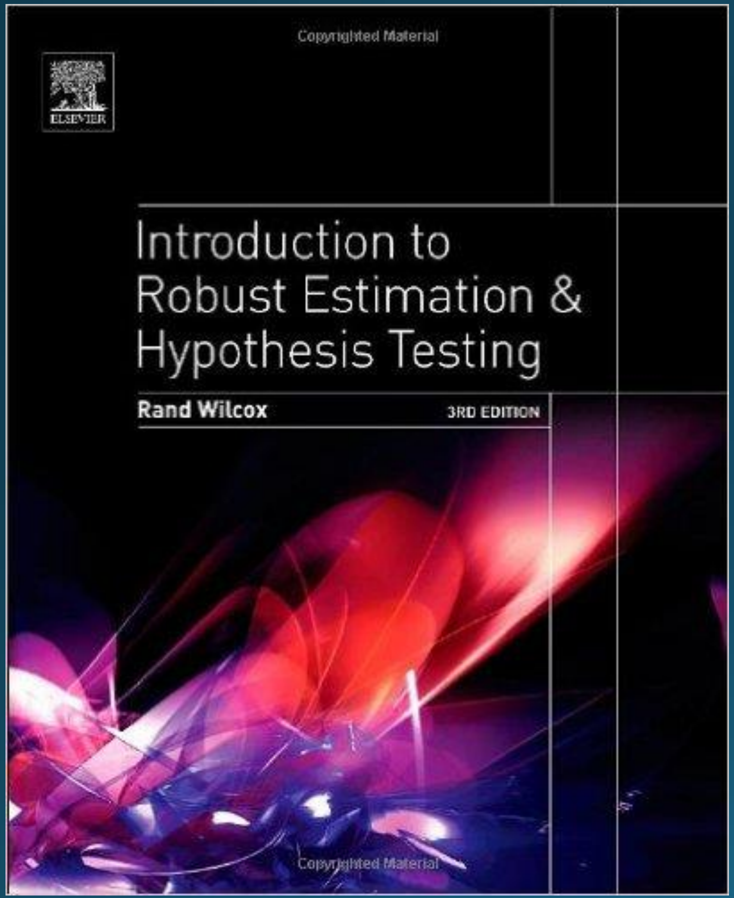
Robust statistics , credible intervals and correction for multiple comparisons for EEG data

Cyril Pernet, PhD
Edinburgh Imaging
Centre for Clinical Brain Sciences



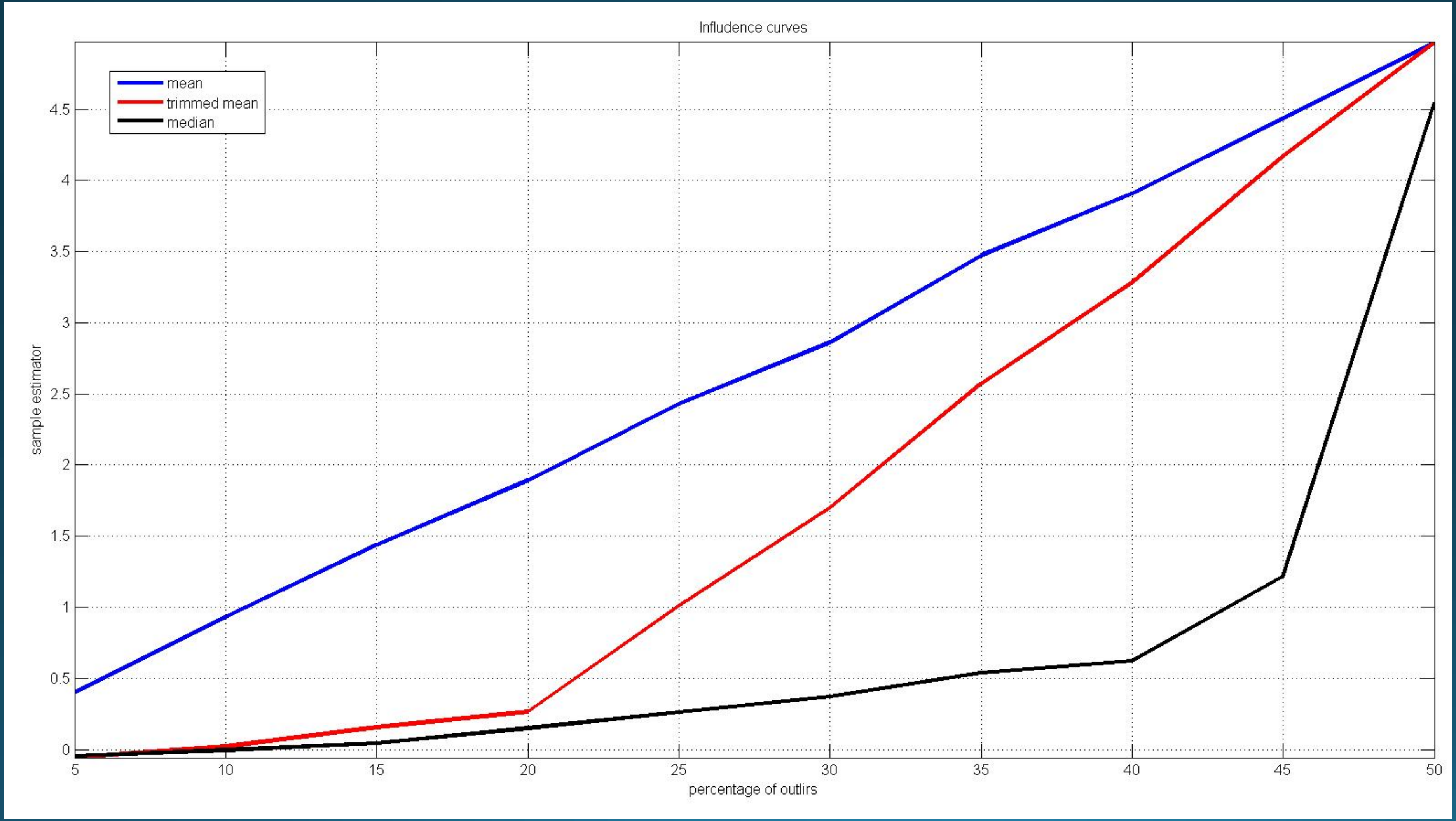
Wilcox, R (2012). *Introduction to robust estimation and hypothesis testing. 3rd Ed. Elsevier*

Issues with standard stats

- Standard stats are all instantiations of a GLM using an Ordinary Least Squares solution → implies looking at the mean
- The breakdown point of an estimator is the proportion of incorrect observations (e.g. arbitrarily large observations) an estimator can handle before giving an incorrect estimate
- For data x1 to xn – the mean has a breakdown point of 0 ! because we can make the mean large changing a single xi (e.g. mean([1 2 2 3 3 3 2 2 1]) = 2.1 & mean([1 2 2 3 3 3 2 2 1000])=113.11).
- Robust estimators: median, trimmed mean, M-estimators

http://en.wikipedia.org/wiki/Robust_statistics

Using the median and trimmed mean



Yes but my data are Gaussian

- Are you sure?
- Micceri (1989). The Unicorn, The Normal Curve, and Other Improbable Creatures. Psych Bul. 105, 156-166
- If the data are Gaussian, the median, the trimmed mean is the same as the mean ! So no reason not to use alternative techniques.

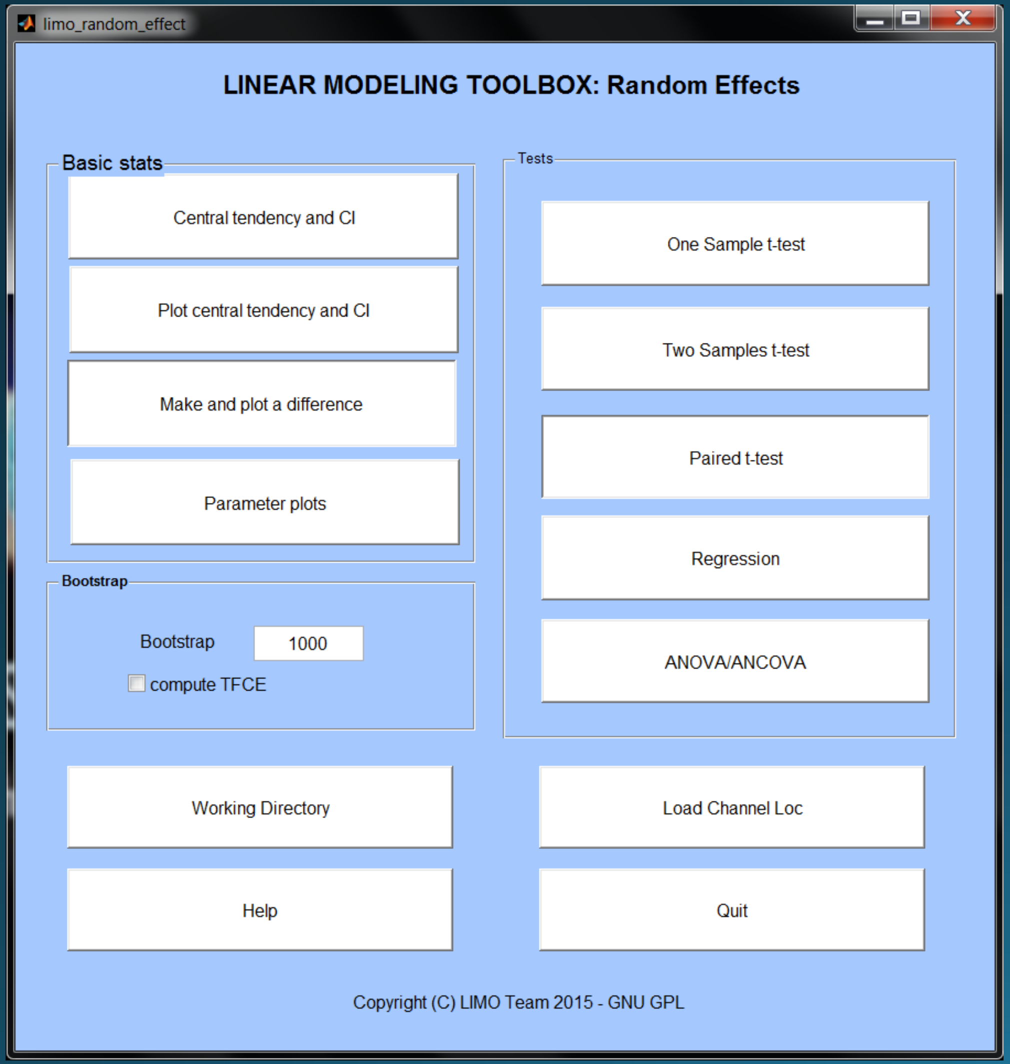
LIMO EEG toolbox

- 1st level GLM using temporally stable weighted least squares (WLS – trials have spatially varying weights)
- 2nd level relies on 20% trimmed mean (weights of 0 for bad subjects) for t-tests, 1-way ANOVA, and (soon) Repeated Measures ANOVA. It relies on Iterative Reweighted Least Squares (IRLS) for regressions and N-way ANOVA/ANOVA (all subjects have weights from 0 to 1 that change in space and time).

http://en.wikipedia.org/wiki/Robust_statistics

Robust tests (LIMO EEG toolbox)

LIMO EEG TOOLBOX



- One sample trimmed mean test
- Yuen t-tests (paired / 2 samples)
- IRLS Regression
- 1 way robust ANOVA (generalized Welch's method)
- IRLS for N-ways ANOVA
- Hoteling T square for repeated measures (soon to be robust)

One sample t-test

$$t = \frac{Mean}{std/\sqrt{n}}$$

p = 2 * tcdf(abs(t), df)

df = n -1

$$t = \frac{Trimmed\ Mean}{\sqrt{WinVar/(1-2*trimming\ percentage)*\sqrt{n}}}$$

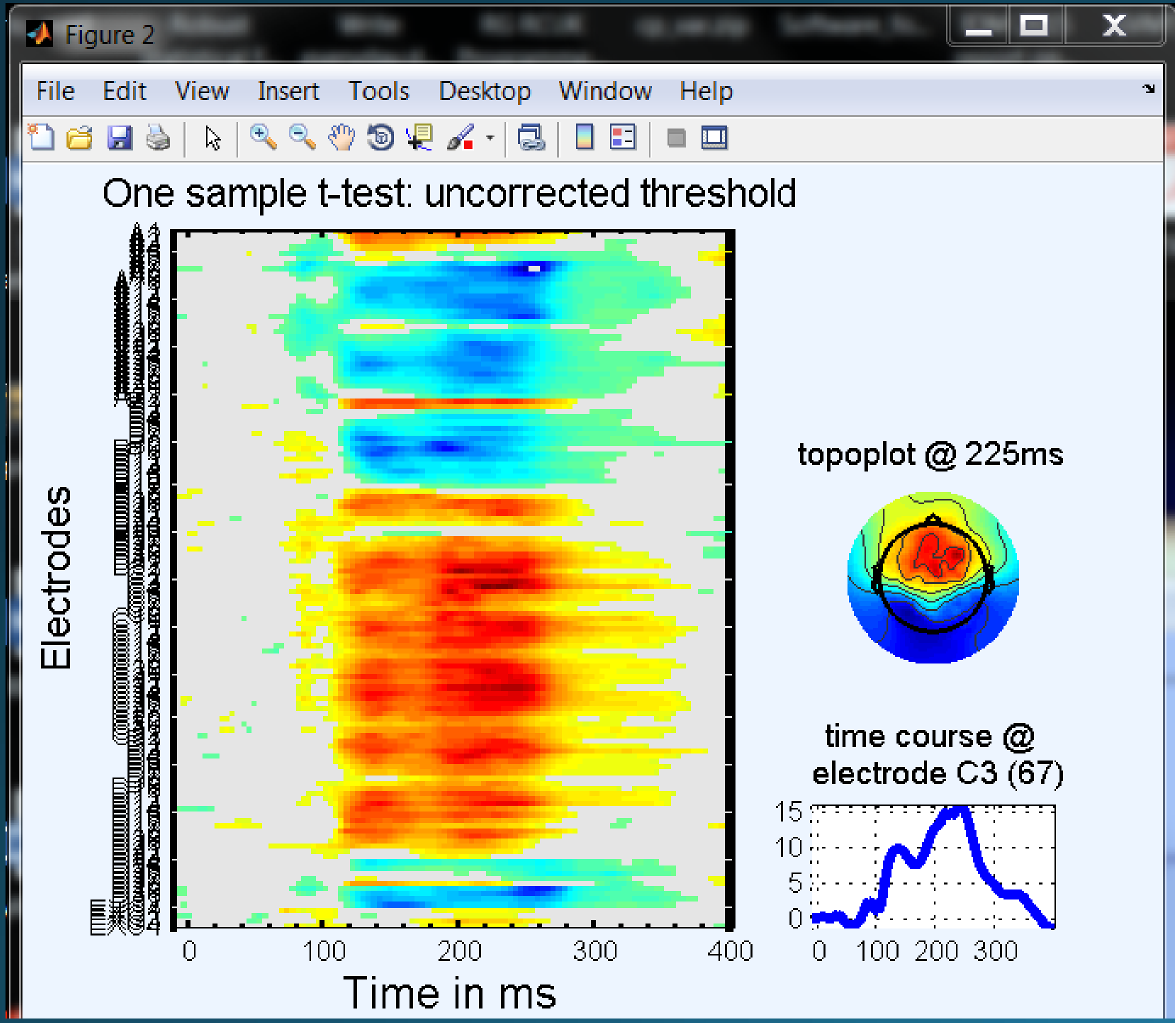
$$p = 2*(1-tcdf(abs(t), df))$$

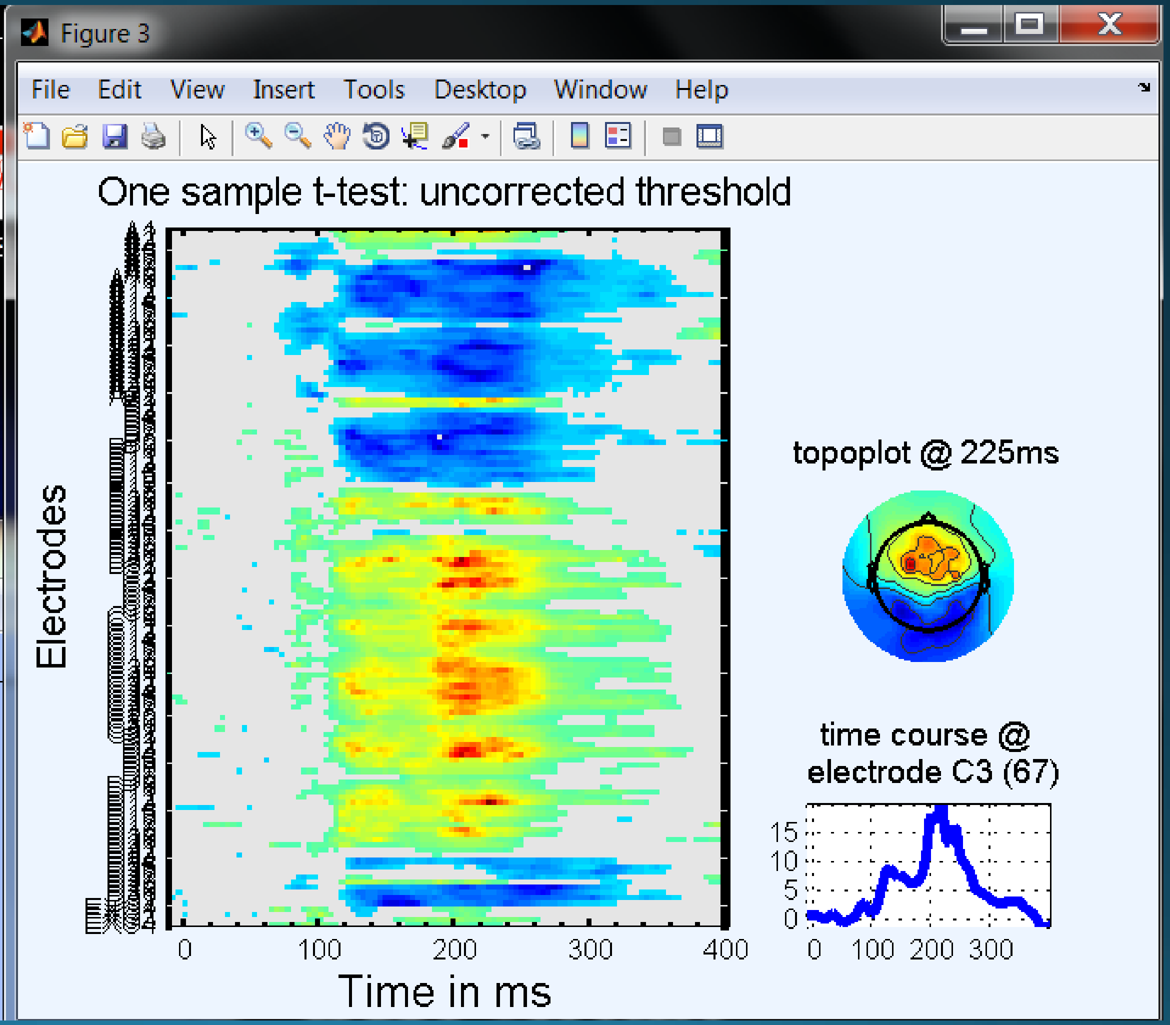
df = n-2*floor((trimming percentage/100)*n)-1

limo_ttest.m

limo_trimci.m

Test standard vs. robust t-test





Paired t-test

$$t = \frac{Mean\ (diffeence)}{std\ (difference)/\sqrt{n}}$$

p = 2 ∗ tcdf(abs(t), df) with df = n -1

limo_ttest.m

$$t = \frac{Difference\ of\ trimmed\ means}{\sqrt{\frac{\left(WinVar1 * (n-1)\right) + \left(WinVar2 * (n-1)\right) - (2 * (n-1) * WinCov)}{(n-2) * n\ trim}}}$$

$p = 2 * (1 - tcdf(abs(t), df)\ with$ df =($(n-2) * n\ trim$)-1

limo_yuend_ttest.m

Two-samples t-test

$$t = \frac{mean(gp1) - mean(gp2)}{\sqrt{\frac{var(gp1)}{n1} + \frac{var(gp2)}{n2}}}$$

p = 2 * tcdf(abs(t), df)

$$df = \frac{(s1+s2)^2}{\frac{s1}{n1-1} + \frac{s2}{n2-1}}$$

limo_ttest.m

$$t = \frac{Difference\ of\ trimmed\ means}{\sqrt{\frac{(n1-1)*WinVar1}{n1\ trim\ *(n1\ trim\ -1)} + \frac{(n2-1)*WinVar2}{n2\ trim\ *(n2\ trim\ -1)}}}$$

$$p = 2 * (1 - tcdf(abs(t), df)$$

$$\mathrm{df} = \frac{(Yuen\ s1 + Yuen\ s2)^2}{\frac{Yuen\ s1}{n1\ trim\ -1} + \frac{Yuen\ s2}{n2\ trim\ -1}}$$

limo_yuen_ttest.m

IRLS

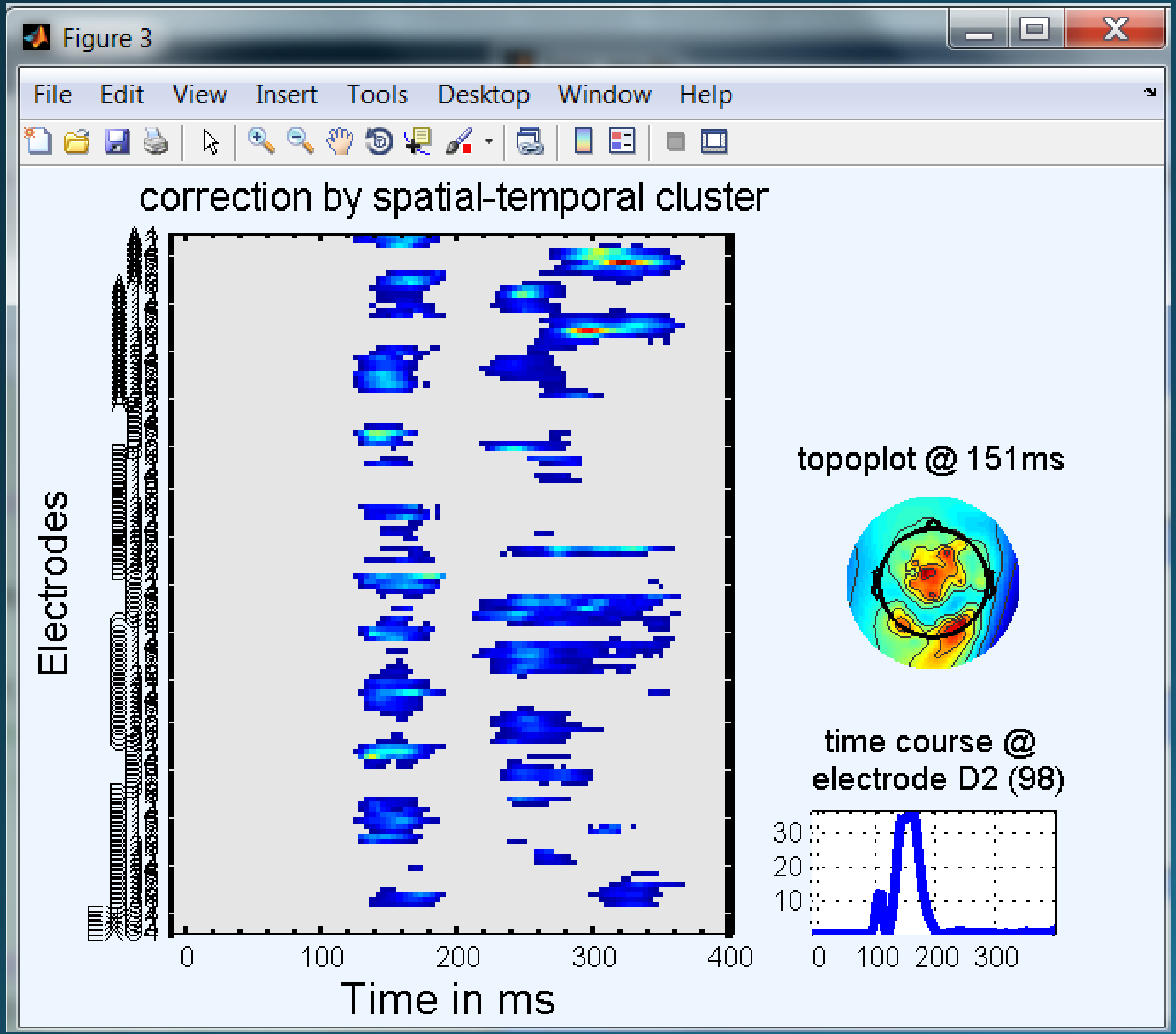
- **limo_irls.m**

- Start by OLS to obtain residuals
- Check outliers in standardized residuals (MAD)
- Compute weights (bisquare function)
- Recompute on weighted data
- Check residuals again until E(e) = 0

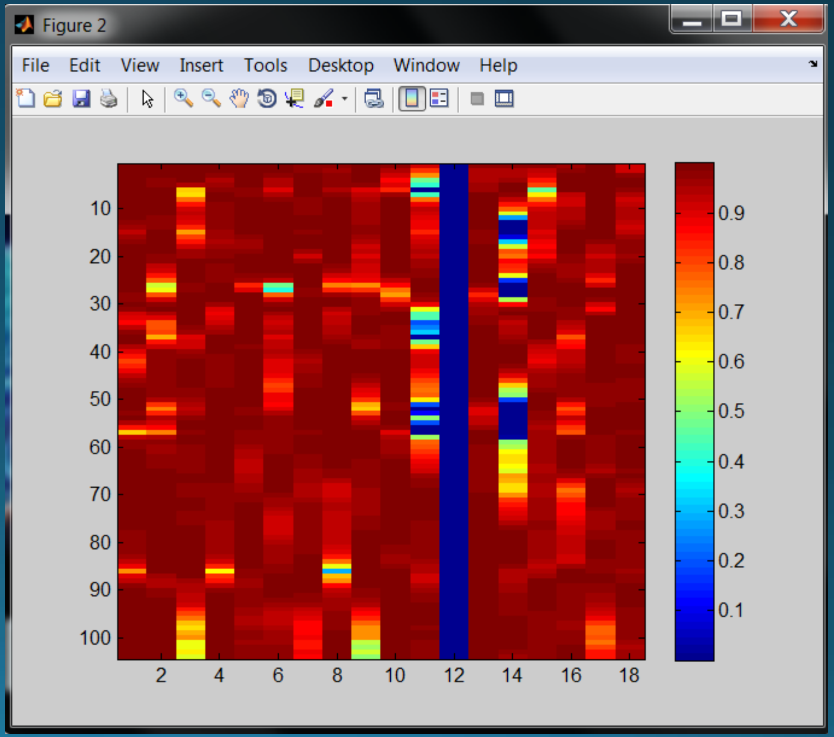
  → for eeg, iterate until max(abs(oldRes-newRes)) < (0.0001)

$$Wy = WX\beta + We, \quad E(e) = 0, \quad Cov(e) = \sigma^2 I$$

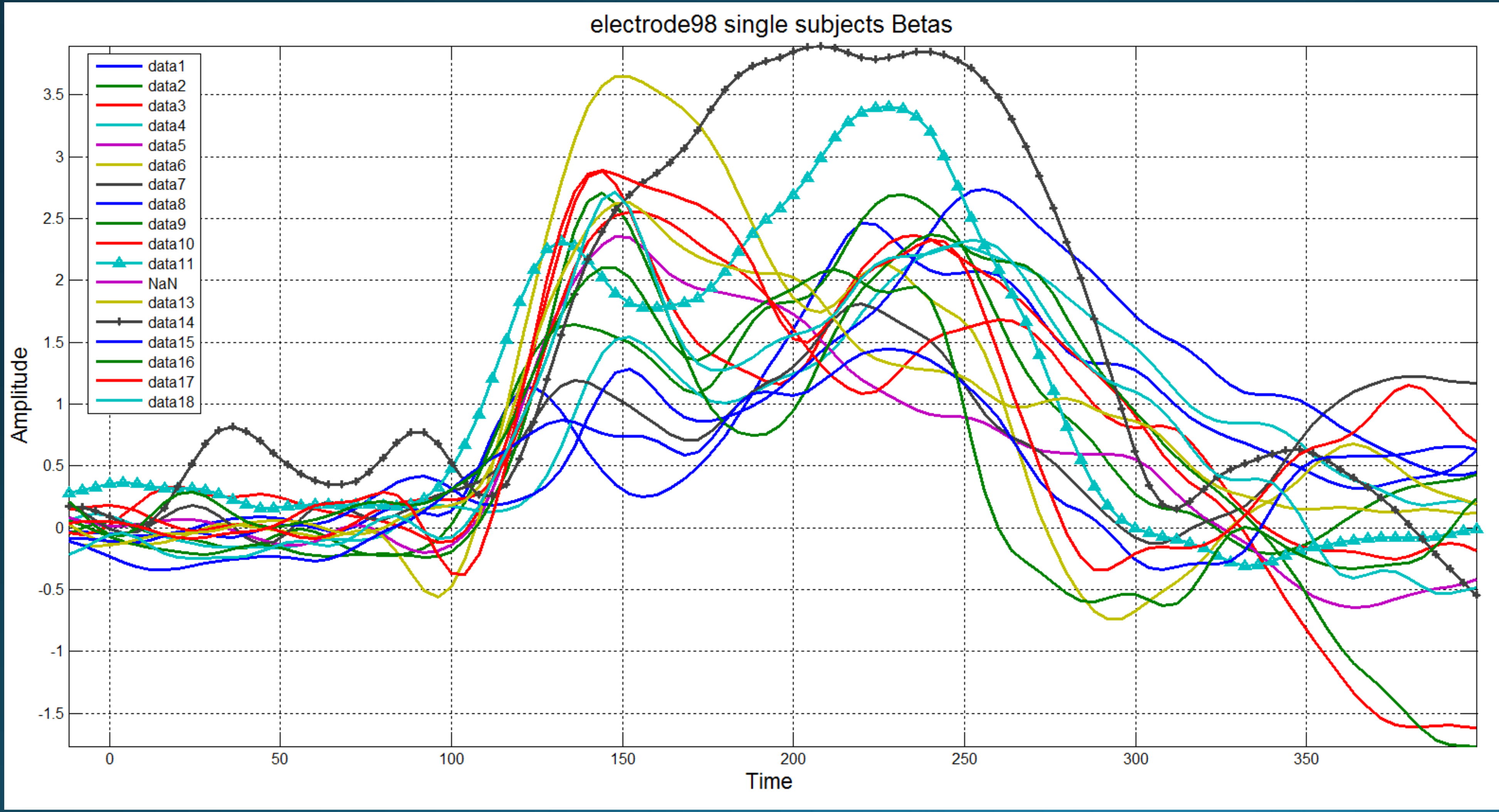
Check the weights of trials/subjects




```
>> load LIMO
>> size(LIMO.design.weights)
>> imagesc(squeeze(LIMO.design.weights(98,:,:)))
```

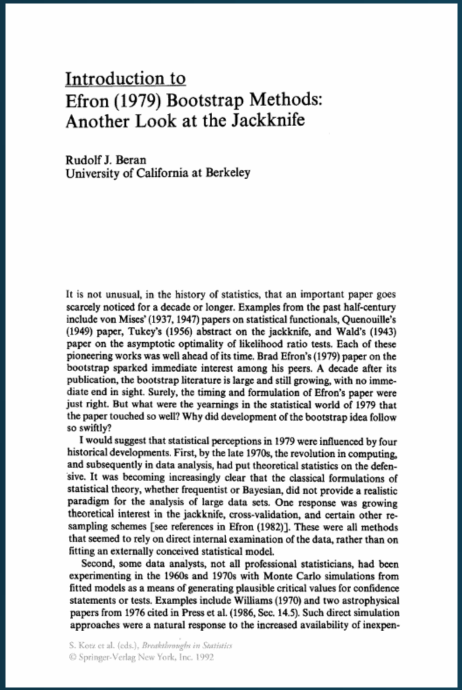


Check the weights of trials/subjects



Use central tendency tools to check what's going on

Building CI using bootstrap



Introduction to
Efron (1979) Bootstrap Methods:
Another Look at the Jackknife

Rudolf J. Beran
University of California at Berkeley

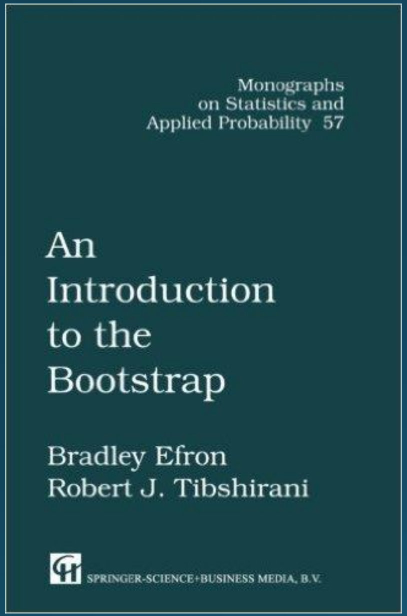
It is not unusual, in the history of statistics, that an important paper goes scarcely noticed for a decade or longer. Examples from the past half-century include von Mises' (1937, 1947) papers on statistical functionals, Quenouille's (1949) paper, Tukey's (1956) abstract on the jackknife, and Wald's (1943) paper on the asymptotic optimality of likelihood ratio tests. Each of these pioneering works was well ahead of its time. Brad Efron's (1979) paper on the bootstrap sparked immediate interest among his peers. A decade after its publication, the bootstrap literature is large and still growing, with no immediate end in sight. Surely, the timing and formulation of Efron's paper were just right. But what were the yearnings in the statistical world of 1979 that the paper touched so well? Why did development of the bootstrap idea follow so swiftly?

I would suggest that statistical perceptions in 1979 were influenced by four historical developments. First, by the late 1970s, the revolution in computing, and subsequently in data analysis, had put theoretical statistics on the defensive. It was becoming increasingly clear that the classical formulations of statistical theory, whether frequentist or Bayesian, did not provide a realistic paradigm for the analysis of large data sets. One response was growing theoretical interest in the jackknife, cross-validation, and certain other resampling schemes [see references in Efron (1982)]. These were all methods that seemed to rely on direct internal examination of the data, rather than on fitting an externally conceived statistical model.

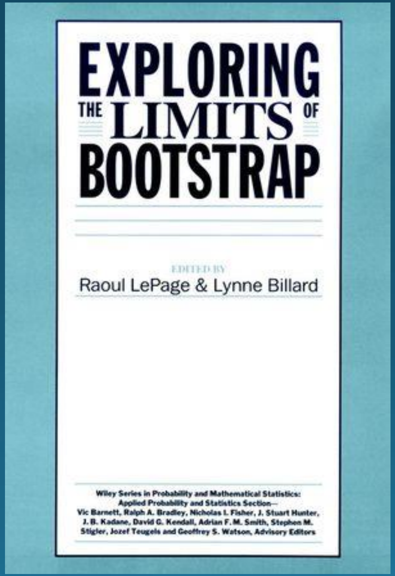
Second, some data analysts, not all professional statisticians, had been experimenting in the 1960s and 1970s with Monte Carlo simulations from fitted models as a means of generating plausible critical values for confidence statements or tests. Examples include Williams (1970) and two astrophysical papers from 1976 cited in Press et al. (1986, Sec. 14.5). Such direct simulation approaches were a natural response to the increased availability of inexpen-

S. Kotz et al. (eds.), *Breakthroughs in Statistics*
© Springer-Verlag New York, Inc. 1992

Efron , B. (1979). Bootstrap methods; another look at the jackknife . *Ann. Statist* . **7** , 1 – 26



Efron , B. , and Tibshirani , R. (1993). *An Introduction to the Bootstrap* . Chapman & Hall , New York



LePage, R & Billard L (Ed)
Exploring the Limits of Bootstrap, 1992

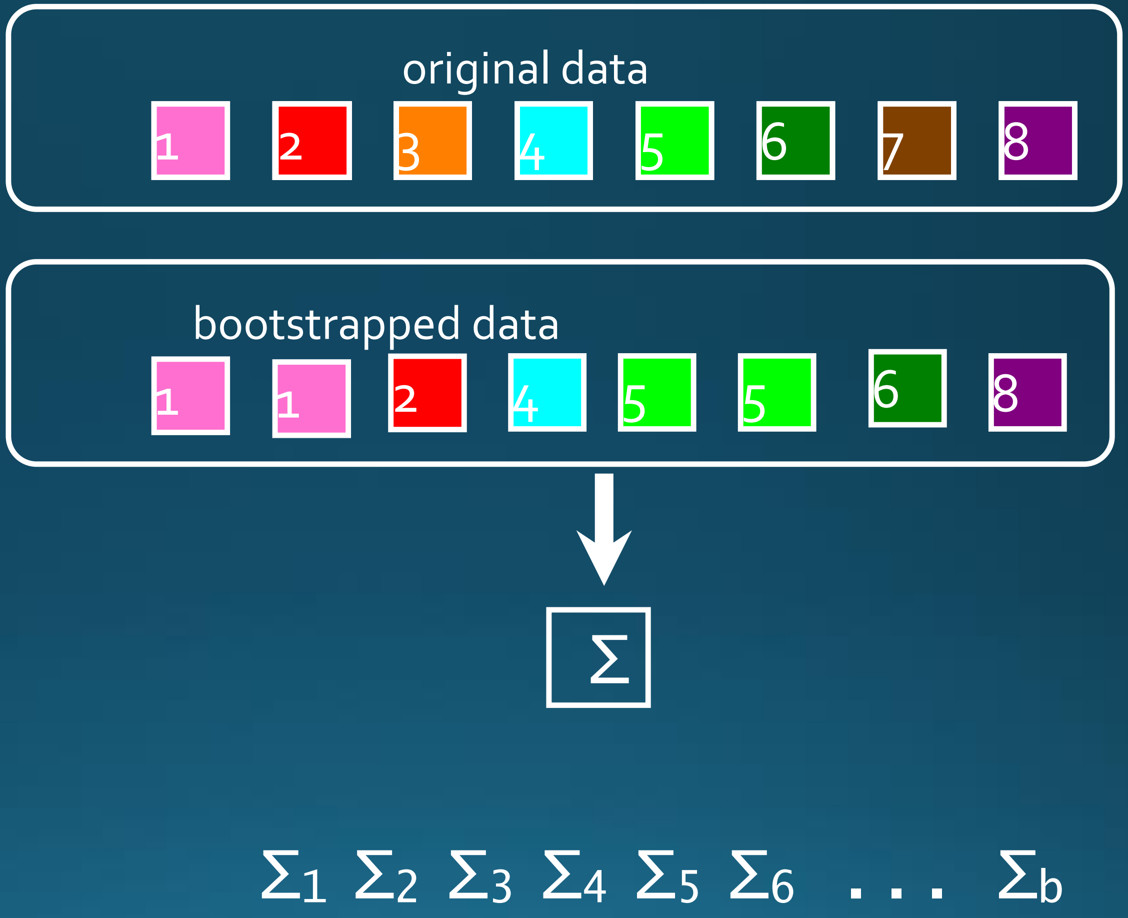
Bootstrap: central idea

- Statistics rely on estimators (e.g. the mean) and measures of accuracy for those estimators (standard error and confidence intervals)
- "The bootstrap is a computer-based method for assigning measures of accuracy to statistical estimates." Efron & Tibshirani, 1993
- The bootstrap is a type of resampling procedure along with jack-knife and permutations.

- Bootstrap is particularly effective at estimating accuracy (bias, SE, CI) but it can also be applied to many other problems – in particular to estimate distributions.

General recipe

(1) sample WITH replacement n observations (under H1 for CI of an estimate, under H0 for the null distribution)



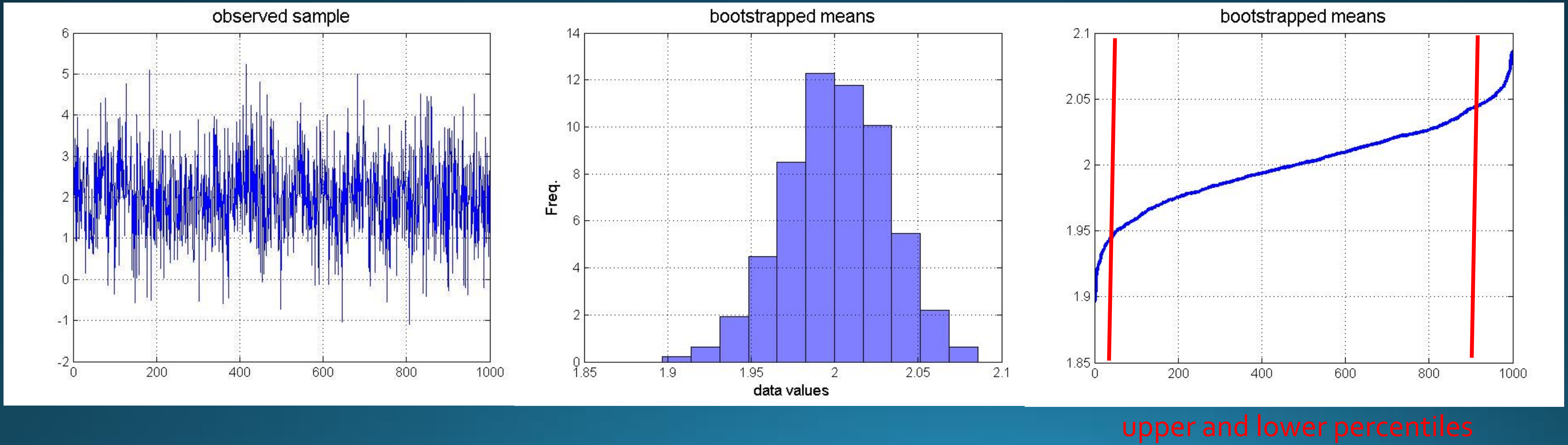
(2) compute estimate
e.g. sum, trimmed mean

(3) repeat (1) & (2) b times

(4) get bias, std, confidence interval, p-value

Percentile boot Confidence Interval

- Let ϑ *be an estimator, and* we want the 1-alpha CI(ϑ)
- *Bootstrap the data computing* ϑ^* *to obtain a distribution of this parameter and take the 1-alpha/2 upper and lower percentile*



The Annals of Statistics
1981, Vol. 9, No. 1, 130–134

THE BAYESIAN BOOTSTRAP

BY DONALD B. RUBIN

Educational Testing Service

The Bayesian bootstrap is the Bayesian analogue of the bootstrap. Instead of simulating the sampling distribution of a statistic estimating a parameter, the Bayesian bootstrap simulates the posterior distribution of the parameter; operationally and inferentially the methods are quite similar. Because both methods of drawing inferences are based on somewhat peculiar model assumptions and the resulting inferences are generally sensitive to these assumptions, neither method should be applied without some consideration of the reasonableness of these model assumptions. In this sense, neither method is a true bootstrap procedure yielding inferences unaided by external assumptions.

Bayesian bootstrap

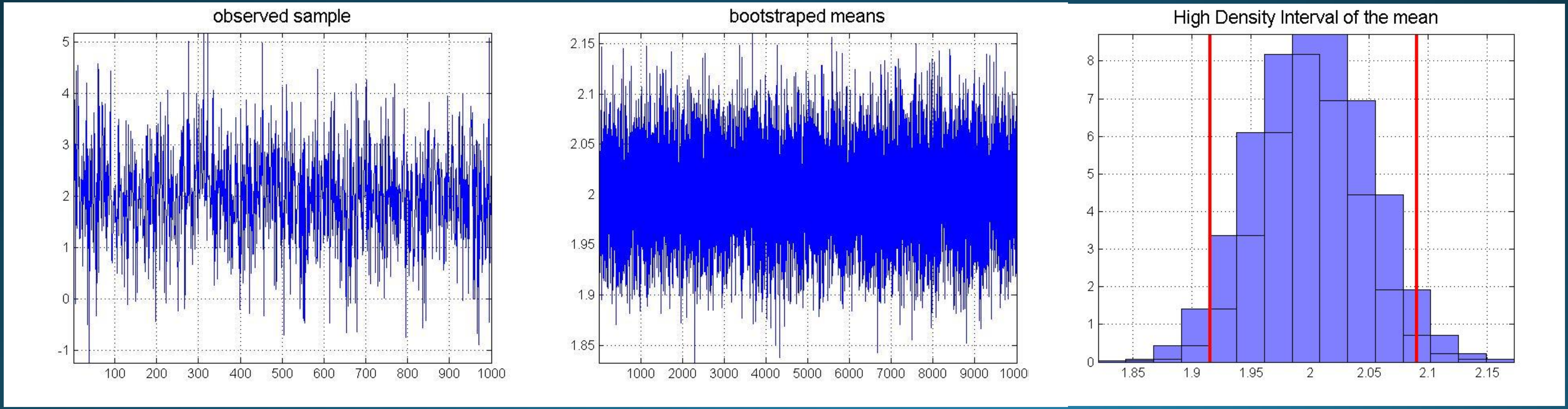
- In the bootstrap, we sample each x_i with replacement, with a probability $1/n$ – *the assumption is that only the observed value are possible values in the parent population*
- In the Bayesian bootstrap, we use a posterior probability distribution for the X_i's.
- Rubin's algorithm: (1) draw u=1:n-1 from uniform
  (2) sort u u(0) =0 and u(n) = 1
  (3) gap = u(i)-u(i-1)
  (4) resample X using prob of xi = gap(i)
  → repeat B times

} Substitute by a Dirichlet

High Density Intervals

- Having the posterior density of means – we can compute the most dense intervals = credible intervals

→ compute the centile distances between bootstrap estimates and take the smallest (i.e. densest)



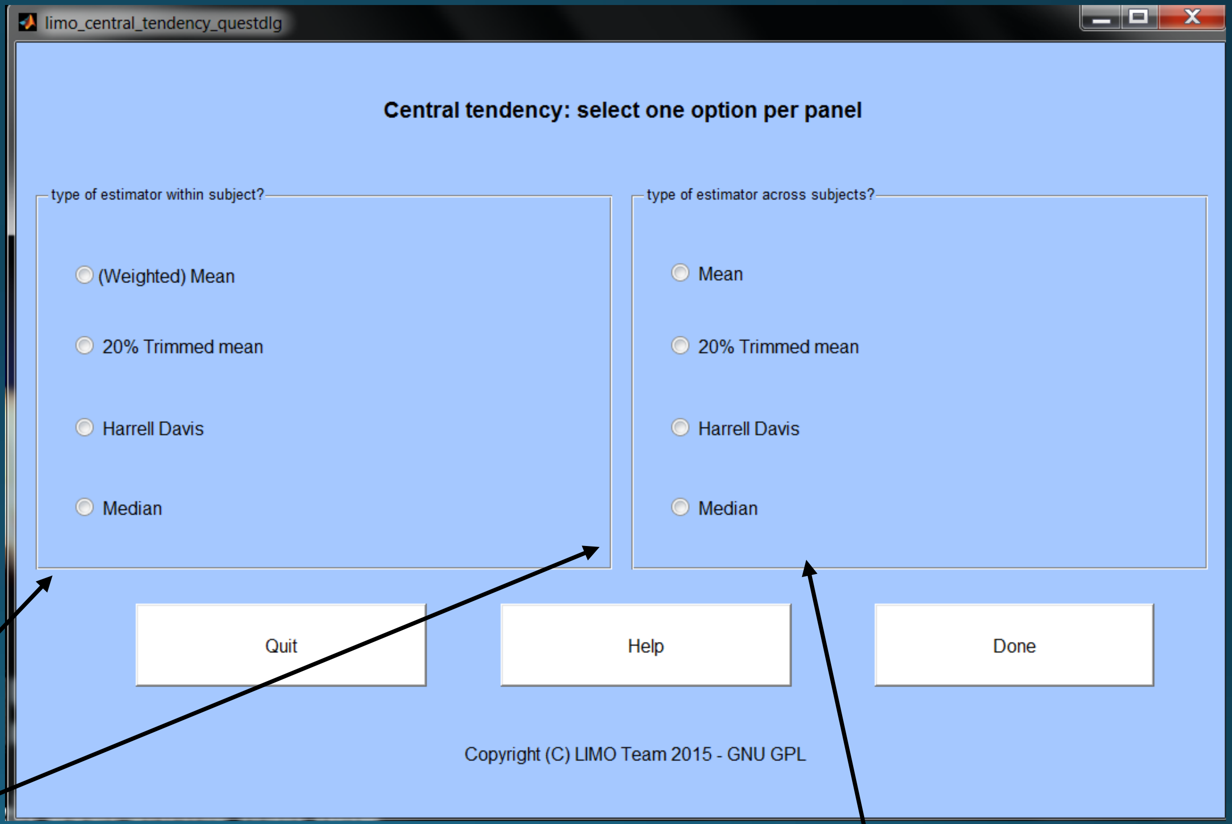
Estimating the mean - revised

- Using posterior densities allows to define the probability of the mean, providing a more natural definition of intervals.

- Frequentist CI: an intervals that fails to cover the population mean 1-alpha percent of the time.
- Bayesian CI: an interval that reflects the probability that mean takes those values 1-alpha percent of the time

Estimators and HDI in one click

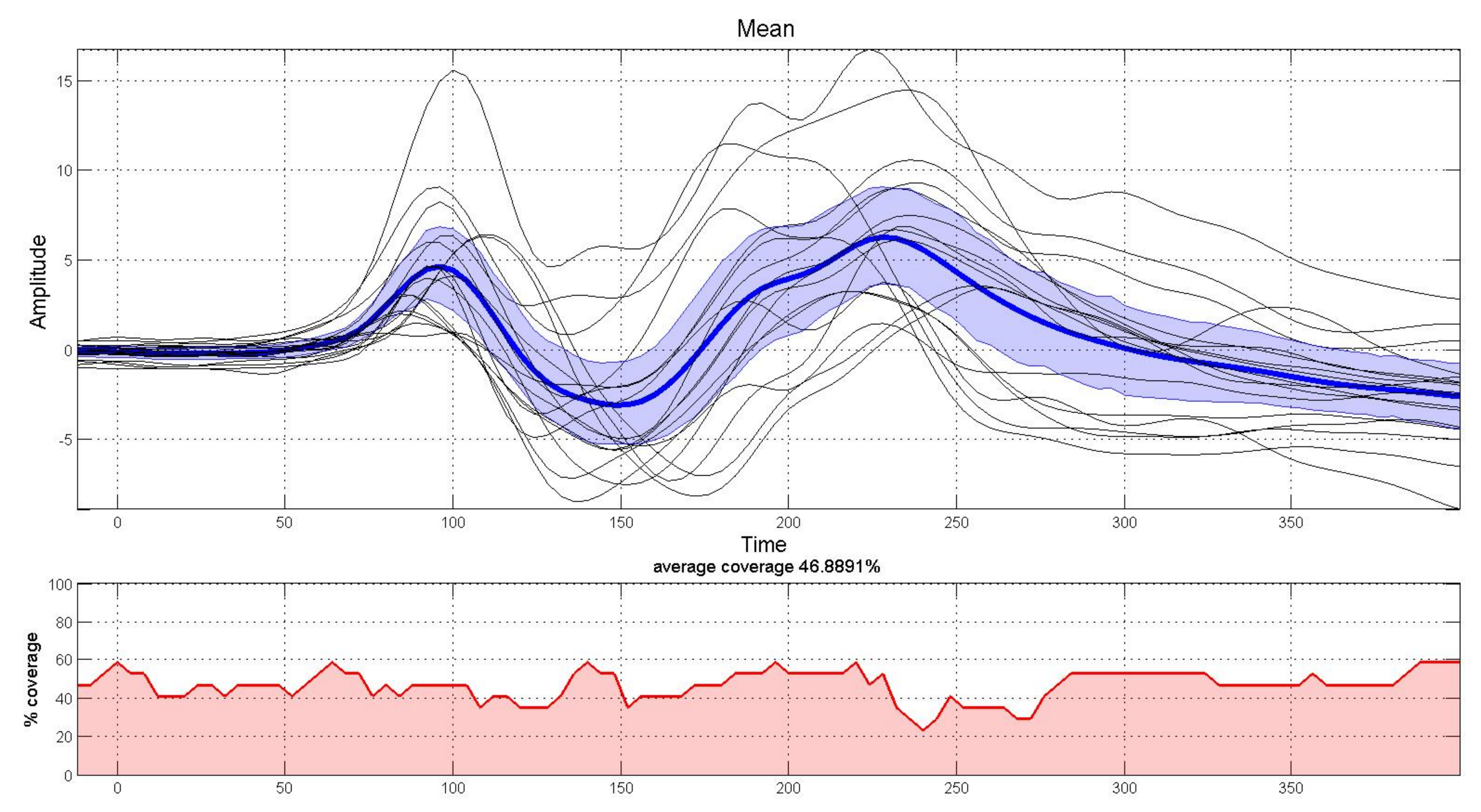
- LIMO EEG 'central tendency and CI' GUI
- Allows computing either on the data or on the betas
- Many different robust estimators 1st and 2nd level.



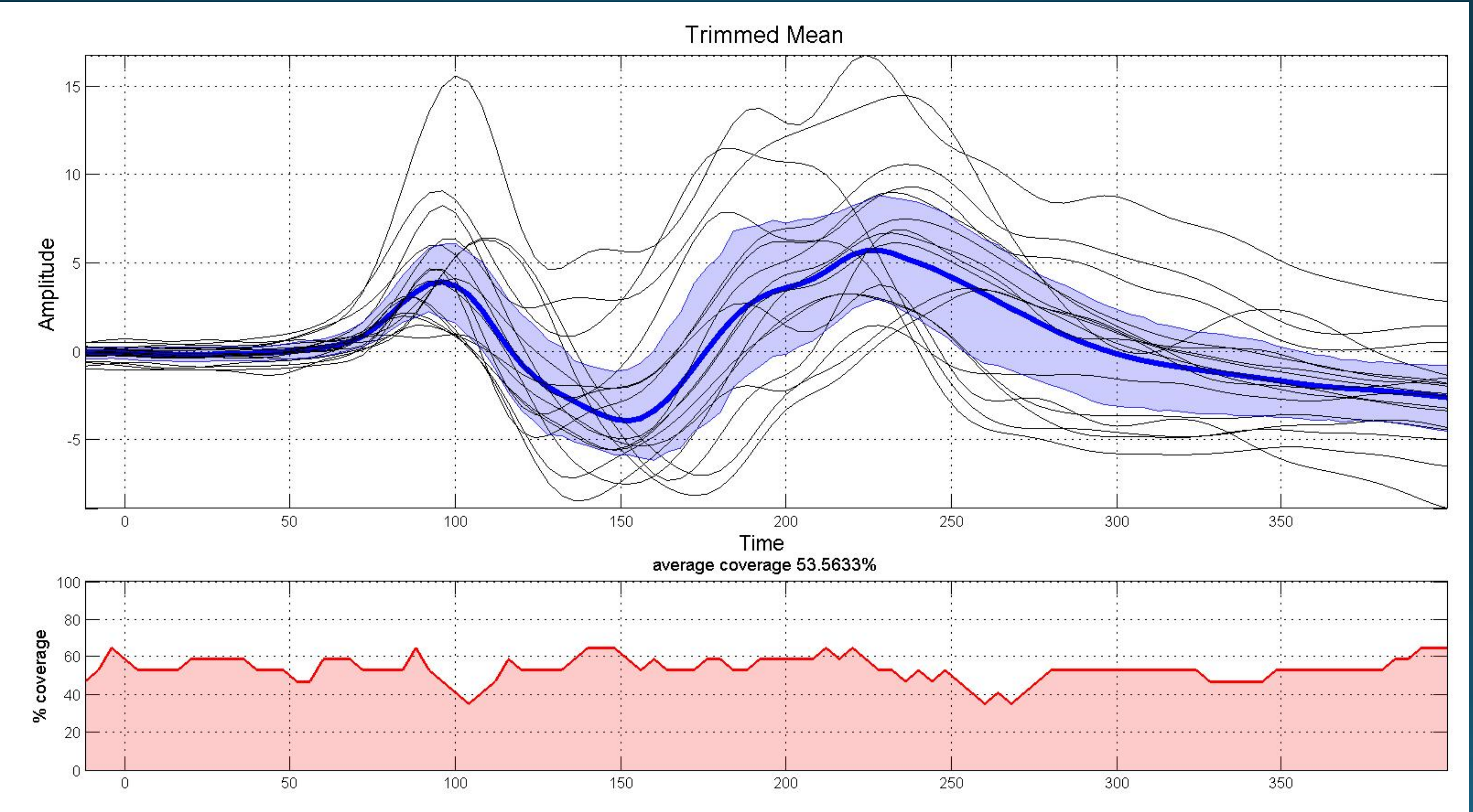
limo_central_tendency_and_ci.m
(2 levels + data handling)

limo_central_estimator.m (estimator and ci)

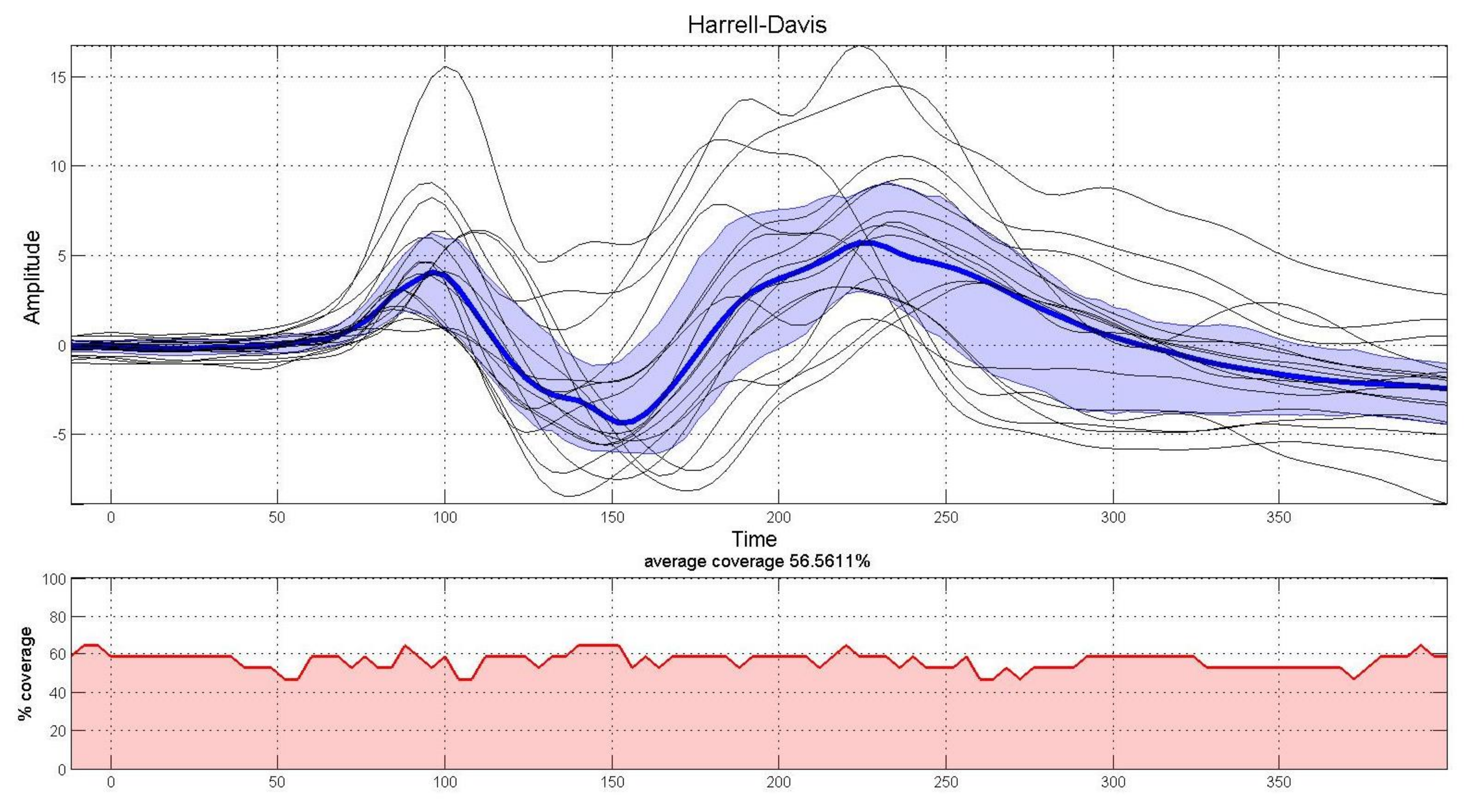
Summary stats do not reflect ERP dynamics



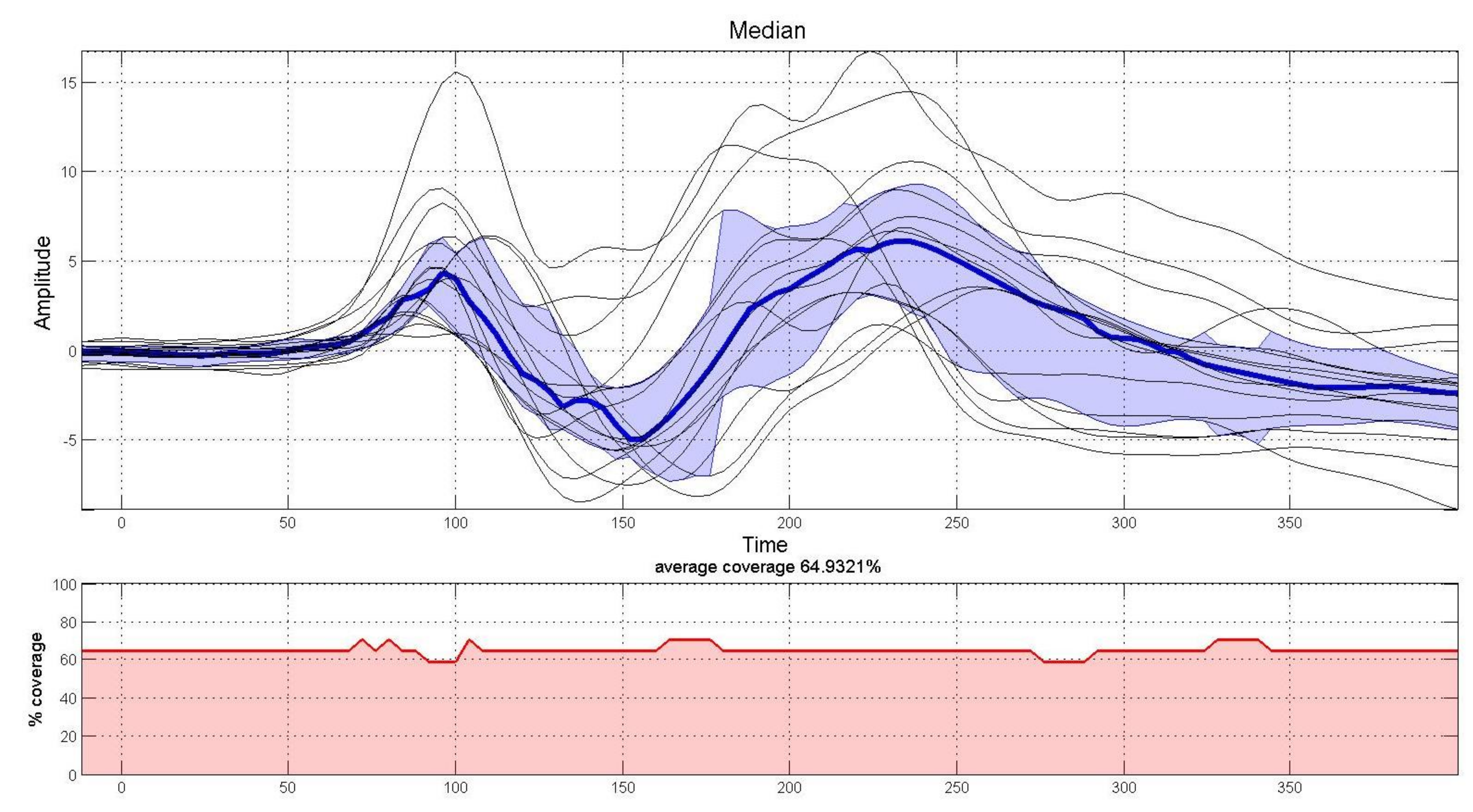
Summary stats do not reflect ERP dynamics



Summary stats do not reflect ERP dynamics



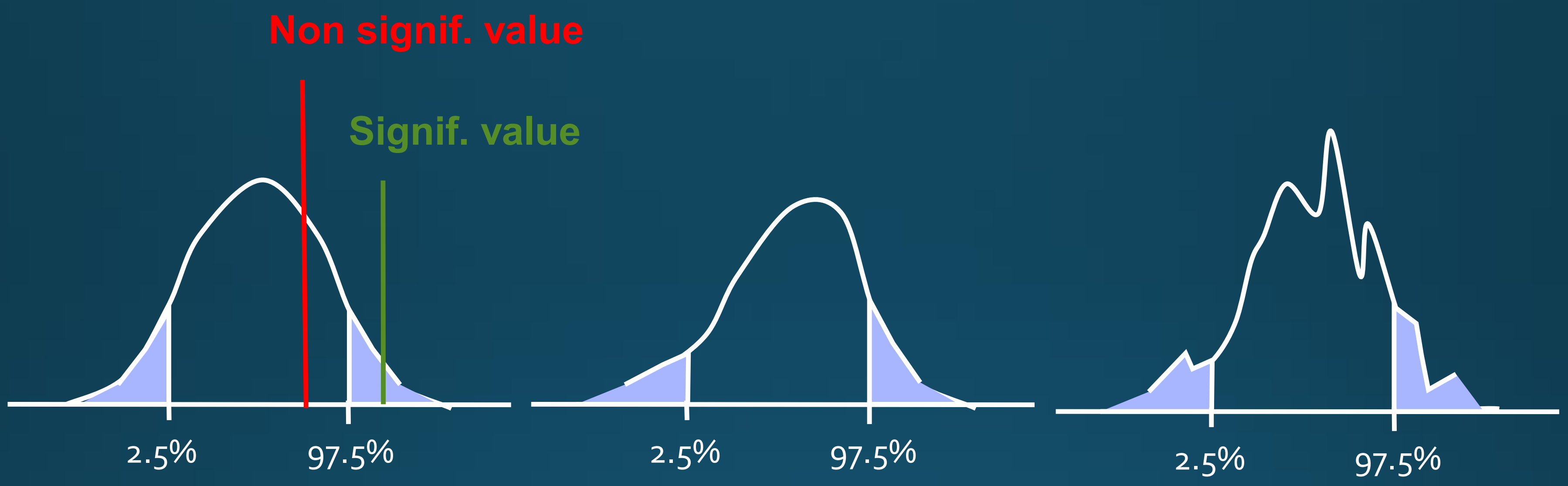
Summary stats do not reflect ERP dynamics



Controlling the FWER using bootstrap

Single subject or group analyses

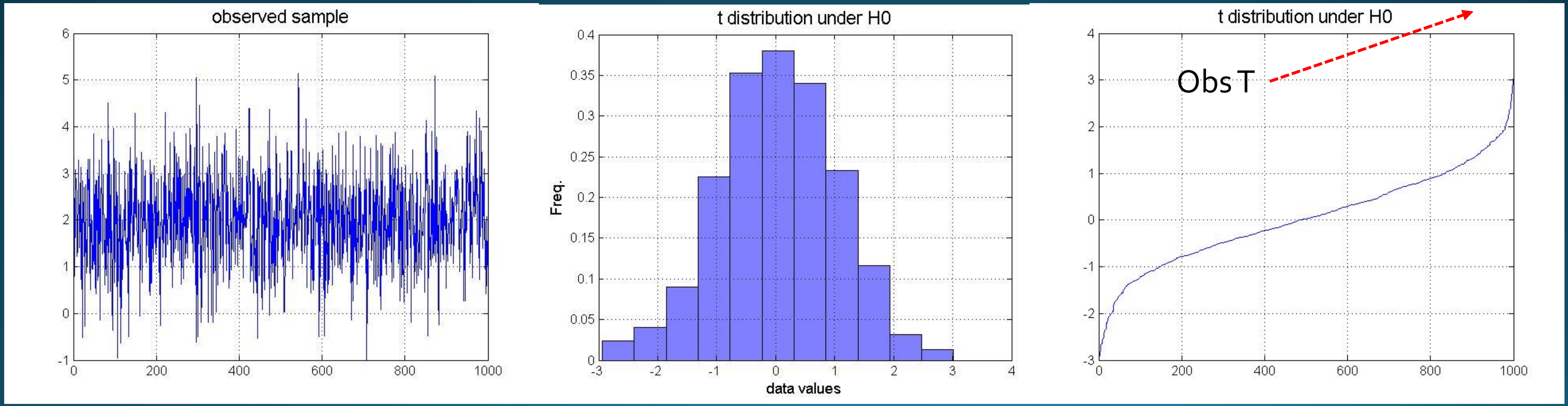
Distributions can take any shape



The bootstrap method allows the bootstrap estimate of the sampling distribution to conform to any shape the data suggest, taking into account the variance and the skewness of the sample. This can be the distribution of estimators (mean, median) or T/F values under Ho or under H1.

Testing the mean with bootstrap

- Let *T be the t-test for the mean*
- *Bootstrap the nullified data computing T* to obtain a distribution and compute the p value*
- *Freq= mean(T>T*) and p = 2*min(Freq,1-Freq)*



Pearson-Newman hypothesis testing

- H0: no effect
- H1: there is an effect

	Results is null	Results is significant
H0 is true	True negative	False positive
H1 is true	False negative	True positive

→ Robust stats reduces false negatives (increase power) by using more stable estimators of distribution parameters

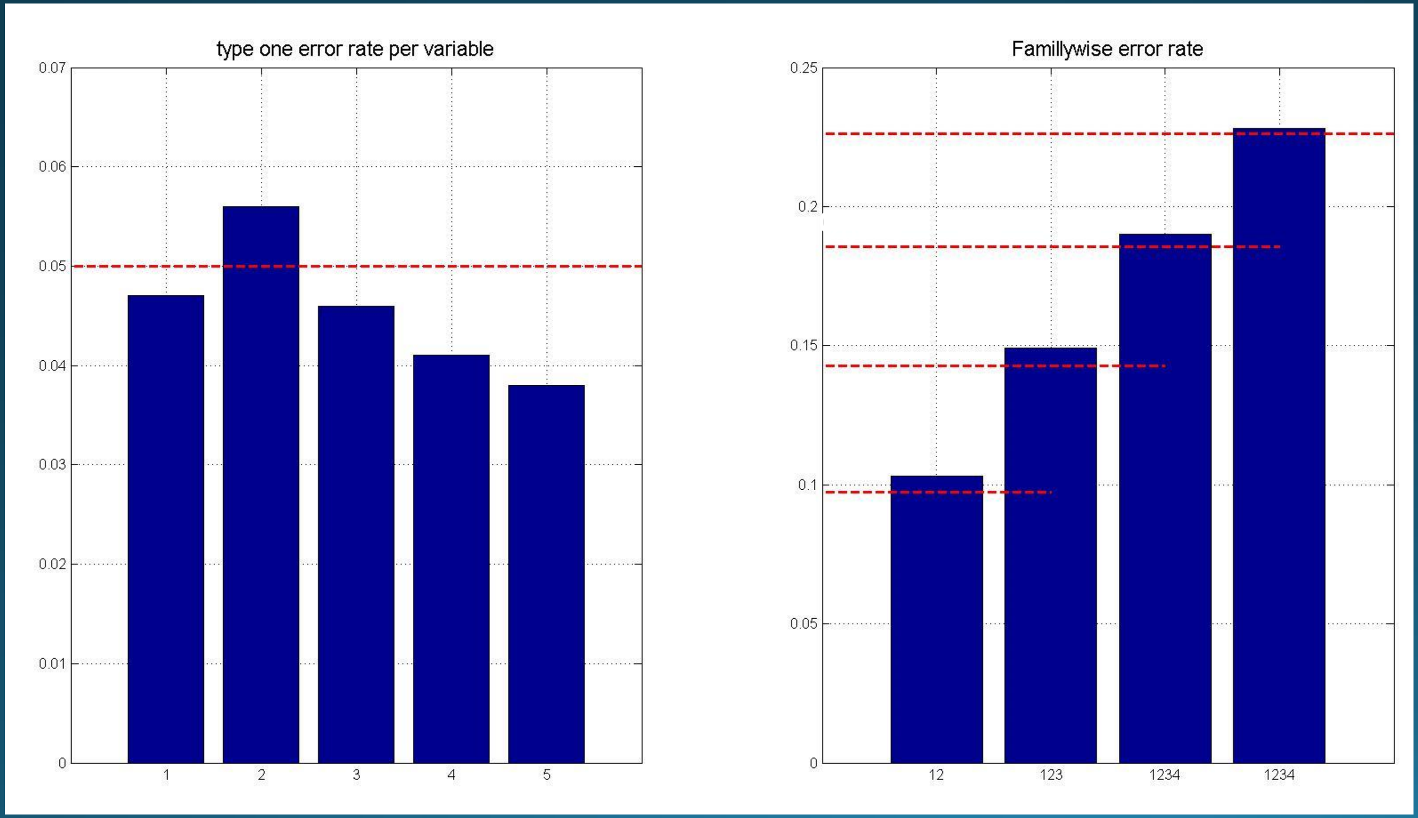
→ Bootstrap controls false positives (i.e. if you choose alpha 0.05 then the test will 'fail' 5% of the time)

What is the problem?

- Assuming tests are independents from each other, the famillywise error rate FWER = 1 - (1 - alpha)^n
- for alpha =5/100, if we do 2 tests we should get about 1-(1-5/100)^2 ~ 9% false positives, if we do 126 electrodes * 150 time frames tests, we should get about 1-(1-5/100)^18900 ~ 100% false positives! i.e. you can't be certain of any of the statistical results you observe

What is the problem?

- Illustration with 5 independent variables from N(0,1)
- Repeat 1000 times and measures type 1 error rate



22%

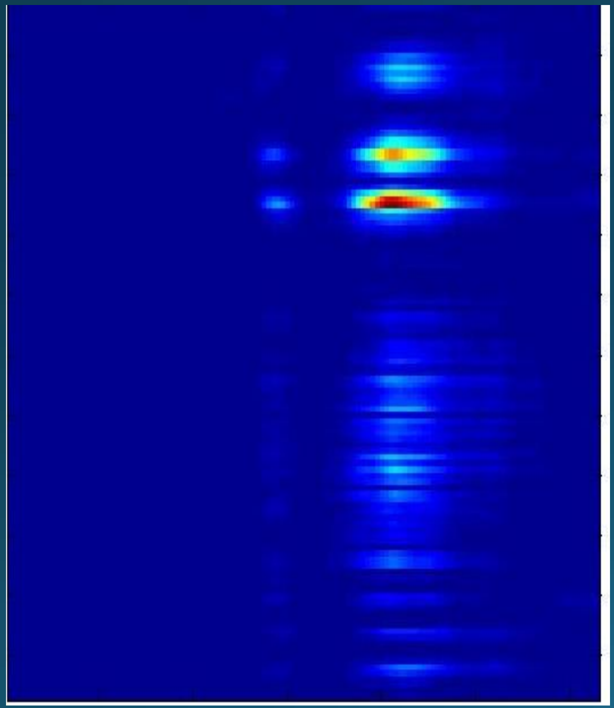
18%

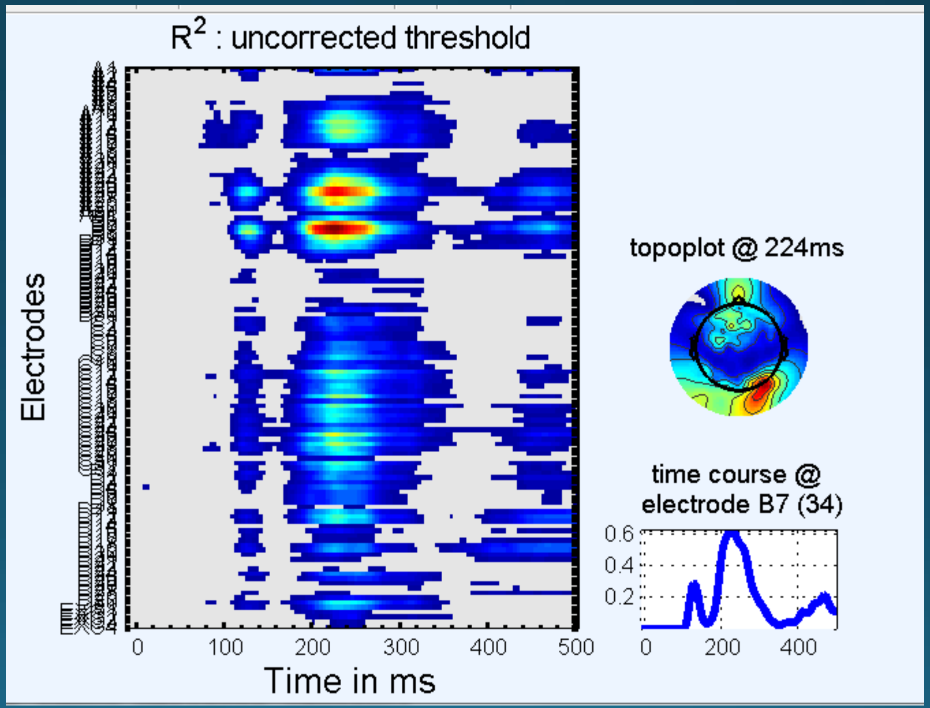
14%

9%

What is the problem?

- Illustration with 18900 independent variables (126 electrodes and 150 time frames)





we know there are false positives – which ones is it?

Family Wise Error rate

- FWER is the probability of making one or more Type I errors in a family of tests, under Ho
- Ho = no effect in any channel/time and/or frequency bins → implies that rejecting a single bin null hyp. is equal to rejecting Ho

$$P(\cup_{i \in V}\{T_i \geq u\} | H_0) \leq \propto$$

We want to find the threshod u such the prob of any false positives under Ho is controlled at value alpha

False Discovery Rate

- In the LIMO EEG toolbox, we control for the false positve rate, i.e. the probability to make alpha percent of errors under Ho (false positive among all results). In EEGLAB/ERPLAB, you have the option to choose a correction based on FDR

	Results is null	Results is significant
Ho is true	True negative	False positive
H1 is true	False negative	True positive

FDR = False positives / All positives
Controls the number of false positves among all positives i.e. it does not control FWER !

Bonferroni Correction

Bonferroni correction allows to keep the FWER at 5% by simply dividing alpha by the number of tests

$$P(T_i \geq u | H0) \leq \frac{\propto}{m}$$

Find u to keep the FWER < α/m

$$\text{FWER} = P(\cup_{i \in V}\{T_i \geq u\} | H_0) \leq \propto$$

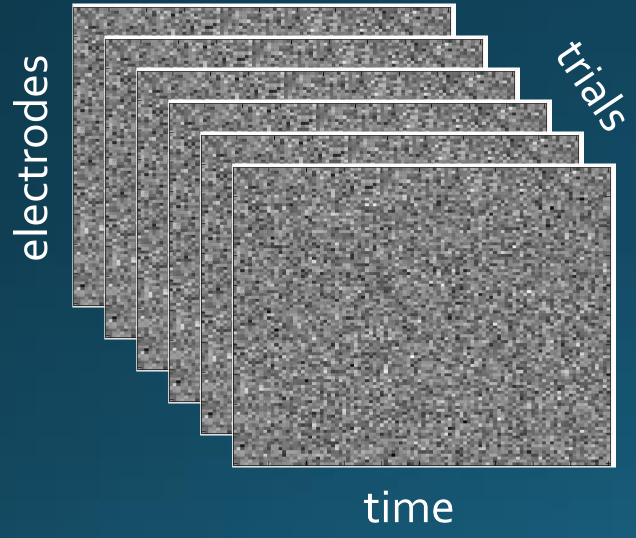
$$\leq \sum P(T_i \geq u | H0)$$

Boole's inequality

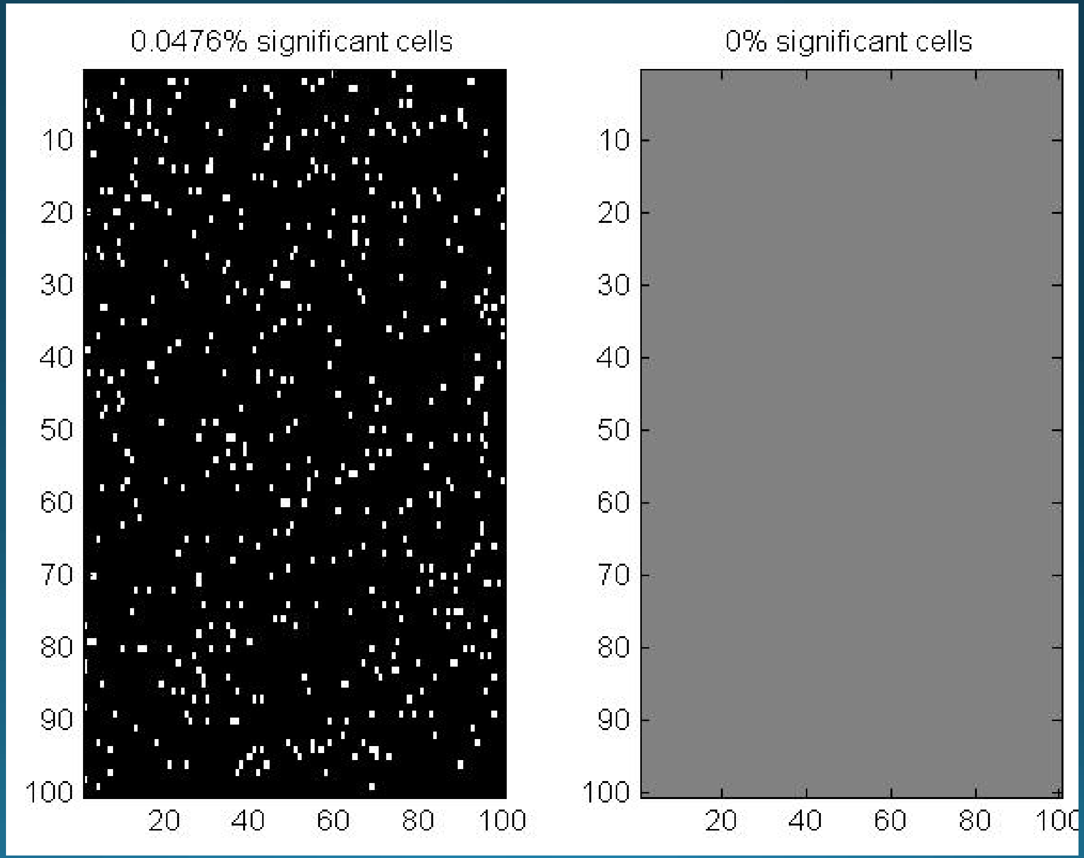
$$\leq \sum_i \frac{\propto}{m} = \propto$$

Bonferroni Correction

- Assumes all tests are independent
- Too conservative



One sample t test > 0 ?

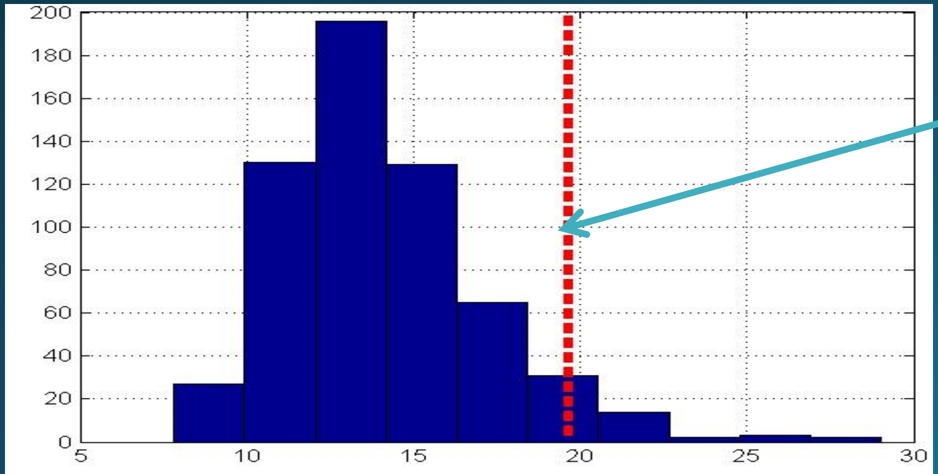


Correcting using the maximum under Ho

Maximum Statistics

- Since the FWER is the prob that any stats > u, then the FWER is also the prob. that the max stats > u
- All we have to do, is thus to find a threshold u such that the max only exceed u alpha percent of the time.

Distribution of max F value under Ho

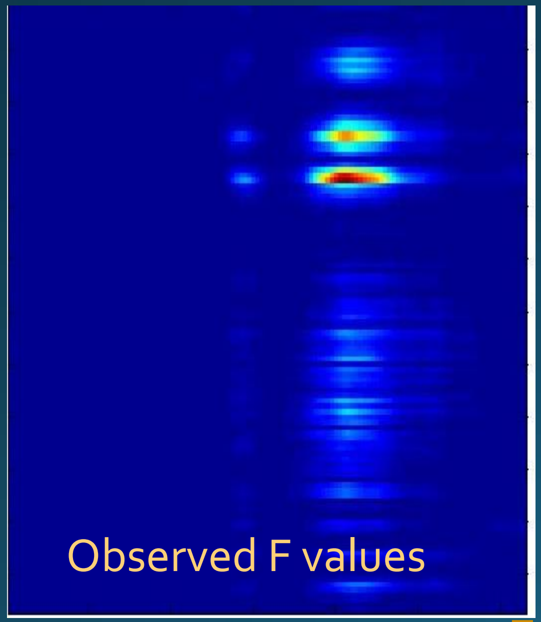


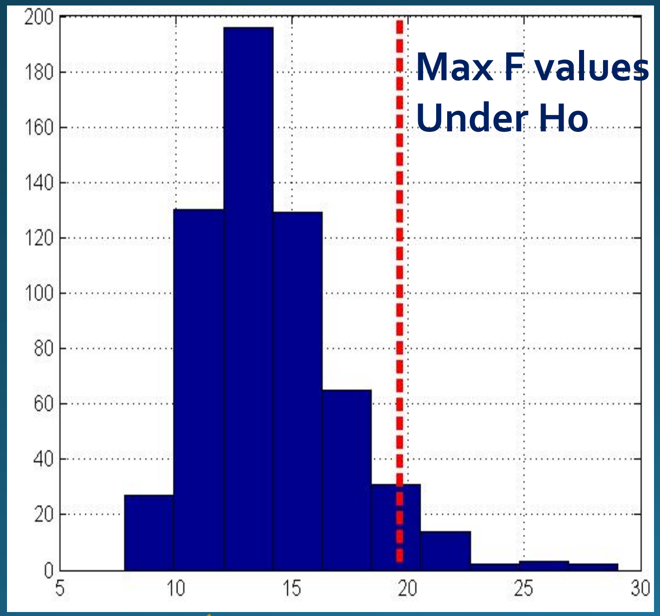
Threshold u such alpha
Percent are above it

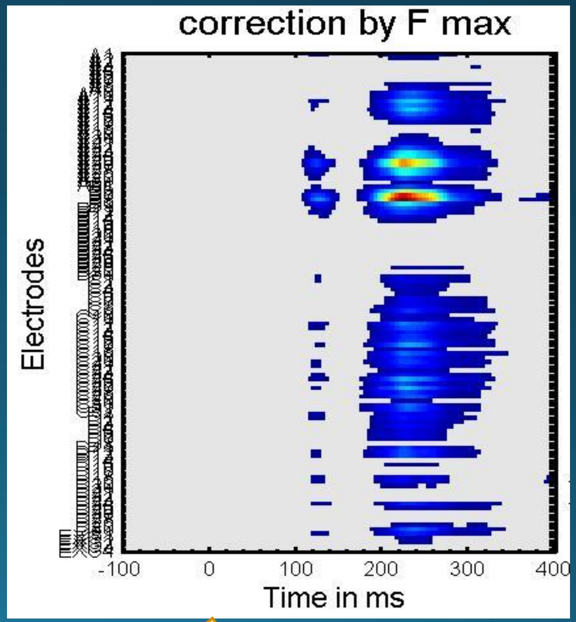
```
[mask,p_val] =
limo_max_correction(A,B,p)

A = observed stats (F,T^2)
B = bootstrapped data
p = alpha value
```

Maximum Statistics

- Estimate the distribution of max under Ho (bootstrap) and simply threshold the observed results a threshold u
- Still assumes all tests are independent



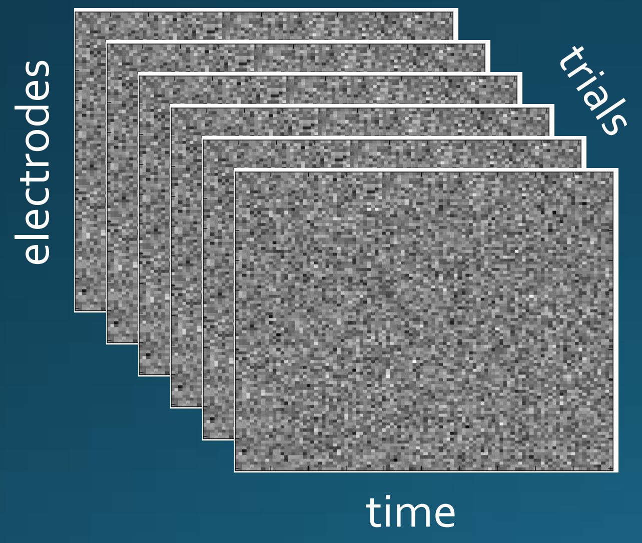




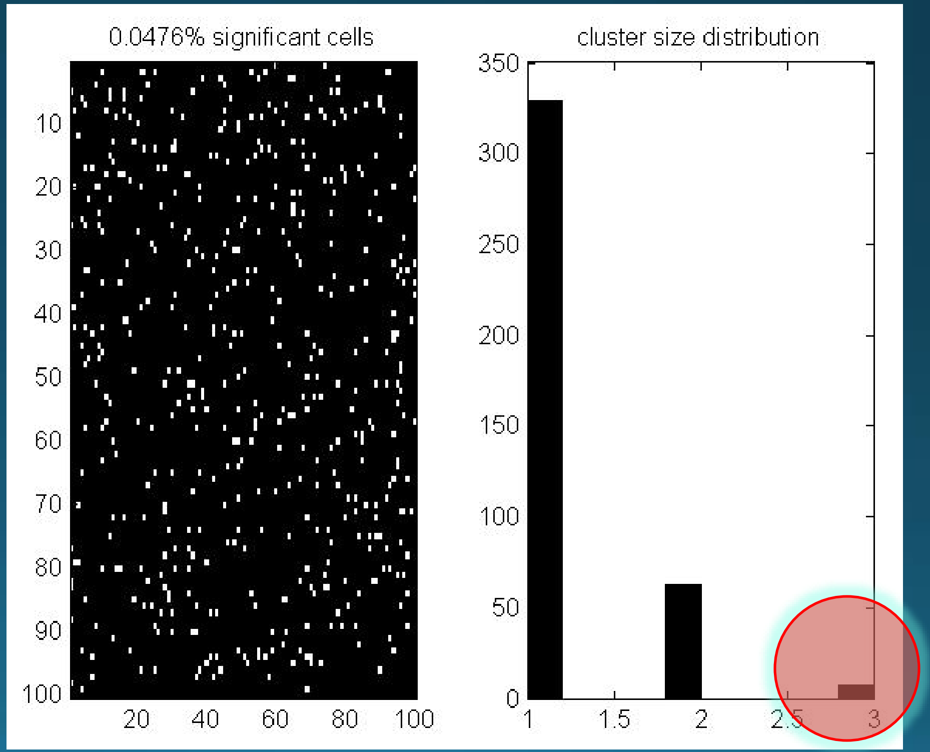
Cluster Mass for MEEG

Let's analyse clusters

- In MEEG, instead of the max, we consider clusters as it is much less likely that statistics are significant in isolation

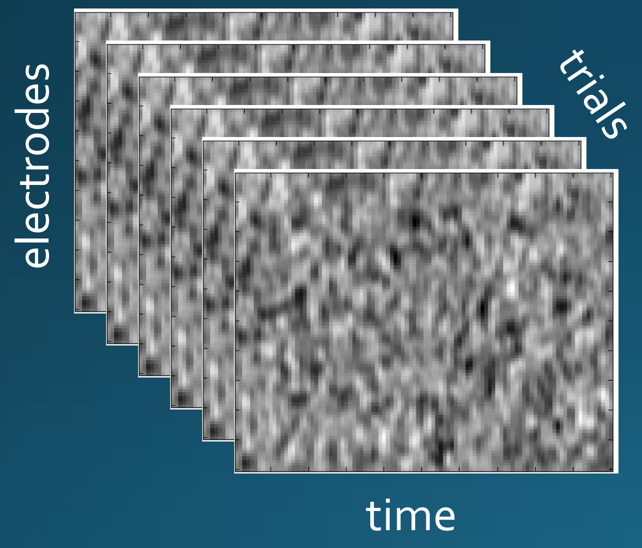


One sample t test > 0 ?

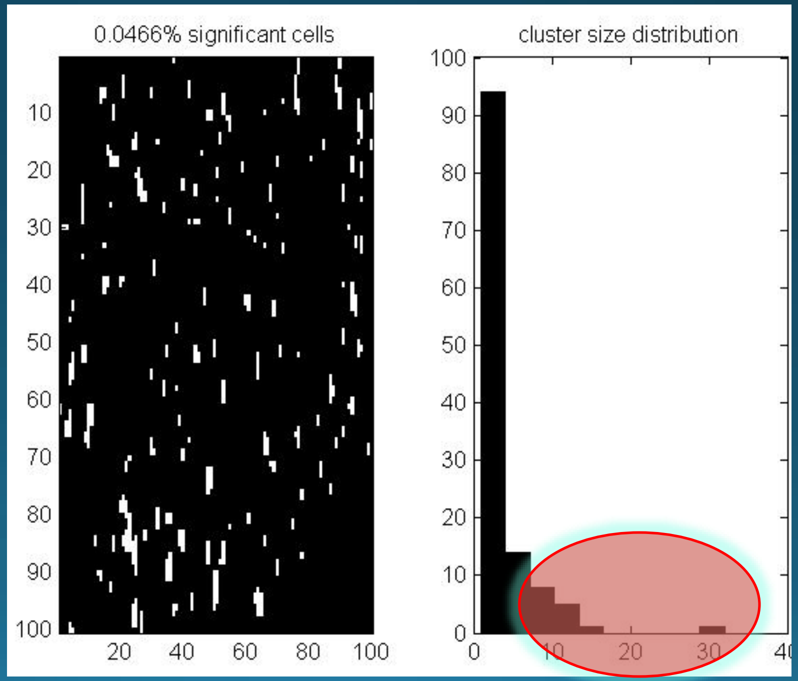


Let's analyse clusters

- In MEEG, instead of the max, we consider clusters as it is much less likely that statistics are significant in isolation because data are smooth in space and time!



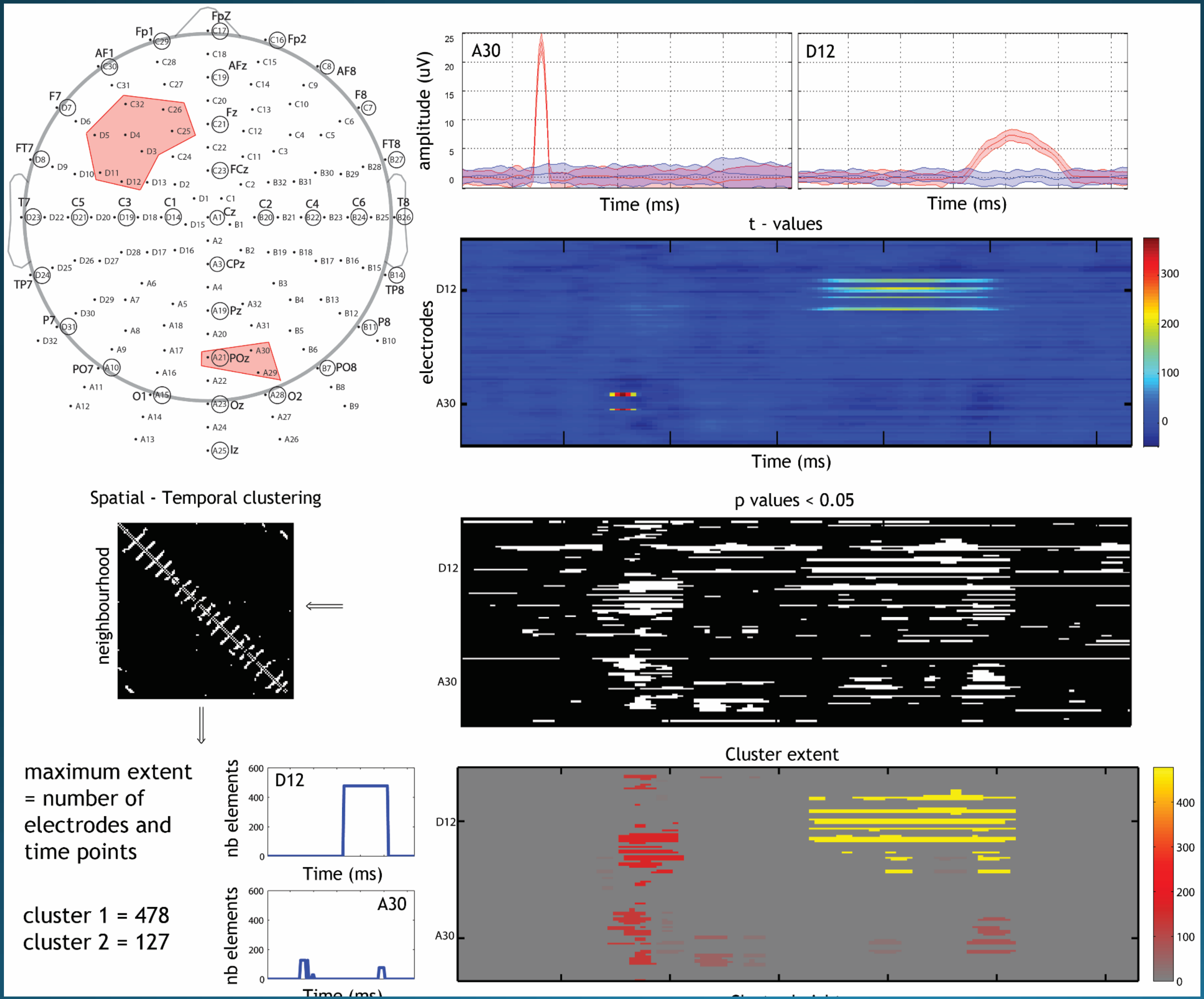
One sample t test > 0 ?

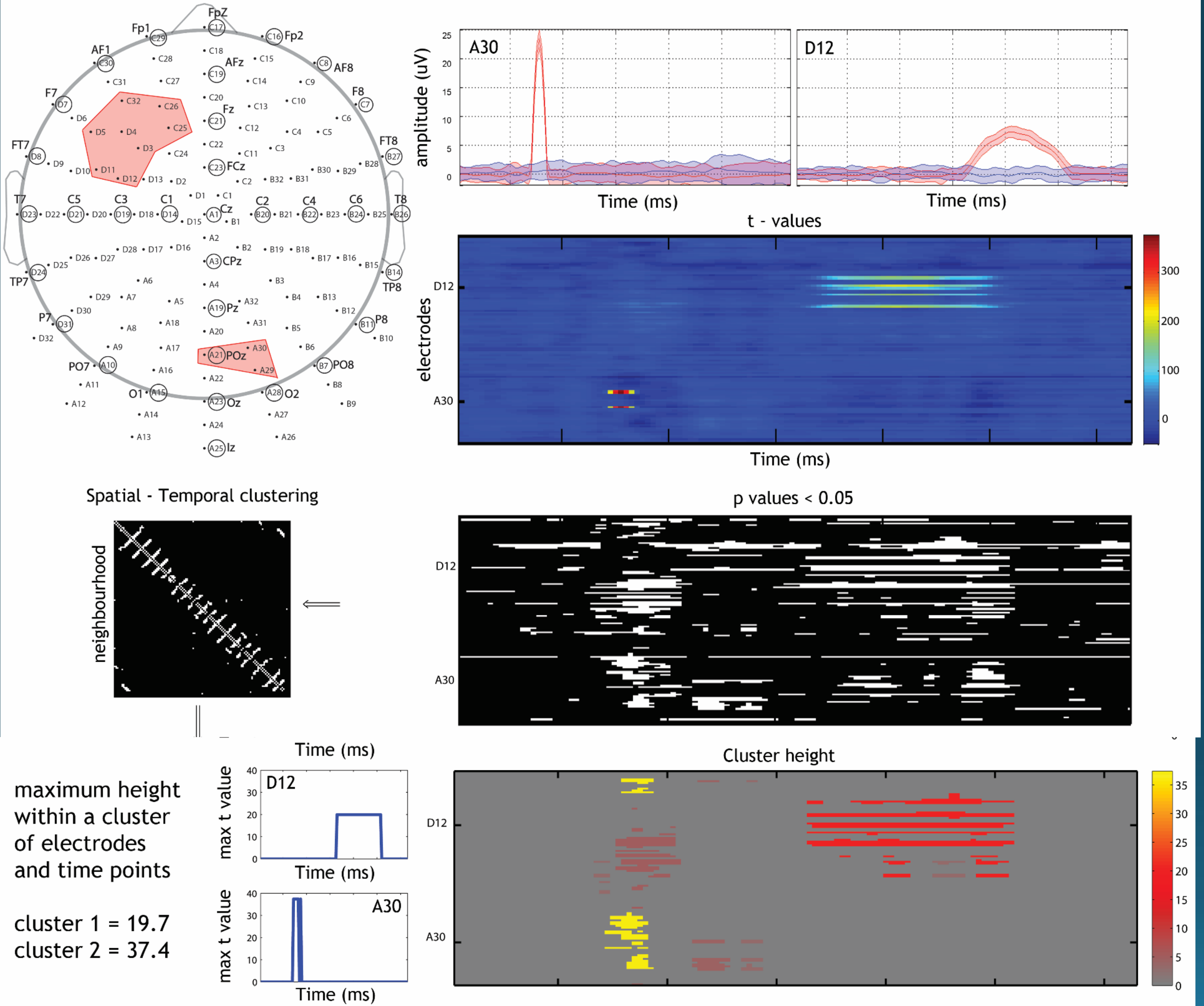


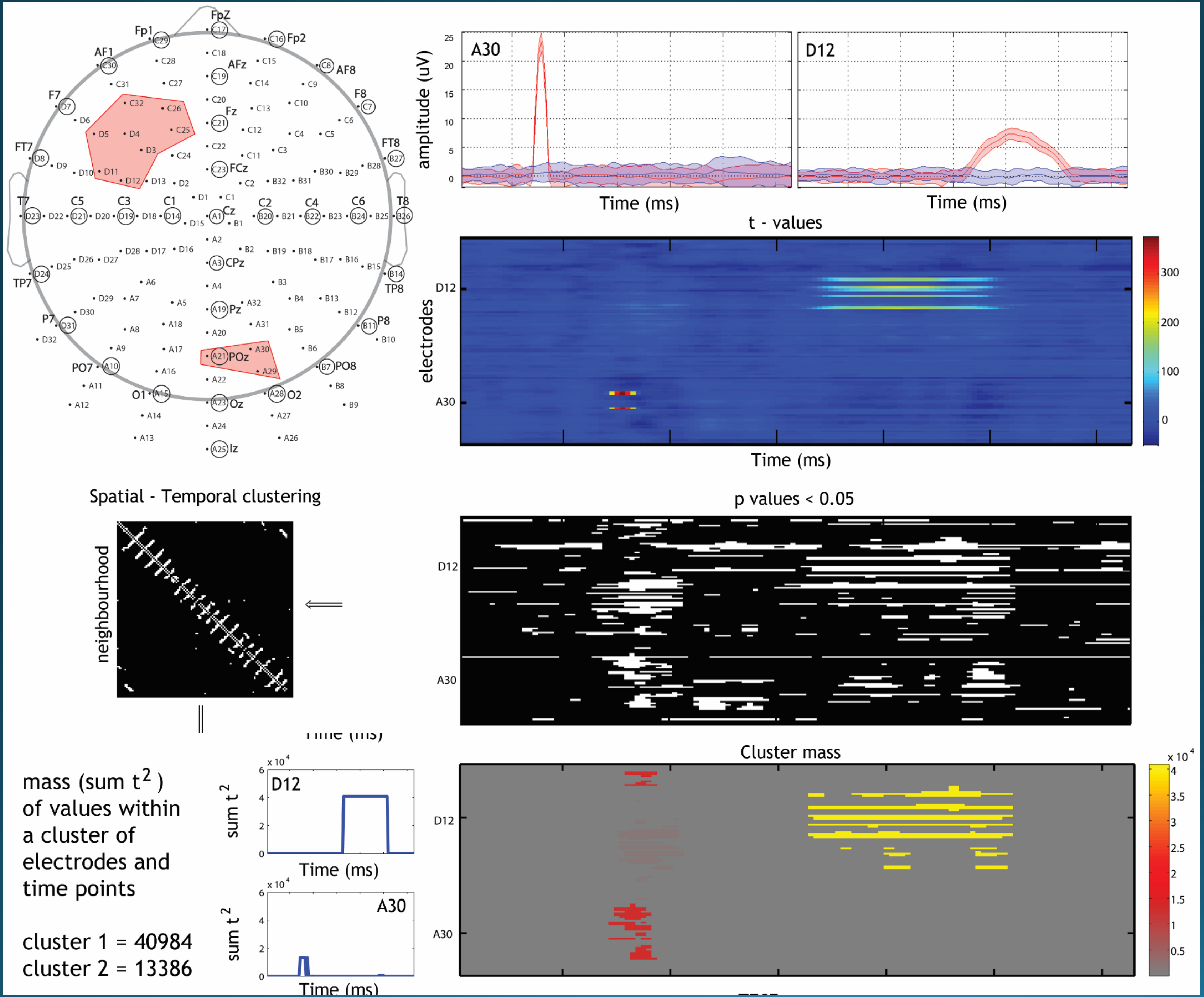
The clustering solution

- Clustering is a good option because it accounts for topological features in the data. Techniques like Bonferroni, FDR, max(stats) control the FWER but independently of the correlation between tests.

- To use clustering we need to consider cluster statistics rather than individual statistics
- Cluster statistics depend on (i) the cluster size, which depends on the data at hand (how correlated data are in space and in time/frequency), and (ii) the strength of the signal (how strong are the t, F values in a cluster) or (iii) a combination of both.

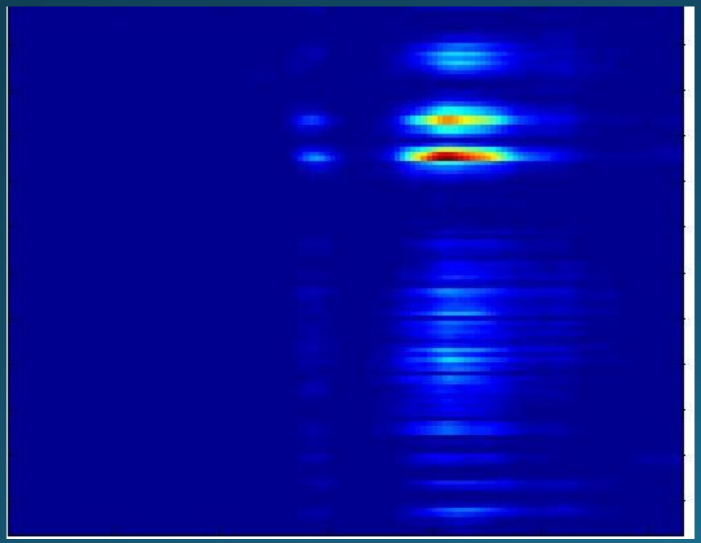

Spatial - Temporal clustering
neighbourhood
t - values
electrodes
amplitude (uV)
Time (ms)
A30
D12
p values < 0.05
Cluster extent
maximum extent = number of electrodes and time points
cluster 1 = 478
cluster 2 = 127
nb elements

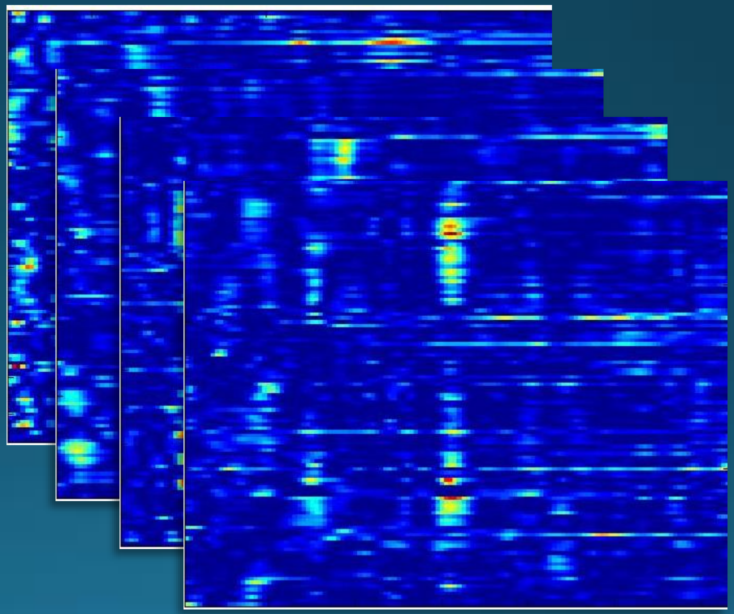

A30
D12
amplitude (uV)
Time (ms)
Time (ms)
t - values
electrodes
D12
A30
Time (ms)
Spatial - Temporal clustering
neighbourhood
p values < 0.05
D12
A30
Time (ms)
Cluster height
D12
A30
max t value
Time (ms)
D12
A30
maximum height
within a cluster
of electrodes
and time points
cluster 1 = 19.7
cluster 2 = 37.4


A30
D12
amplitude (uV)
Time (ms)
t - values
electrodes
Spatial - Temporal clustering
neighbourhood
p values < 0.05
Cluster mass
mass (sum t2) of values within a cluster of electrodes and time points
cluster 1 = 40984
cluster 2 = 13386
sum t2

The clustering solution

- In LIMO EEG, we bootstrap the data under Ho: center the data or break the link between the design matrix and the data and then resample and test. This way we can find u for a single bin, the whole space, or for clusters.



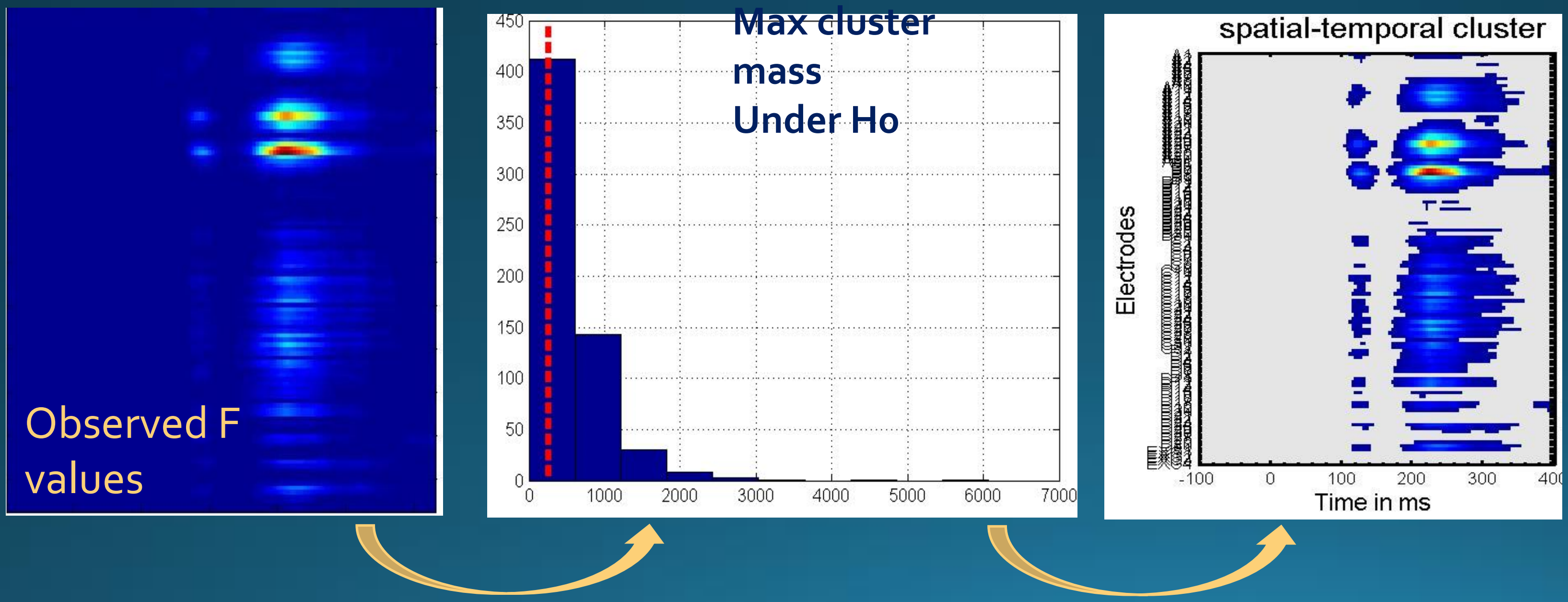


Observed F values → F values under Ho

The clustering solution

- Spatial-Temporal clustering: for each bootstrap, threshold at alpha and record the max(cluster mass), i.e. sum of F values within a cluster. Then threshold the observed clusters based on there mass using this distribution → accounts for correlations in space and time.

```
[mask,cluster_p] = limo_cluster_correction(A,AP,B,BP,neighbouring,method,p)
```



Loss of resolution: inference is about the cluster, not max in time or a specific electrode !

TFCE for MEEG

Threshold Free Cluster Enhancement

- Threshold Free Cluster Enhancement (TFCE): Integrate the cluster mass at multiple thresholds. A TFCE score is thus obtain per cell but the value is a weighted function of the statistics by it's belonging to a cluster. (limo_tfce.m followed by limo_max_correction)

Smith & Nichols 2009 NeuroImage 44

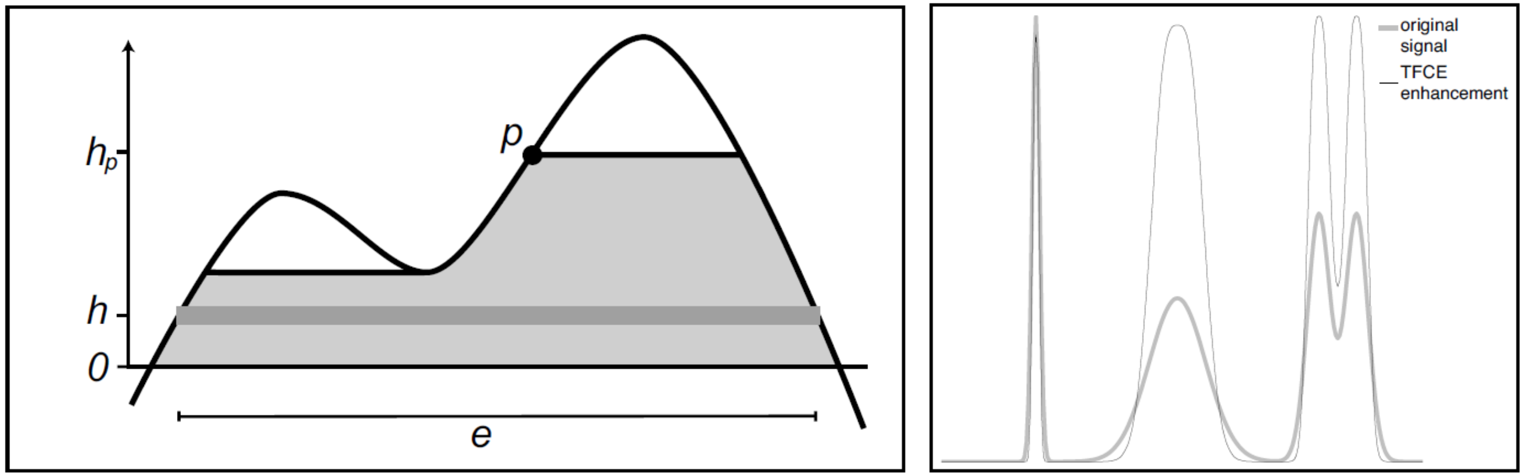
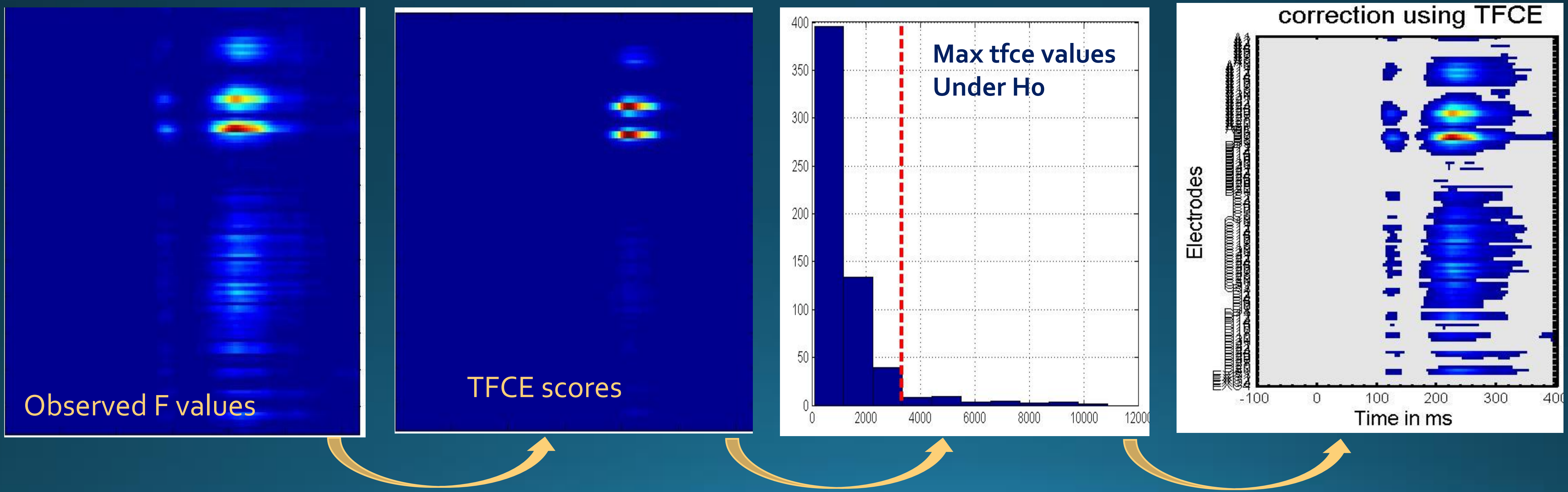


Figure 1: Illustration of the TFCE approach. Left: The TFCE score at voxel p is given by the sum of the scores of all incremental supporting sections (one such is shown as the dark grey band) within the area of "support" of p (light grey). The score for each section is a simple function of its height h and extent e. Right: Example input image and TFCE-enhanced output. The input contains a focal, high signal, a much more spatially extended, lower, signal and a pair of overlapping signals of intermediate extent and height. The TFCE output has the same maximal values for all three cases, and preserves the distinct local maxima in the third case.

Threshold Free Cluster Enhancement

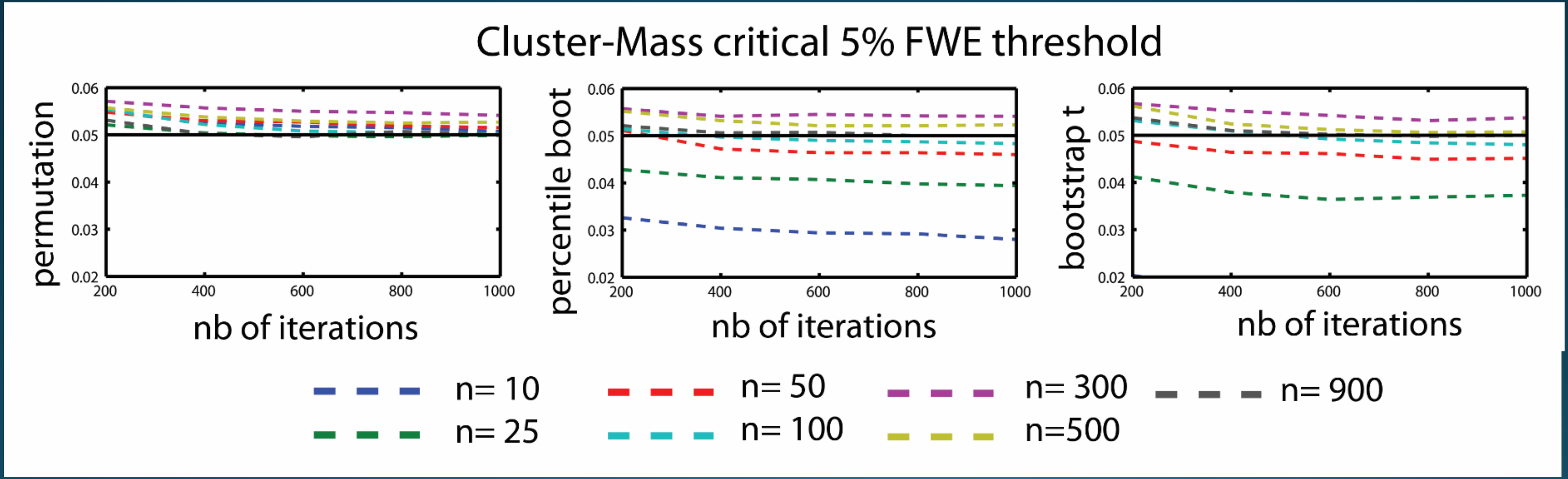
- Threshold Free Cluster Enhancement (TFCE): Integrate the cluster mass at multiple thresholds. A TFCE score is thus obtain per cell but the value is a weighted function of the statistics by it's belonging to a cluster. As before, bootstrap under Ho and get max(tfce).



Excellent resolution: inference is about cells, but we accounted for space/time dependence

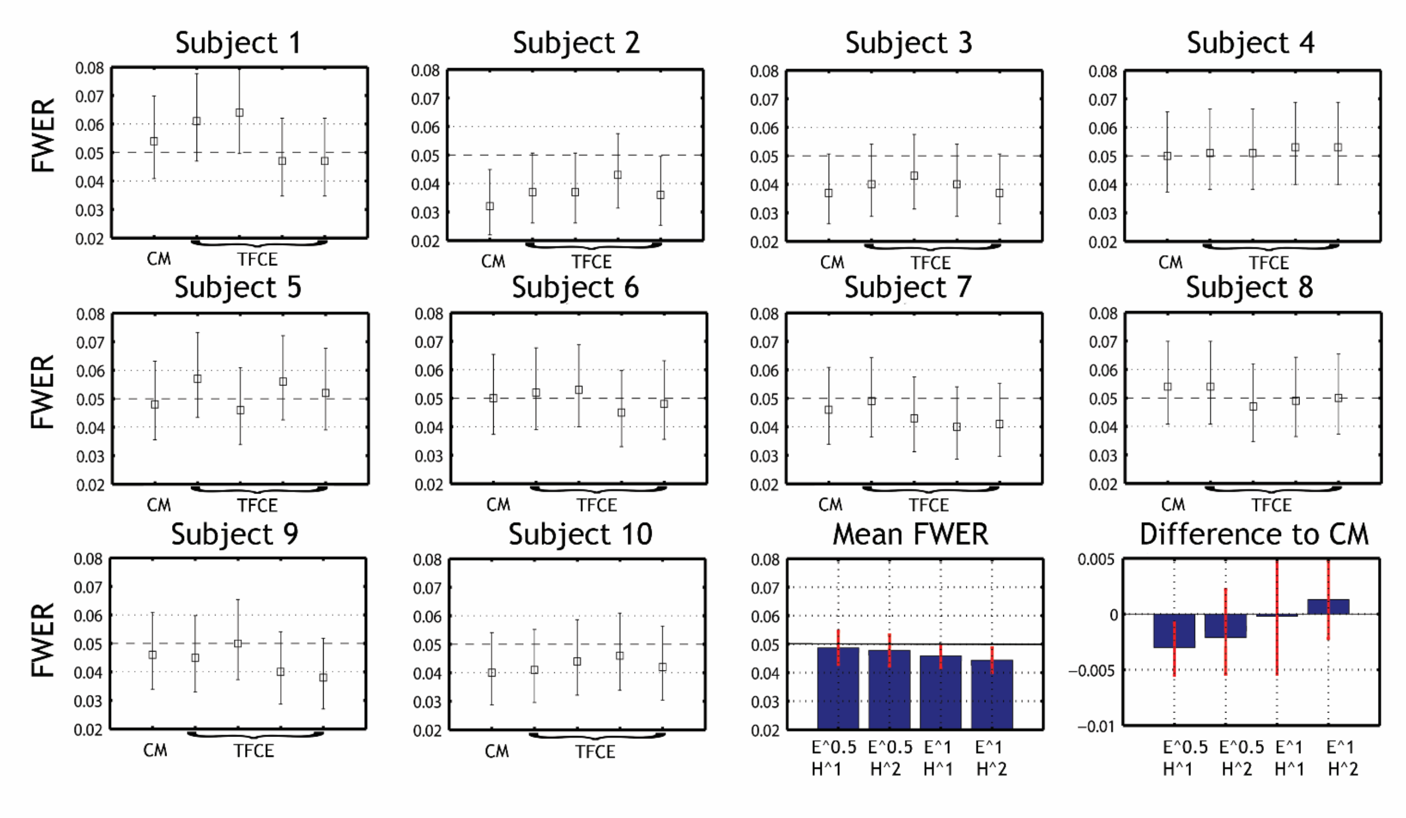
Review of techniques

- All techniques (including permutation not shown here) control well the FWER under Ho with some limitations for small sample sizes



Review of techniques

- All techniques (including permutation not shown here) control well the FWER under Ho with some limitations for small sample sizes



MCC summary

- Simulation work show that overall permutation / bootstrap / cluster-mass / TFCE control well the type 1 FWER.
- a minimum of 800 iterations are necessary to obtain stable results
- for low critical family-wise error rates (e.g. p = 1%), permutations can be too liberal;
- For within subject bootstrap, a min of 50 trials per condition is requested at the risk to be too conservative

Conclusions

- When performing multiple tests, statistical correction MUST be applied.

- All techniques provide a FWER at the specified level but not all techniques have the same power.
- Spatial-temporal clustering and TFCE seem to provide good estimates, with TFCE giving higher spatio-temporal inference resolution, but at the cost of long computing time.

References

- **Maris, E. & Oostenveld, R. (2007).** Nonparametric statistical testing of EEG- and MEG-data. Journal of *Neuroscience Methods*, *164*, 177-190
- **Pernet, C, Chauveau, N., Gaspar, C. & Rousselet, G.A. (2011).** LIMO EEGLIMO: a toolbox for hierarchical LInear MOdeling of ElectroEncephaloGraphic data. *Computational Intelligence and Neuroscience, Volume 2011*
- **Pernet, C., Latinus, M., Nichols, T. & Rousselet, G.A. (2015).** Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study. Journal of *Neuroscience Methods*, *250*, 85-93