Robust statistics

Arnaud Delorme
(with feedback/slides from C. Pernet & G. Roussellet)

Robust statistics

Parametric & non-parametric statistics: use mean and standard deviation (t-test, ANOVA, …)

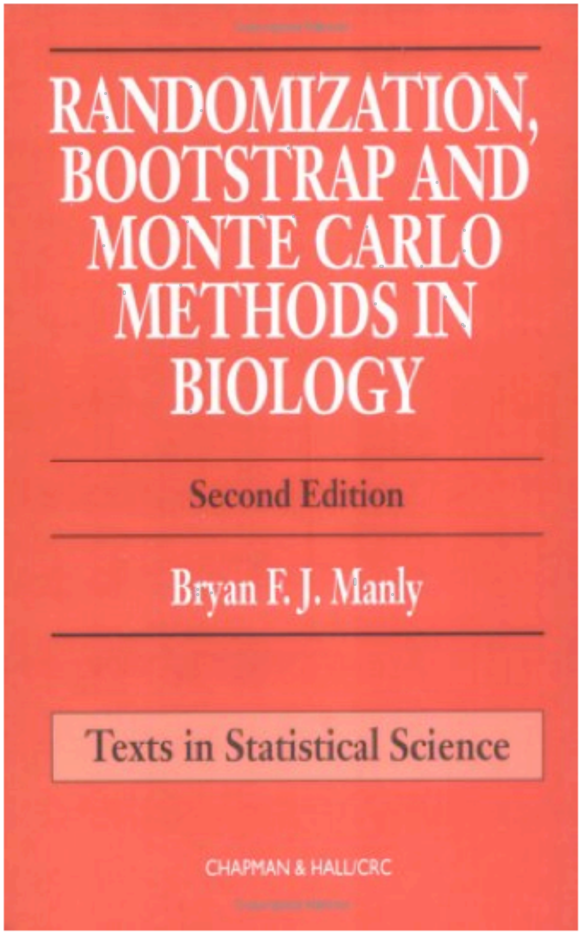
Bootstrap and permutation methods: shuffle/bootstrap data and recompute measure of interest. Use the tail of the distribution to asses significance.

Correction for multiple comparisons: computing statistics on time(/frequency) series requires correction for the number of comparisons performed.

Take-home messages

- *Look at your data! Show your data!*
- *A perfect & universal statistical recipe does not exist*
- *Keep exploring: there are many great options, most of them available in free softwares and toolboxes*

References



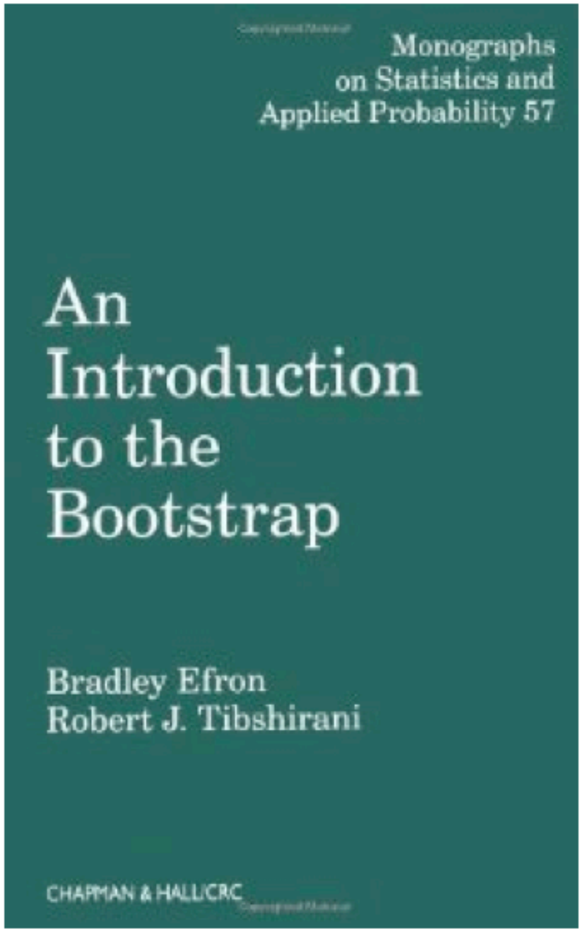
RANDOMIZATION, BOOTSTRAP AND MONTE CARLO METHODS IN BIOLOGY

Second Edition

Bryan F. J. Manly

Texts in Statistical Science

CHAPMAN & HALL/CRC

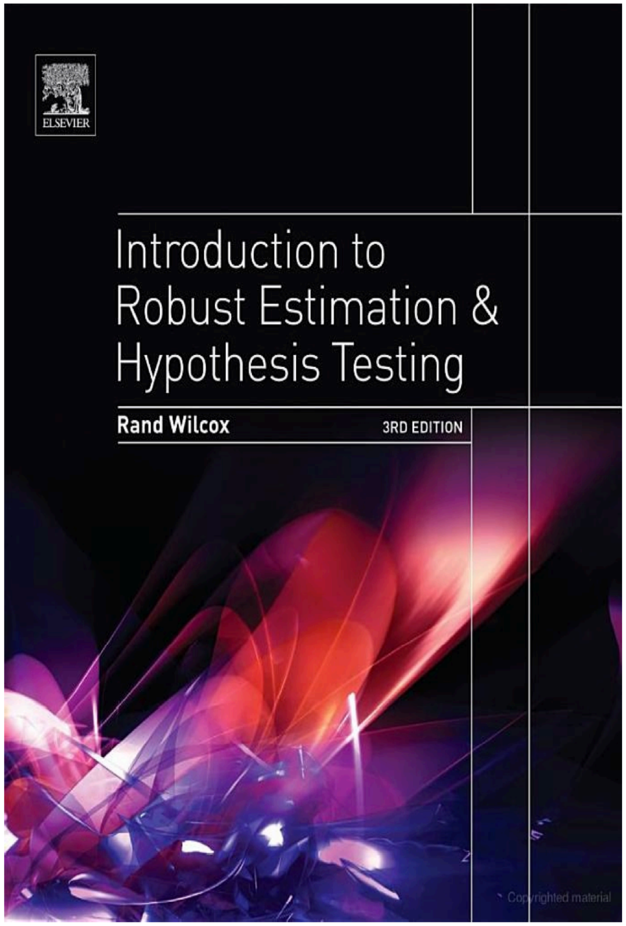


Monographs on Statistics and Applied Probability 57

An Introduction to the Bootstrap

Bradley Efron
Robert J. Tibshirani

CHAPMAN & HALL/CRC



ELSEVIER

Introduction to Robust Estimation & Hypothesis Testing

Rand Wilcox

3RD EDITION

Parametric statistics

Assume gaussian distribution of data

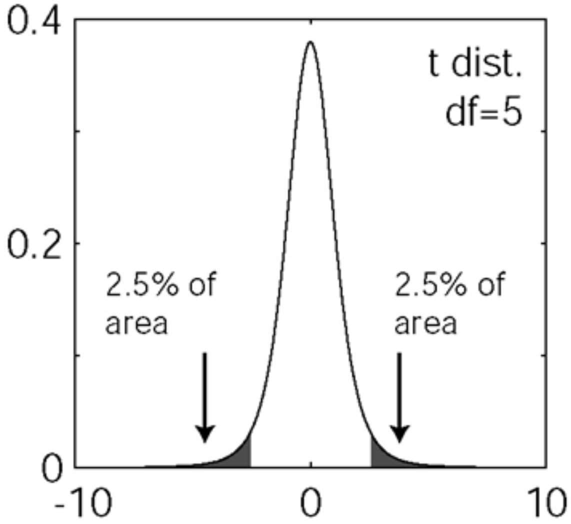
T-test: Compare paired/ unpaired
Samples for continuous data. In EEGLAB, used for grand-average ERPs.

Paired

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

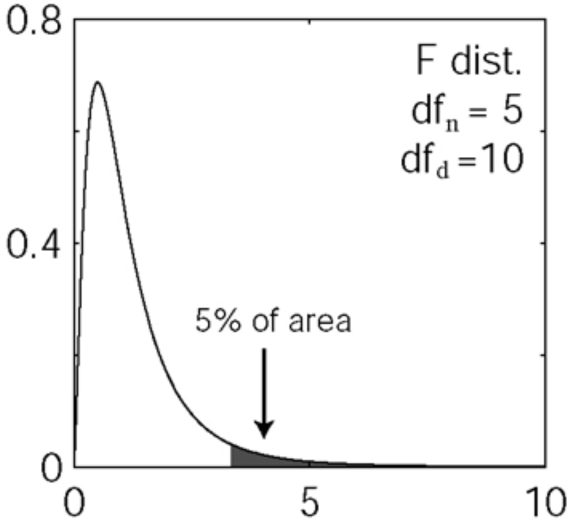
Unpaired

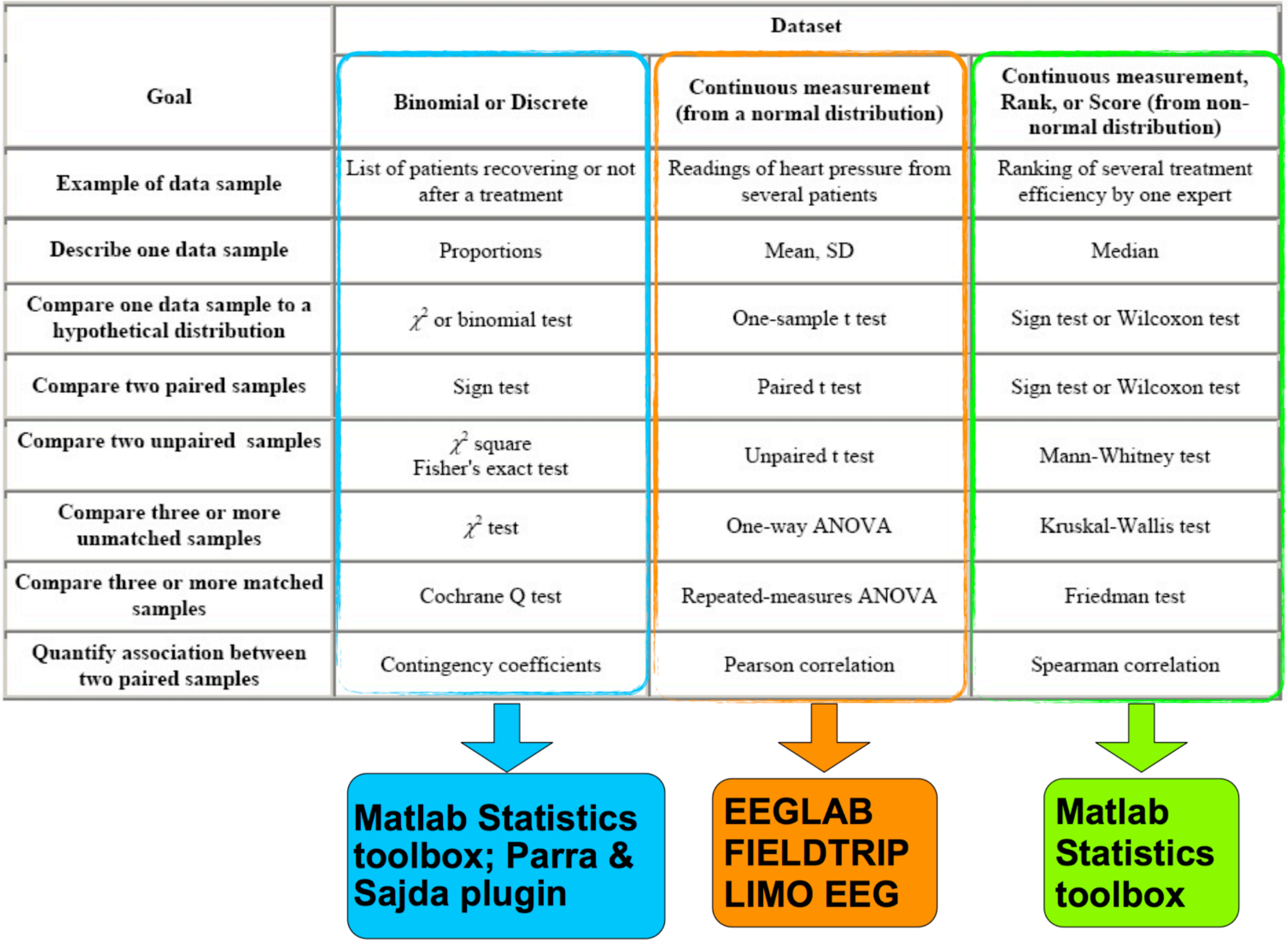
$$t = \sqrt{N}\,\frac{Mean_A - Mean_B}{\sqrt{(SD_A)^2 - (SD_B)^2}}$$



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

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





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

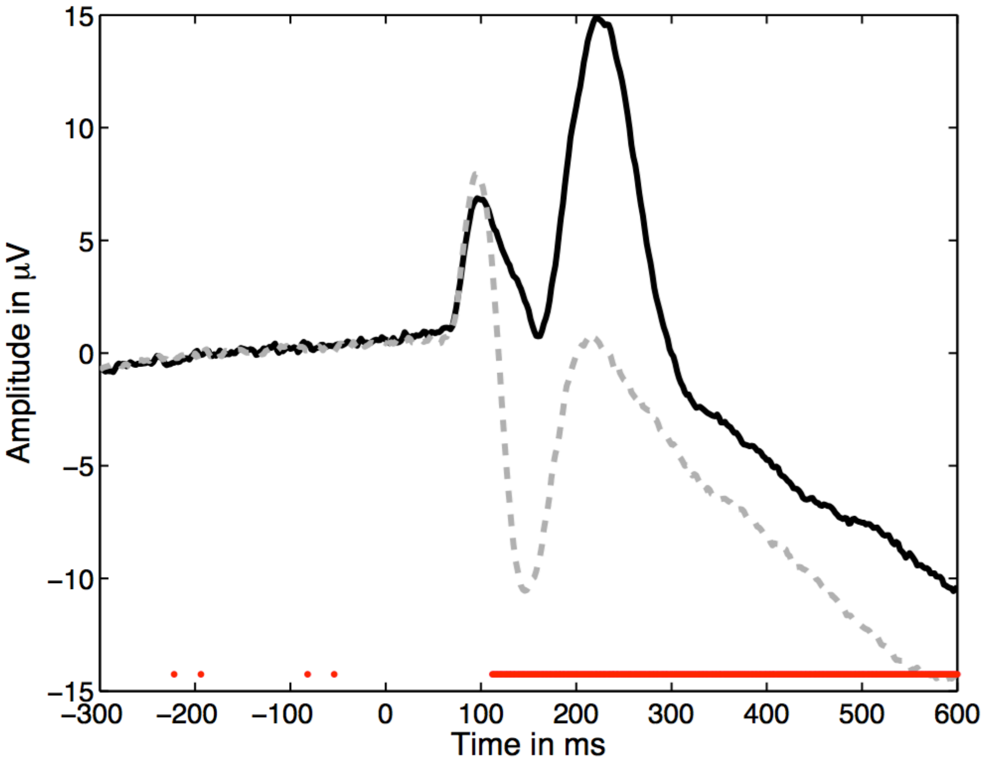
Matlab Statistics toolbox; Parra & Sajda plugin

EEGLAB
FIELDTRIP
LIMO EEG

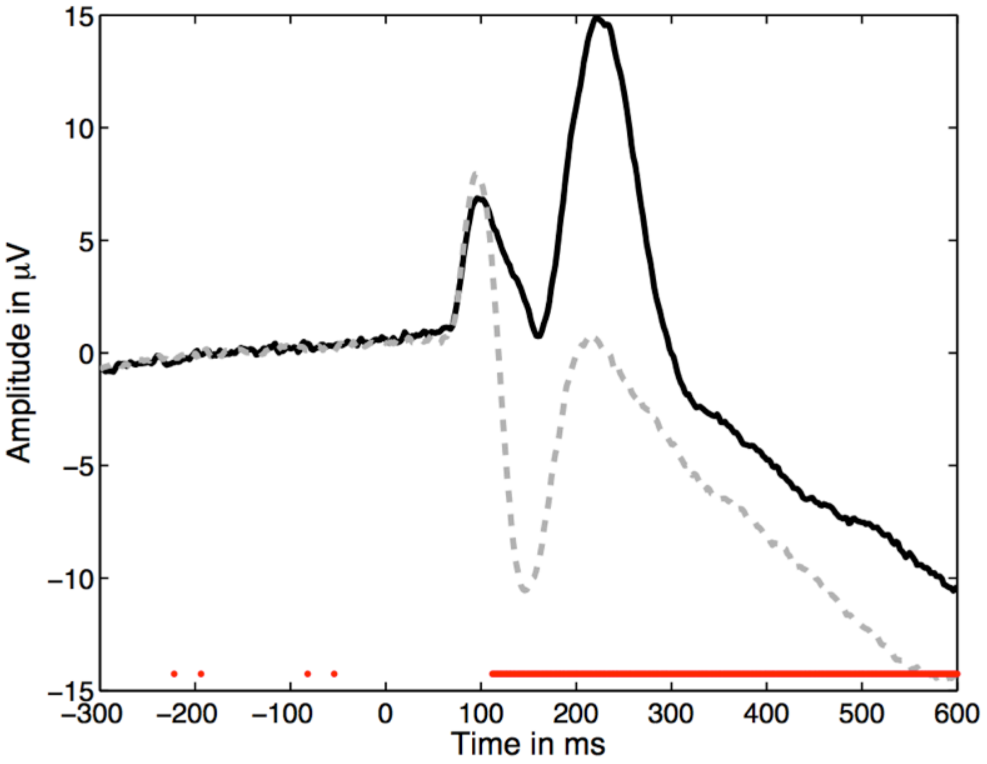
Matlab Statistics toolbox

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

Why the standard figure is not good enough



Why the standard figure is not good enough

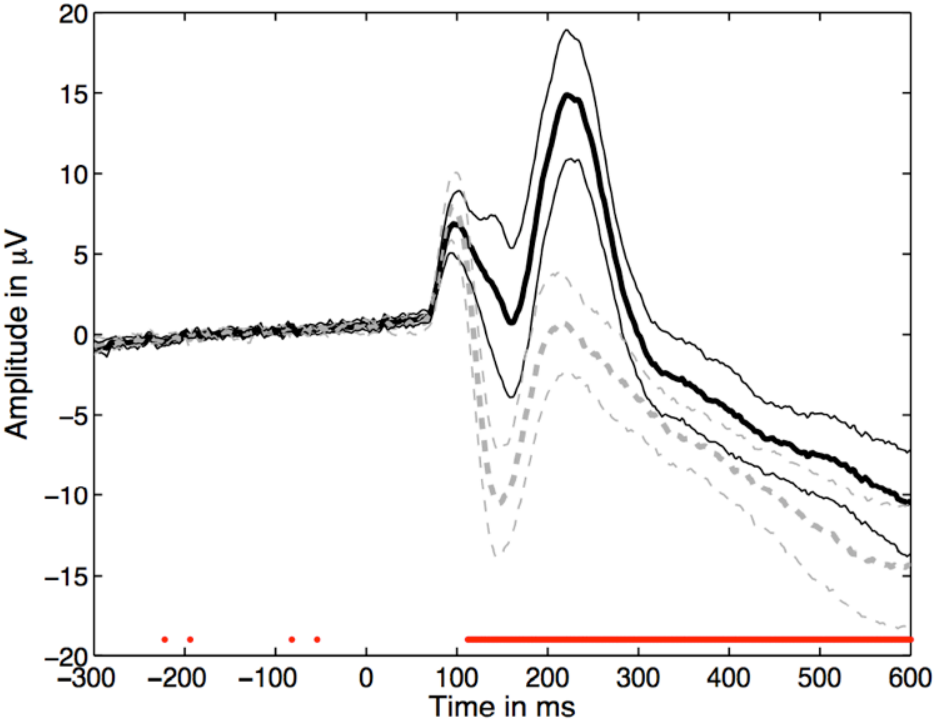


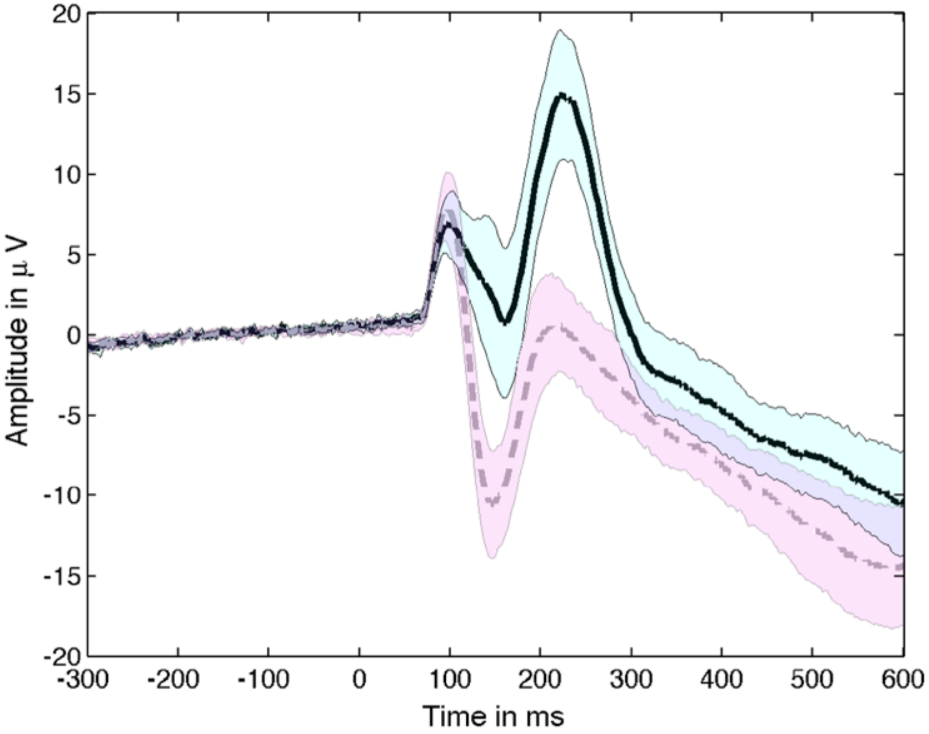
Significant effect?
- interesting?
- how many subjects?
- effect size?

Non-significant effect?
- not there?
- lack of power?
- how many subjects?
- effect size?

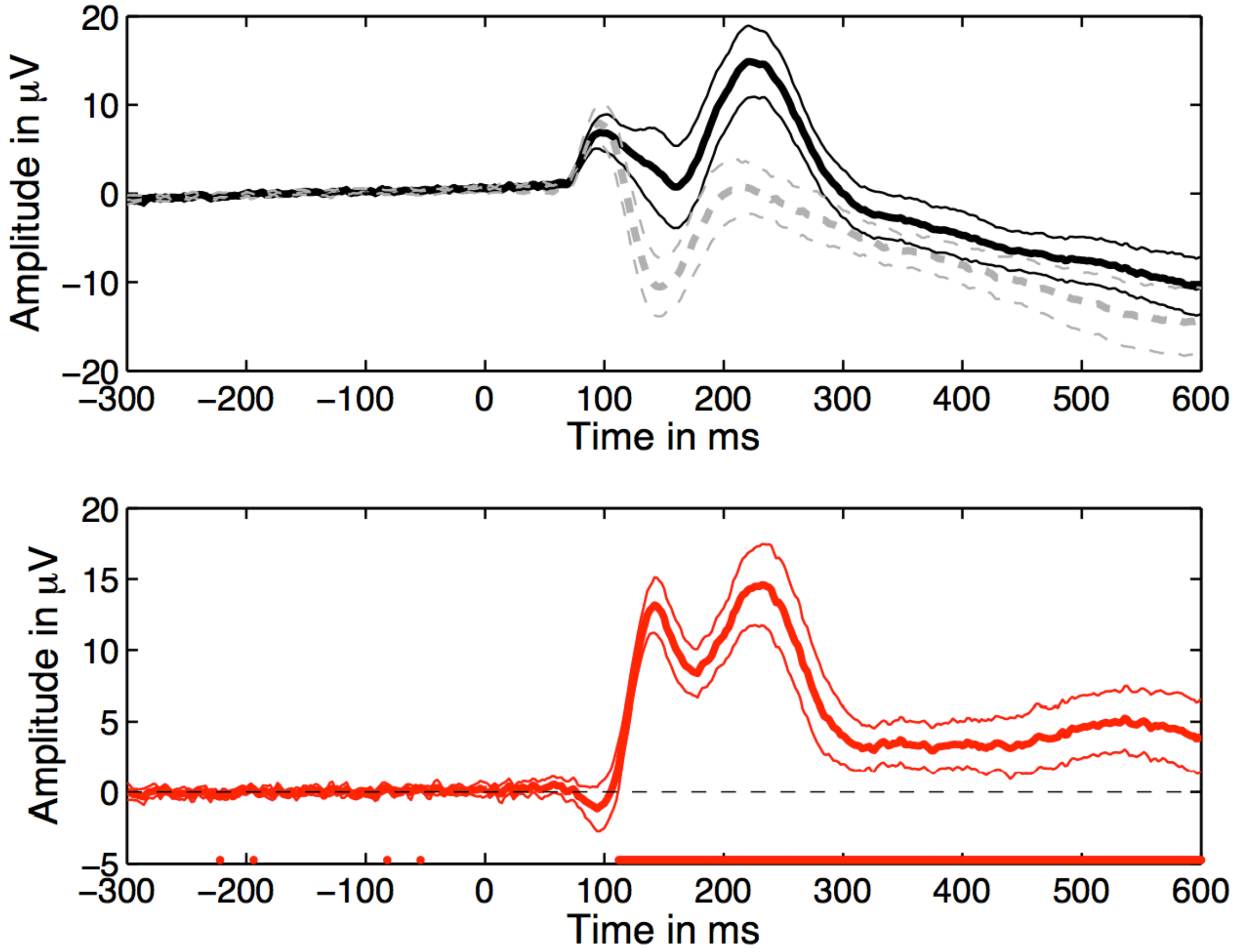
Interpretations should be limited to what was measured: group differences in means

Add confidence intervals

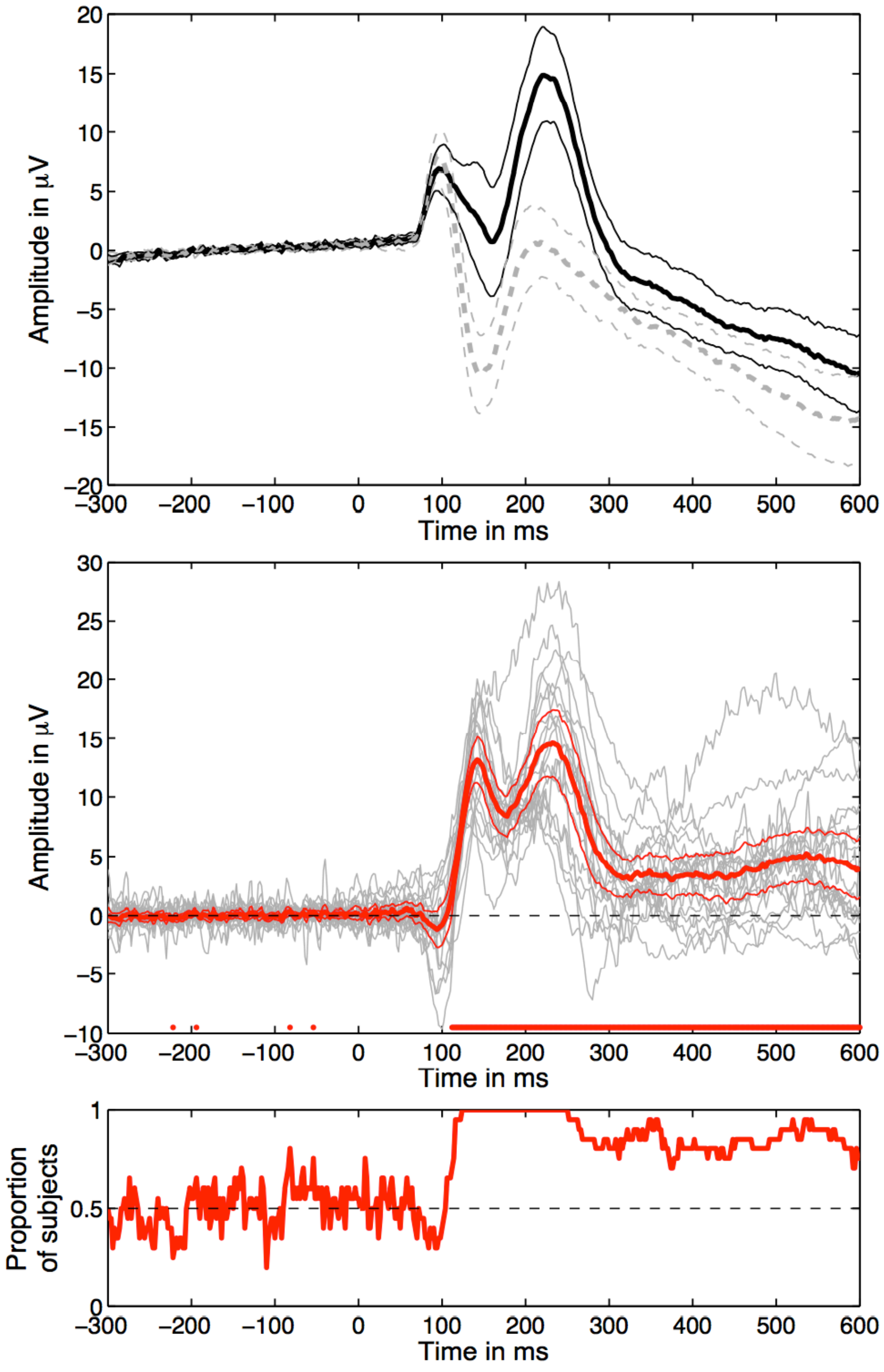




Add plot of the difference



How many subjects show an effect?

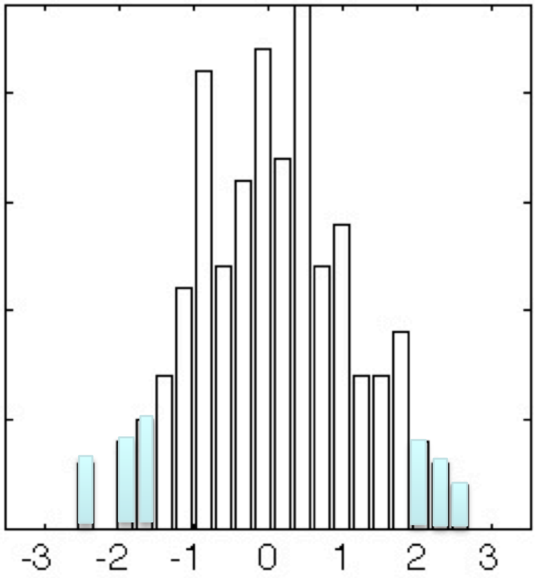


Robust measures of central tendency (location)

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

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

Trimmed means



- 20% trimmed means provide high power under normality and high power in the presence of outliers
- Rand Wilcox, 2012, Introduction to Robust Estimation and Hypothesis Testing, Elsevier **ERP application:** Rousselet, Husk, Bennett & Sekuler, 2008, *J. Vis.* + Desjardins 2013

Non-parametric statistics

Paired t-test → Wilcoxon

Unpaired t-test → Mann-Whitney

One way ANOVA → Kruskal Wallis

Values

Ranks

BOTH ASSUME NORMAL DISTRIBUTIONS

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

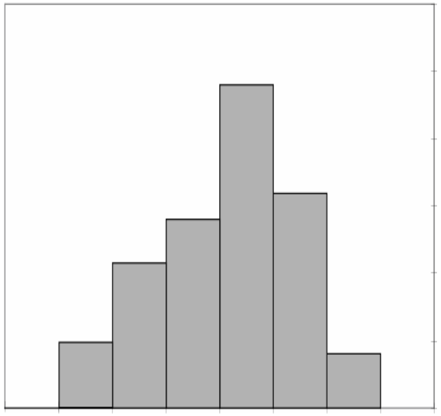
1. Randomization approach
2. Bootstrap approach

Bootstrap: central idea

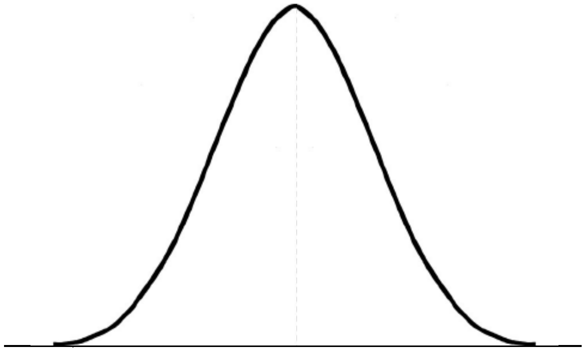
- “The bootstrap is a computer-based method for assigning measures of accuracy to statistical estimates.” Efron & Tibshirani, 1993

- “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



Population

H0: the mean is not 0 for the population

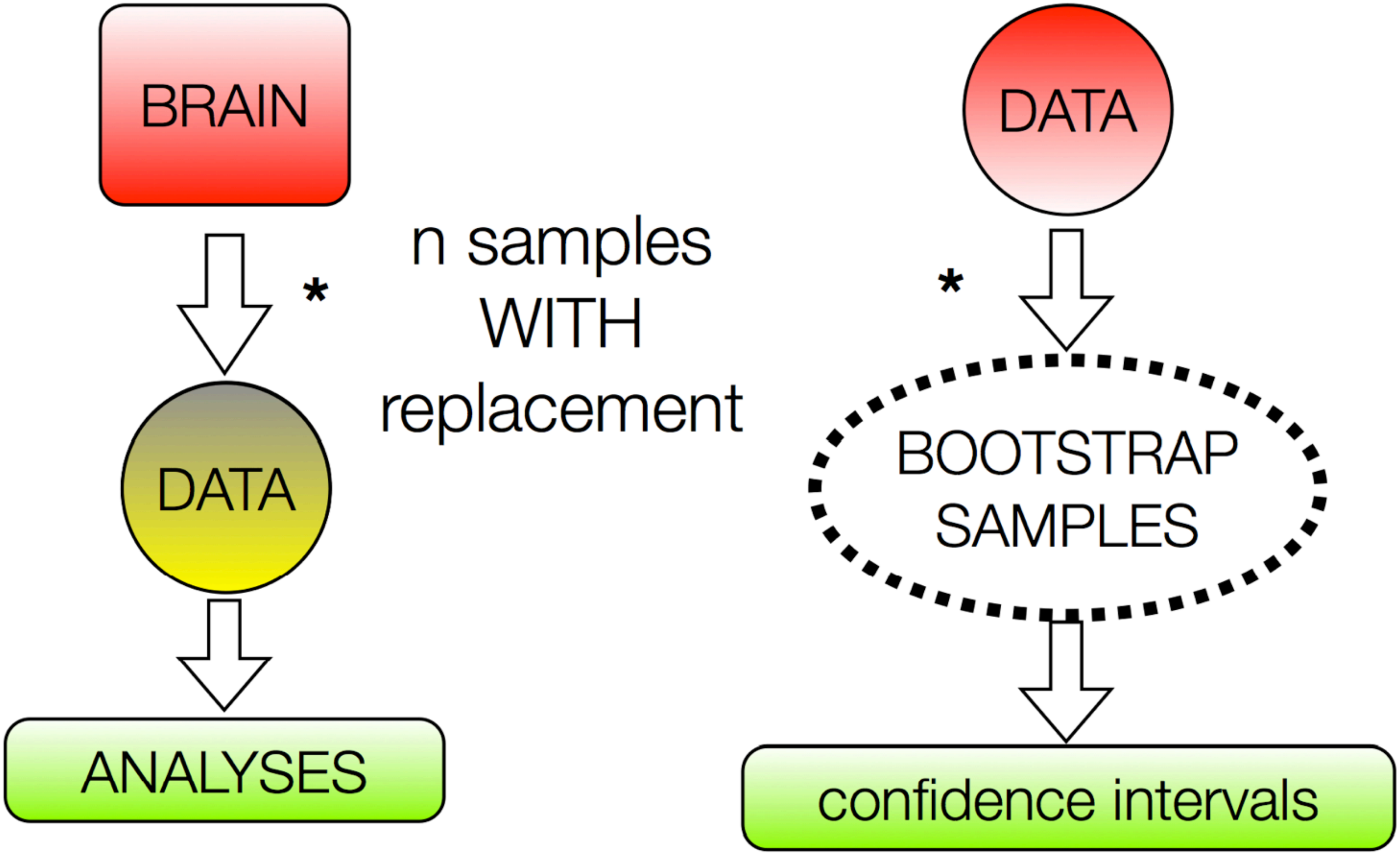
given that we have no other information about the population, the sample is our best single estimate of the population

Percentile bootstrap: general recipe

- sample = X1, ..., Xn
- resample n observations with replacement
- compute estimate
- repeat B times

with B large enough the B estimates provide a good approximation of the distribution of the estimate of the sample

Bootstrap philosophy



BRAIN

DATA

ANALYSES

*

n samples WITH replacement

DATA

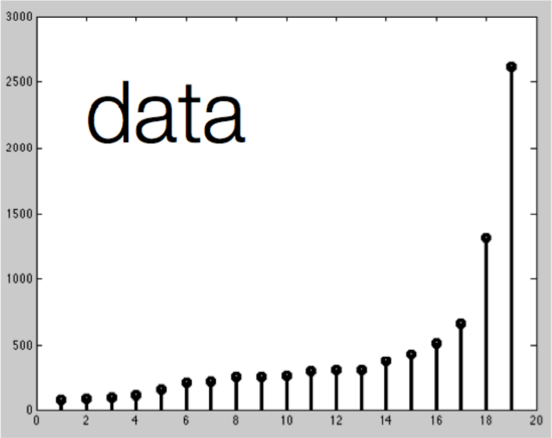
*

BOOTSTRAP SAMPLES

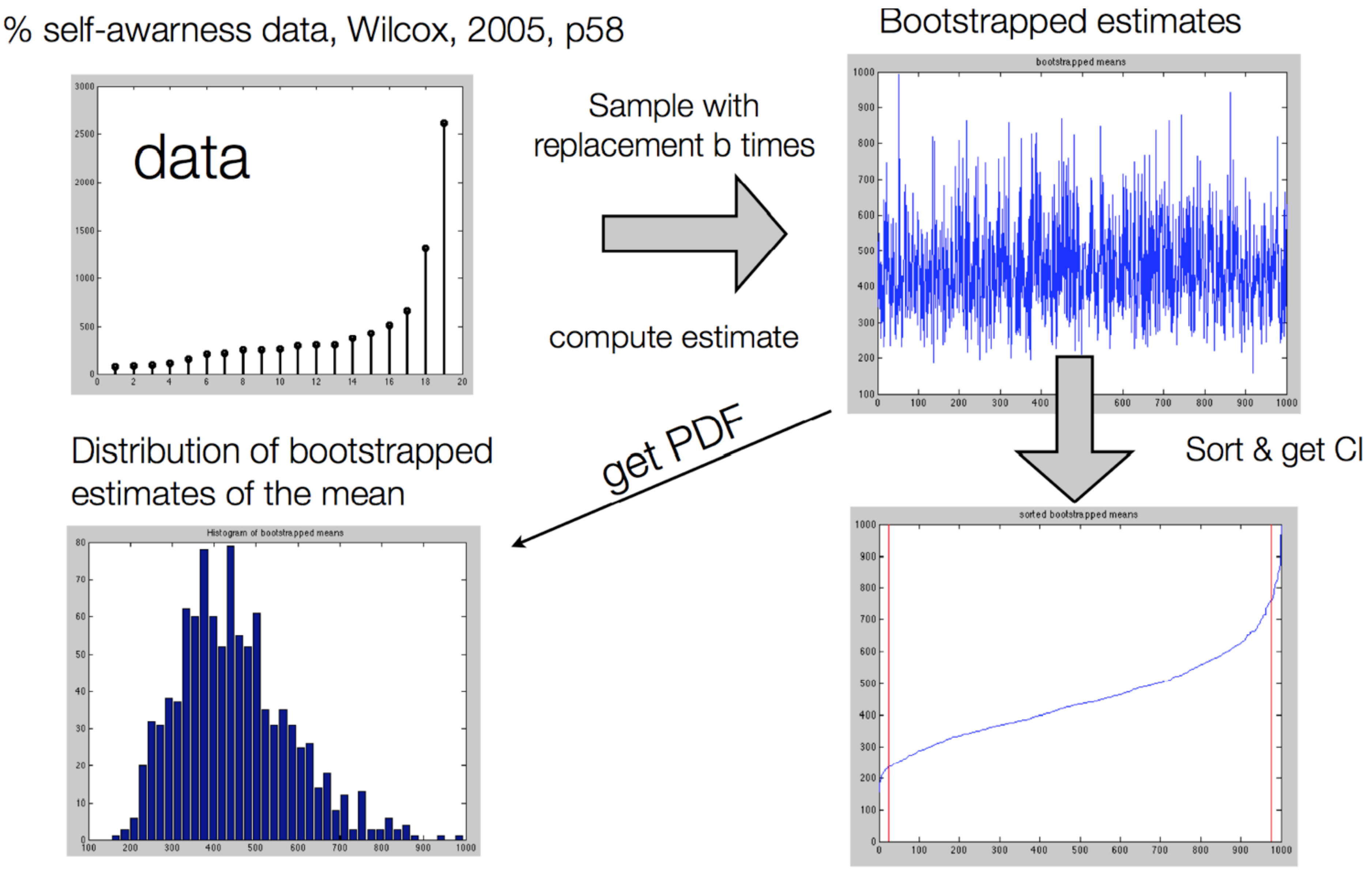
confidence intervals

Percentile bootstrap estimate of confidence intervals

% self-awarness data, Wilcox, 200



Percentile bootstrap estimate of confidence intervals



% self-awarness data, Wilcox, 2005, p58

data

Sample with replacement b times

compute estimate

Bootstrapped estimates

bootstrapped means

Sort & get CI

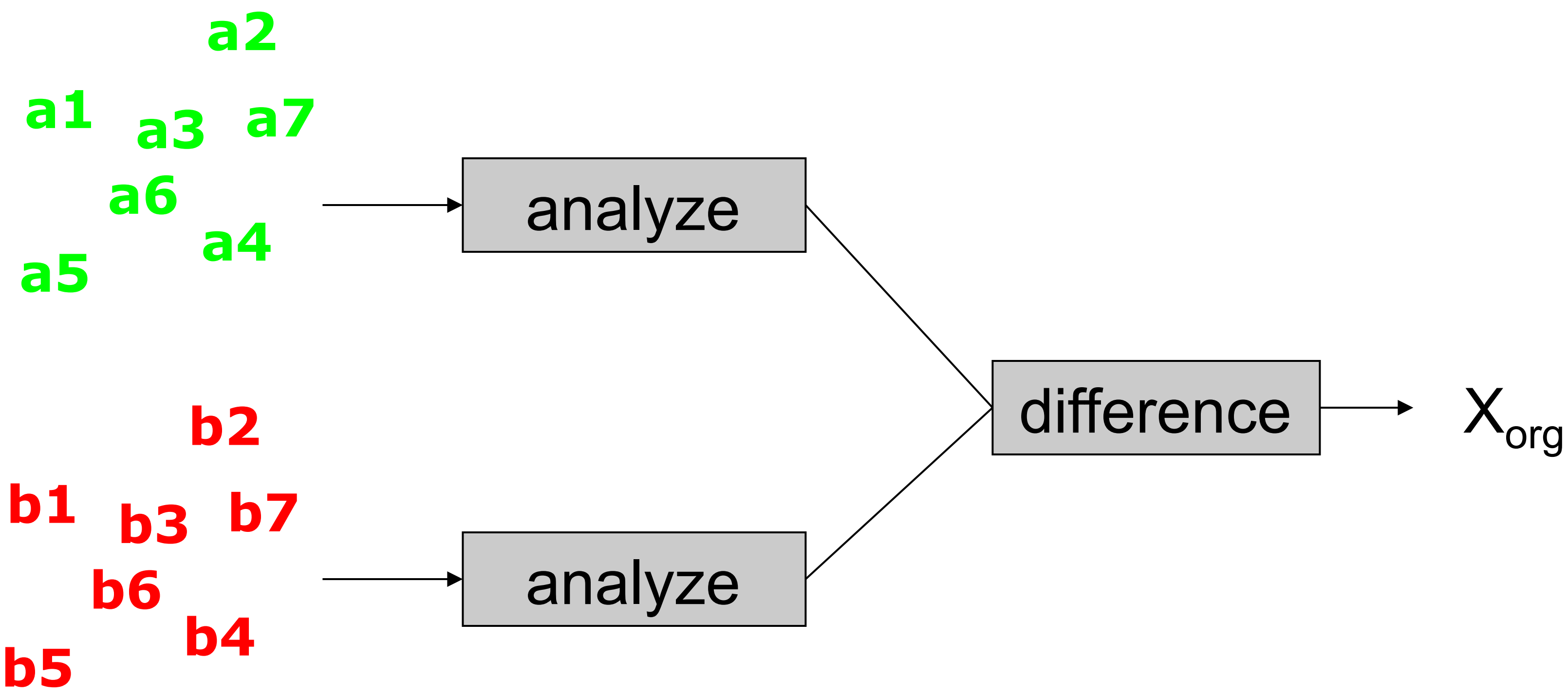
sorted bootstrapped means

get PDF

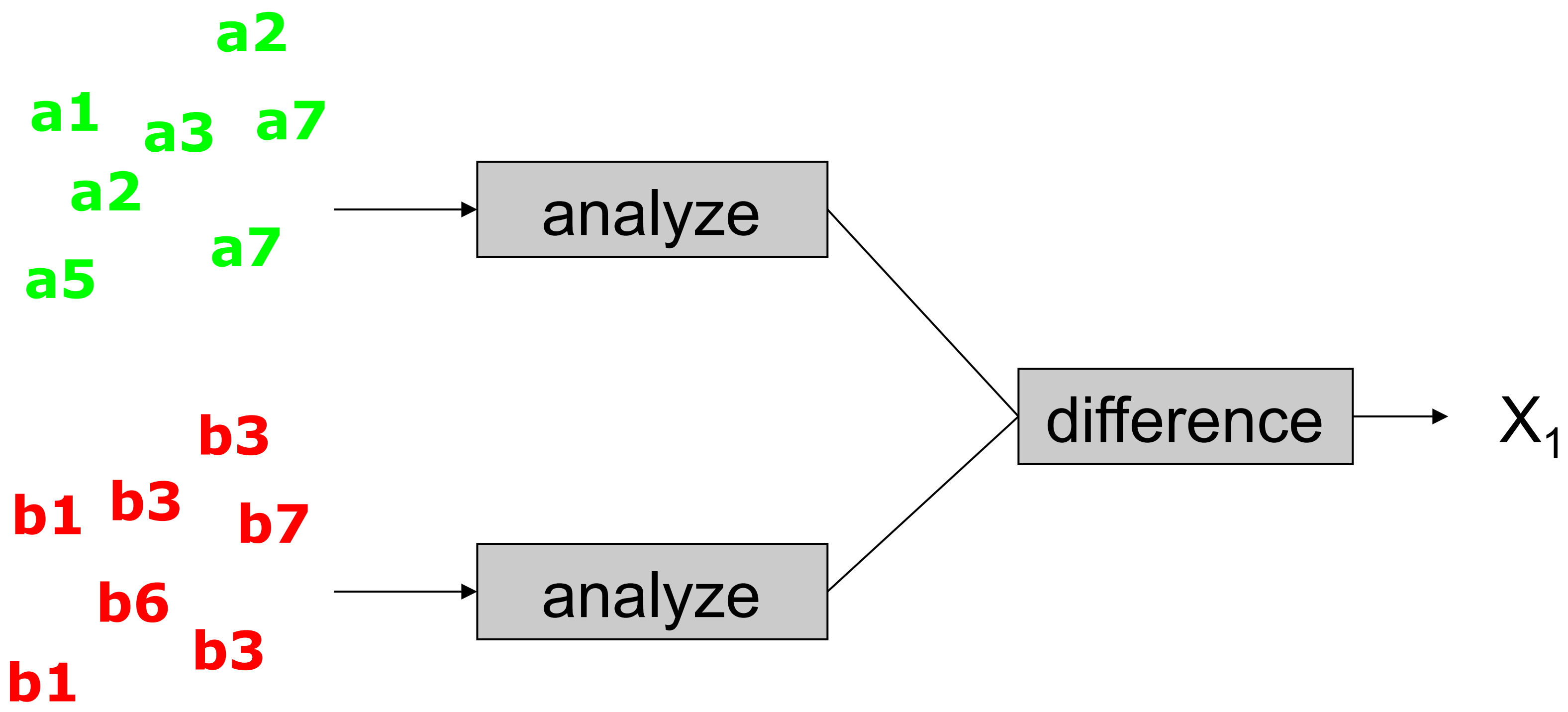
Distribution of bootstrapped estimates of the mean

Histogram of bootstrapped means

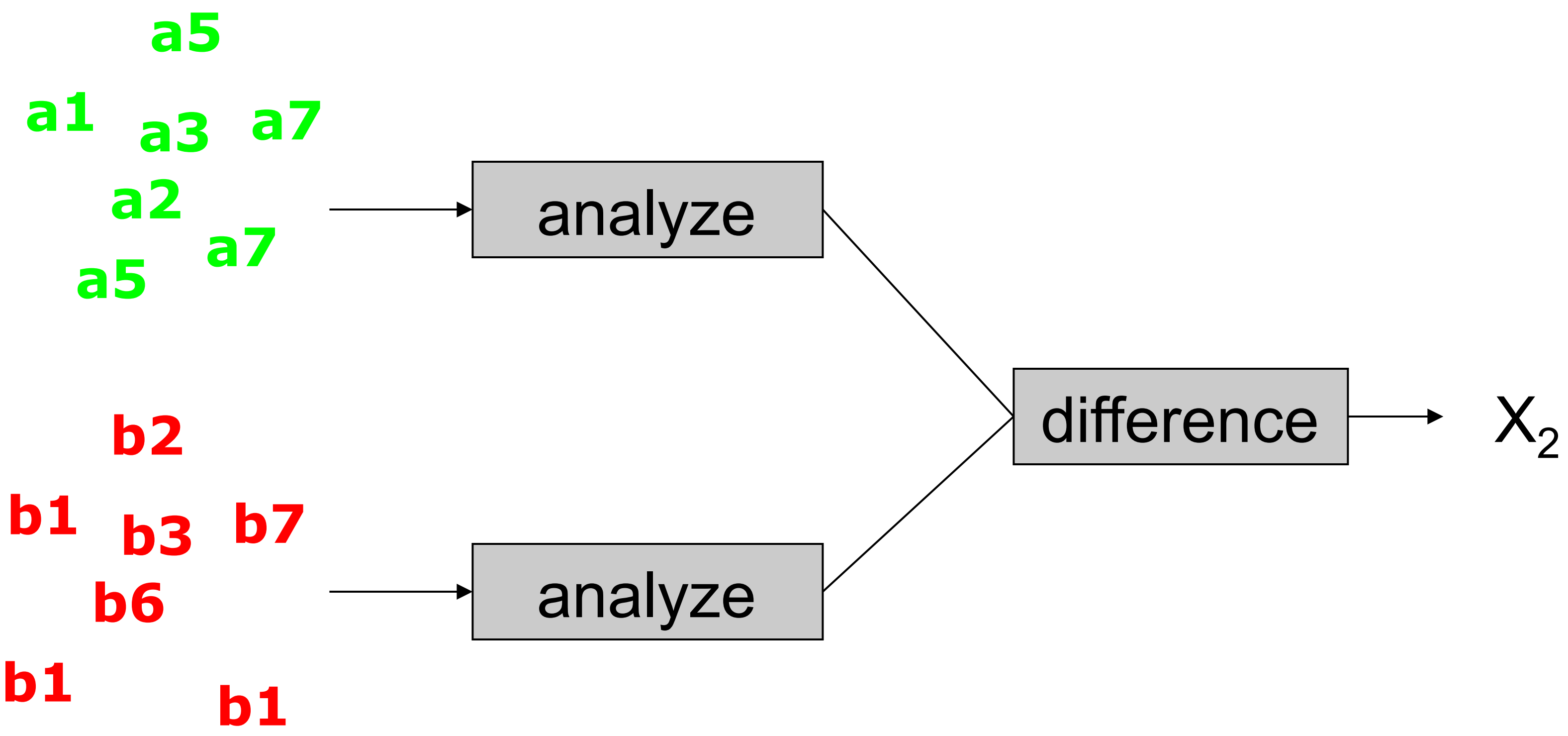
Confidence interval for the difference
Bootstrap approach 1



Confidence interval for the difference Bootstrap approach 1



Confidence interval for the difference
Bootstrap approach 1

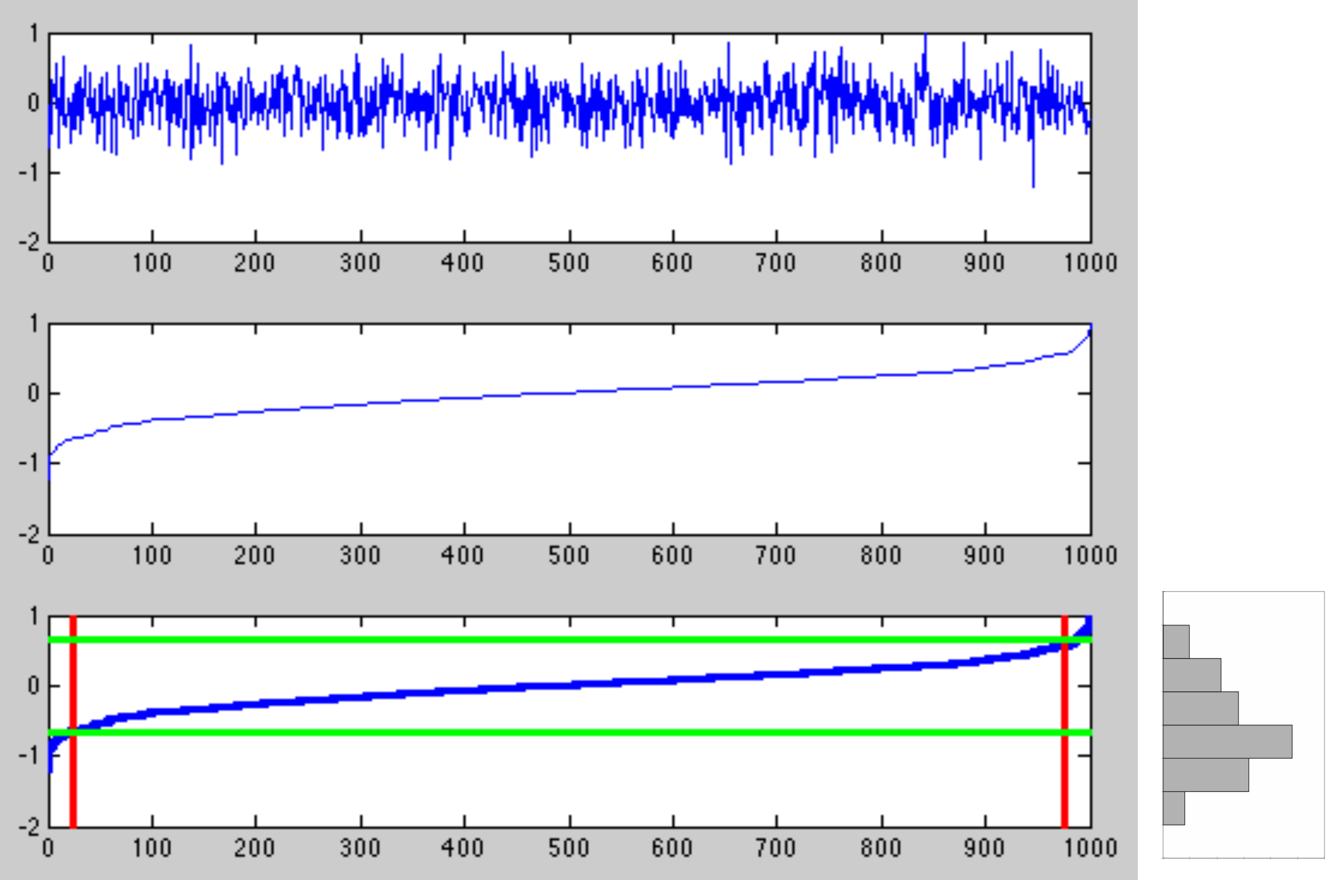


Confidence interval for the difference
Bootstrap approach 1

Permutation /bootstrap

Sorted values

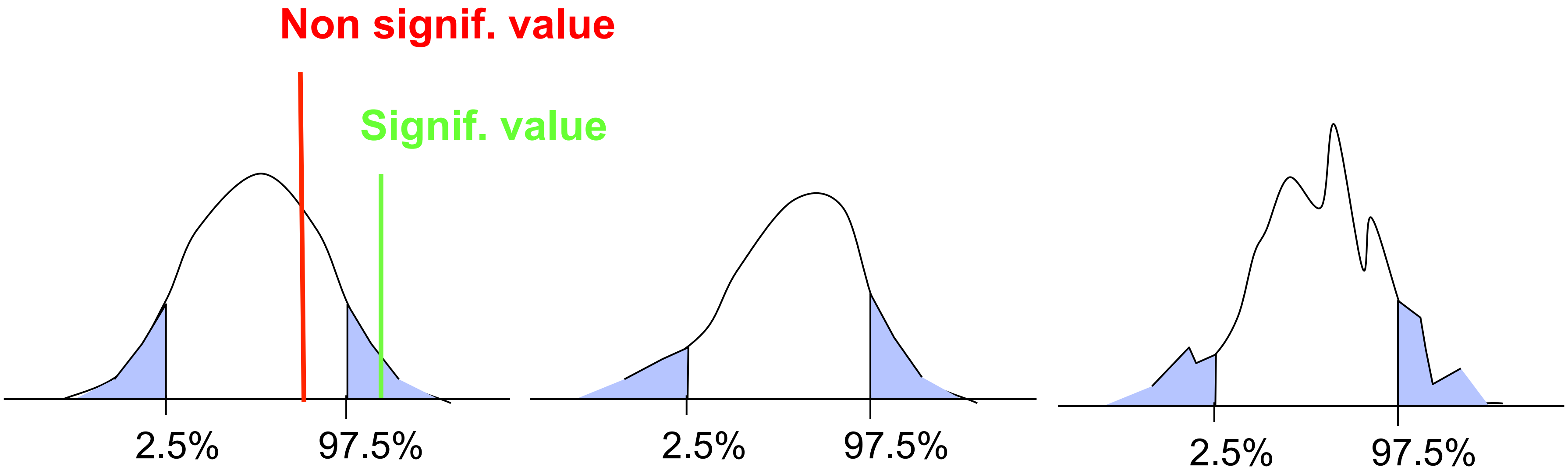
Thresholds



2.5%

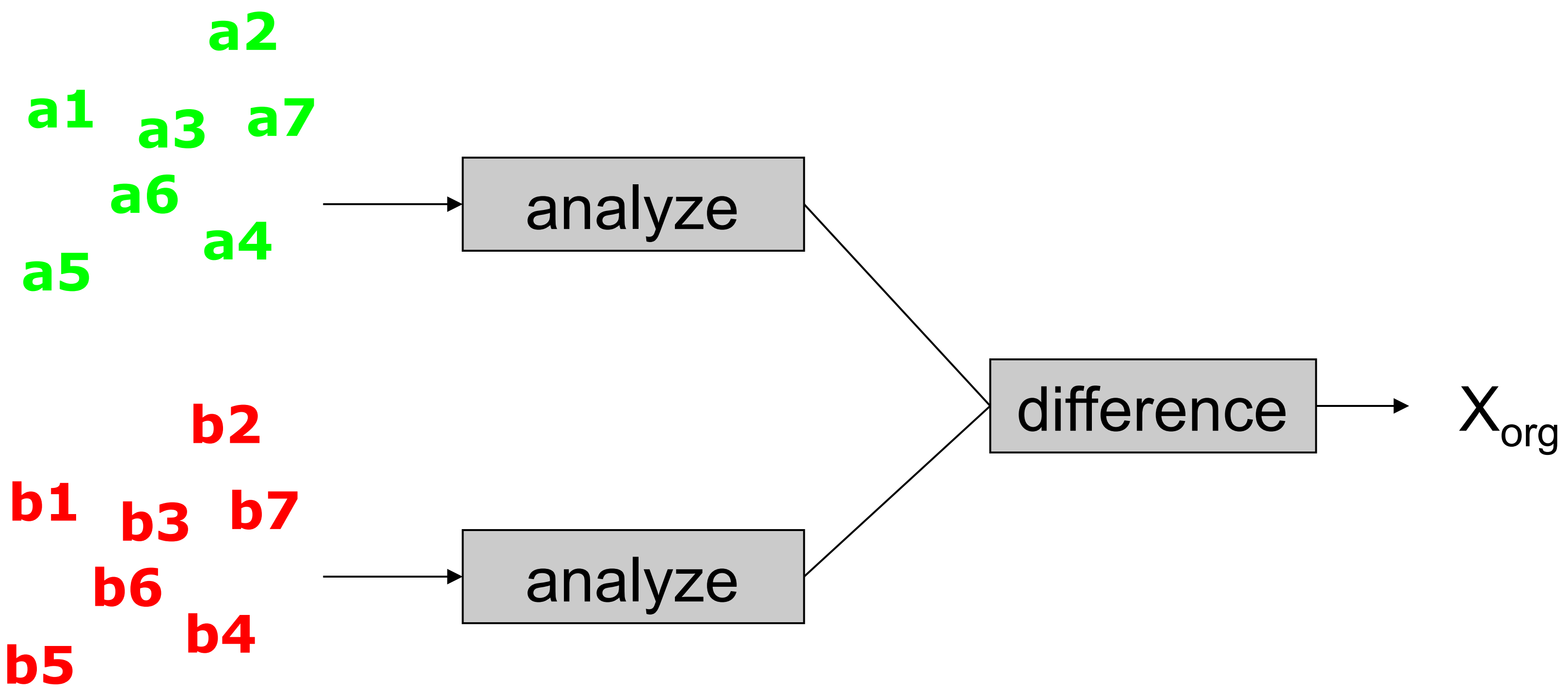
97.5%

Distribution can take any shape

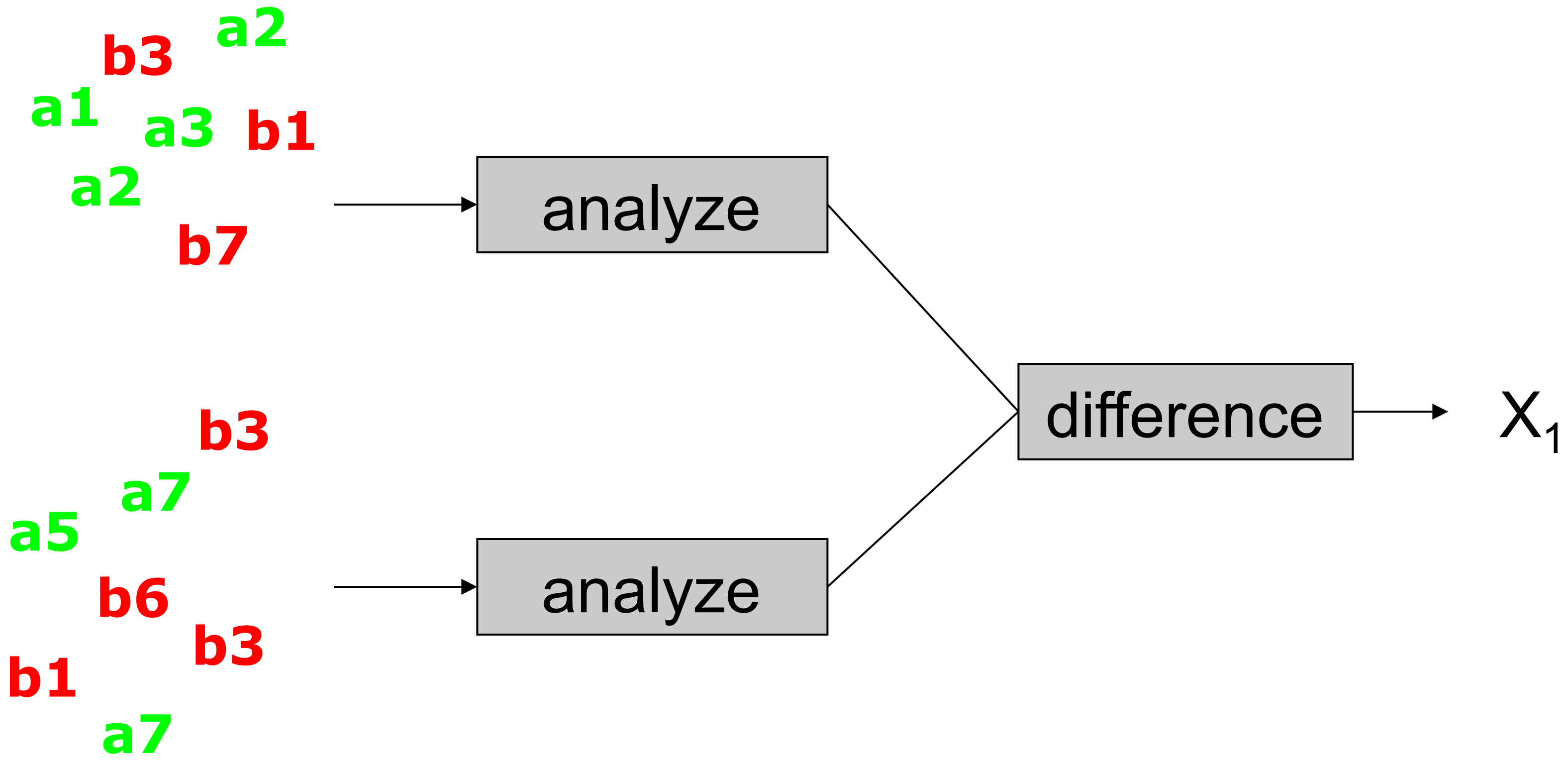


Once you have the 95% confidence interval for the difference: significance only involve assessing if 0 is included in the tails.

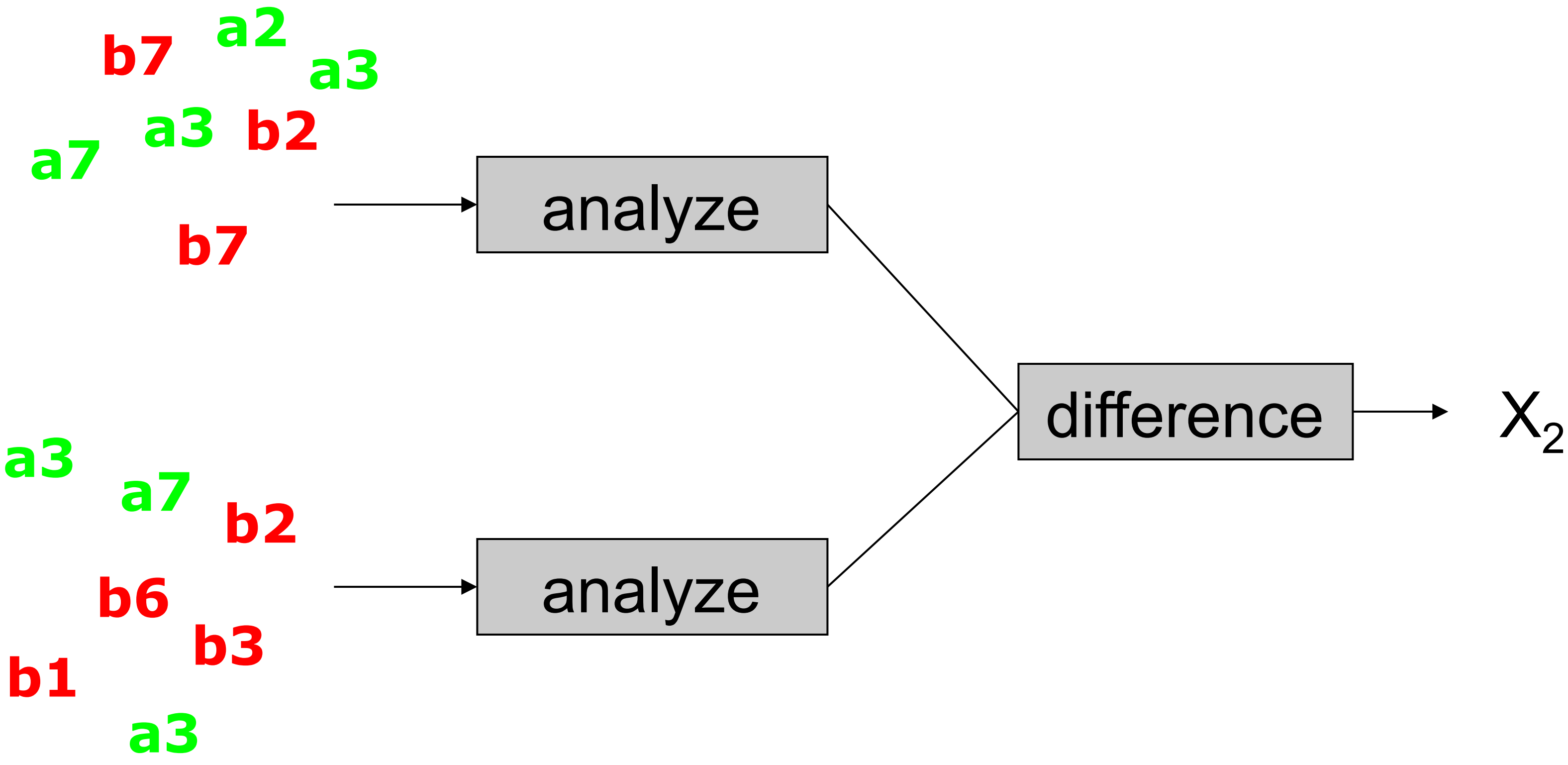
Confidence interval for the difference Bootstrap approach 2



Confidence interval for the difference
Bootstrap approach 2



Confidence interval for the difference
Bootstrap approach 2

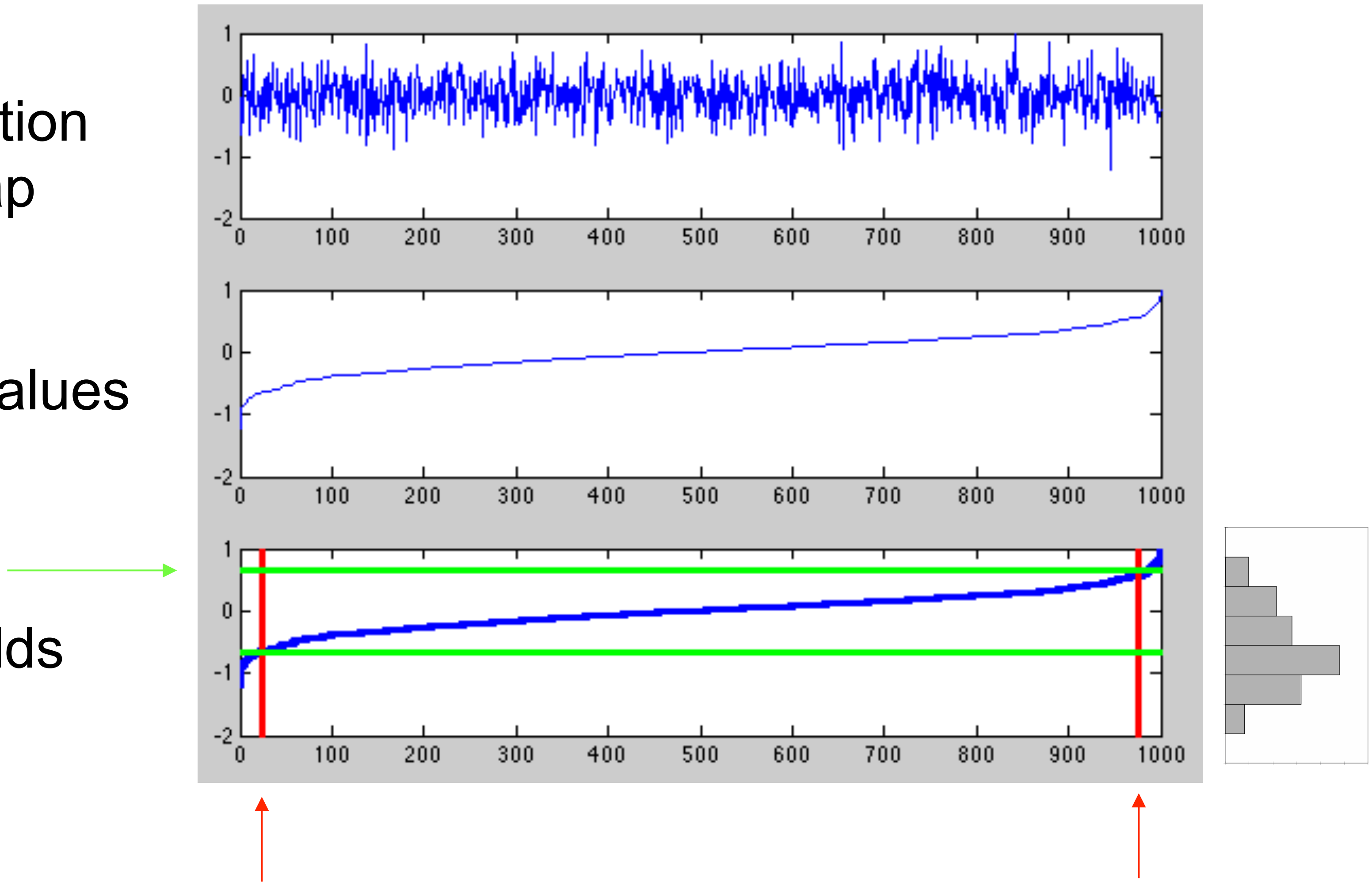


Confidence interval for the difference
Bootstrap approach 2

Permutation
/bootstrap

Sorted values

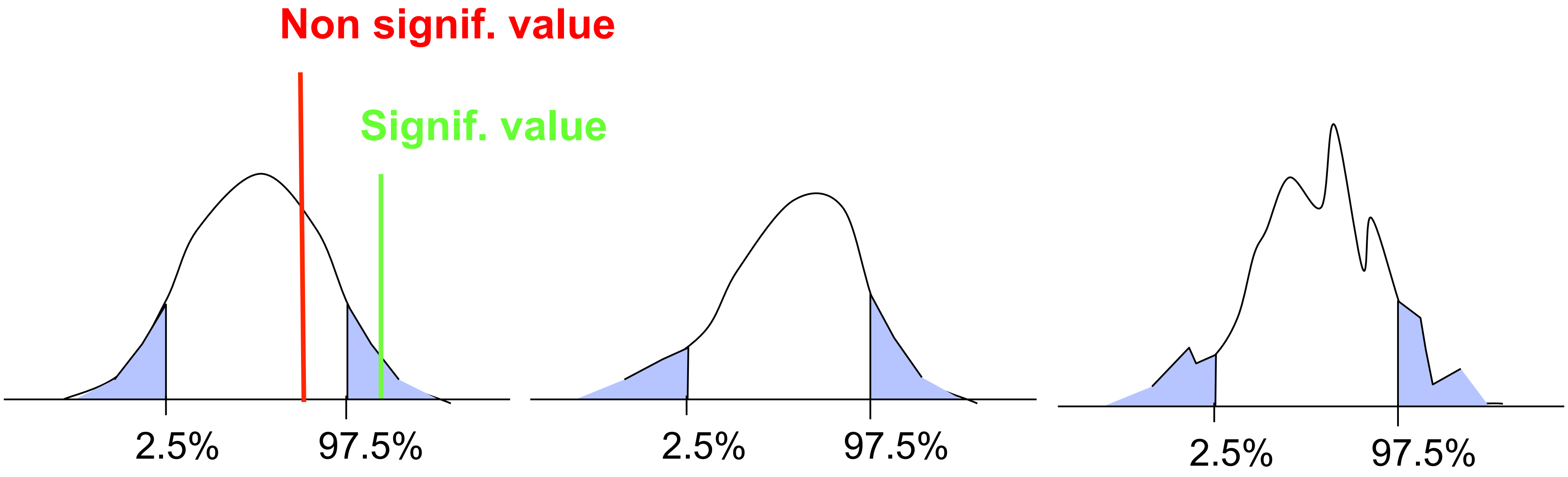
Thresholds



2.5%

97.5%

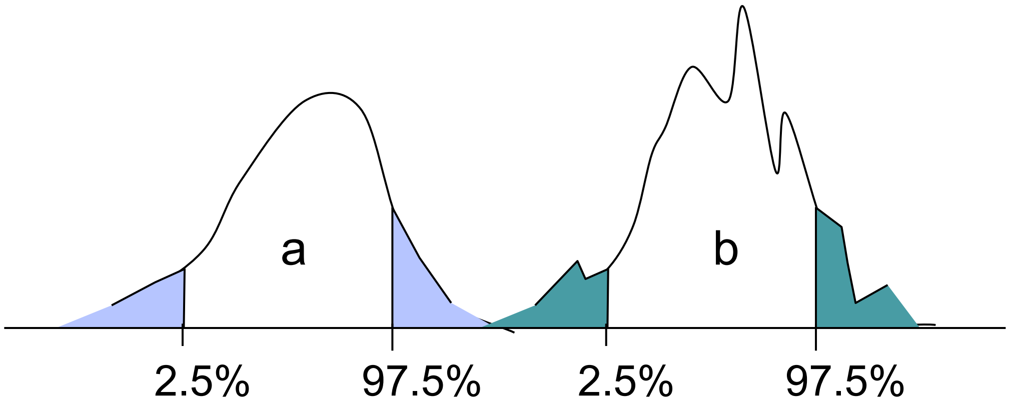
Distribution can take any shape



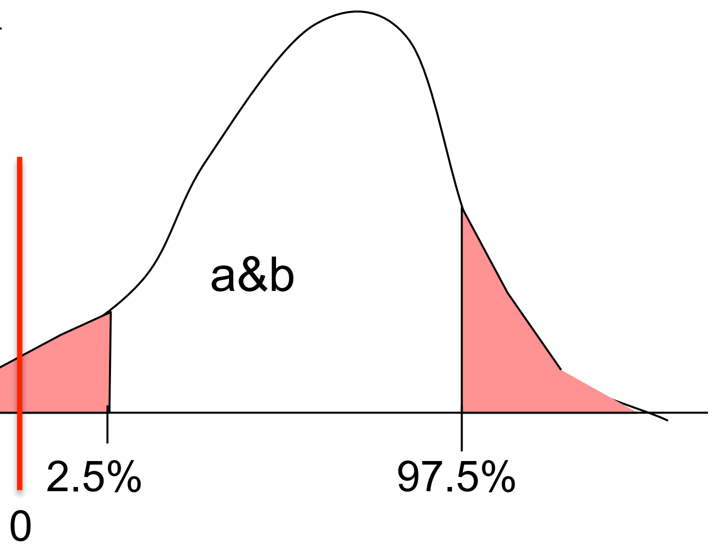
Once you have the 95% confidence interval for the difference: significance only involve assessing if 0 is included in the tails.

Difference between the two bootstrap approches

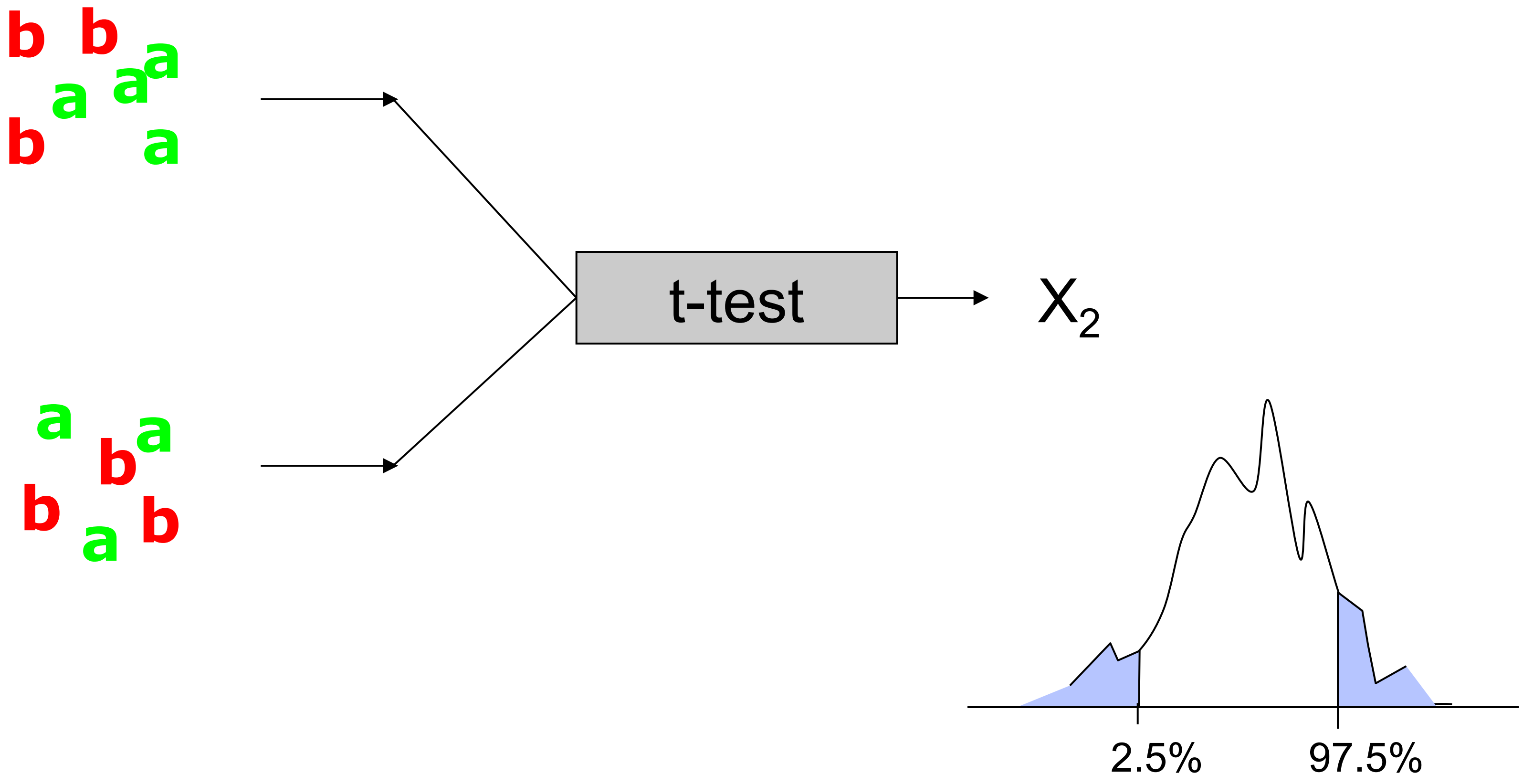
- Bootstrap 1 is testing against H1: the two samples originate from the different distributions.



- Bootstrap 2 is testing against H0: the two samples originate from the same distribution.



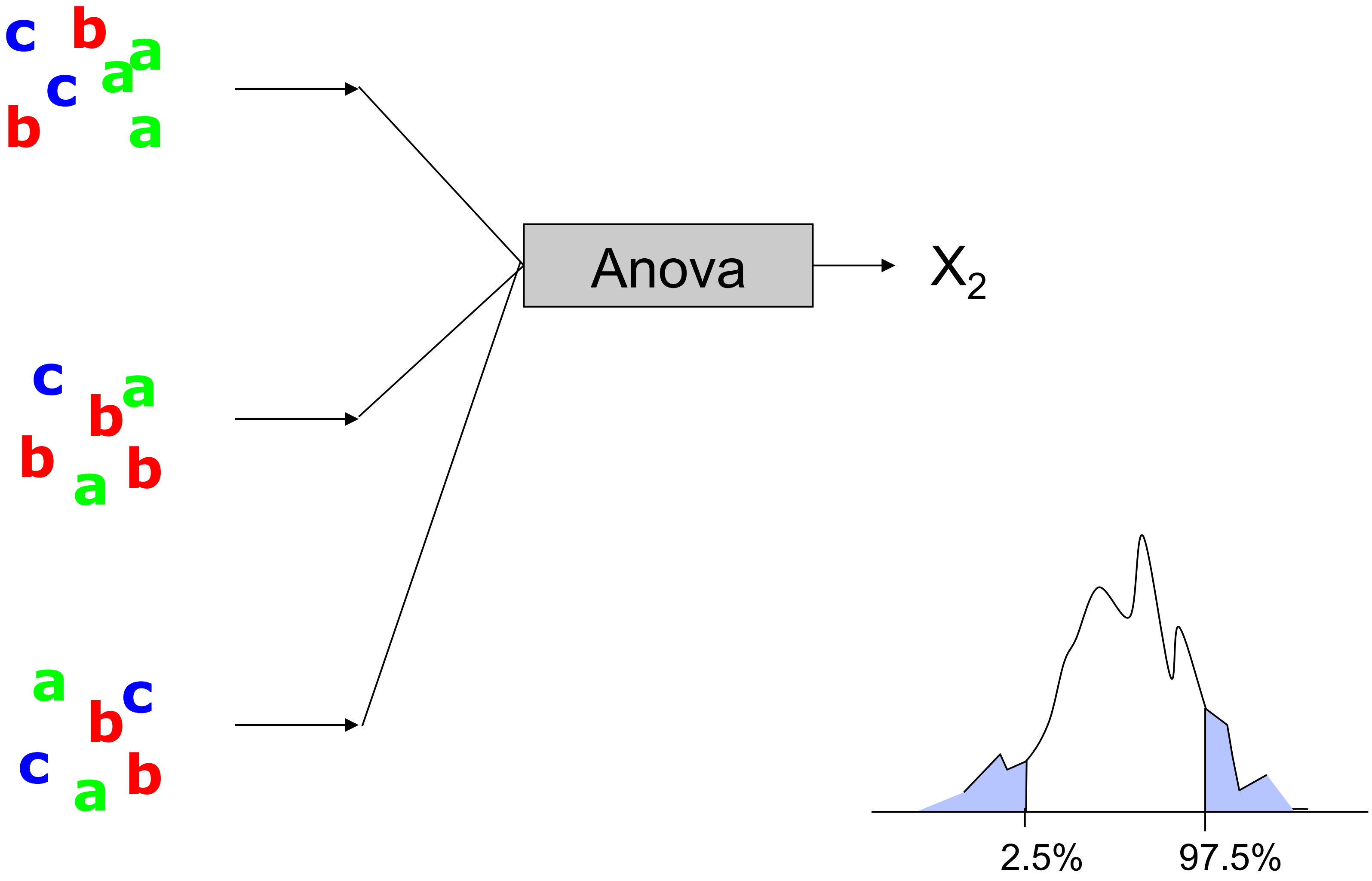
Measure for the bootstrap



t-test

X_2

Measure for the bootstrap

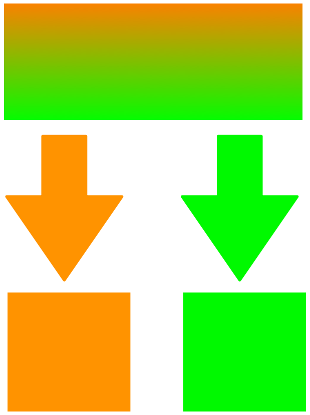


Anova

X_2

Bootstrap versus permutation

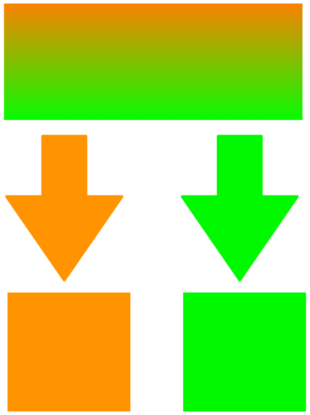
Permutation



each element only
get picked once

Draws are dependent of each others

Bootstrap



each element can
get picked several
times

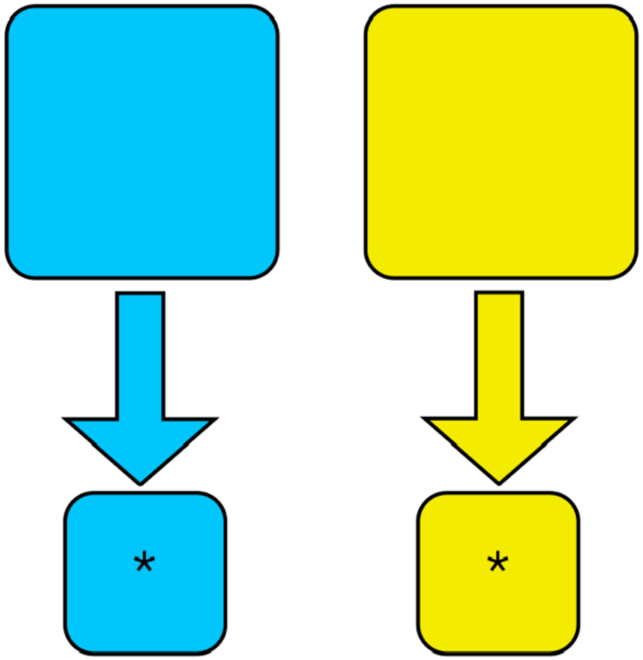
Draws are independent of each others

Use bootstrap!

Resampling strategies:
follow the data acquisition process

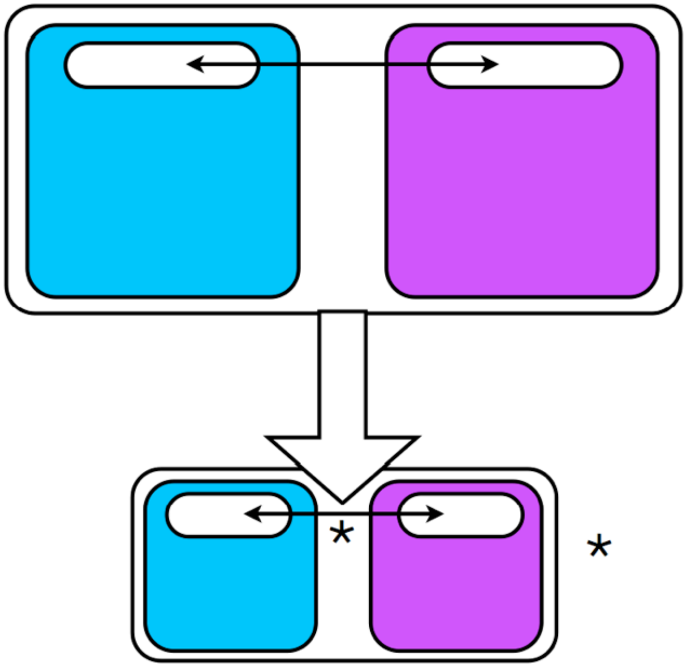
Independent sets:

- 2 conditions in single-subject analyses
- 2 groups of subjects, e.g. patients vs. controls



Dependent sets:

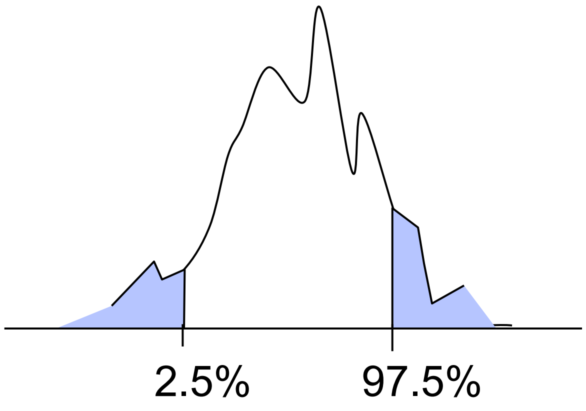
- 2 conditions in group analyses
- Correlations
- Linear regression



Husband	Wifes
22	25
32	25
50	51
25	25
33	38
27	30
45	60
47	54
30	31
44	54
23	23
39	34
24	25
22	23
16	19
73	71
27	26
36	31
24	26
60	62
26	29
23	31
28	29
36	35

Diff= -1.88

Are the two groups different: that’s an unpaired test (comparing the mean or median of husband and the mean or median of wife)

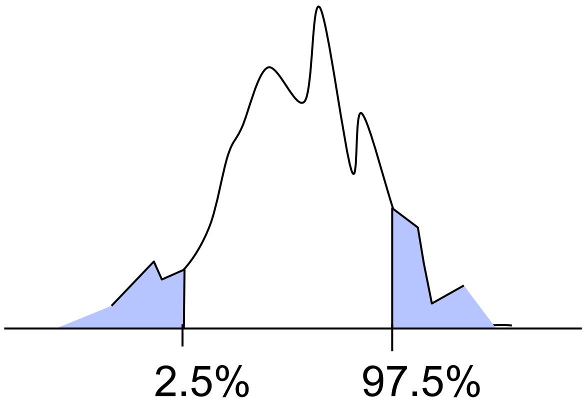


Husband	Wifes
22	25
32	25
50	51
25	25
33	38
27	30
45	60
47	54
30	31
44	54
23	23
39	34
24	25
22	23
16	19
73	71
27	26
36	31
24	26
60	62
26	29
23	31
28	29
36	35

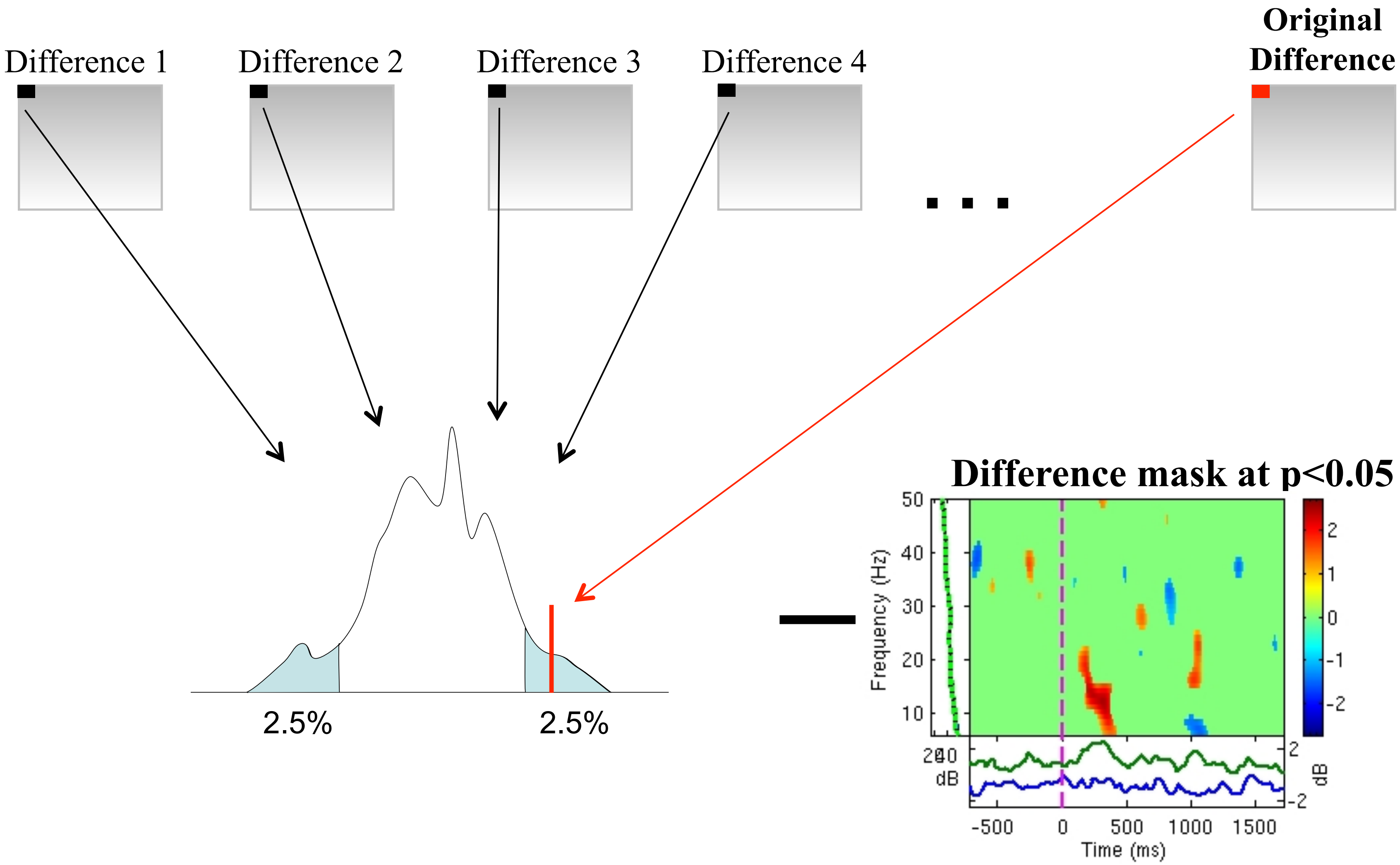
Median

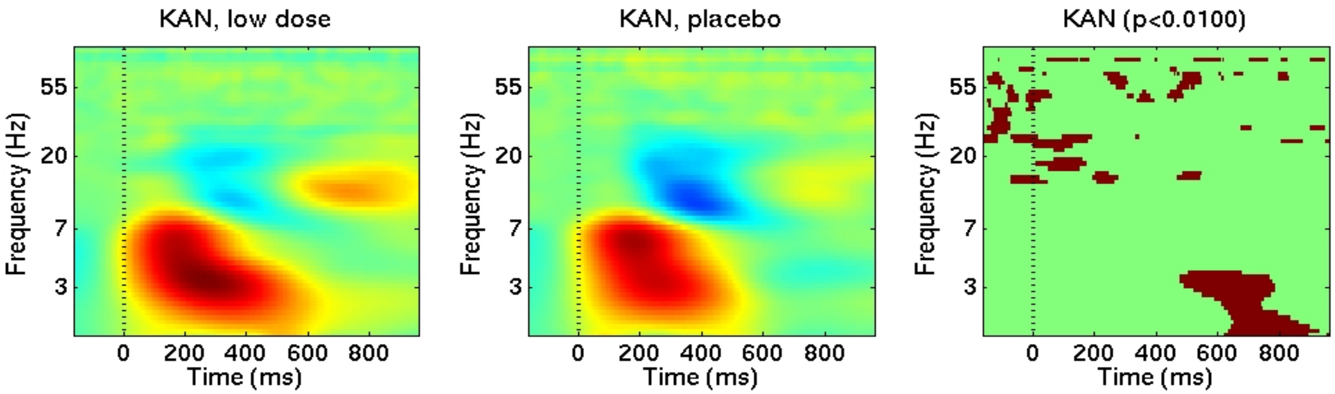
Are the two groups different: that’ s an unpaired test (comparing the mean or median of husband and the mean or median of wife)

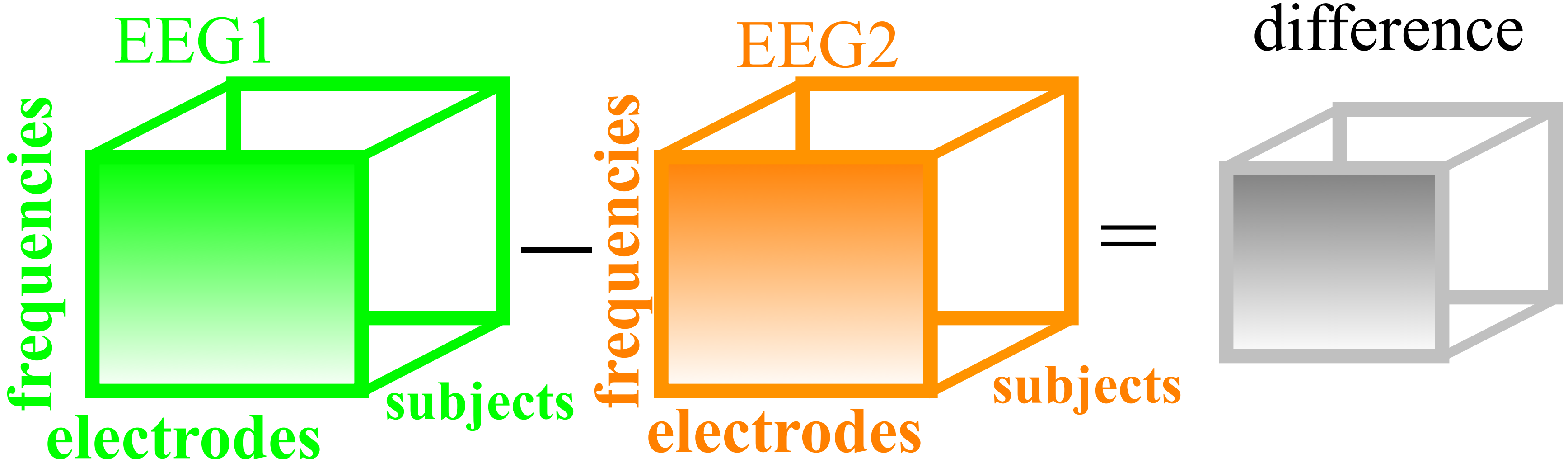
Are husbands older than wifes: that’ s a paired test. Compute difference between the two and change sign to bootstrap.



Assessing significance







EEG1

EEG2

difference

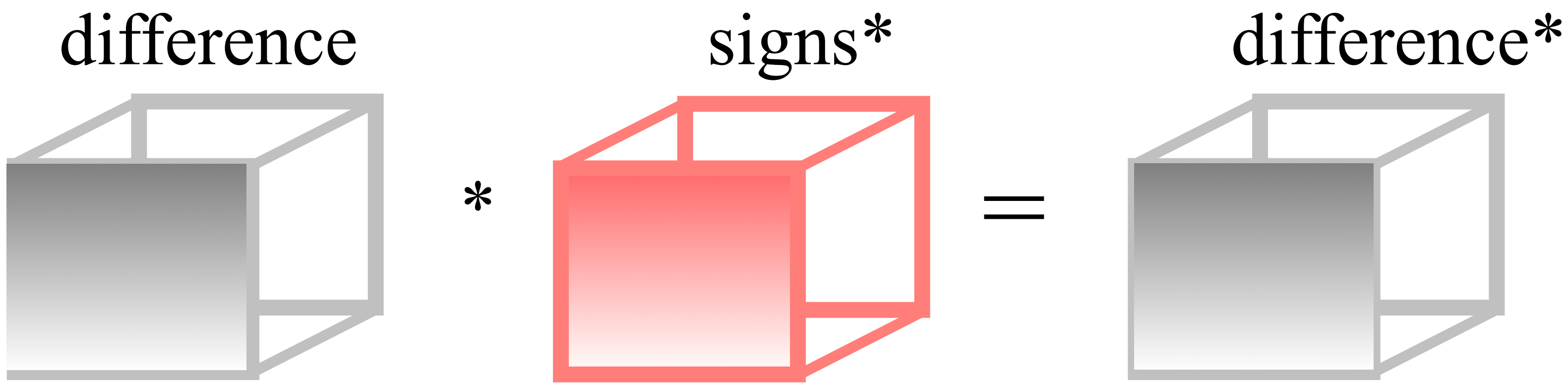
frequencies

electrodes

subjects

—

=



difference

signs*

difference*

*

=

Correcting for multiple comparisons

- Bonferoni correction: divide by the number of comparisons (Bonferroni CE. Sulle medie multiple di potenze. Bollettino dell'Unione Matematica Italiana, 5 third series, 1950; 267-70.)

- Holms correction: sort all p values. Test the first one against α/N, the second one against α/(N-1)

- Max method

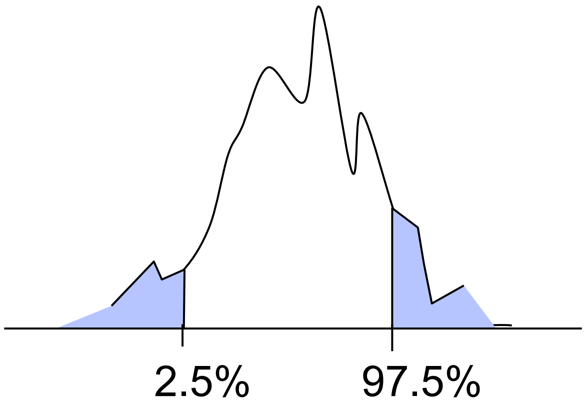
- False detection rate

- Clusters

Max procedure

- for each permutation or bootstrap loop, simply take the MAX of the absolute value of your estimator (e.g. mean difference) across electrodes and/or time frames and/or temporal frequencies.

- compare absolute original difference to this distribution



FDR procedure

Procedure:

- Sort all p values (column C1)

C3

- Create column C2 by computing $j*\alpha/N$

- Subtract column C1 from C2 to build column C3

- Find the highest negative index in C3 and

find the corresponding p-value in C1 (*p_fdr)*

- Reject all null hypothesis whose p-value are less than or equal to *p_fdr*

	C1
Index "j"	Actual
1	0.001
2	0.002
3	0.01
4	0.03
5	0.04
6	0.045
7	0.05
8	0.1
9	0.2
10	0.6

FDR procedure

Procedure:

- Sort all p values (column C1)

C3

- Create column C2 by computing $j*\alpha/N$

- Subtract column C1 from C2 to build column C3

- Find the highest negative index in C3 and

find the corresponding p-value in C1 *(p_fdr)*

- Reject all null hypothesis whose p-value are less than or equal to *p_fdr*

	C1	C2
Index "j"	Actual	j*0.05/10
1	0.001	0.005
2	0.002	0.01
3	0.01	0.015
4	0.03	0.02
5	0.04	0.025
6	0.045	0.03
7	0.05	0.035
8	0.1	0.04
9	0.2	0.045
10	0.6	0.05

FDR procedure

Procedure:

- Sort all p values (column C1)
C3

- Create column C2 by computing $j*\alpha/N$

- Subtract column C1 from C2 to build column C3

- Find the highest negative index in C3
  and
find the corresponding p-value in C1 *(p_fdr)*

- Reject all null hypothesis whose p-value are less than or equal to *p_fdr*

	C1	C2	C3
Index "j"	Actual	j*0.05/10	C2-C1
1	0.001	0.005	-0.004
2	0.002	0.01	-0.008
3	0.01	0.015	-0.005
4	0.03	0.02	0.01
5	0.04	0.025	0.015
6	0.045	0.03	0.015
7	0.05	0.035	0.015
8	0.1	0.04	0.06
9	0.2	0.045	0.155
10	0.6	0.05	0.55

FDR procedure

Procedure:

- Sort all p values (column C1)
C3

- Create column C2 by computing *j*α/N*

- Subtract column C1 from C2 to build column C3

- Find the highest negative index in C3
and
find the corresponding p-value in C1 *(p_fdr)*

- Reject all null hypothesis whose p-value are less than or equal to *p_fdr*

	C1	C2	C3
Index "j"	Actual	j*0.05/10	C2-C1
1	0.001	0.005	-0.004
2	0.002	0.01	-0.008
3	0.01	0.015	-0.005
4	0.03	0.02	0.01
5	0.04	0.025	0.015
6	0.045	0.03	0.015
7	0.05	0.035	0.015
8	0.1	0.04	0.06
9	0.2	0.045	0.155
10	0.6	0.05	0.55

Bonferoni

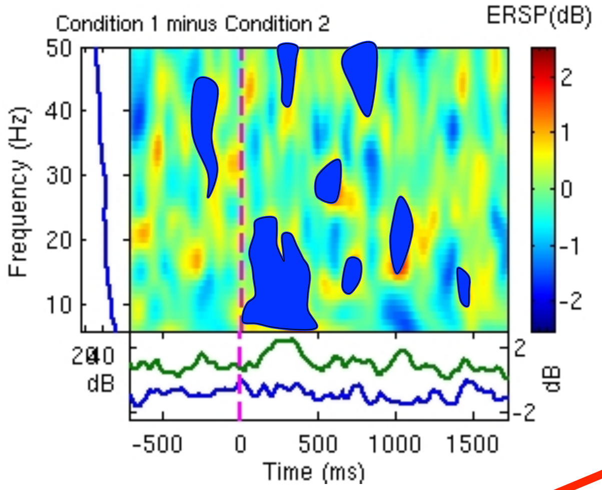
Holms

FDR

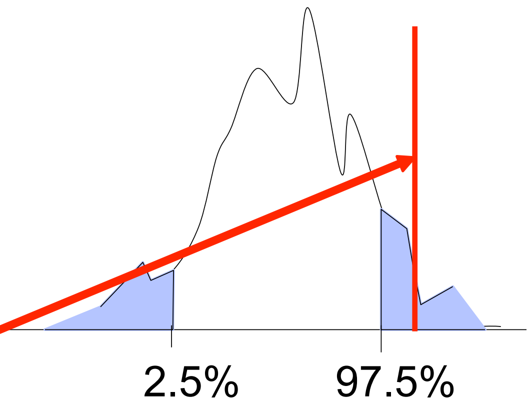
Uncorrected

Cluster correction for multiple comparisons

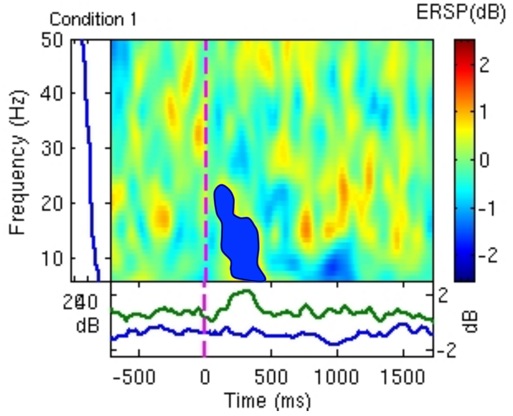
Original difference



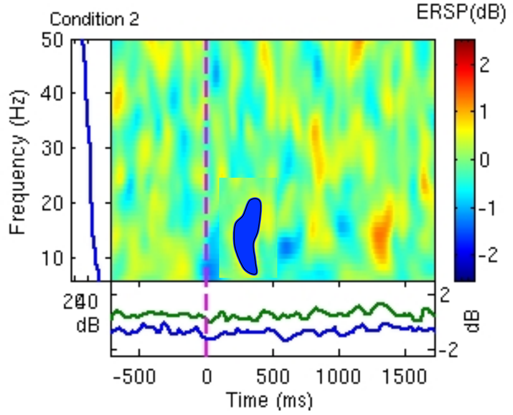
44 pixels



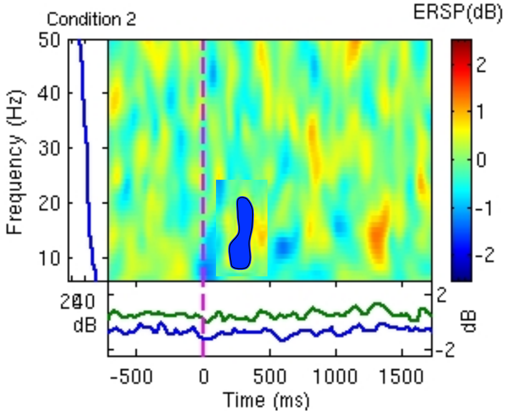
Difference bootstrap 1



Difference bootstrap 2

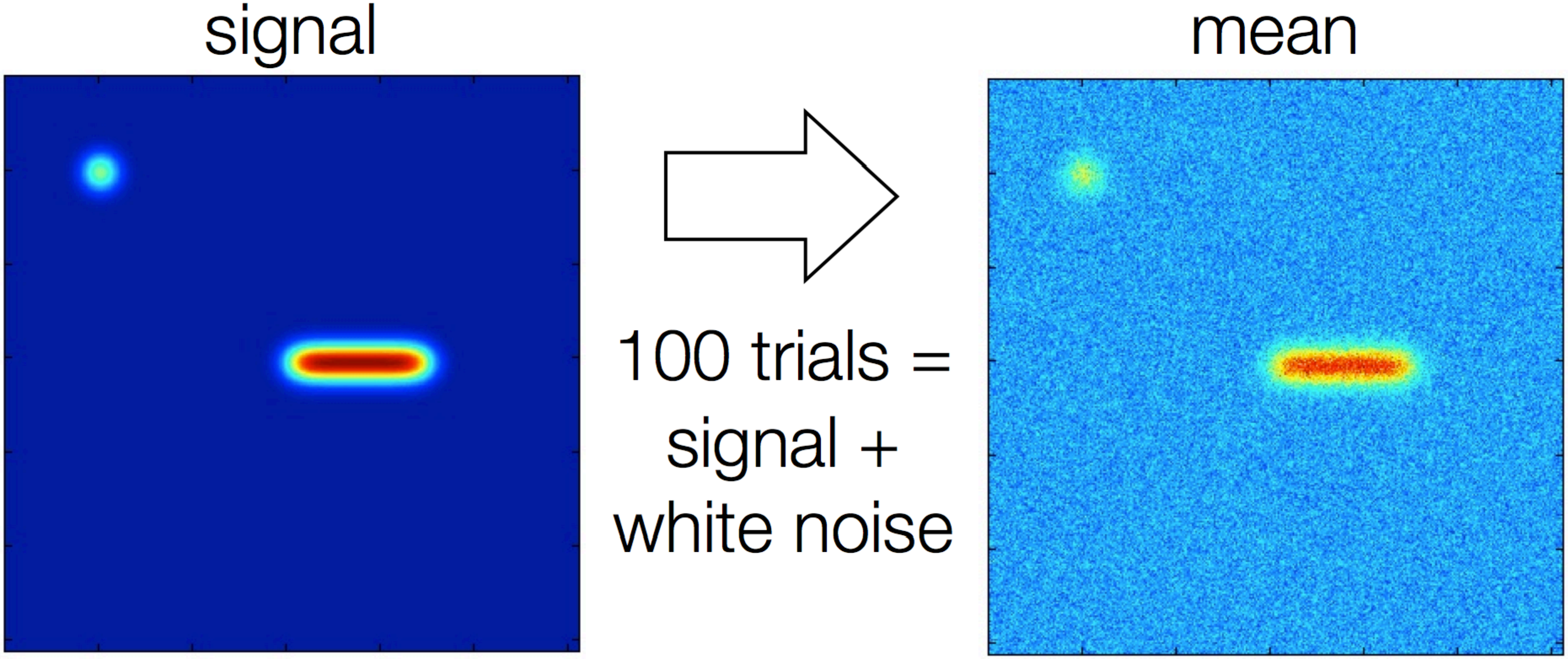


Difference bootstrap 3



....

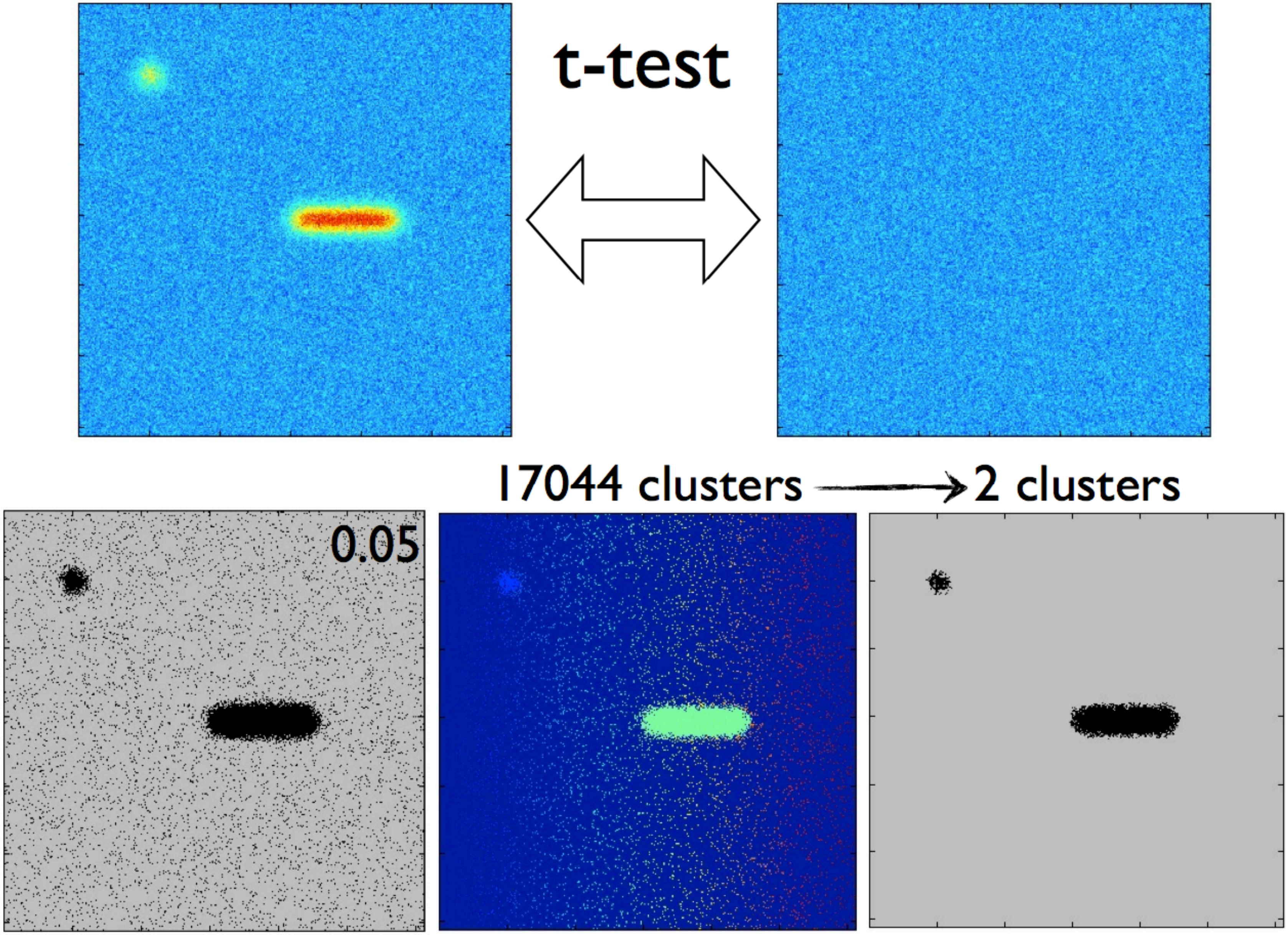
Control for multiple comparisons cluster method



signal

mean

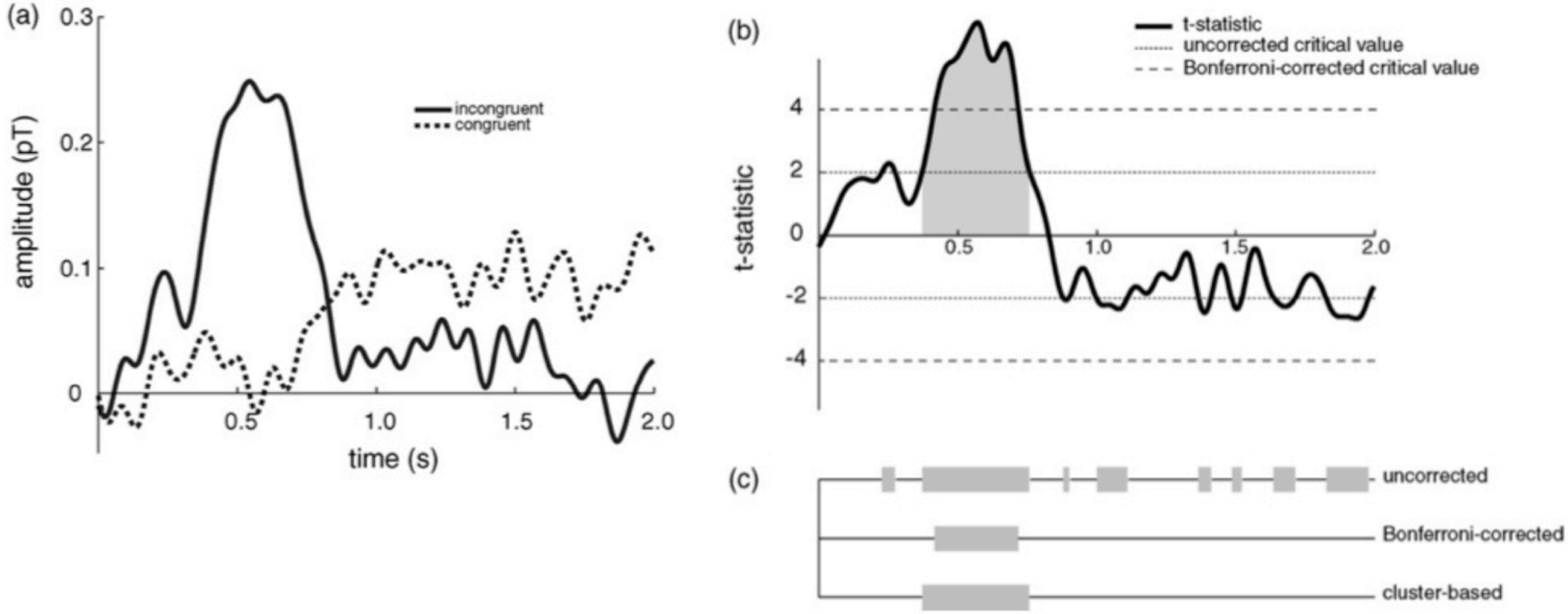
100 trials = signal + white noise



t-test

17044 clusters → 2 clusters

Control for multiple comparisons cluster method



Maris & Oostenveld, J. Neurosci. Methods 2007

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

statcond function in EEGLAB

```
a = { rand(1,10) rand(1,10)+0.5 }; % pseudo 'paired' data vectors

[t df pvals] = statcond(a , 'mode', 'perm'); % perform paired t-test
pvals = 5.2807e-04 % standard t-test probability value

% Note: for different rand() outputs, results will differ.
[t df pvals surog] = statcond(a, 'mode', 'perm', 'naccu', 2000);
pvals = 0.0065 % nonparametric t-test using 2000 permuted data sets

a = { rand(2,11) rand(2,10) rand(2,12)+0.5 };
[F df pvals] = statcond(a , 'mode', 'perm'); % perform an unpaired ANOVA

pvals =
    0.00025 % p-values for difference between columns
    0.00002 % for each data row
```

statcond function in EEGLAB

```
a = { rand(3,4,10) rand(3,4,10) rand(3,4,10); ...
      rand(3,4,10) rand(3,4,10) rand(3,4,10)+0.5 };

% pseudo (2,3)-condition data array, each entry containing
% ten (3,4) data matrices
[F df pvals] = statcond(a , 'mode', 'perm');
                                   % paired 2-way ANOVA

% Output:
pvals{1} % a (3,4) matrix of p-values; effects across columns
pvals{2} % a (3,4) matrix of p-values; effects across rows
pvals{3} % a (3,4) matrix of p-values; interaction effects across
   rows and columns
```

Exercice

- Experiment with the statcond function
  - Create 2 random vectors of values
  - Add “signal” to one of the variable
  - Use statcond EEGLAB function and compare permutation and parametric results
  - Repeat 100 times and plot the histogram of p-values