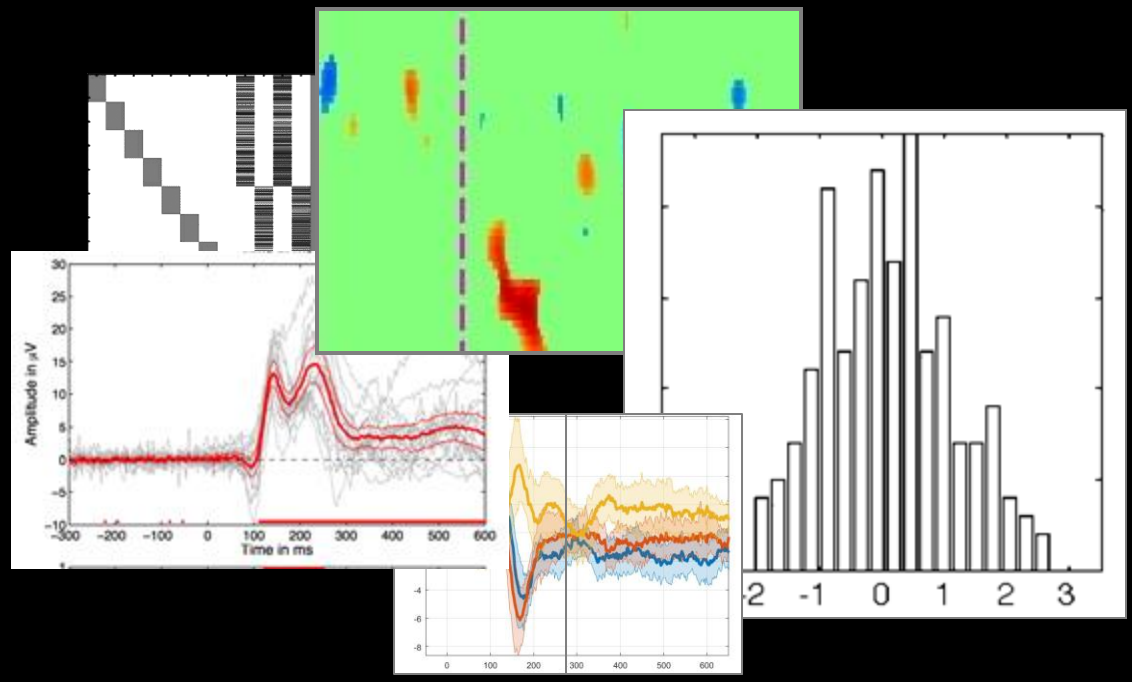
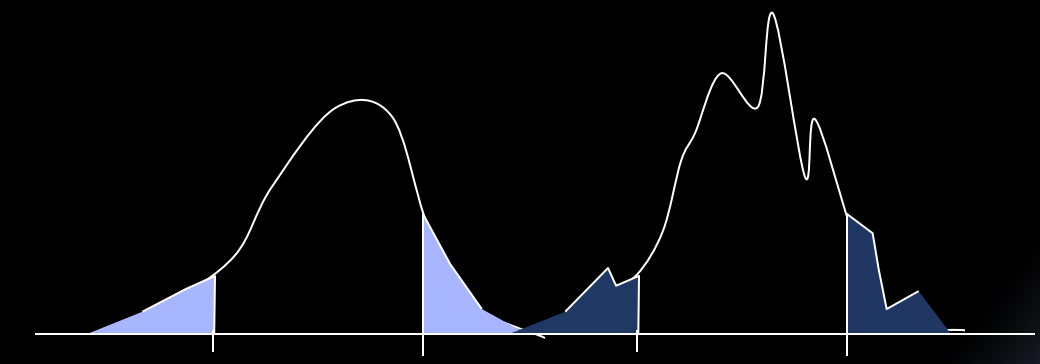
Robust statistics applied to EEG data

Arnaud Delorme



Robust statistics outline

- Parametric & non-parametric statistics
- How to increase robustness
- Bootstrap and permutation methods
- Correction for multiple comparisons
- Statistical analysis using GLMs



Take-home messages

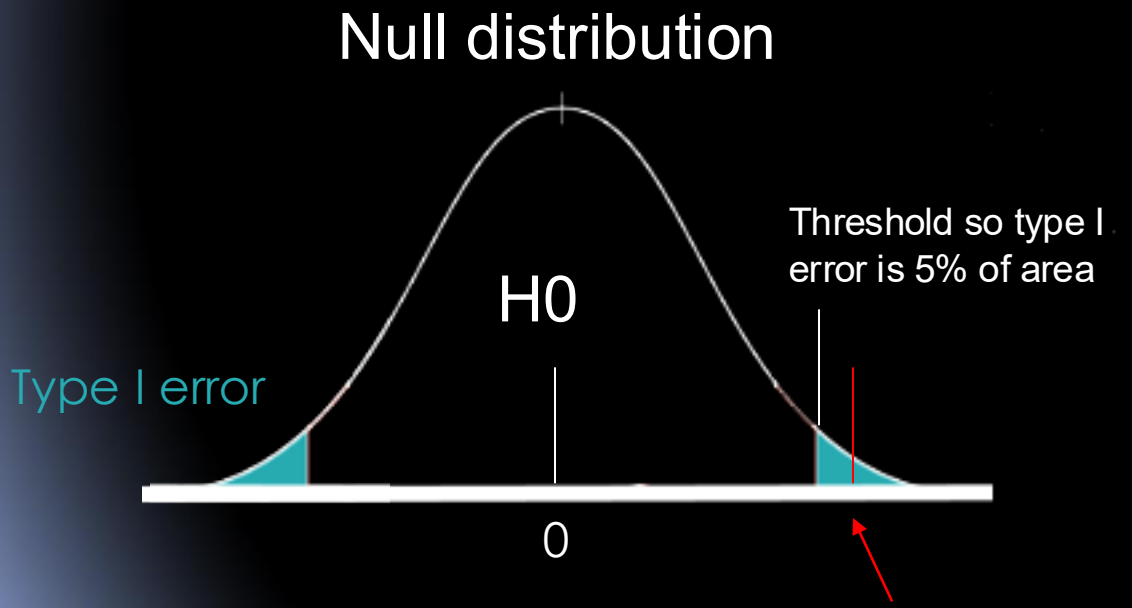
- *Look at your data! Show your data!*
- *A perfect & universal statistical recipe does not exist*
- *Keep exploring*

Inferential statistics

NULL hypothesis (2 conditions):

- e.g. no difference between 2 sets of values
- Compare actual difference between average of 2 sets with the null distribution
- If in the tail, significant

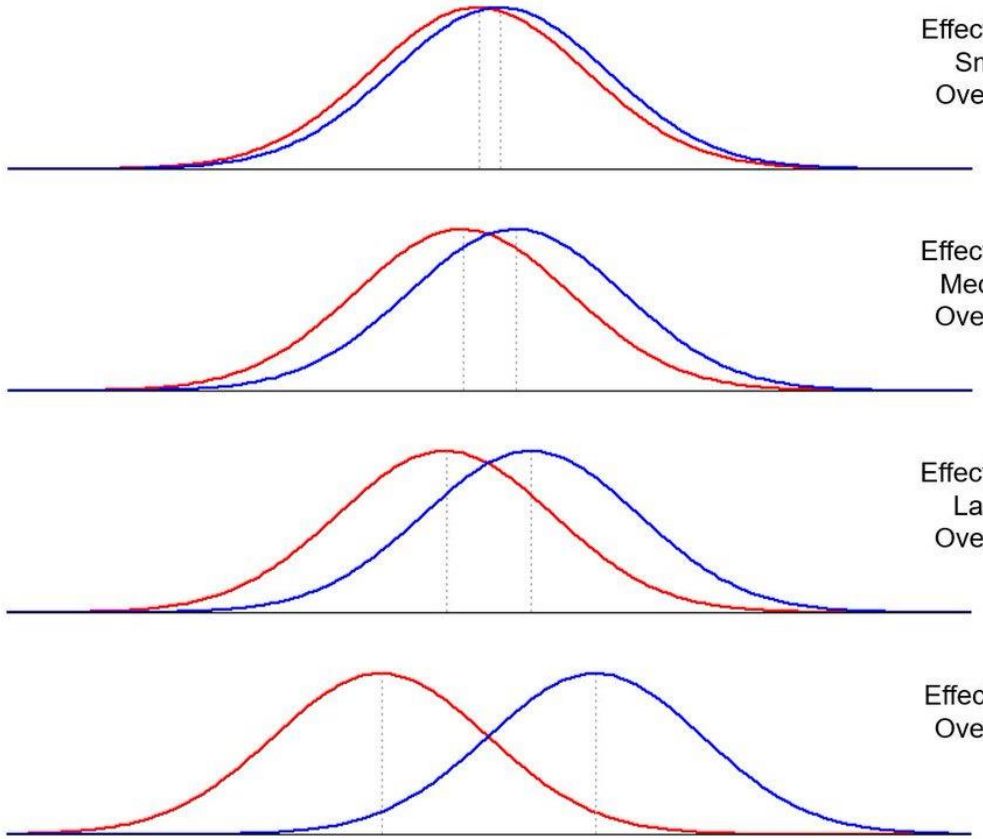
Type I error: we incorrectly reject the null hypothesis (p-value)
Type II error: we incorrectly accept the null hypothesis
Power is 1 – type II error (usually 0.8 or 0.9)
Effect size is difference of the mean divided by the pooled standard deviation (Cohen's d)



Decision Made ↓	Null true	Null false
Reject Null	Type I error (α)	Correct decision (1 – b) POWER!
Fail to reject null	Correct decision	Type II error (b)

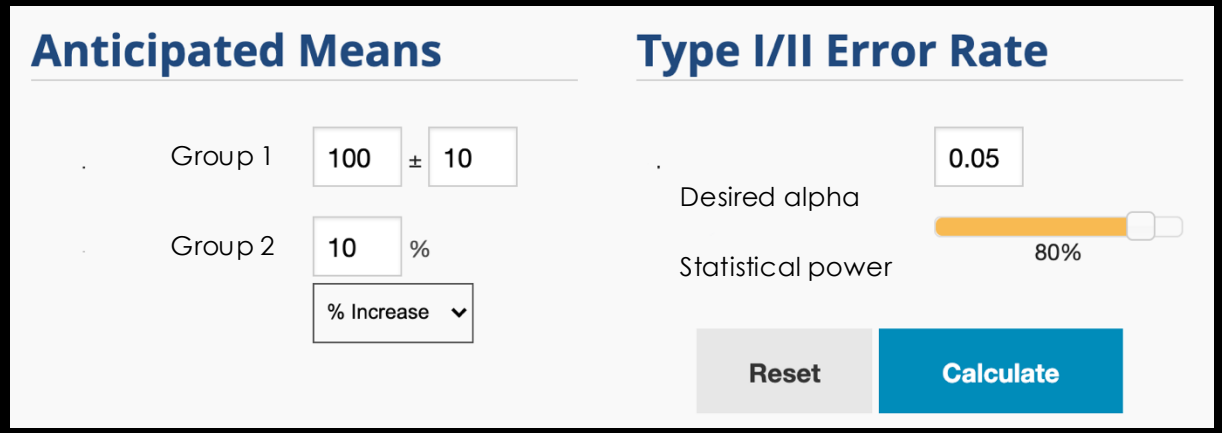
Yai p-value below 0.05, done!

Understanding Effect Sizes



Sample size: power calculation

Calculate the sample size you will need for your experiment



https://clincalc.com/stats/samplesize.aspx
Sampsizepwr function in Matlab; *Sempower* package in R

Sample Size	
Group 1	16
Group 2	16
Total	**32**

For group analysis in neuroimaging, sample sizes usually range from 16 to 32

Parametric statistics

T-test: Compare paired/unpaired Samples for continuous data.

Paired

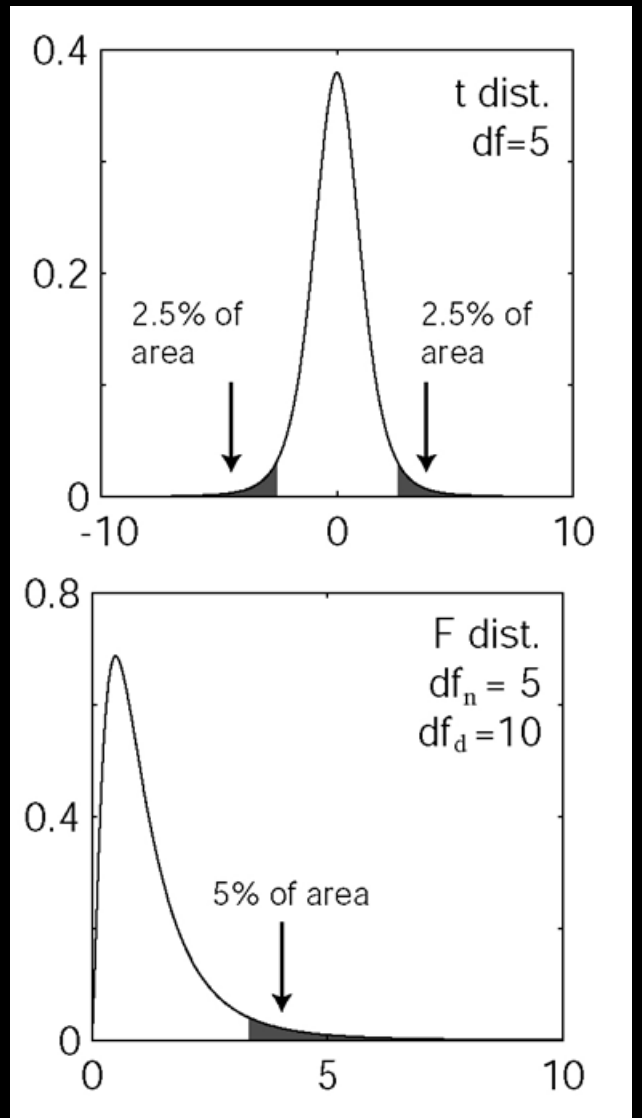
$$t = \frac{Mean\_difference}{Standard\_deviation}\sqrt{N-1}$$

Unpaired

$$t = \sqrt{N}\,\frac{Mean_A - Mean_B}{\sqrt{(SD_A)^2 - (SD_B)^2}}$$

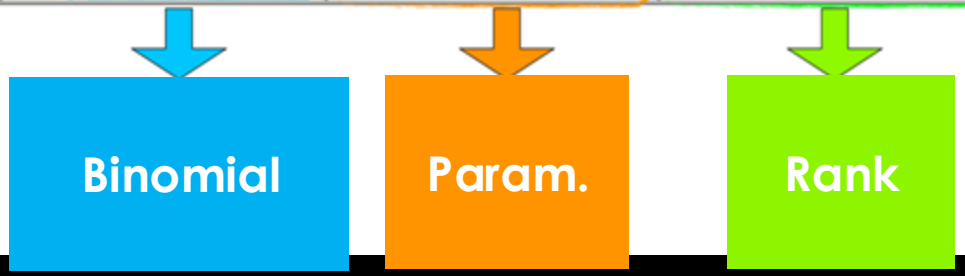
ANOVA: compare several groups (can test interaction between two factors for the repeated measure ANOVA)

$$F = \frac{Variance_{interGroup} \Big/ N_{Group} - 1}{Variance_{WithinGroup} \Big/ N - N_{Group}}$$



Assume gaussian distribution of data

Goal	Dataset		
	Binomial or Discrete	Continuous measurement (from a normal distribution)	Continuous measurement, Rank, or Score (from non-normal distribution)
Example of data sample	List of patients recovering or not after a treatment	Readings of heart pressure from several patients	Ranking of several treatment efficiency by one expert
Describe one data sample	Proportions	Mean, SD	Median
Compare one data sample to a hypothetical distribution	χ^2 or binomial test	One-sample t test	Sign test or Wilcoxon test
Compare two paired samples	Sign test	Paired t test	Sign test or Wilcoxon test
Compare two unpaired samples	χ^2 square Fisher's exact test	Unpaired t test	Mann-Whitney test
Compare three or more unmatched samples	χ^2 test	One-way ANOVA	Kruskal-Wallis test
Compare three or more matched samples	Cochrane Q test	Repeated-measures ANOVA	Friedman test
Quantify association between two paired samples	Contingency coefficients	Pearson correlation	Spearman correlation



Delorme, A. (2005) Statistical Methods. Encyclopedia of Medical Device and Instrumentation, vol 6, pp 240-264. Wiley interscience.

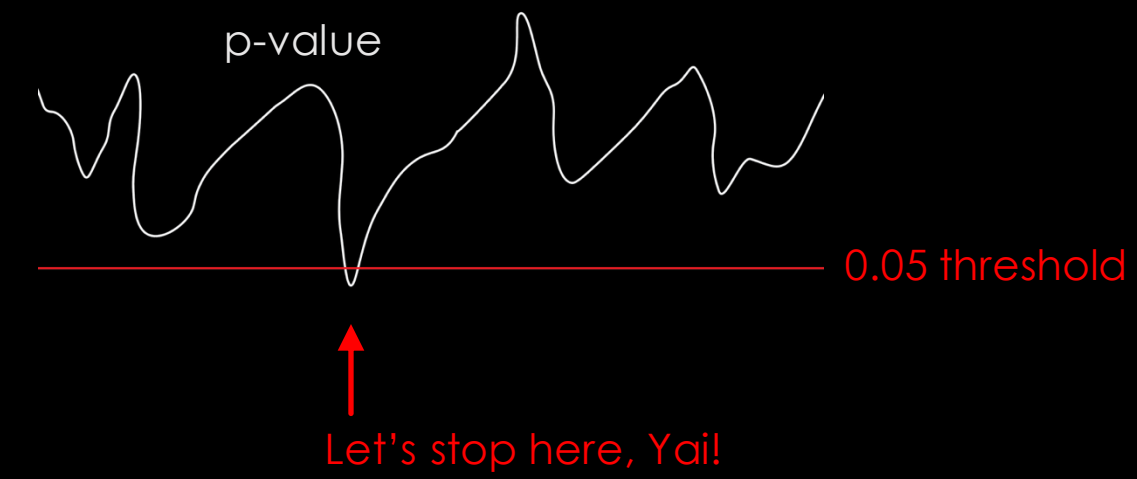
Non-parametric statistics

Values		Ranks
Paired t-test	→	Wilcoxon
Unpaired t-test	→	Mann-Whitney
One way ANOVA	→	Kruskal Wallis

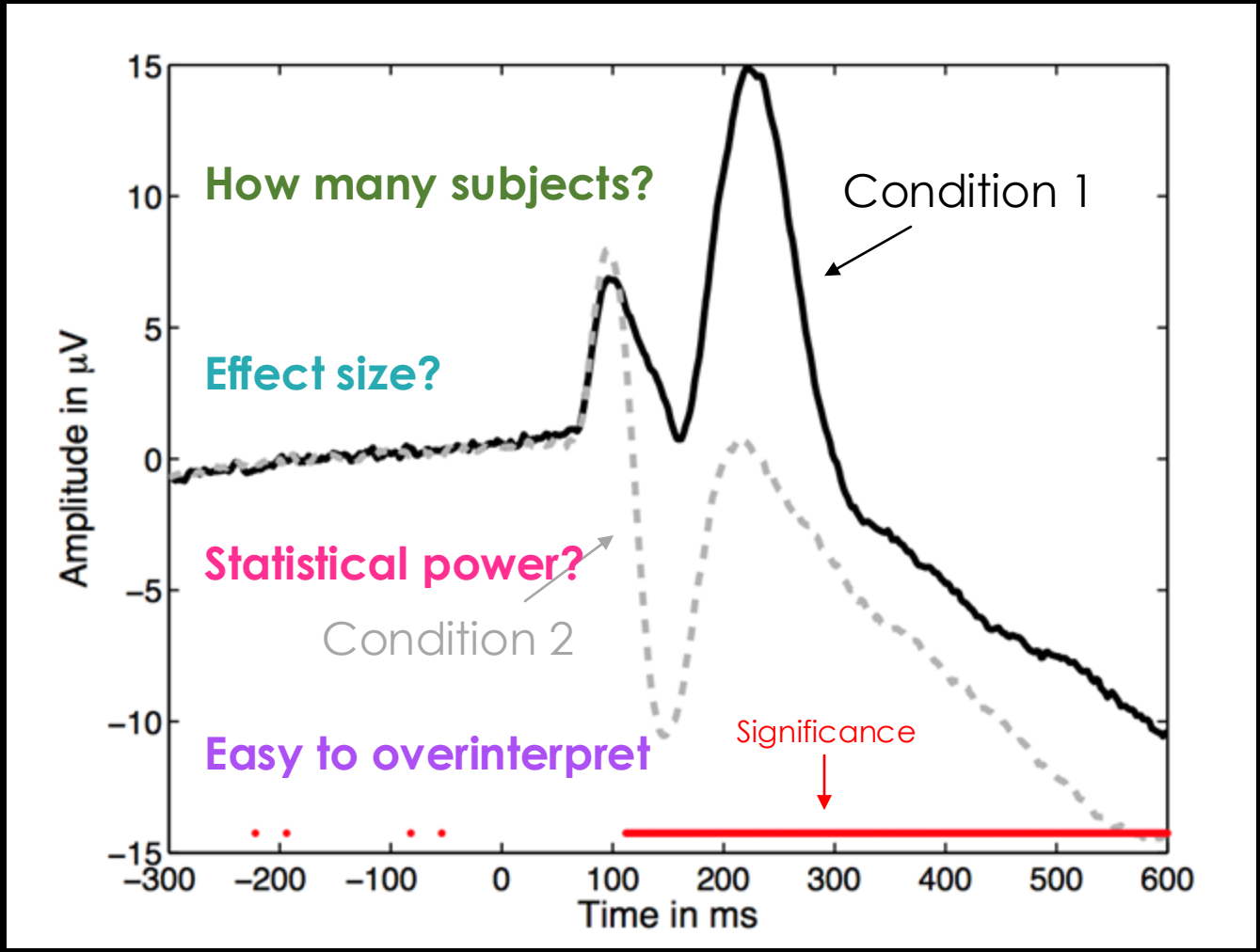
BOTH ASSUME NORMAL DISTRIBUTIONS
Need to check the normality of the distribution/rank

Improving robustness

Optional stopping
(do not do it)

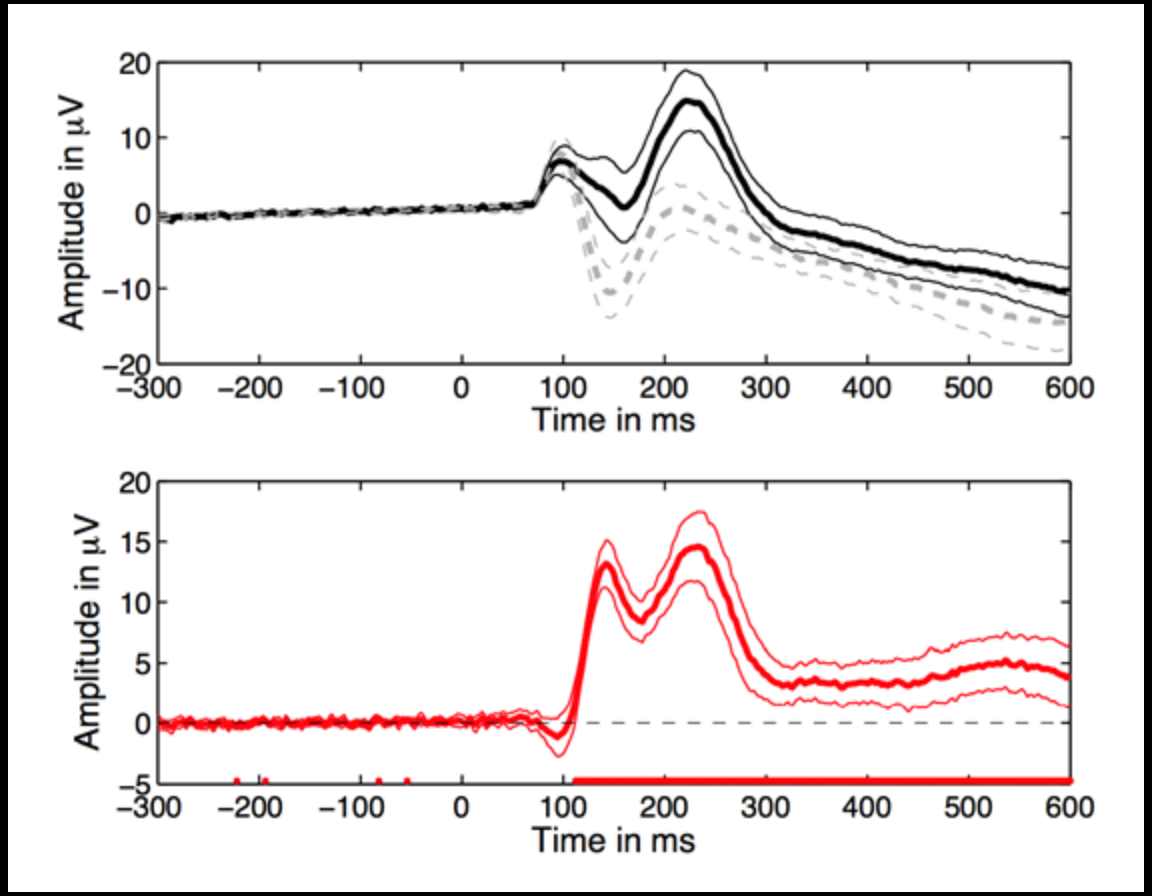


Why the standard figure is not good enough

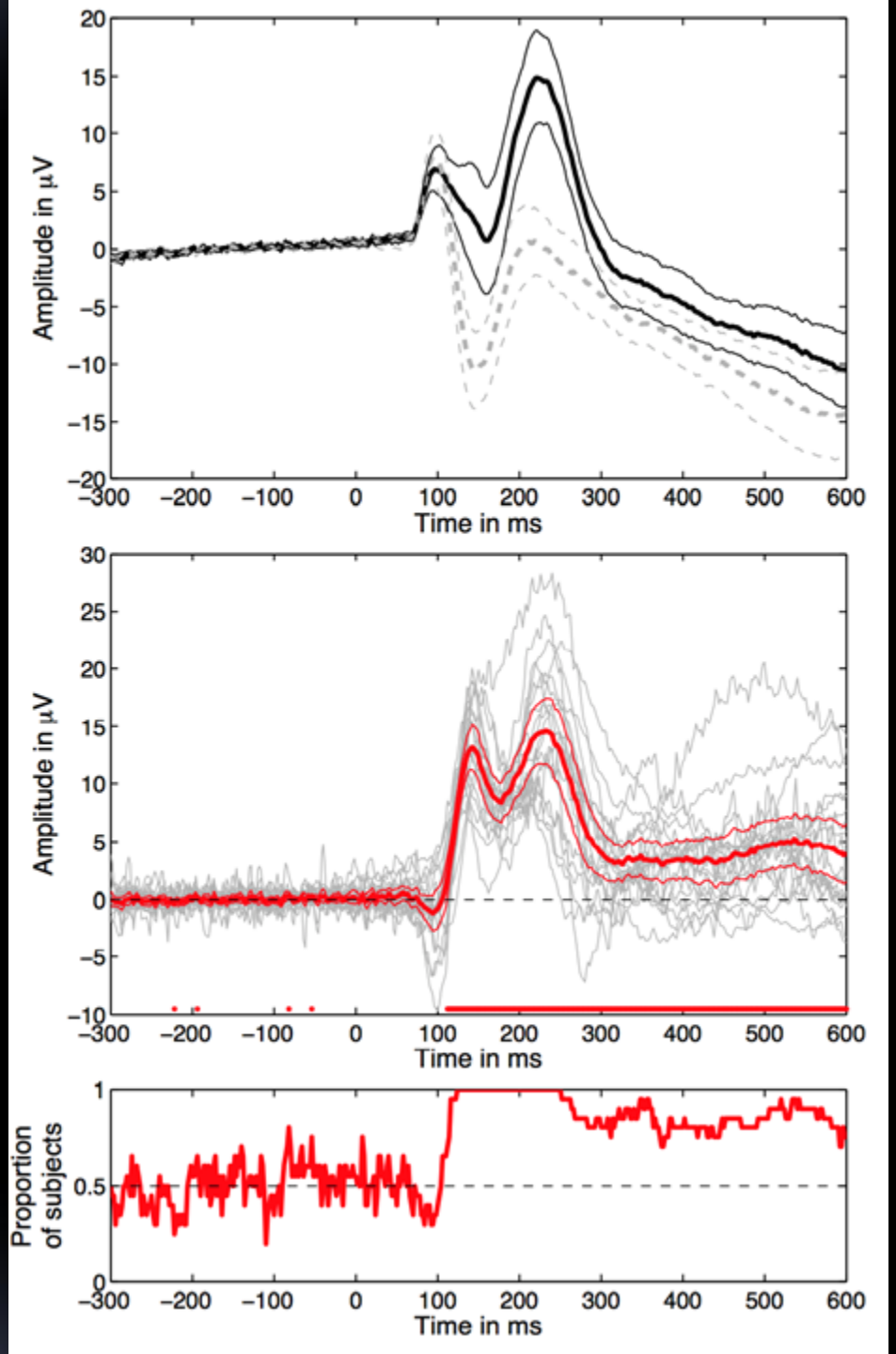


Credit: G. Rousselet

Add confidence intervals and plot of the difference



Credit: G. Rousselet



How many subjects show an effect in the right direction?

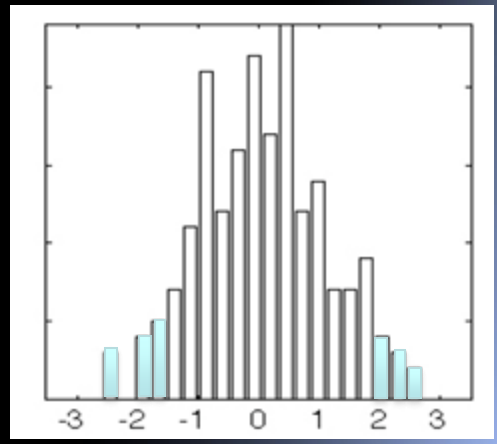
Robust measures of central tendency

- Non-robust estimator
  - Mean: mERP = mean(EEG.data,3)

  (EEG.data is an array of *channels* x *times* x *trials*)

- Robust estimators of central tendency
  - Median: mdERP = median(EEG.data,3)
  - Trimmed mean: tmERP = trimmean(EEG.data,20,'round',3)

    20% trimmed means provide high statistical power in the presence of outliers



Rand Wilcox, 2012, Introduction to Robust Estimation and Hypothesis Testing, Elsevier

Problems

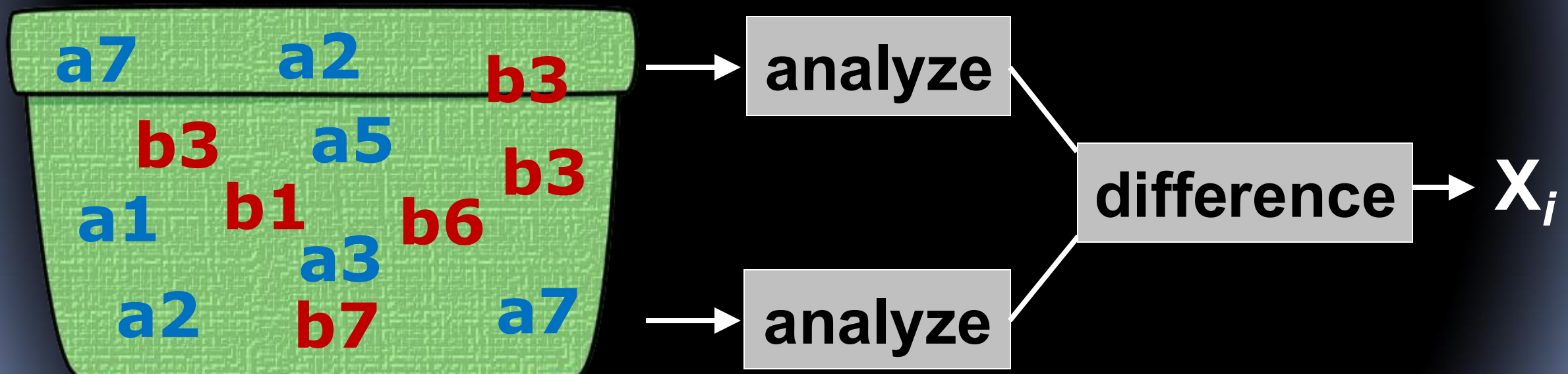
- Not resistant against outliers
- For ANOVA and t-test non-normality is an issue when distributions differ or when variances are not equal.
- Slight departure from normality can have serious consequences

---

Solutions

- Randomization approach
- Bootstrap approach

The Bootstrap approach

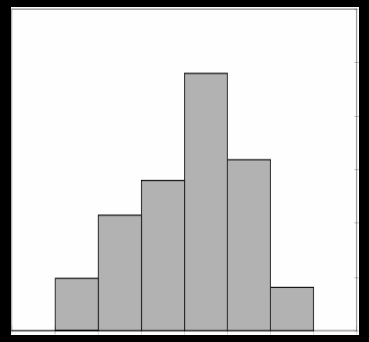


Bootstrap: central idea

- "The bootstrap is a computer-based method for assigning measures of accuracy to statistical estimates." Efron & Tibshirani, 1993

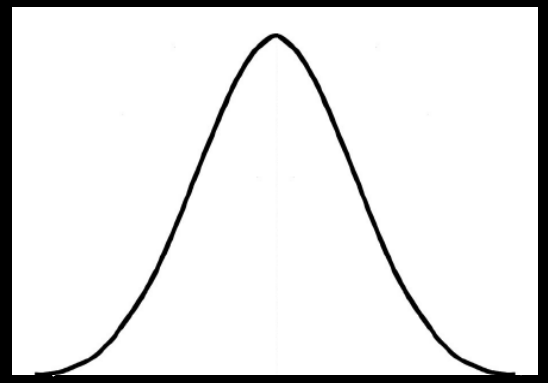
- "The central idea is that it may sometimes be better to draw conclusions about the characteristics of a population strictly from the sample at hand, rather than by making perhaps unrealistic assumptions about the population." Mooney & Duval, 1993

Sample and population

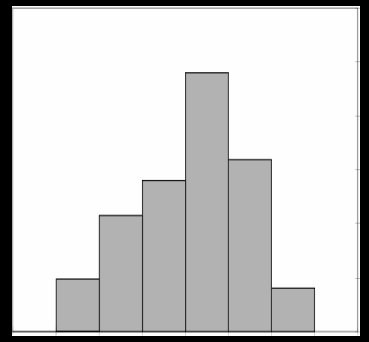


Sample

Mean and
Standard deviation



PDF of population when
using parametric statistics

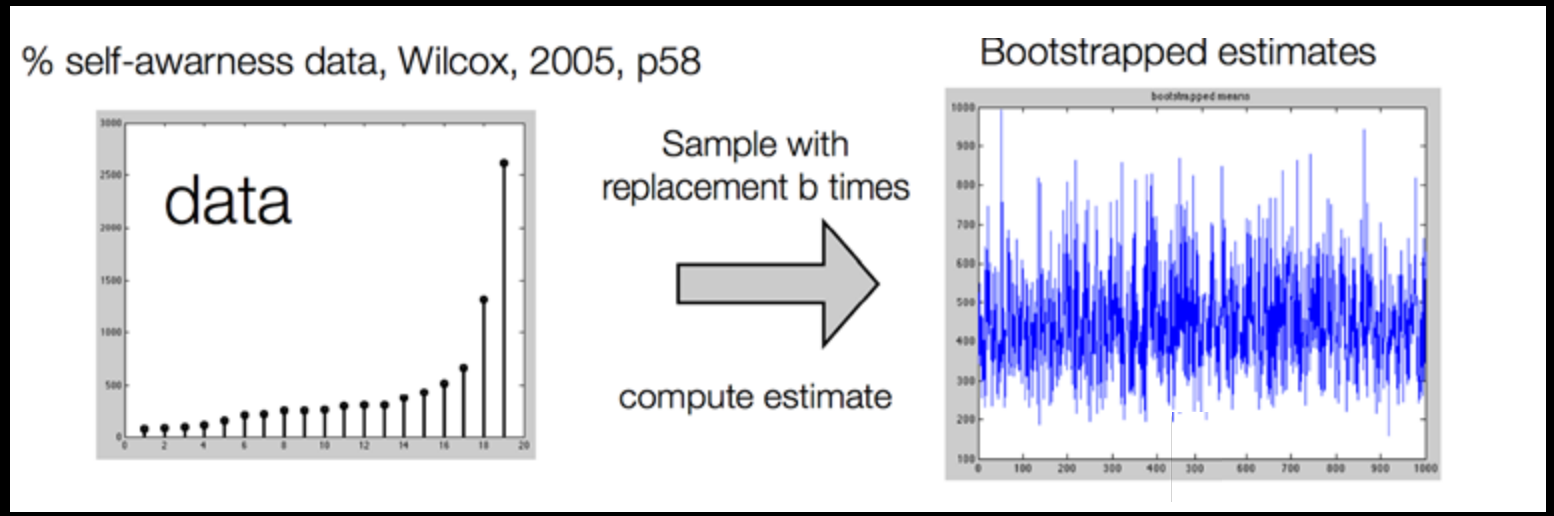


PDF of population when
using bootstrap statistics

Given that we have no other information about the population,
the sample is our best single estimate of the population.

PDF: Probability density distribution

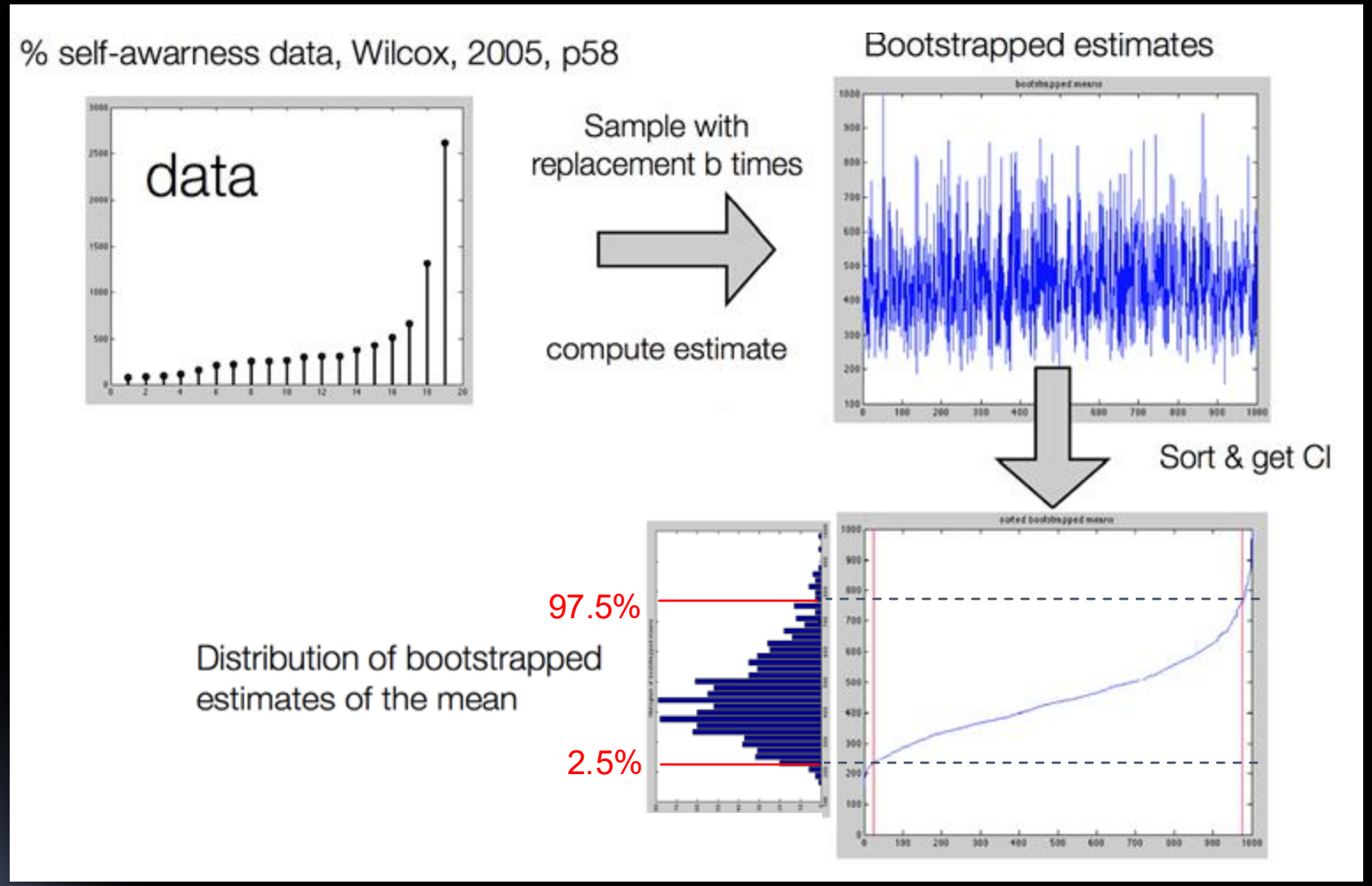
Percentile bootstrap estimate of confidence intervals



Parametric statistics

CI_{95} = mean ±1.96*SD

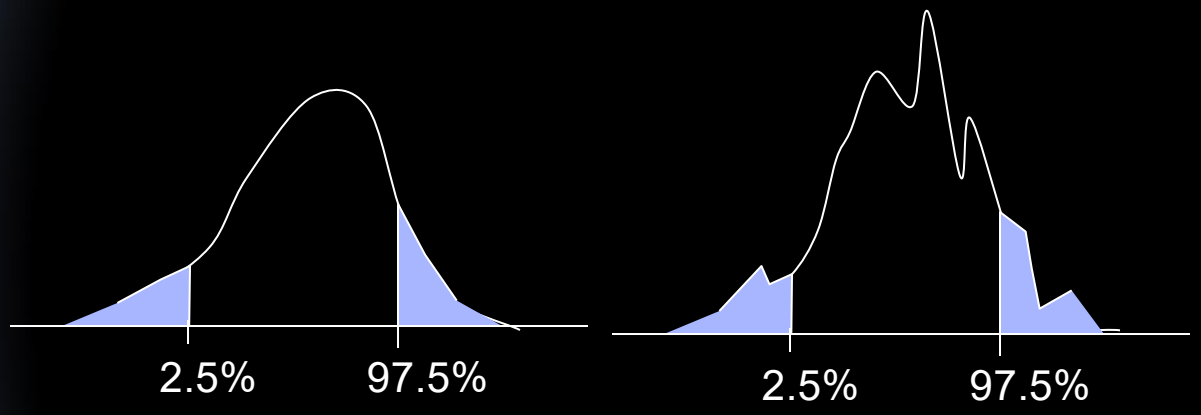
Percentile bootstrap estimate of confidence intervals



Parametric statistics

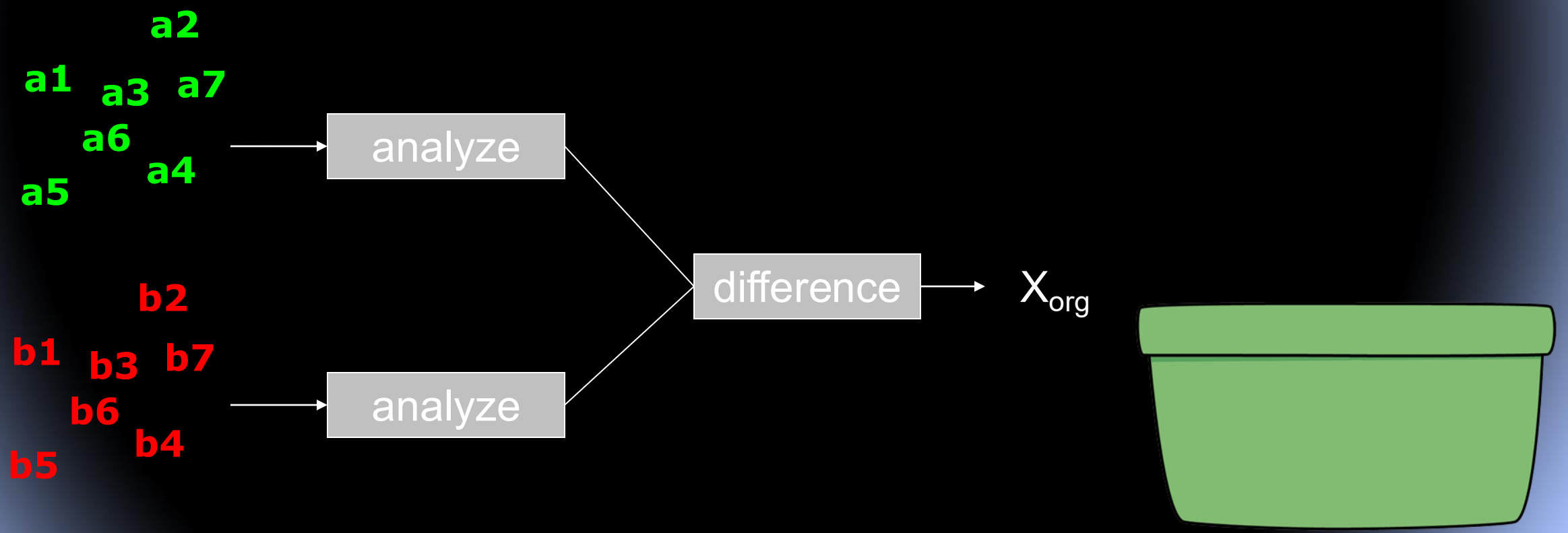
CI_{95} = mean ±1.96*SD

Distribution can take any shape

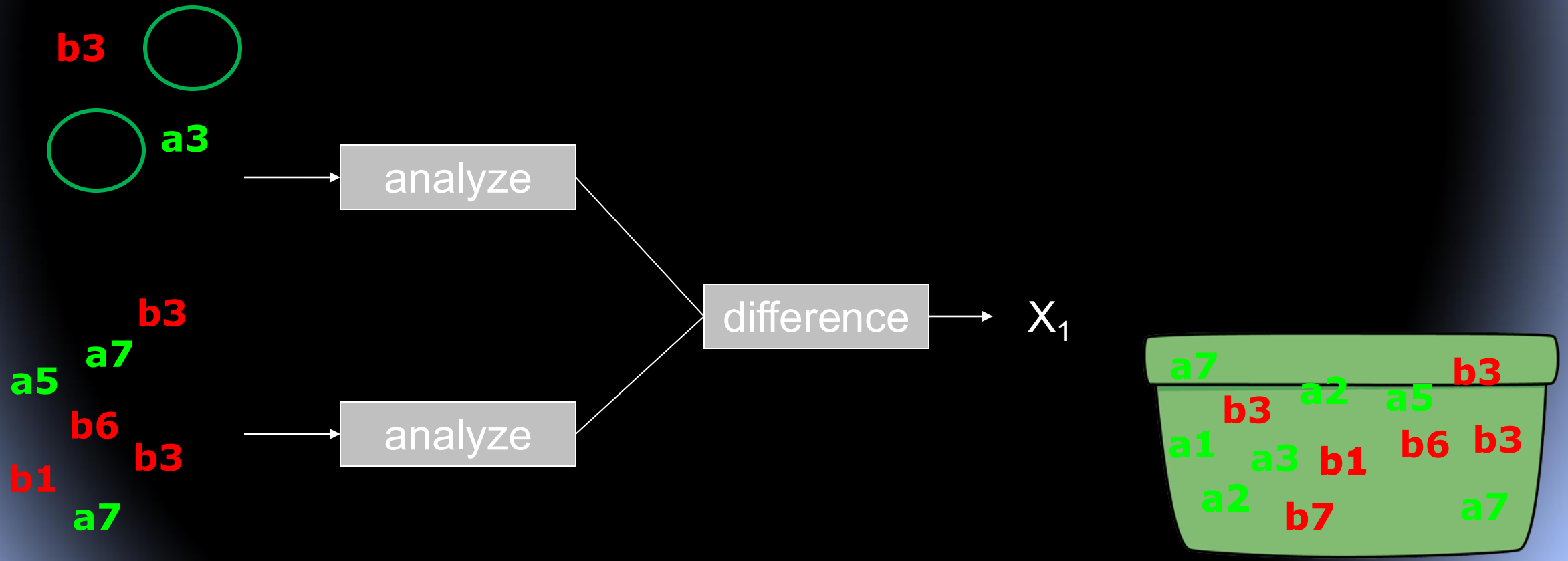


Once you have the 95% confidence interval, you can perform inferential statistics.

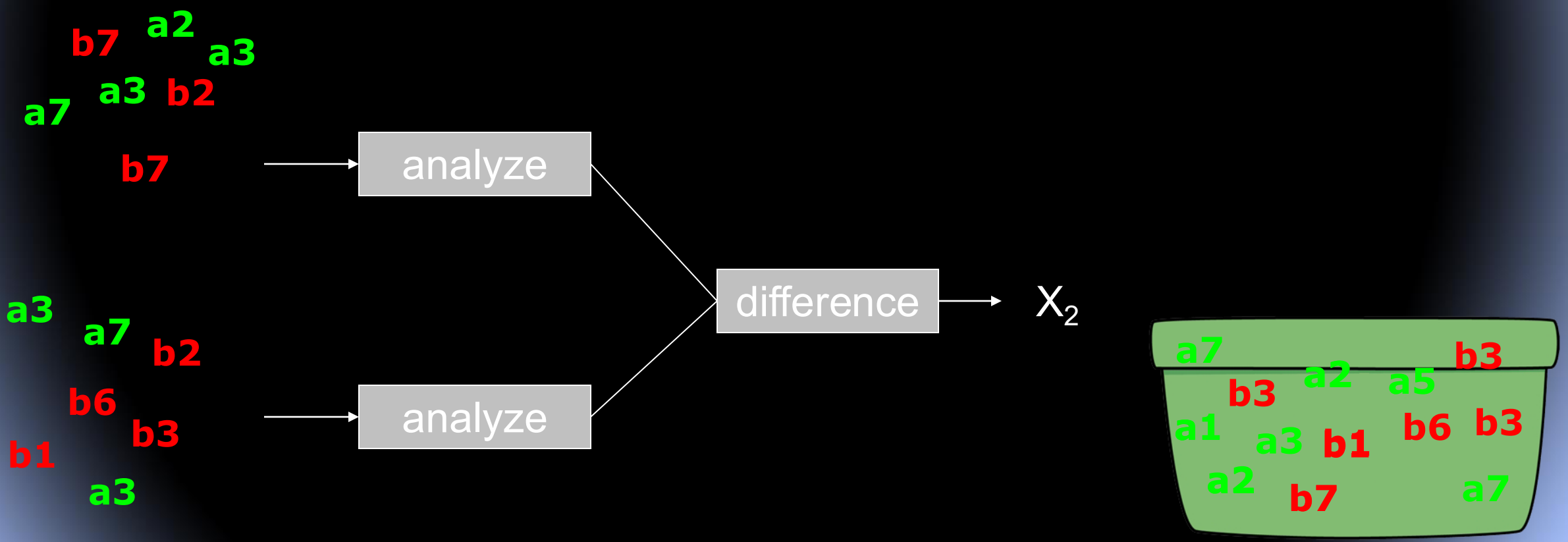
Confidence interval for the difference
Bootstrap approach H0



Bootstrap approach iteration 1

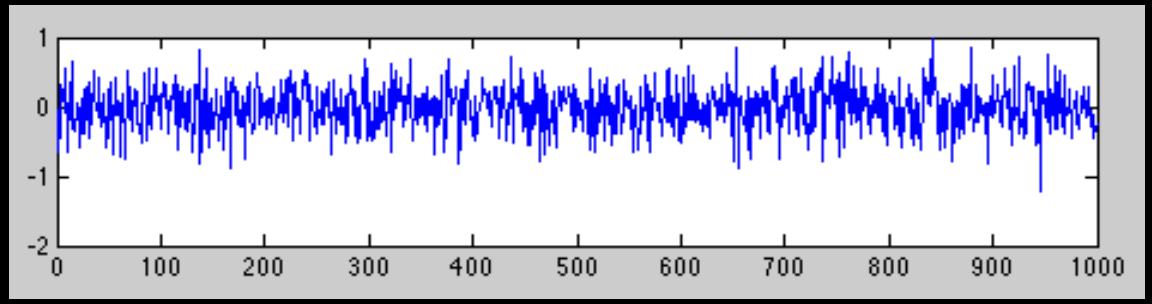


Bootstrap approach iteration 2



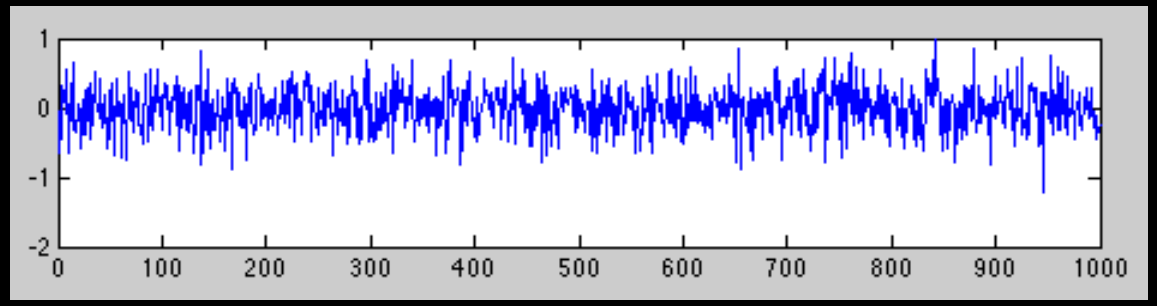
Inferences based on percentile bootstrap method H0

1000 bootstraps

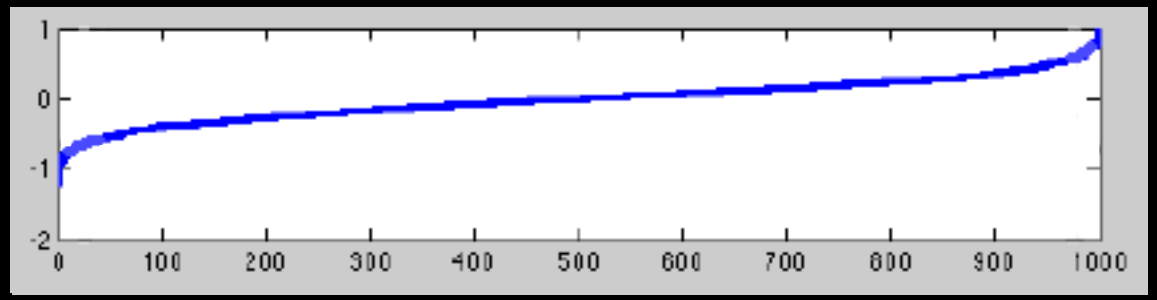


Inferences based on percentile bootstrap method H0

1000 bootstraps

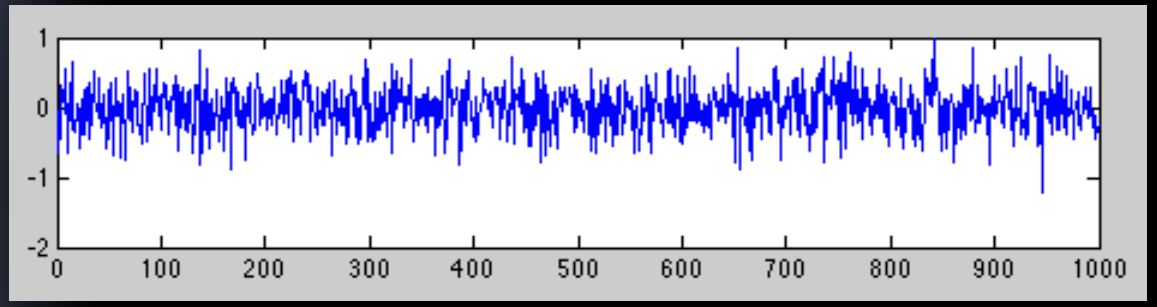


Sorted values

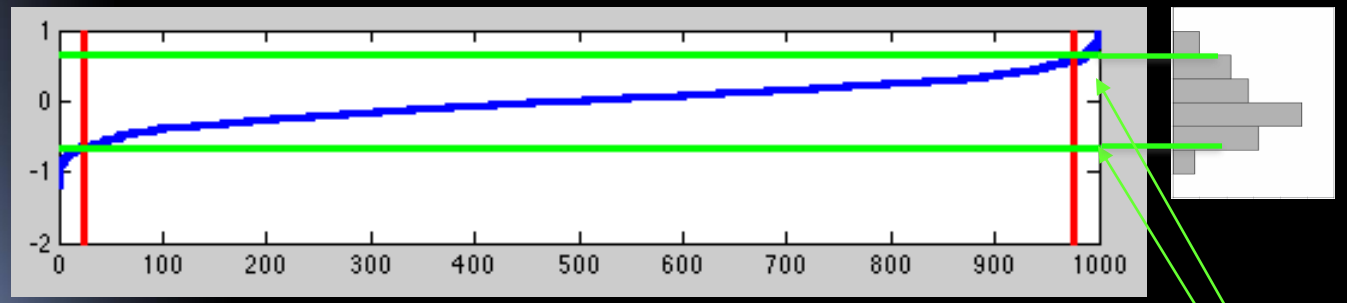


Inferences based on percentile bootstrap method H0

1000 bootstraps



Sorted values



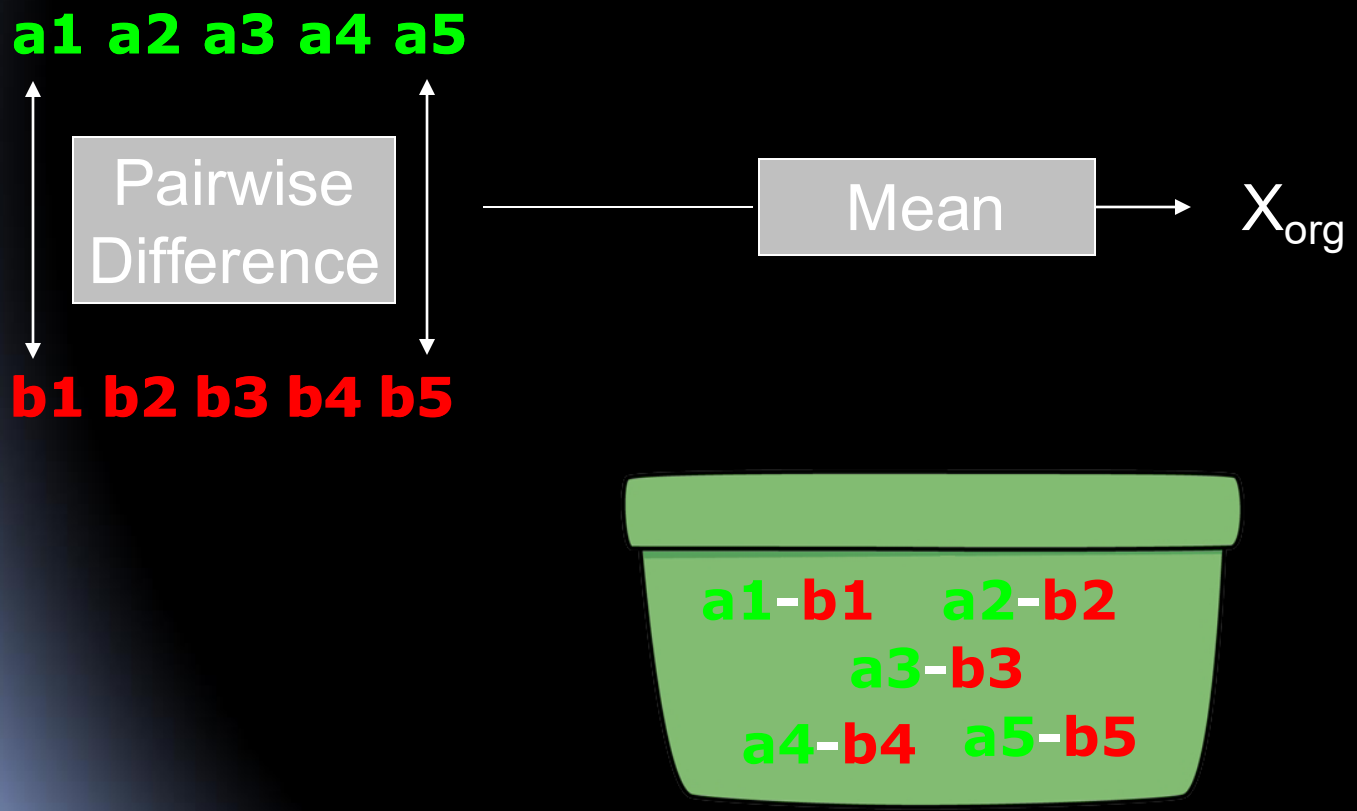
2.5%

97.5%

Thresholds

Doing the same using a Gaussian distribution for the population
→ parametric statistics

Confidence interval for the difference
Bootstrap approach H0 (paired)



Confidence interval for the difference
Bootstrap approach H0 (paired)

Bootstrap iteration 1

a5-b5 a1-b1
a3-b3
a3-b3 a5-b5

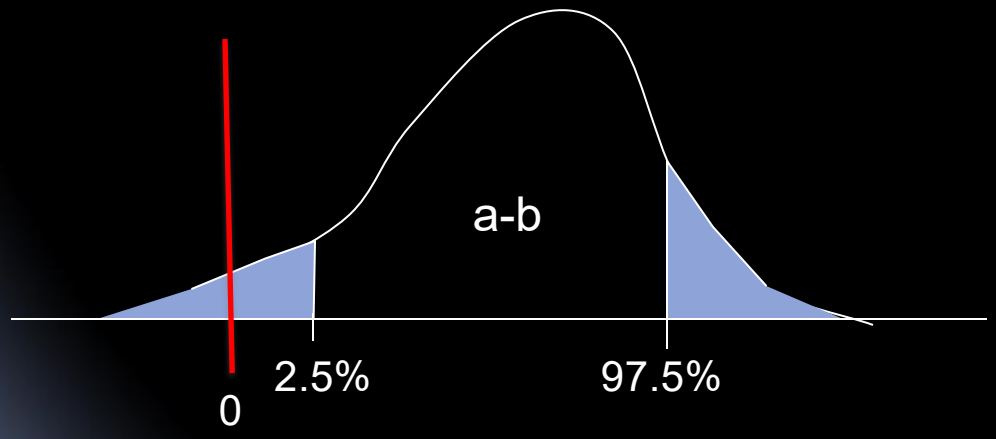
Mean → X_1

Bootstrap iteration 2

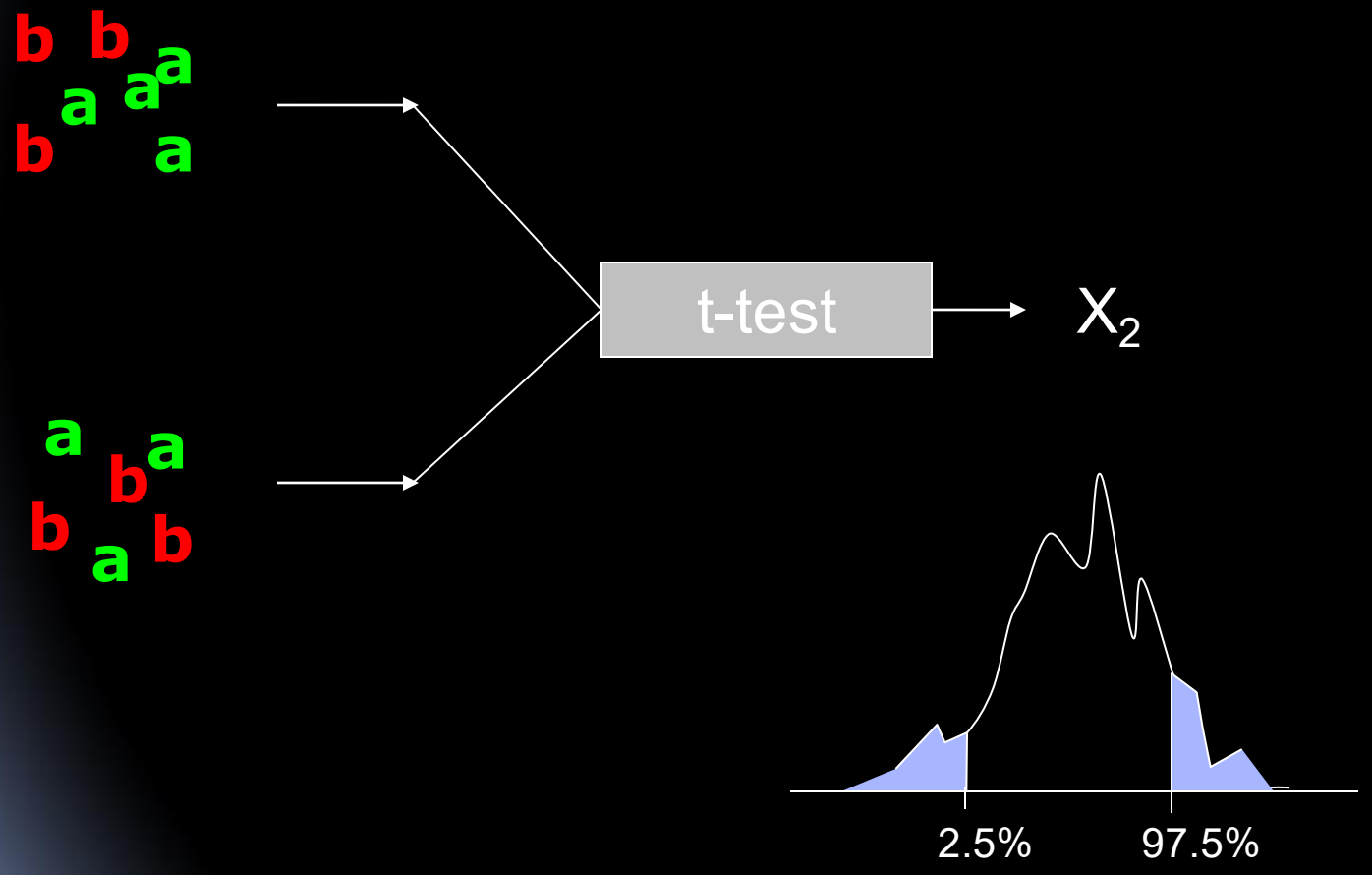
a3-b3 a2-b2
a2-b2
a4-b4 a4-b4

Mean → X_2

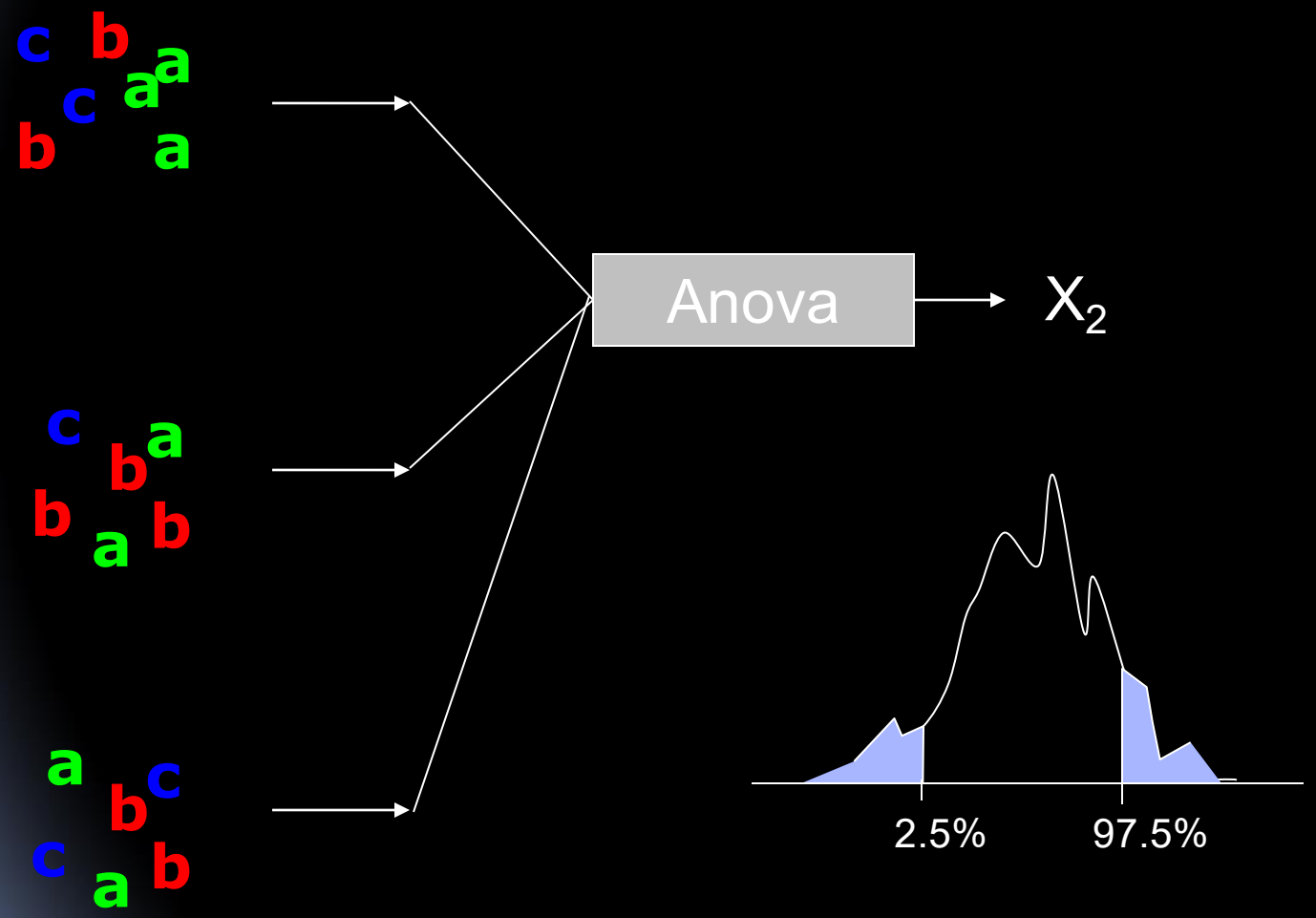
…



Measures for the bootstrap



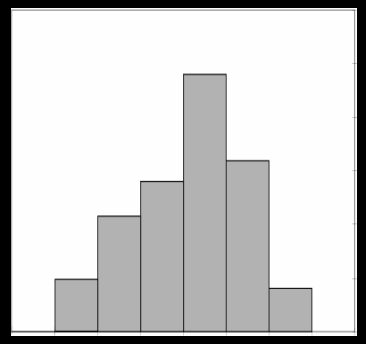
Measures for the bootstrap



Bootstrap versus permutation

Bootstrap: independent draws

Permutation: dependent draws

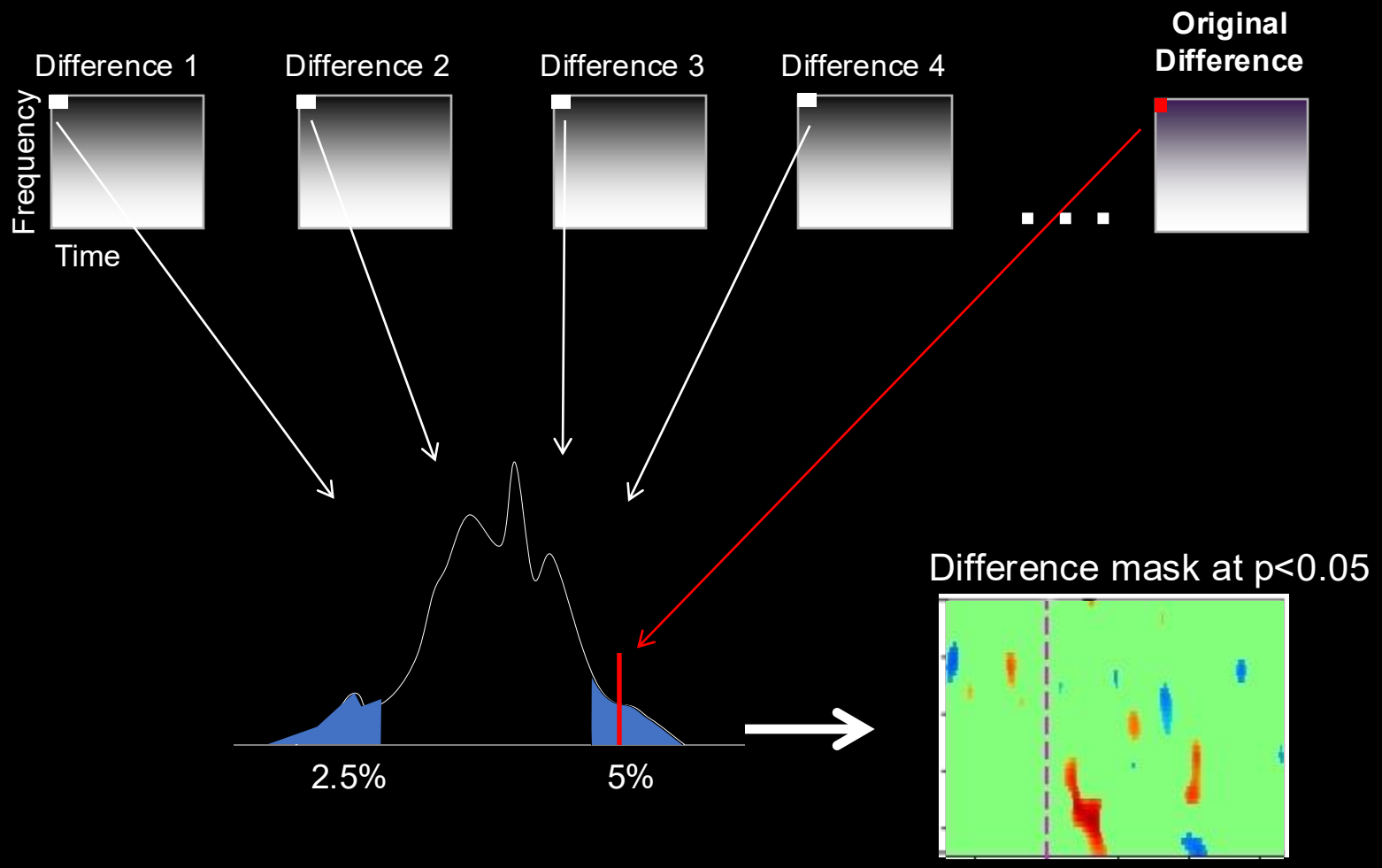


Surrogate statistics: our sample is our best estimate of the population

Use bootstrap when possible!

Corrections for multiple comparisons

Assessing significance



Correcting for multiple comparisons

- Bonferoni correction

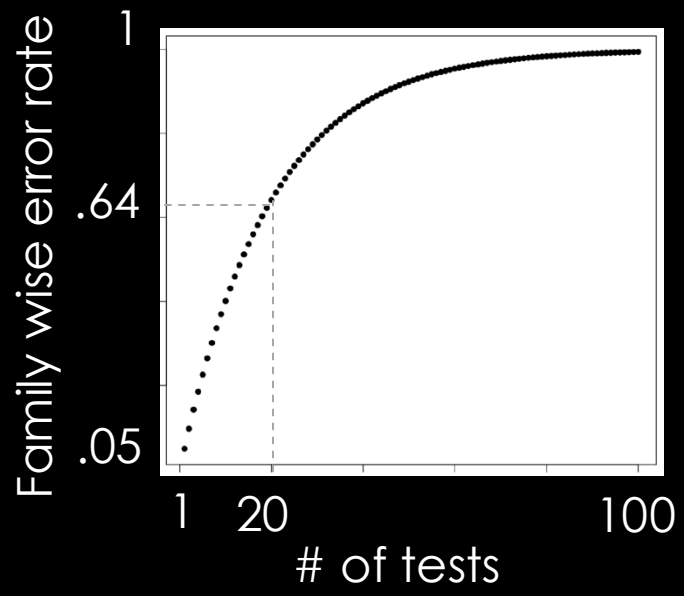
  → divide p-value threshold by the number of comparisons

- Holm-Bonferoni correction
- False Discovery Rate
- Max method
- Clusters

Family-wise error rate

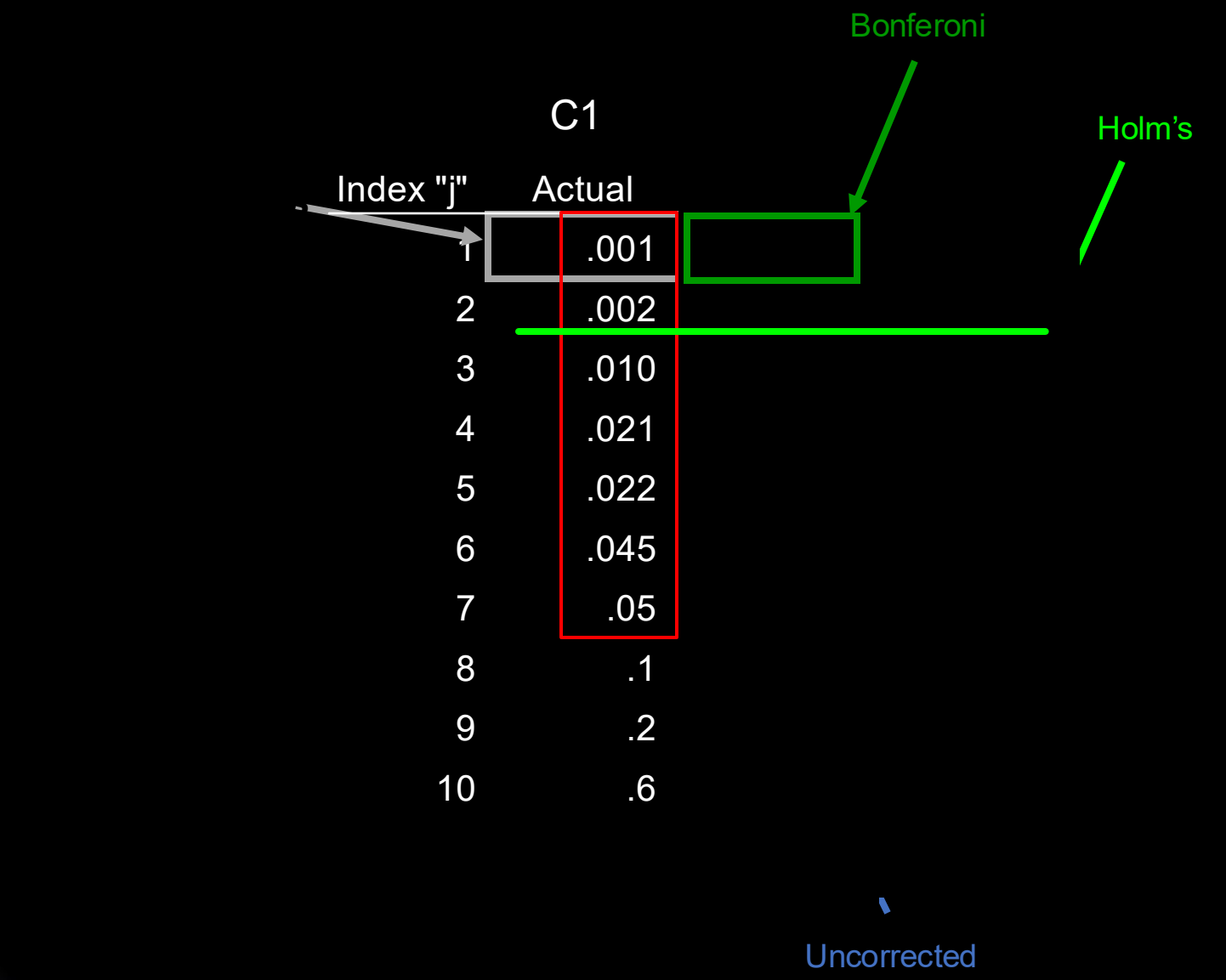
- Family-wise error rate (FWER) is the probability of making **at least ONE** errors when performing multiple hypotheses tests. With α=0.05 set as the p-value threshold:

  - 1 test → 5%
  - 2 tests → 10%
  - 20 tests → 64%



$$FWER = 1 - (1 - \alpha)^{n_{tests}}$$

Holm-Bonferoni's procedure



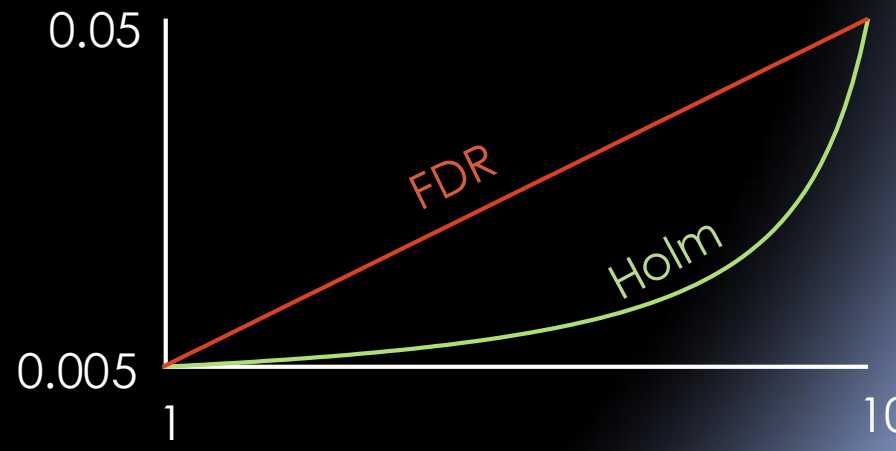
FDR procedure

Bonferoni

	C1	C2
Index "j"	Actual	j*0.05/10
1	.001	.005
2	.002	.010
3	.010	.015
4	.021	.020
5	.022	.025
6	.045	.030
7	.05	.035
8	.1	.040
9	.2	.045
10	.6	.05

FDR

Uncorrected



Holm-Bonfferoni at p=0.05

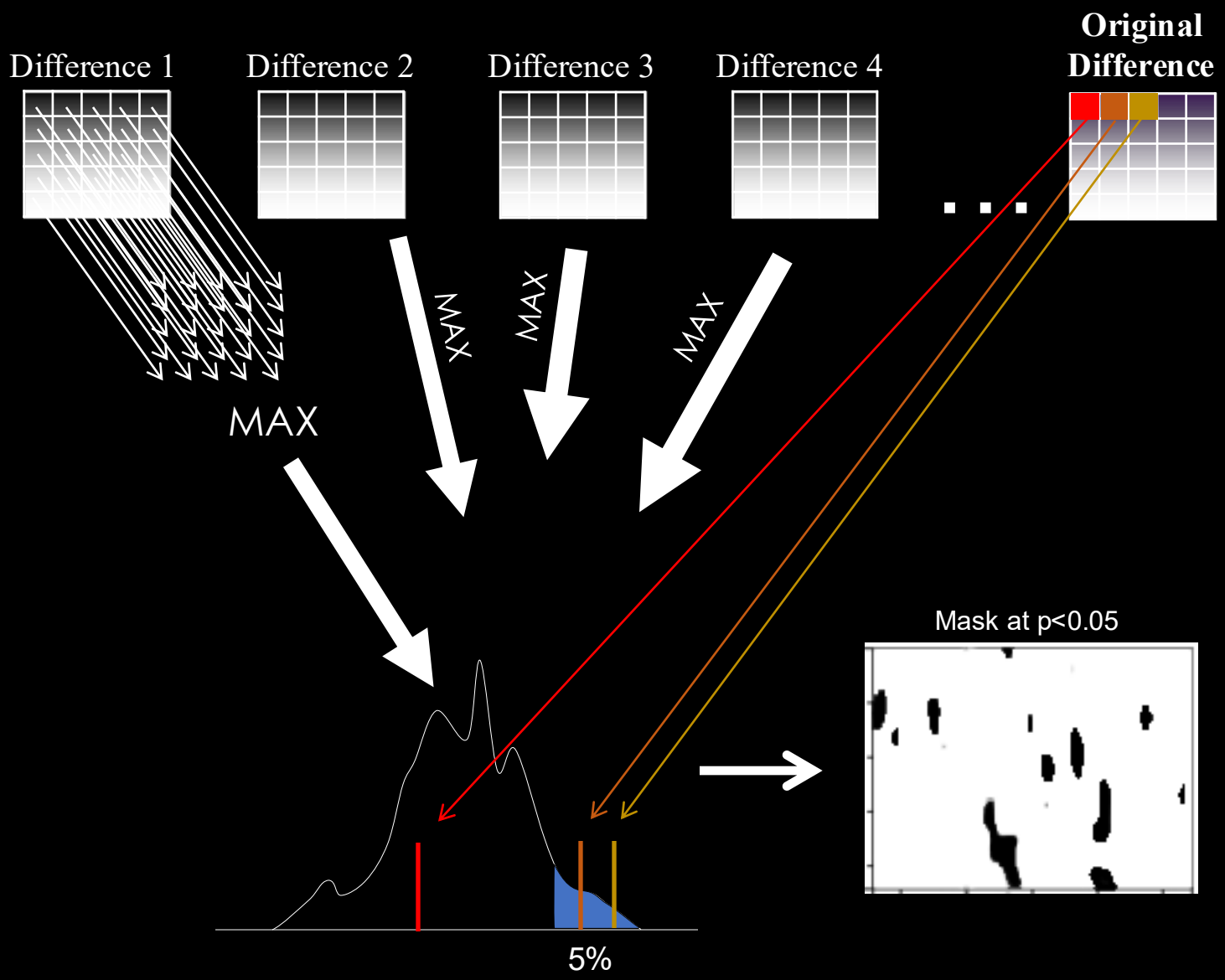
→ less than 5% chance of having **one** false positive (family-wise error rate of 5%)

FDR procedure at 0.05

→ at most 0.05% false positives

→ 40% change of observing a false positive for 100 values (family-wise error rate of 40%)

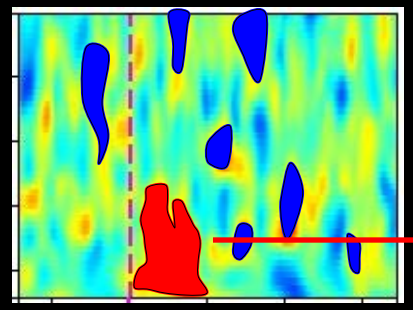
Max procedure (maxT)



maxT vs minP

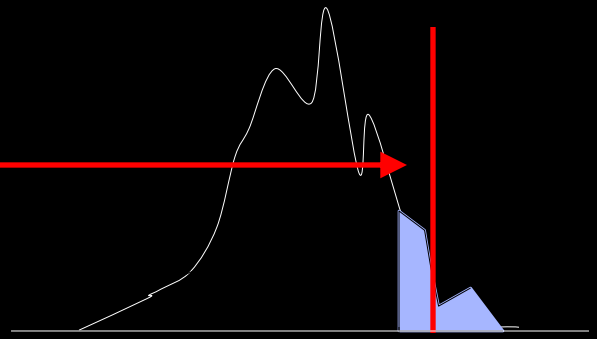
Cluster correction for multiple comparisons

Original difference



44 pixels

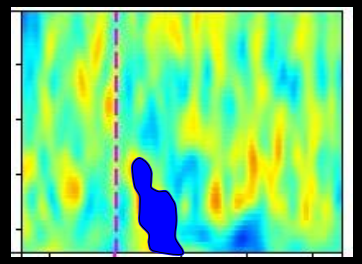
Area of significant regions in number of pixels



5%

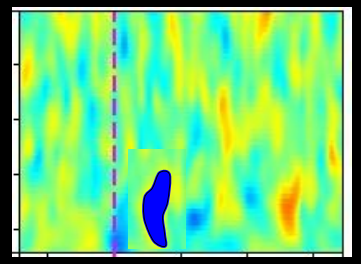
Difference bootstrap 1

Region of 35 pixels



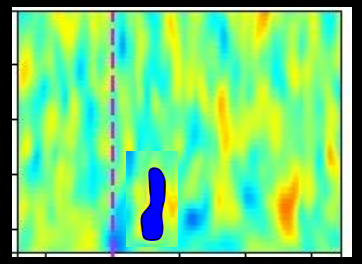
Difference bootstrap 2

23 pixels



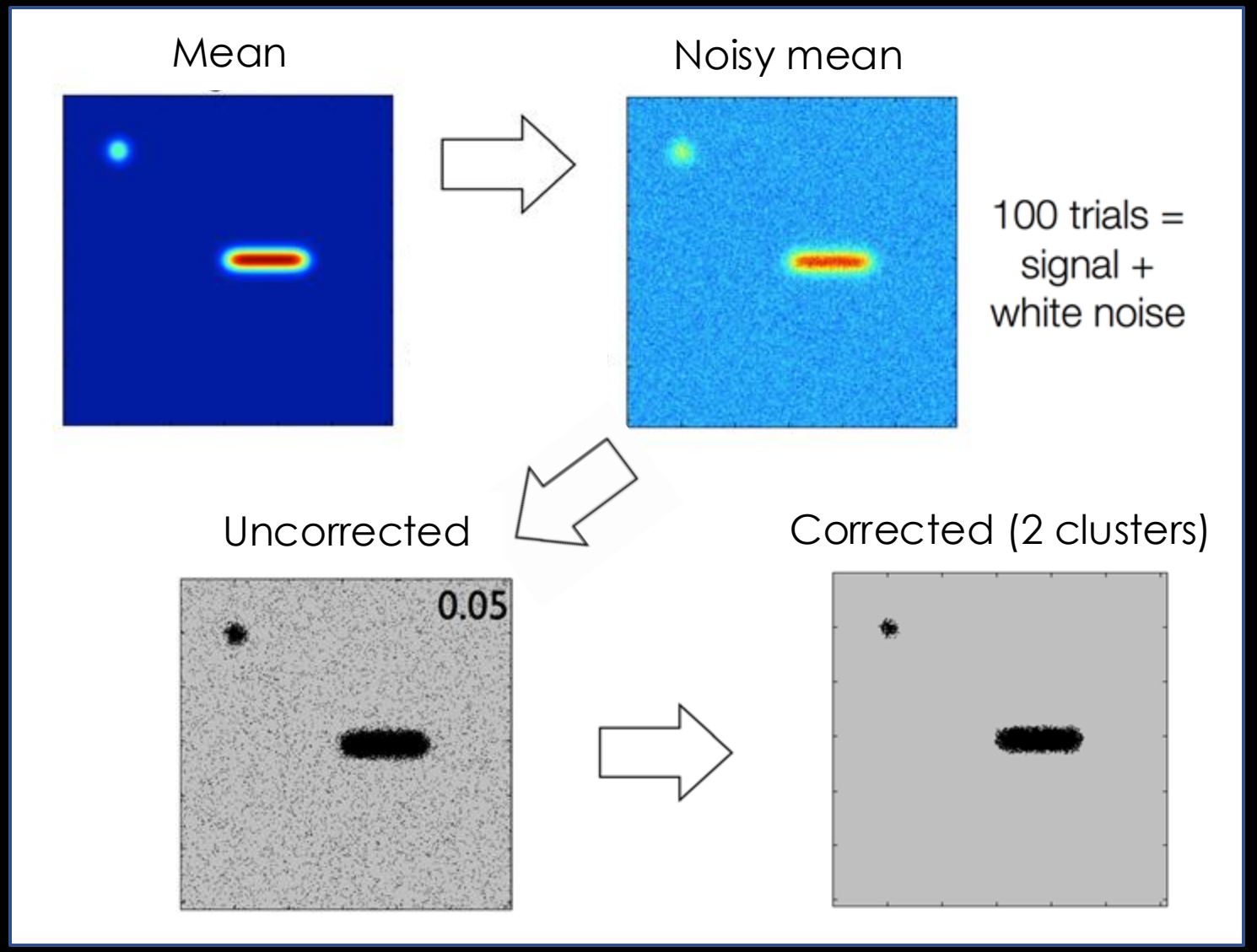
Difference bootstrap 3

22 pixels

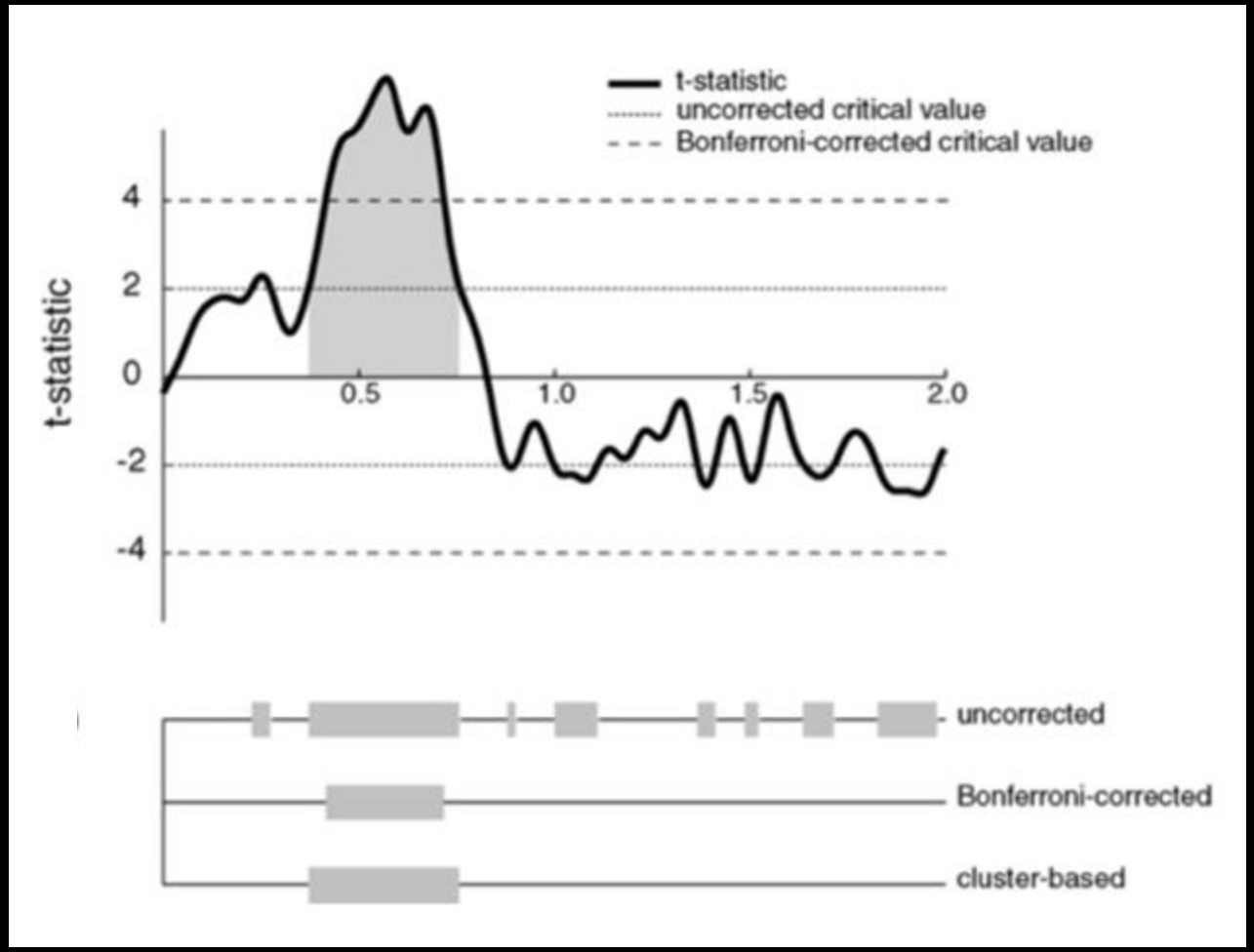


....

Control for multiple comparisons cluster method example

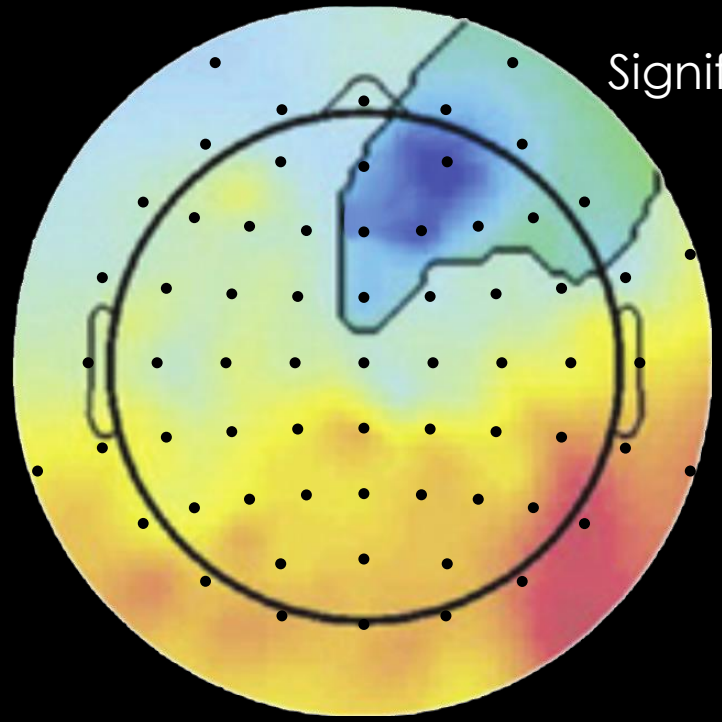


Cluster correction in 1 dimension

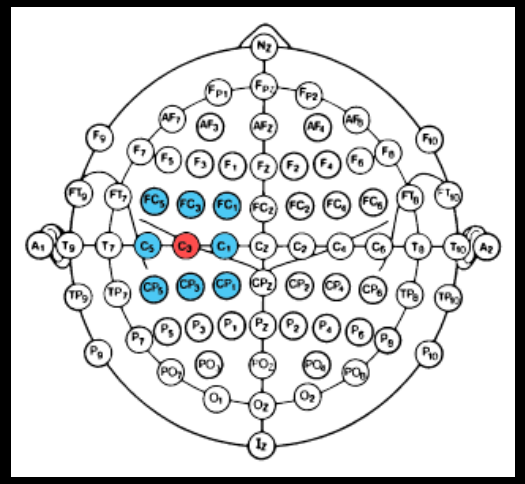


Maris and Oostenveld, J. Neurosci. Methods, 2007

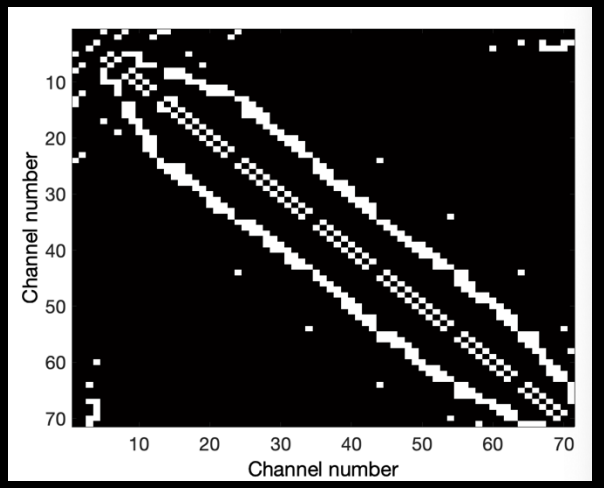
Significant region



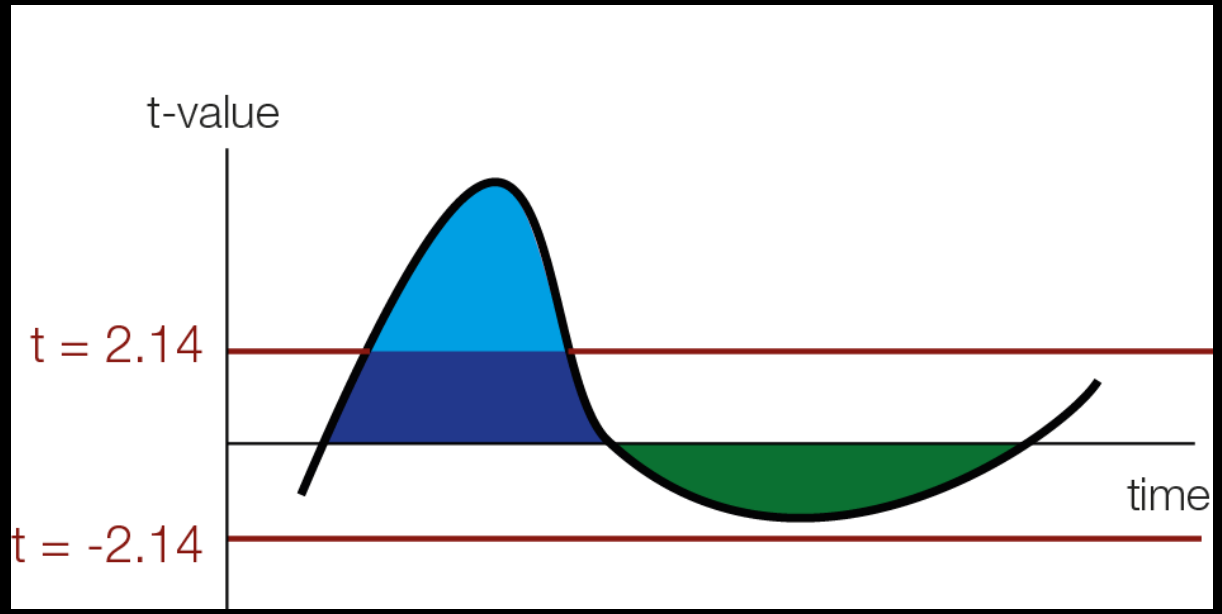
Channel neighbors



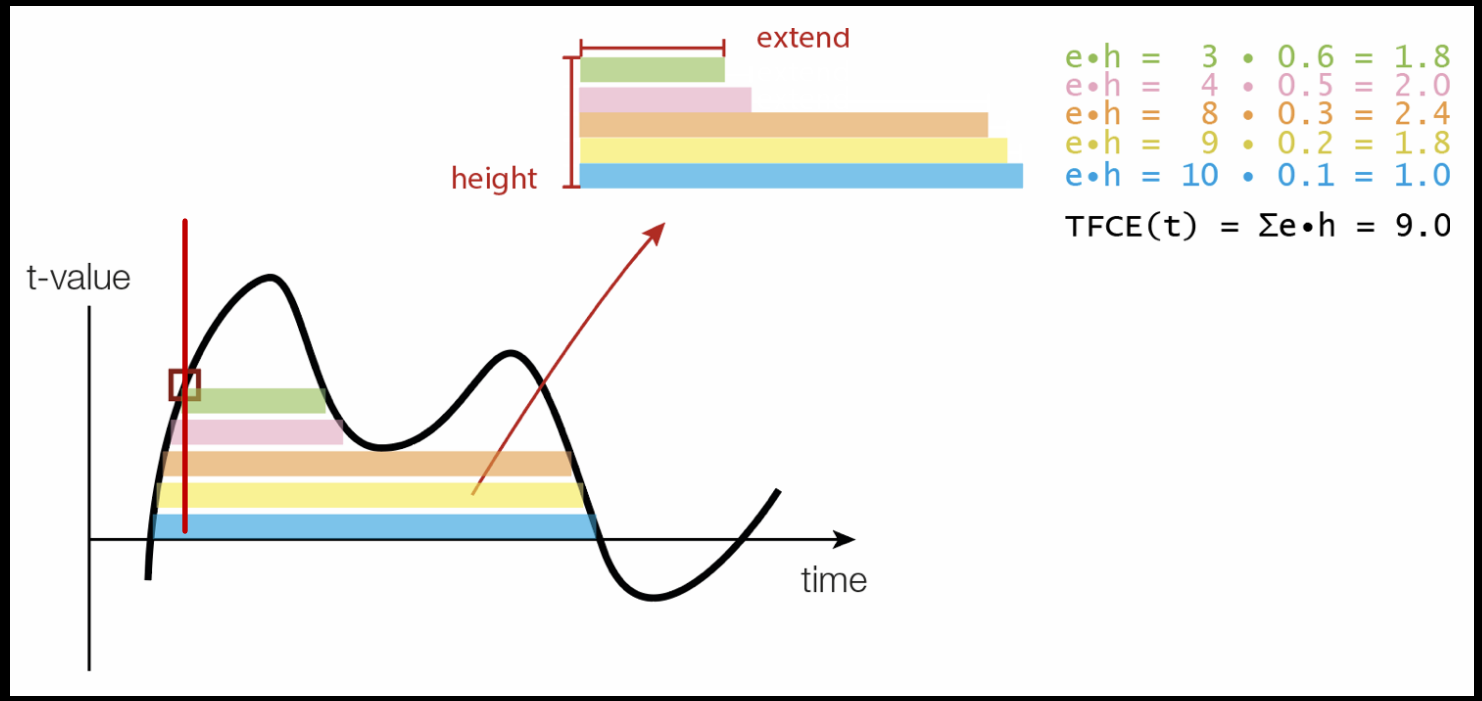
Channel neighbors matrix



TFCE – threshold free cluster enhancement



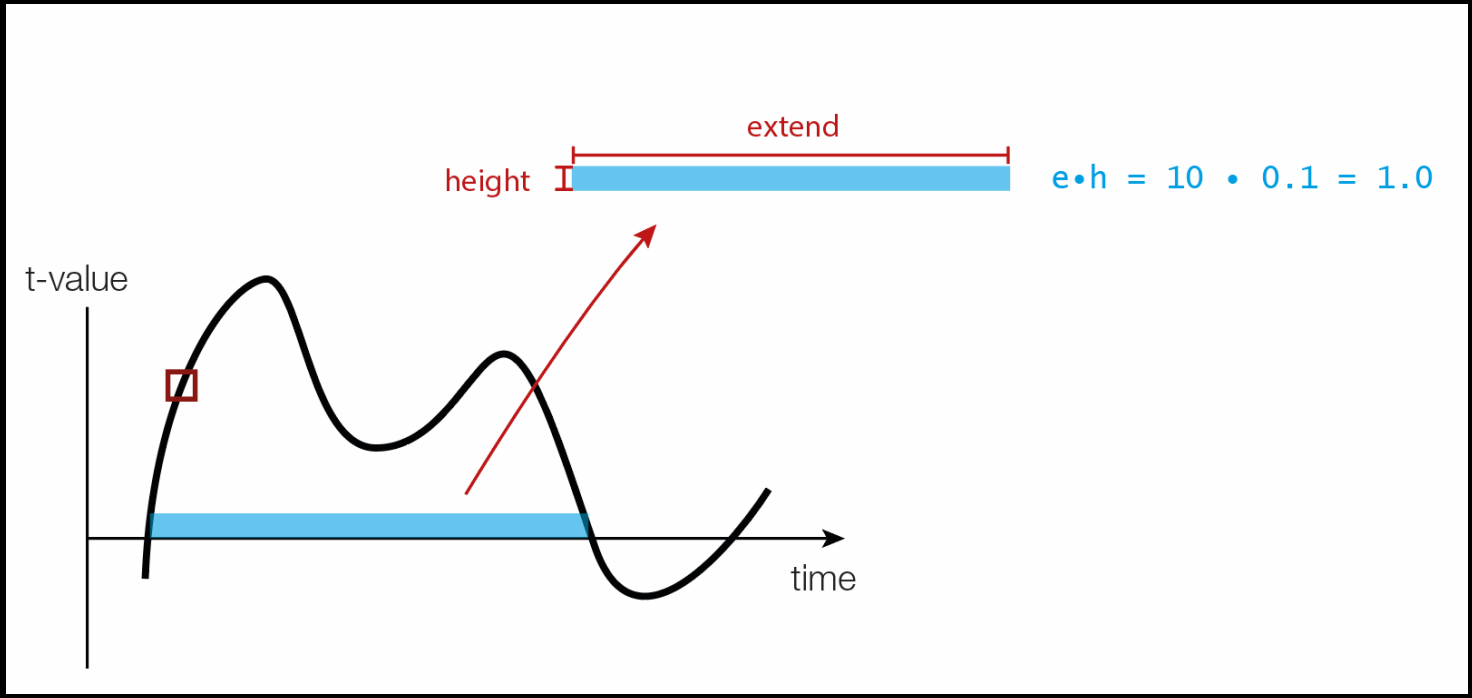
TFCE – threshold free cluster enhancement



Credit: Benedick Ehinger

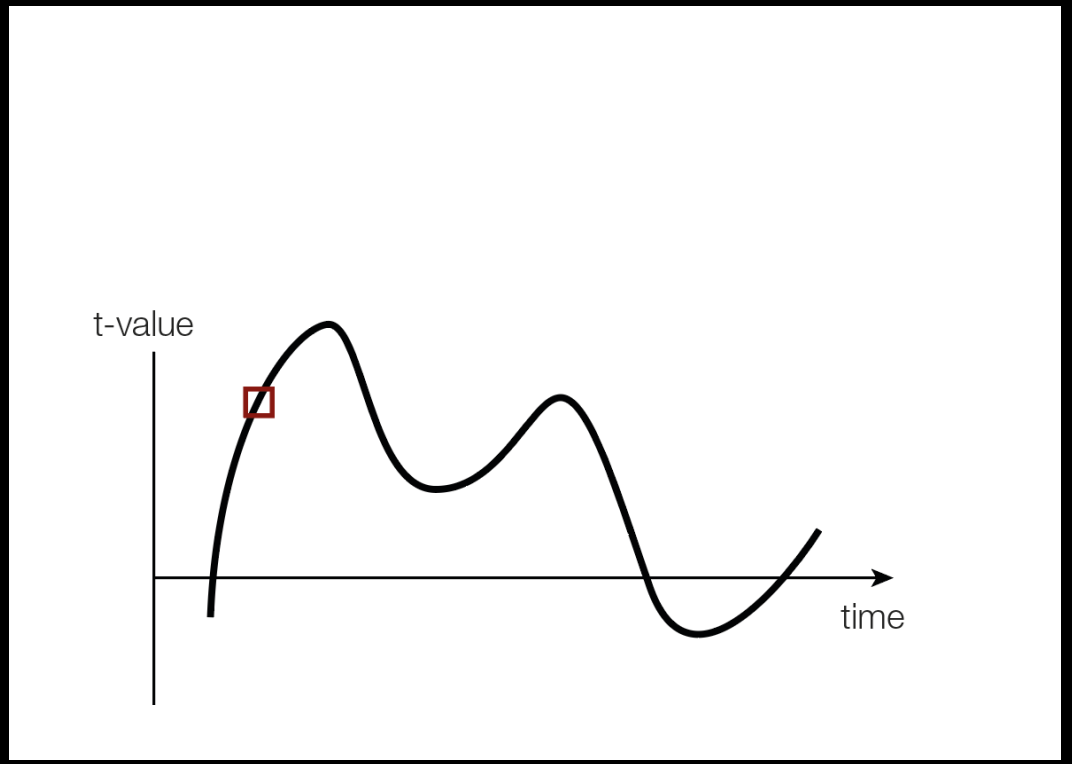
Smith SM, Nichols TE. Threshold free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage. 2009.

TFCE – threshold free cluster enhancement



Credit: Benedick Ehinger

Smith SM, Nichols TE. Threshold free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage. 2009.



Smith SM, Nichols TE. Threshold free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage. 2009.

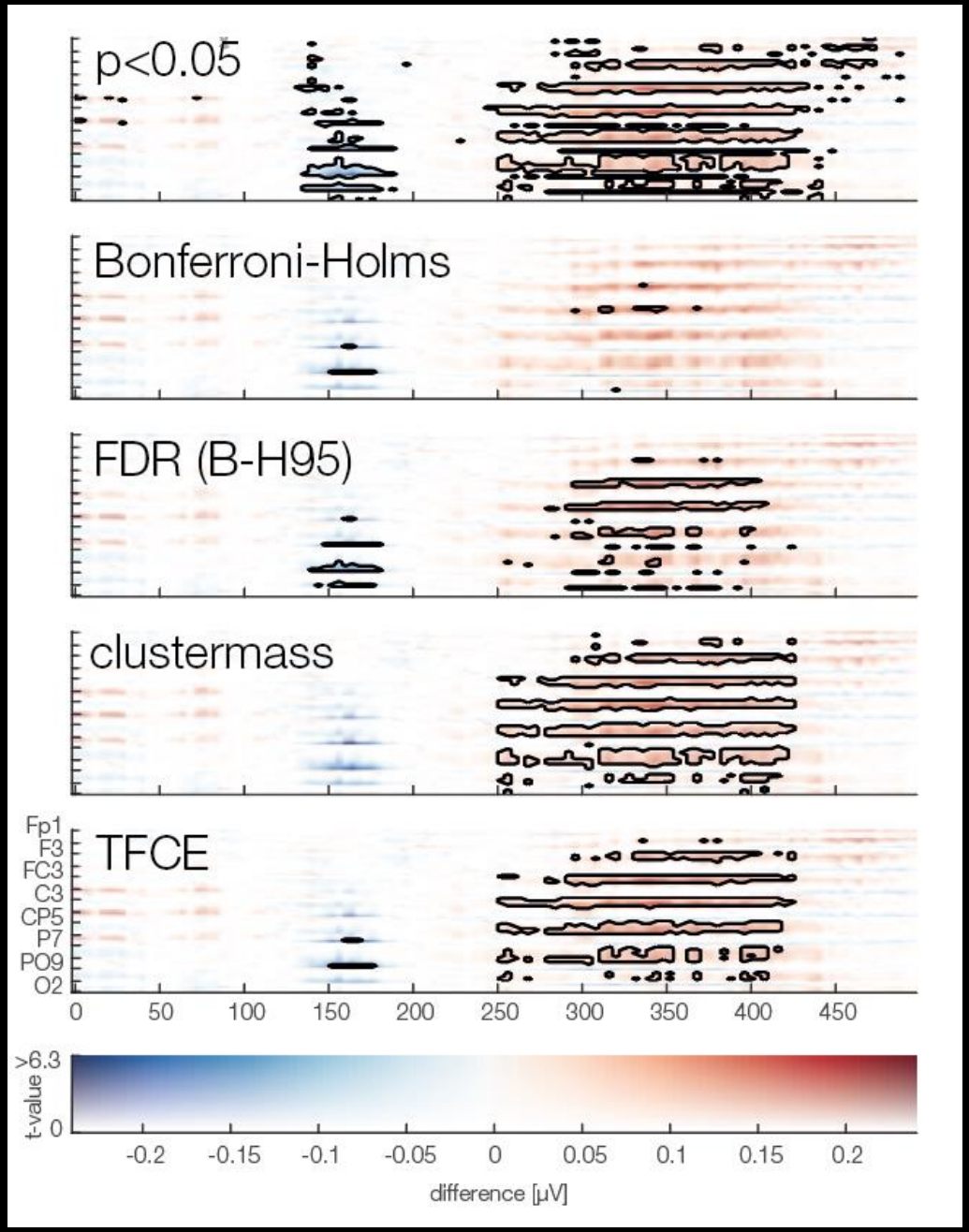
Uncorrected

Bonferroni-Holms
+ Fast
- Samples are not independent, therefore too strong of a correction

FDR
+ Fast
+ Interpretation clear
- Does not control FWER

Cluster
+ Effective use of prior knowledge
+ Appropriately control FWER
- Statistical interpretation limited
- Computationally expensive

TFCE
+ Effective use of prior knowledge
+ No initial threshold
+ Appropriately control FWER
- Statistical interpretation limited
- Very computationally expensive



Ehinger & Dimigen. PeerJ. 2019 Oct 24;7:e7838. doi: 10.7717/peerj.7838.

References

Delorme, A. 2006. Statistical methods. *Encyclopedia of Medical Device and Instrumentation*, vol 6, pp 240-264. Wiley interscience.

Genovese et al. 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15: 870-878

Nichols & Hayasaka, 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, 12:419-446

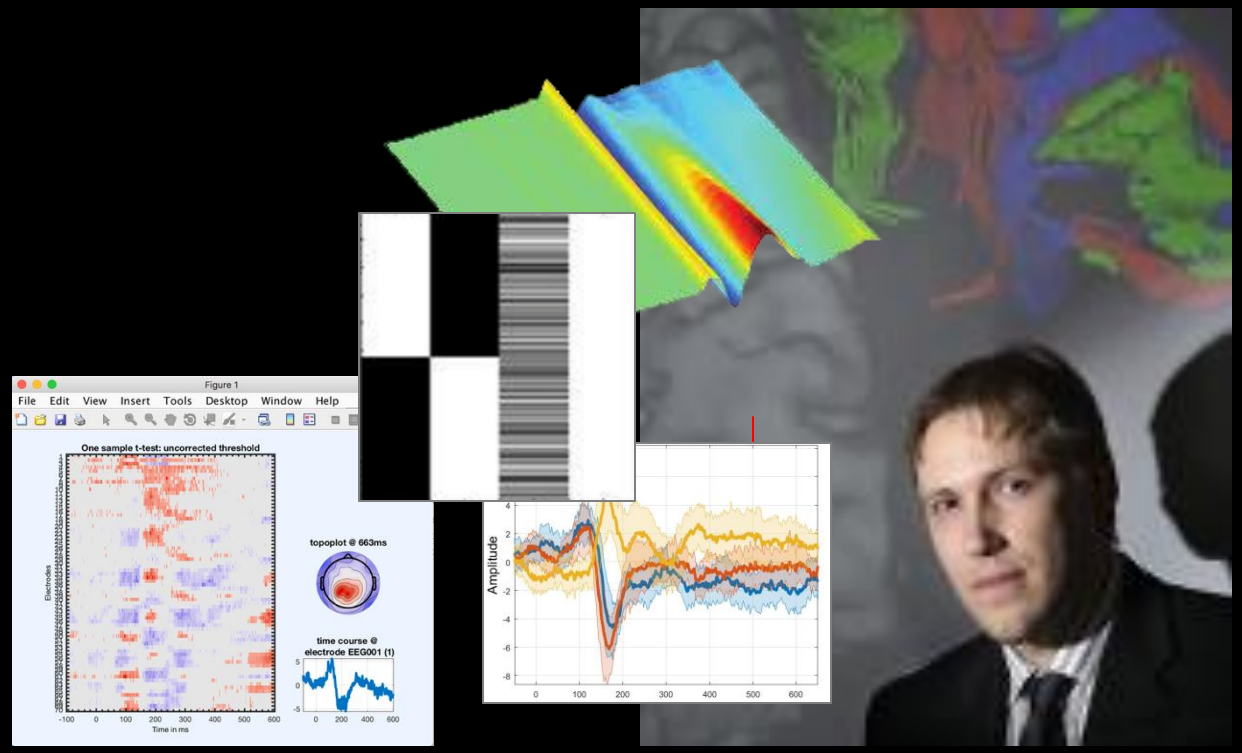
Maris, 2004. Randomization tests for ERP topographies and whole spatiotemporal data matrices. *Psychophysiology*, 41: 142-151

Maris et al. 2007. Nonparametric statistical testing of coherence differences. *Journal of Neuroscience Methods*, 163: 161-175

Groppe, D.M., Urbach, T.P., & Kutas, M. (2011) *Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review*. Psychophysiology, 48(12) pp. 1711-1725.

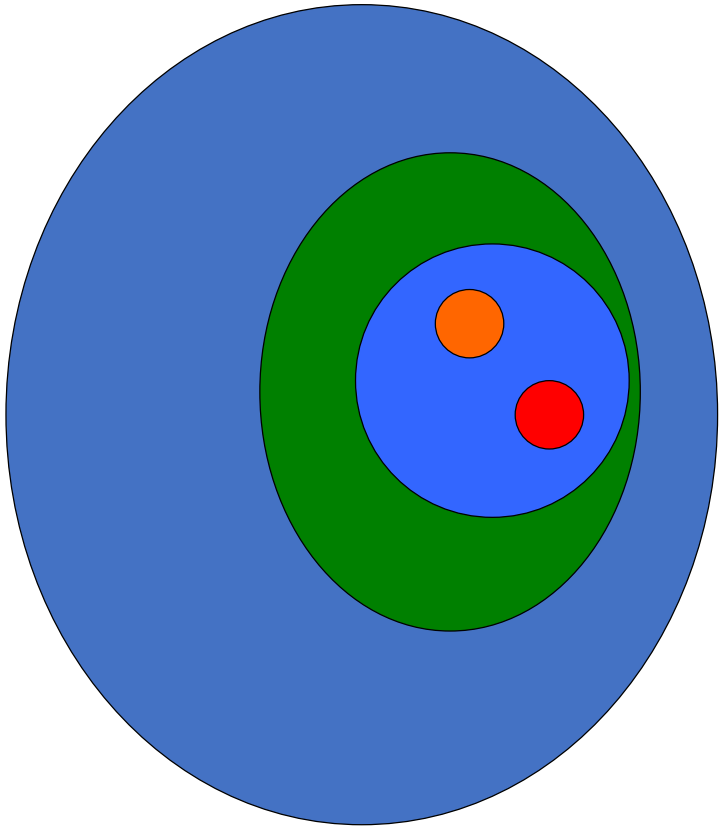
Thanks to G. Rousselet

General Linear Modeling in EEG



Cyril Pernet

The GLM family



T-tests

Simple regression

ANOVA

Multiple regression

General linear model

- Mixed effects/hierarchical
- Timeseries models (e.g., autoregressive)
- Robust regression
- Penalized regression (LASSO, Ridge)

Generalized linear models

- Non-normal errors
- Binary/categorical outcomes (logistic regression)

One-step solution

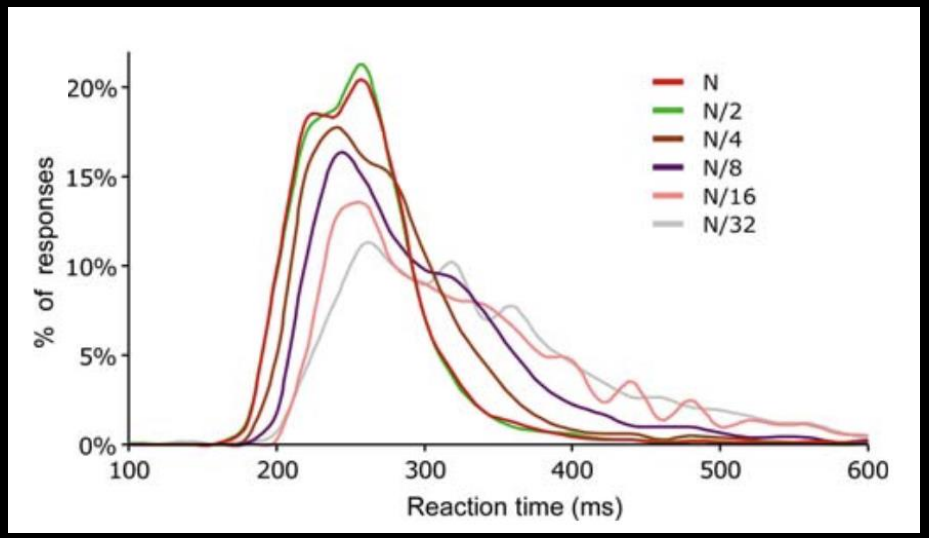
Iterative solutions (e.g., IWLS)

Slide from Tor Wager

A regression is a linear model

Varying factor: Luminance of image

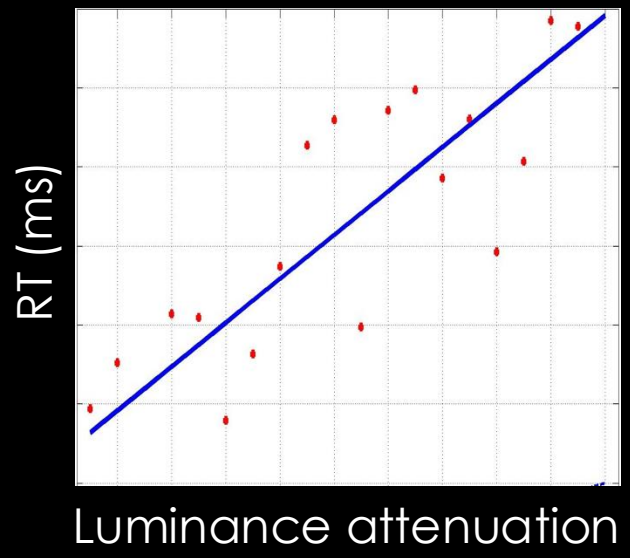
Outcome: Reaction time



Mace, M., Delorme, A., Richard, G., Fabre-Thorpe, M. (2010) Spotting animals in natural scenes: efficiency of humans and monkeys at very low contrasts. *Animal Cognition*, 13(3):405-18.

A regression is a linear model

- Given an experimental measure x (e.g. luminance)
- We collect data RT (e.g. reaction time)
- Model: $RT = \beta_0 + x\beta_1 + \varepsilon$
- Do some maths / run a software to find β_1 and β_0
- $\hat{RT} = 23.6 + 2.7x$



A regression is a linear model

For each trial

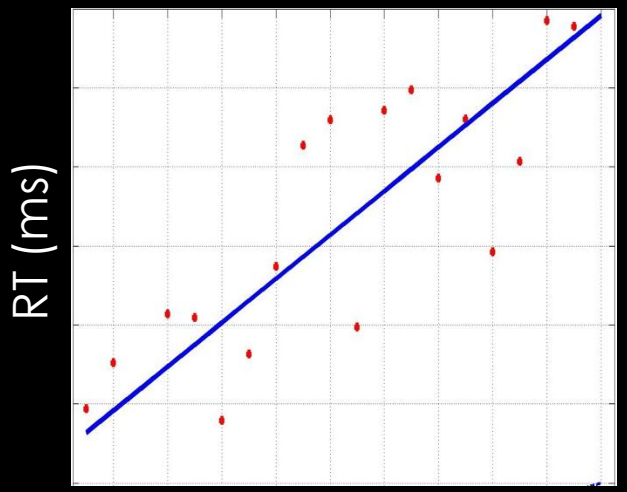
Luminance level

$RT_1 = \beta_0 + 10*\beta_1 + \varepsilon_1$

$RT_2 = \beta_0 + 5*\beta_1 + \varepsilon_2$

$RT_3 = \beta_0 + 7*\beta_1 + \varepsilon_3$

…



To test for significance compare the original regression model $RT_i = \beta_0 + c_i*\beta_1 + \varepsilon_i$ with the simplified model $RT_i = \beta_0 + \varepsilon_i$

Compare the fit

Test if 0 included in confidence interval

An ANOVA is a linear model

Varying factor: Type of image

Outcome: Reaction time (go/no-go)

$$RT_{i,j} = \beta_0 + \beta_i + \varepsilon_{i,j}$$

that is to say the data (e.g. RT) = a constant term (grand mean β_0) + the effect of a treatment (β_1 for fishes 1 and β_2, β_3 for birds and reptiles) and the error term ($\varepsilon_{i,j}$)

For trial 4 (for example first trial of birds) we have

$$RT_{2,1} = \beta_0 + 0*\beta_1 + 1*\beta_2 + 0*\beta_3 + \varepsilon_{2,1}$$

This is a GLM that is equivalent to an ANOVA

For trial 13 (for example second trial of birds) we have

$$RT_{2,2} = \beta_0 + 0*\beta_1 + 1*\beta_2 + 0*\beta_3 + \varepsilon_{2,2}$$

Statistics: if there is an effect of treatment then error of the simplified model $RT_{i,j} = \beta_0 + \varepsilon_{i,j}$ should be lower than the original model $RT_{i,j} = \beta_0 + \beta_i + \varepsilon_{i,j}$

Compare the fit

A GLM can do both a Regression and an ANOVA (ANCOVA)

Varying factor: Type of image **AND** luminance
Outcome: Reaction time (go/no-go)

For example, for trial (first bird with lumiance $c_{2,1}$) we have

$$RT_{2,1} = \beta_0 + \underbrace{0*\beta_1 + 1*\beta_2 + 0*\beta_3 + 0*\beta_3}_{\text{Categorical var. ANOVA}} + \underbrace{c_{2,1}*\beta_4}_{\text{Continous var. REGRESSION}} + \varepsilon_{2,1}$$

The design matrix

Y	Gp
8	1
9	1
7	1
5	2
7	2
3	2
3	3
4	3
1	3
6	4
4	4
9	4

$y(1..3) = 1x\beta1+0x\beta2+0x\beta3+0x\beta4+c+error$

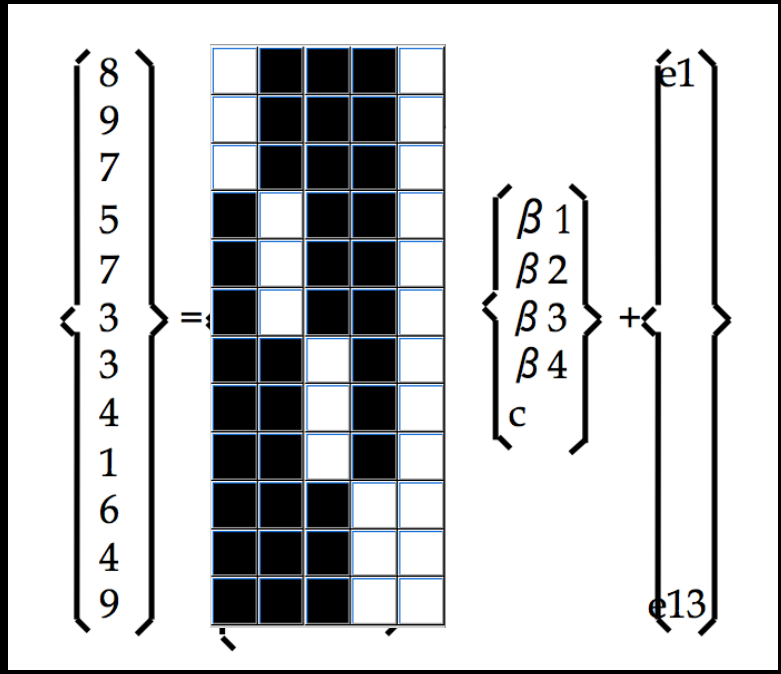
$y(4..6) = 0x\beta1+1x\beta2+0x\beta3+0x\beta4+c+error$

$y(7..9) = 0x\beta1+0x\beta2+1x\beta3+0x\beta4+c+error$

$y(10..12) = 0x\beta1+0x\beta2+0x\beta3+1x\beta4+c+error$

Design matrix

$G_1\ G_2\ G_3\ G_4\ C$

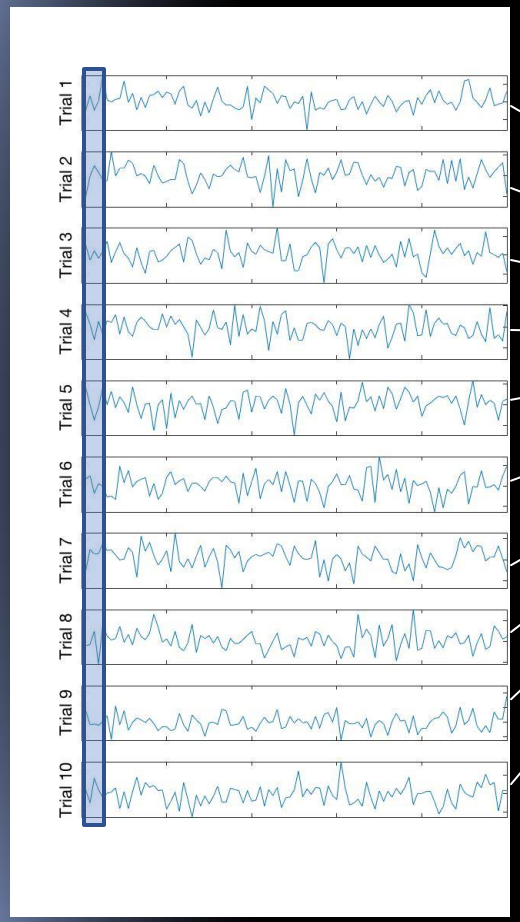


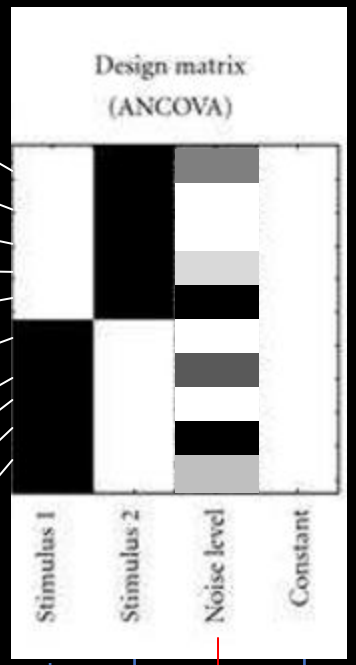
Y = D * β + ε

Measures — Model/ Design matrix — Unknown — Errors

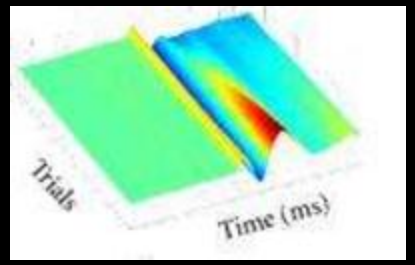
Linear Modeling of EEG data: level 1

Electrode 1

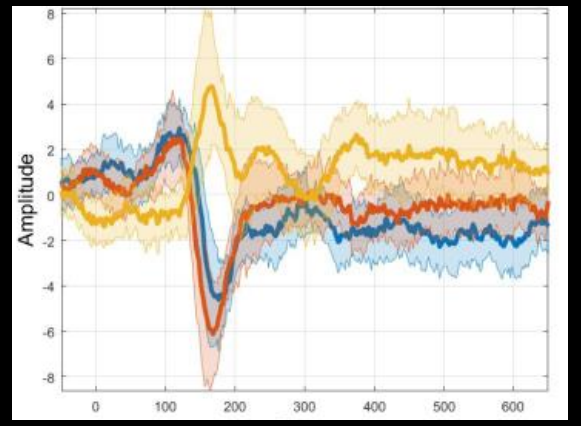




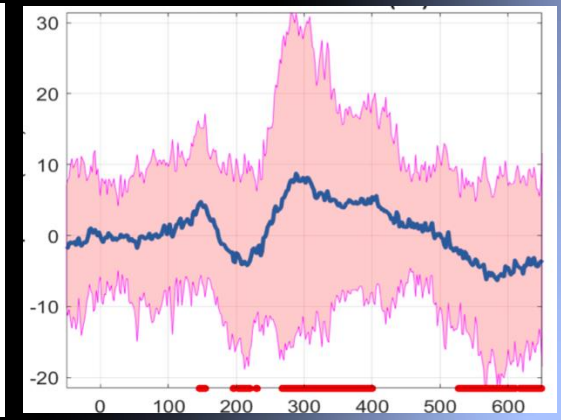
Continuous var.



Categorical var.



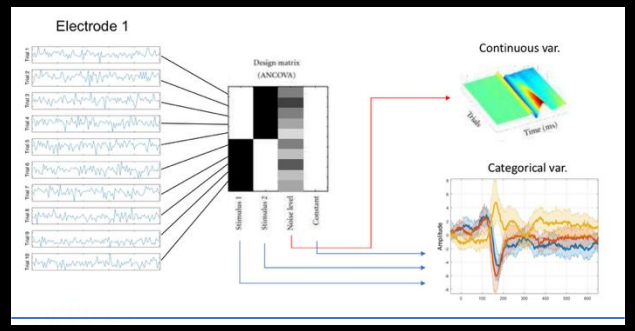
Electrode difference
Between conditions



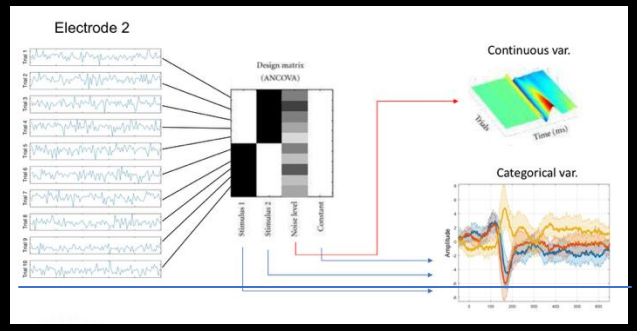
Significance: bootstrap trials to get confidence interval of βs

Linear Modeling of EEG data: level 1

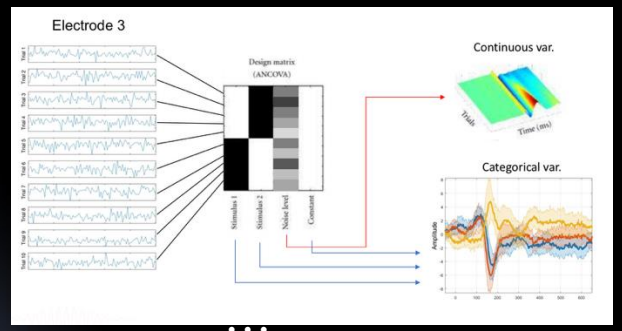
Electrode 1



Electrode 2

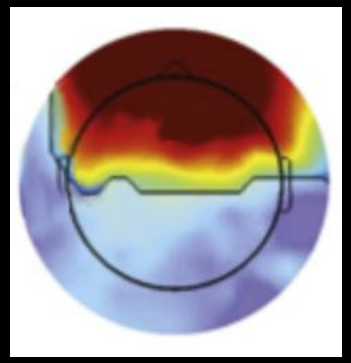


Electrode 3

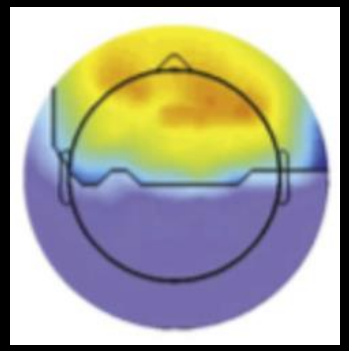


...

Scalp topography of
beta difference
at a given latency



Scalp topography of
potential difference
(masked using beta signif.)

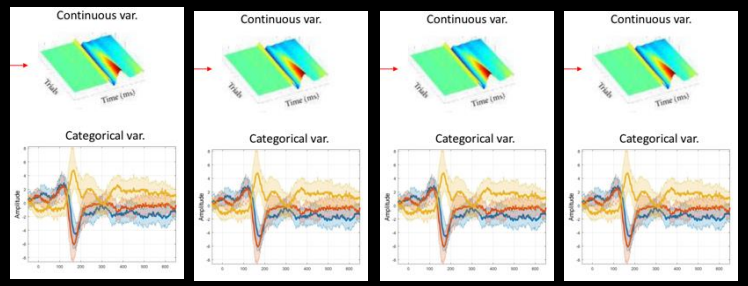


Limit of the regions masked for significance

Linear Modeling of EEG data: level 2

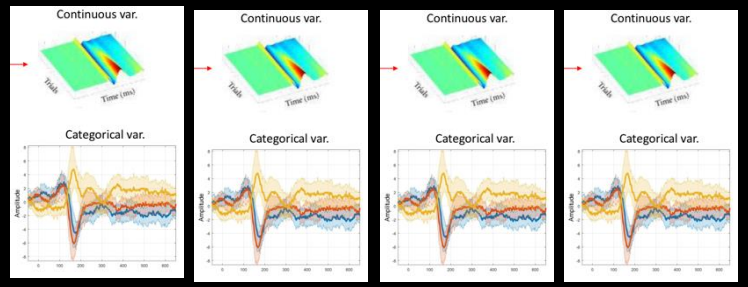
Participant 1

chan1 chan2 chan3 chan4



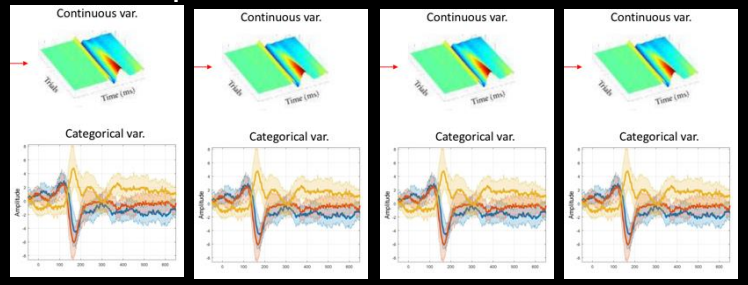
…

Participant 2



…

Participant 3



…

…

Level 2

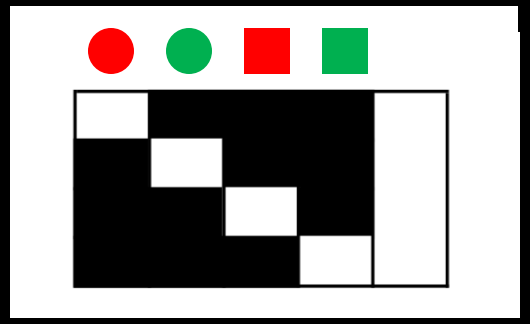
Standard stats.
2nd level-GLM

GLM: ordinary least square (OLS) vs. weighted least square (WLS)

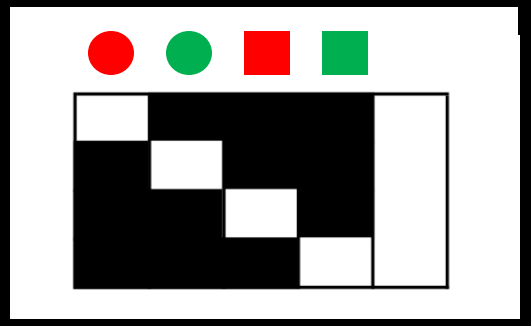
Linear Modeling of EEG data: level 2

Level 1

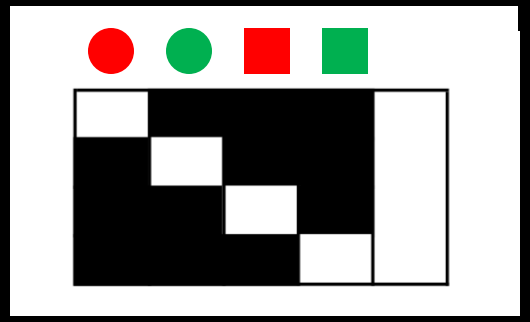
Participant 1



Participant 2



Participant 3



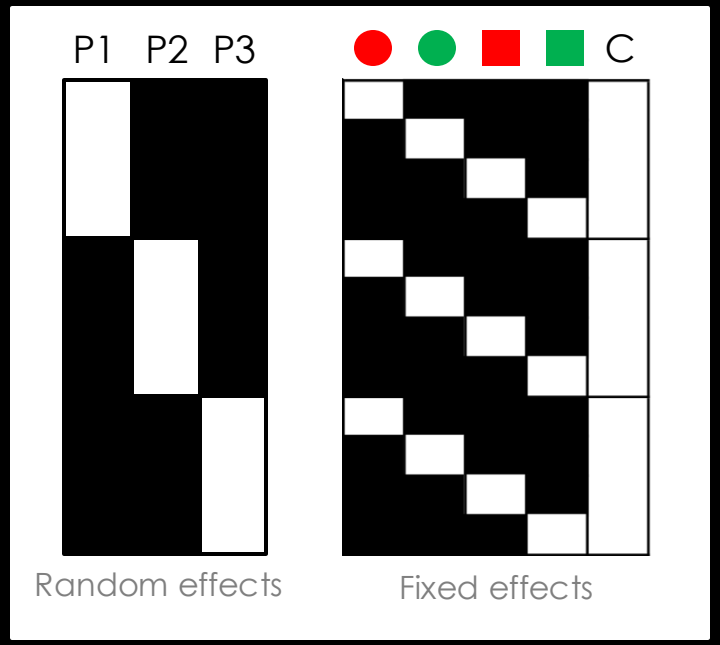
Level 2

2-way ANOVA:

- Main effect 1 (shape)
- Main effect 2 (color)
- Interaction

Linear Modeling of EEG data: level 2

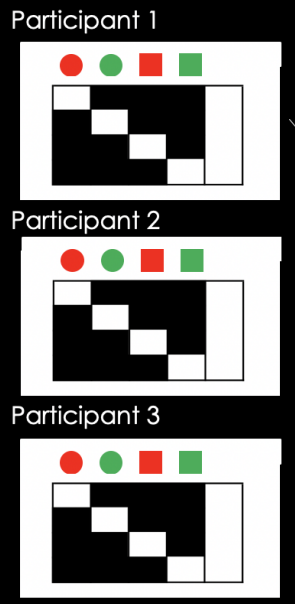
Mixed effect model
(still a GLM)



VS

Hierarchical GLM

Level 1



Level 2

2-way ANOVA:
- Main effect 1 (shape)
- Main effect 2 (color)
- Interaction

The all-powerful mixed model

$$Y = \text{gender} + \text{age} + \text{spectral\_power\_cz\_10hz} + (1 \mid \text{subject})$$

Categorical var. Continuous vars. Random effect

$$
\begin{aligned}
Y_{trial\_1,s\_1} &= gender_{s\_1} + age_{s\_1} + spectral\_power\_cz\_10hz_{trial\_1,s\_1} + Constant_{s\_1} + Error_{trial\_1,s\_1} \\
Y_{trial\_2,s\_1} &= gender_{s\_1} + age_{s\_1} + spectral\_power\_cz\_10hz_{trial\_2,s\_1} + Constant_{s\_1} + Error_{trial\_2,s\_1} \\
Y_{trial\_3,s\_1} &= gender_{s\_1} + age_{s\_1} + spectral\_power\_cz\_10hz_{trial\_3,s\_1} + Constant_{s\_1} + Error_{trial\_3,s\_1} \\
&\dots \\
Y_{trial\_1,s\_2} &= gender_{s\_2} + age_{s\_2} + spectral\_power\_cz\_10hz_{trial\_1,s\_2} + Constant_{s\_2} + Error_{trial\_1,s\_2} \\
Y_{trial\_2,s\_2} &= gender_{s\_2} + age_{s\_2} + spectral\_power\_cz\_10hz_{trial\_2,s\_2} + Constant_{s\_2} + Error_{trial\_2,s\_2} \\
Y_{trial\_3,s\_2} &= gender_{s\_2} + age_{s\_2} + spectral\_power\_cz\_10hz_{trial\_3,s\_2} + Constant_{s\_2} + Error_{trial\_3,s\_2} \\
&\dots
\end{aligned}
$$

The all-powerful mixed model

Y = gender + age + spectral_power_cz_10hz + (1 | subject)

Categorical var.

Continuous vars.

Random effect

MATLAB: model = fitglme(df, y ~ pred1 + (1 | subject)');

Python: model = Lmer('y ~ pred1 + (1 | subject)', data=df)

R: model <- glmer(y ~ pred1 + (1 | subject), data = df)

The all-powerful mixed model

Y = gender + age + spectral_power_cz_10hz + (1 | subject)

Categorical var.

Continuous vars.

Random effect

MATLAB: model = fitglme(df, y ~ pred1 + (1 | subject)', 'Distribution', 'Binomial');

Python: model = Lmer('y ~ pred1 + (1 | subject)', data=df, family='binomial')

R: model <- glmer(y ~ pred1 + (1 | subject), data = df, family = binomial)

The all-powerful mixed model

Y = gender + age + spectral_power_cz_10hz + (1 | subject)

Categorical var.

Continuous vars.

Random effect

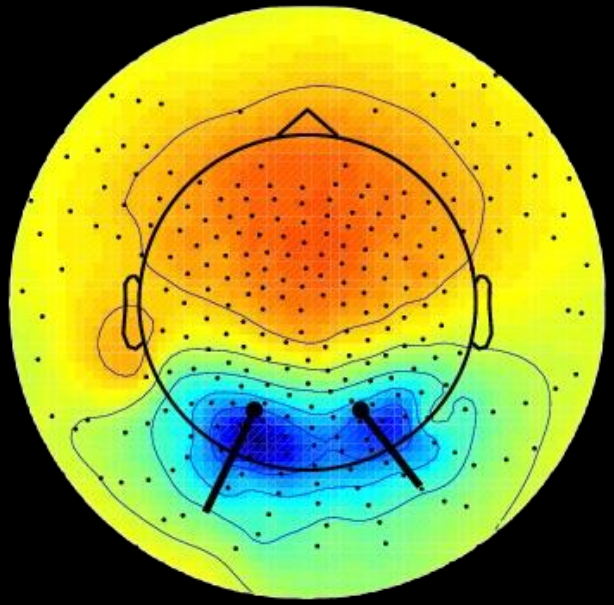
MATLAB: model = fitglme(df, y ~ pred1 + (1 | subject)', 'Distribution', 'Binomial');

Python: model = Lmer('y ~ pred1 + (1 | subject)', data=df, family='binomial')

R: model <- glmer(y ~ pred1 + (1 | subject), data = df, family = binomial)

Correction for multiple comparisons

MATLAB: limo_tfce() – general/bootstrap
Python: mne.stats.spatio_temporal_cluster_1samp_test() – only t-values and sign test permutation



The End