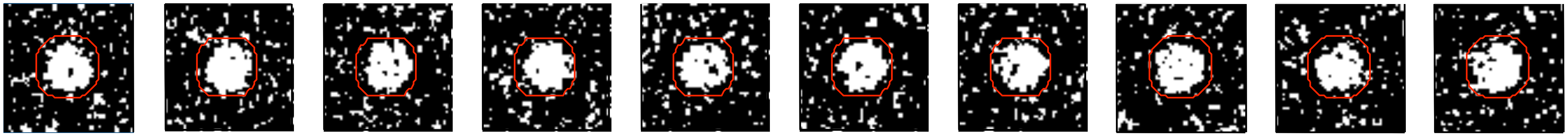
Signal



Signal+Noise

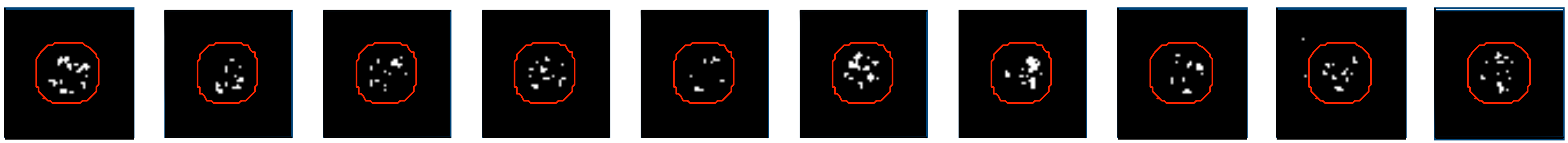


uncorrected voxelwise control of FP rate at 10%



percentage of all null pixels that are False Positives

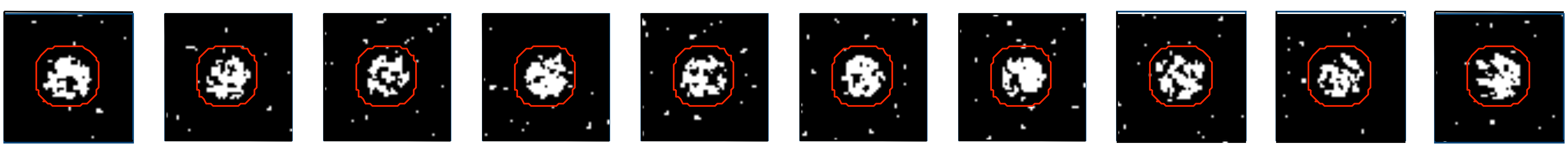
control of FamilyWise Error rate at 10%



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.