Contemporary Statistical Methods Useful for EEG Analysis

David Groppe
Marta Kutas's lab
University of California, San Diego

Arnaud Delorme
Swartz Center for
Computational Neuroscience
University of California, San Diego

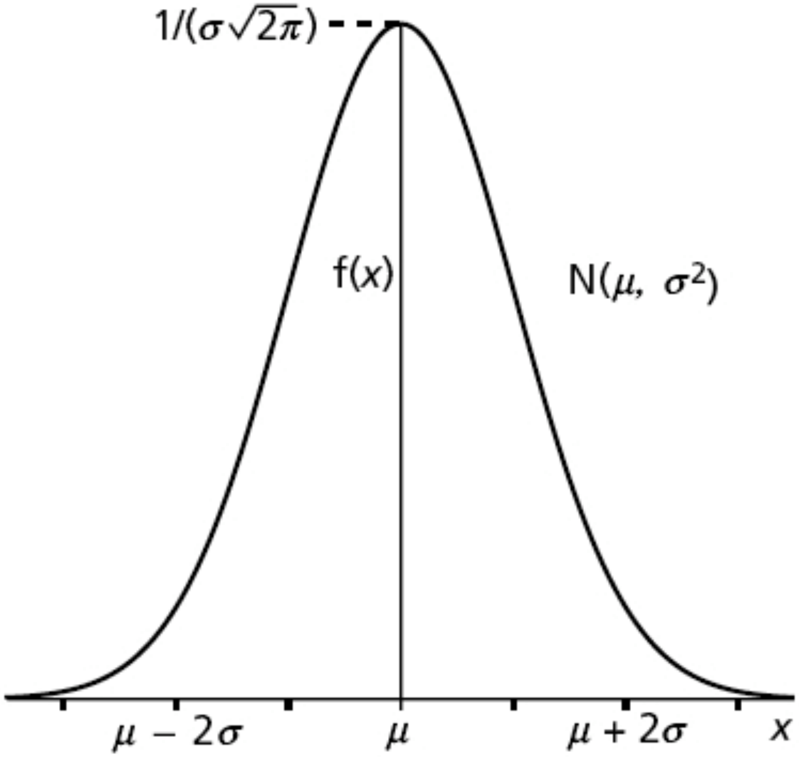
12th EEGLAB Workshop
Nov. 19, 2010

Presentation Outline

- **“Classic” Analytical Inferential Statistics**
  - Parametric & non-parametric
- **Resampling-Based Inferential Statistics**
  - Randomization/permutation tests
  - Bootstrap statistics
- **Correcting for Multiple Comparisons**
  - Permutation test based control of family-wise error
  - Benjamini methods for control of false discovery rate
  - Evaluating multiple comparison correction on simulated ERP data

Analytic Parametric Statistics:

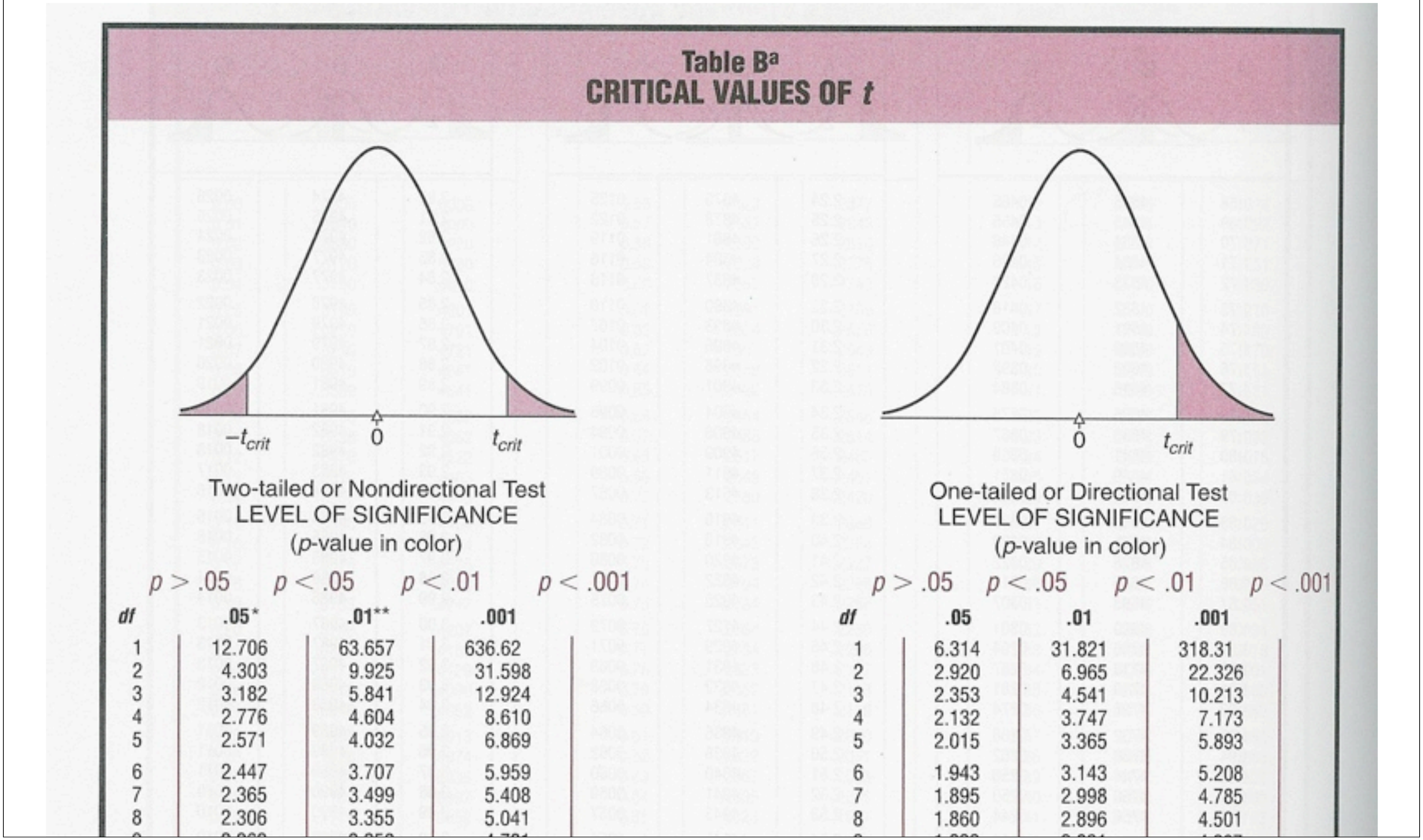
Assume Data Come from a Particular Distribution



Gaussian Distribution

Analytic Parametric Statistics:

Critical Values Analytically Derived



$-t_{crit}$
0
t_{crit}
0
t_{crit}

Table B[a]
CRITICAL VALUES OF *t*

Two-tailed or Nondirectional Test
LEVEL OF SIGNIFICANCE
(*p*-value in color)

	$p > .05$	$p < .05$	$p < .01$	$p < .001$
df		**.05***	**.01****	**.001**
1		12.706	63.657	636.62
2		4.303	9.925	31.598
3		3.182	5.841	12.924
4		2.776	4.604	8.610
5		2.571	4.032	6.869
6		2.447	3.707	5.959
7		2.365	3.499	5.408
8		2.306	3.355	5.041

One-tailed or Directional Test
LEVEL OF SIGNIFICANCE
(*p*-value in color)

	$p > .05$	$p < .05$	$p < .01$	$p < .001$
df		**.05**	**.01**	**.001**
1		6.314	31.821	318.31
2		2.920	6.965	22.326
3		2.353	4.541	10.213
4		2.132	3.747	7.173
5		2.015	3.365	5.893
6		1.943	3.143	5.208
7		1.895	2.998	4.785
8		1.860	2.896	4.501

Analytic Parametric Statistics:

Popular Parametric Tests

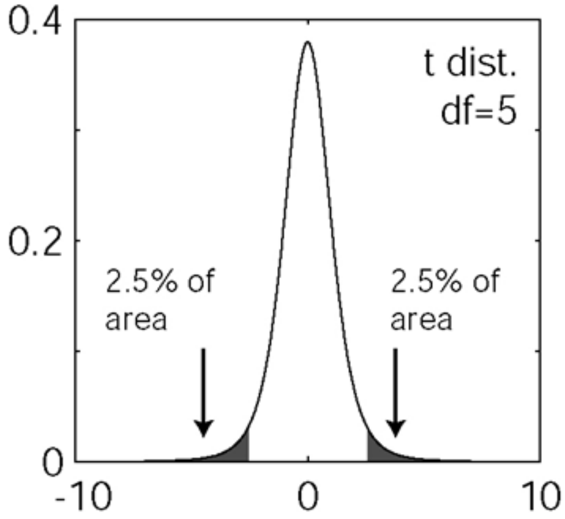
T-test: Compare paired/ unpaired Samples for continuous data. In EEGLAB, used for grand-average ERPs.

Paired

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

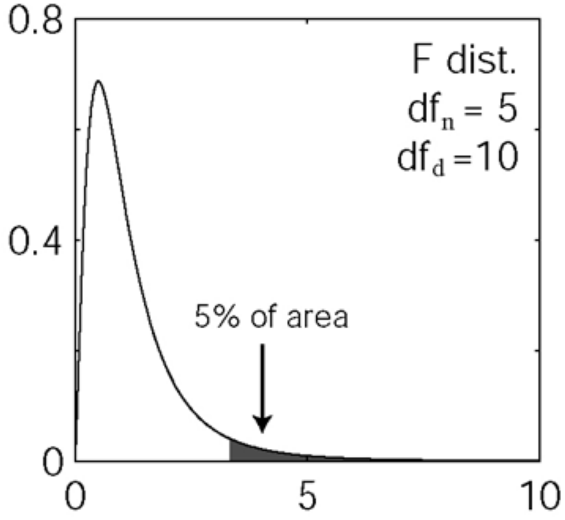
Unpaired

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



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

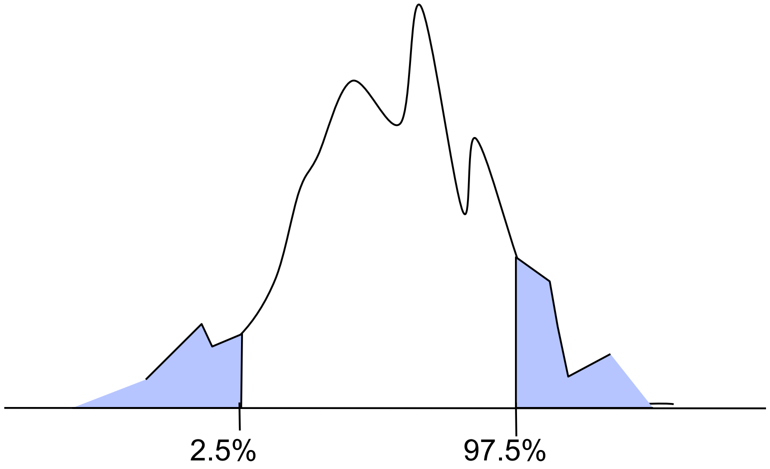
$$F = \frac{Variance_{interGroup} \big/ N_{Group} - 1}{Variance_{WithinGroup} \big/ N - N_{Group}}$$



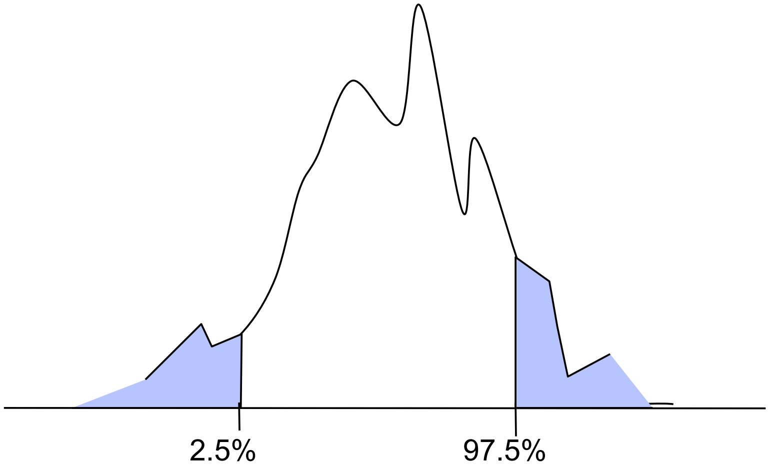
Analytic Non-Parametric Statistics:

Minimal Distribution Assumptions

Population A



Population B



Mann-Whitney *U* Test: Null hypothesis is that the distribution of Population A and B are the same

Analytic Non-Parametric Statistics:

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

Problems with Analytic Statistics:

1. No analytic solution for some situations (e.g., comparing the mean of two groups that differ in variance)
2. Often, data don’t fit parametric assumptions
3. Non-parametric tests may lack power and rank transformation can make it tricky to do things like derive confidence intervals

Presentation Outline

- **"Classic" Analytical Inferential Statistics**
    - Parametric & non-parametric
- **Resampling-Based Inferential Statistics**
    - Randomization/permutation tests
    - Bootstrap statistics
- **Correcting for Multiple Comparisons**
    - Permutation test based control of family-wise error
    - Benjamini methods for control of false discovery rate
    - Evaluating multiple comparison correction on simulated ERP data

Resampling-Based Statistics:

Inferential statistics based on “simulating” an experiment a large number of times with the observed data

Observed Data

Group A	Group B
8	5
4	3
6	4

Resampling-Based Statistics:

Inferential statistics based on “simulating” an experiment a large number of times with the observed data

Observed Data

Group A	Group B
8	5
4	3
6	4

Resample

“Simulated Replication”

Group A	Group B

Resampling-Based Statistics:

Inferential statistics based on “simulating” an experiment a large number of times with the observed data

Observed Data

Group A	Group B
8	5
4	3
6	4

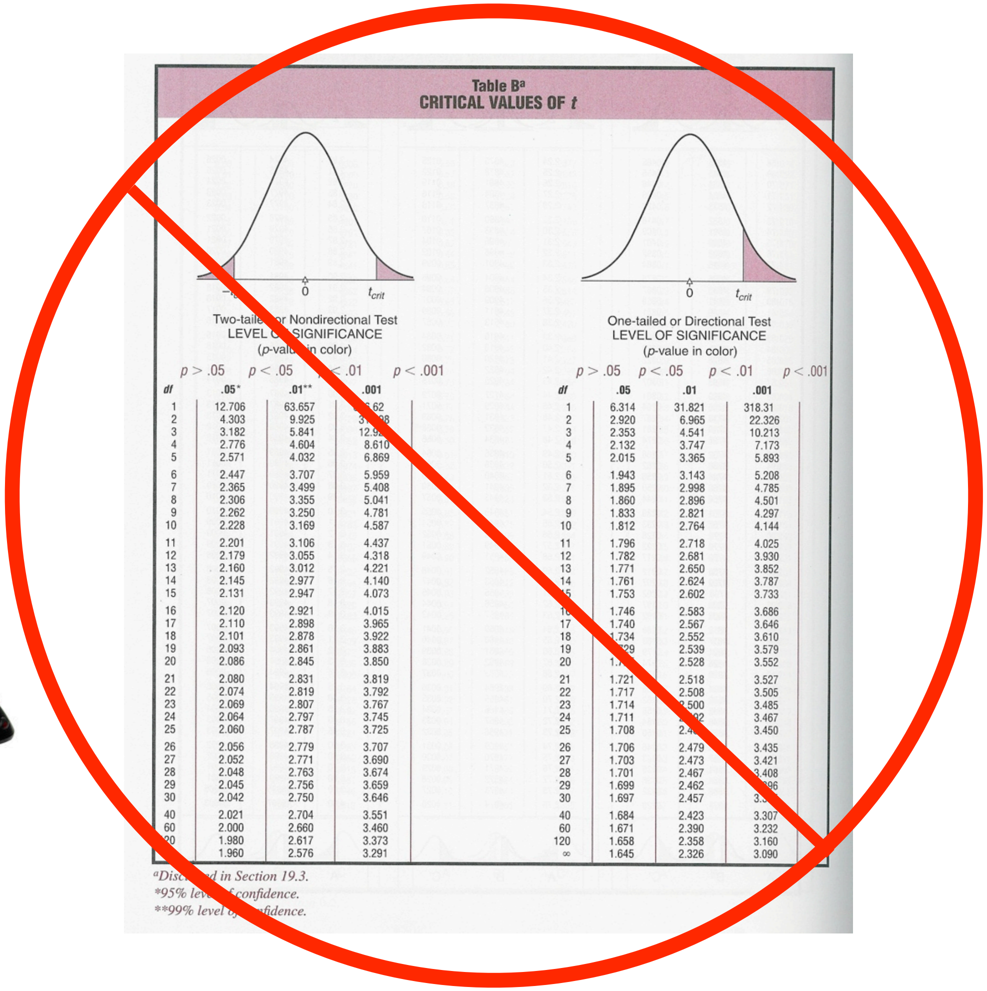
Resample →

“Simulated Replication”

Group A	Group B

Resampling-Based Statistics:

Inferential statistics based on “simulating” an experiment a large number of times with the observed data



−t_crit
0
t_crit
0
t_crit

Table B[a]
CRITICAL VALUES OF t

Two-tailed or Nondirectional Test
LEVEL OF SIGNIFICANCE
(p-value in color)

	$p > .05$	$p < .05$	$p < .01$	$p < .001$
df	**.05***	**.01****	**.001**	
1	12.706	63.657	636.62	
2	4.303	9.925	31.598	
3	3.182	5.841	12.924	
4	2.776	4.604	8.610	
5	2.571	4.032	6.869	
6	2.447	3.707	5.959	
7	2.365	3.499	5.408	
8	2.306	3.355	5.041	
9	2.262	3.250	4.781	
10	2.228	3.169	4.587	
11	2.201	3.106	4.437	
12	2.179	3.055	4.318	
13	2.160	3.012	4.221	
14	2.145	2.977	4.140	
15	2.131	2.947	4.073	
16	2.120	2.921	4.015	
17	2.110	2.898	3.965	
18	2.101	2.878	3.922	
19	2.093	2.861	3.883	
20	2.086	2.845	3.850	
21	2.080	2.831	3.819	
22	2.074	2.819	3.792	
23	2.069	2.807	3.767	
24	2.064	2.797	3.745	
25	2.060	2.787	3.725	
26	2.056	2.779	3.707	
27	2.052	2.771	3.690	
28	2.048	2.763	3.674	
29	2.045	2.756	3.659	
30	2.042	2.750	3.646	
40	2.021	2.704	3.551	
60	2.000	2.660	3.460	
120	1.980	2.617	3.373	
[illegible]	1.960	2.576	3.291	

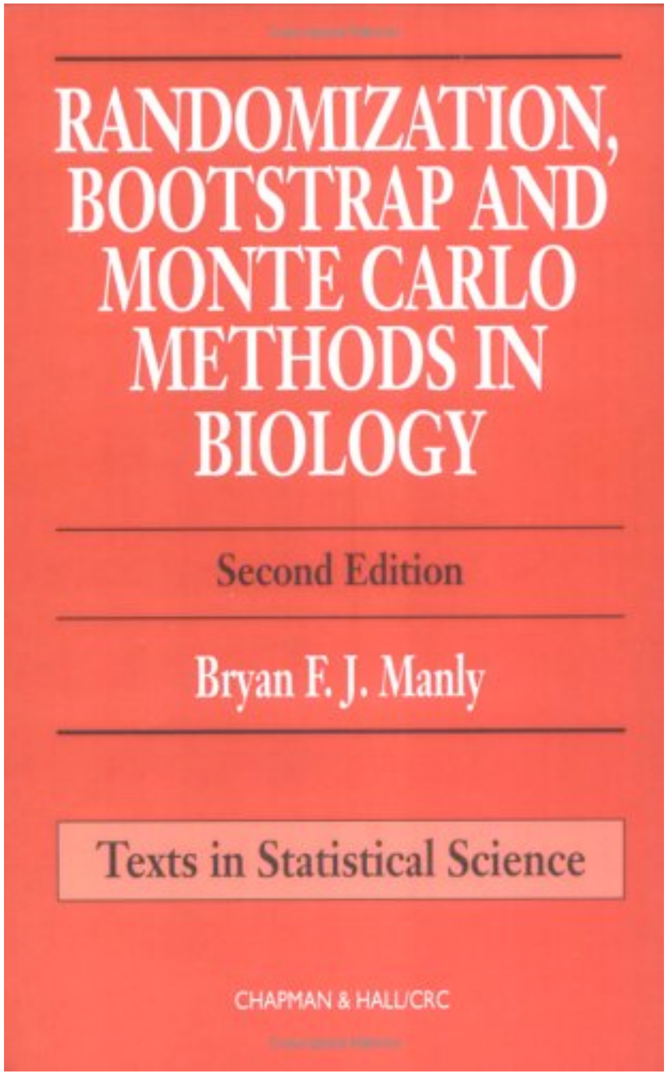
One-tailed or Directional Test
LEVEL OF SIGNIFICANCE
(p-value in color)

	$p > .05$	$p < .05$	$p < .01$	$p < .001$
df	**.05**	**.01**	**.001**	
1	6.314	31.821	318.31	
2	2.920	6.965	22.326	
3	2.353	4.541	10.213	
4	2.132	3.747	7.173	
5	2.015	3.365	5.893	
6	1.943	3.143	5.208	
7	1.895	2.998	4.785	
8	1.860	2.896	4.501	
9	1.833	2.821	4.297	
10	1.812	2.764	4.144	
11	1.796	2.718	4.025	
12	1.782	2.681	3.930	
13	1.771	2.650	3.852	
14	1.761	2.624	3.787	
15	1.753	2.602	3.733	
16	1.746	2.583	3.686	
17	1.740	2.567	3.646	
18	1.734	2.552	3.610	
19	1.729	2.539	3.579	
20	1.725	2.528	3.552	
21	1.721	2.518	3.527	
22	1.717	2.508	3.505	
23	1.714	2.500	3.485	
24	1.711	2.492	3.467	
25	1.708	2.485	3.450	
26	1.706	2.479	3.435	
27	1.703	2.473	3.421	
28	1.701	2.467	3.408	
29	1.699	2.462	3.396	
30	1.697	2.457	3.385	
40	1.684	2.423	3.307	
60	1.671	2.390	3.232	
120	1.658	2.358	3.160	
∞	1.645	2.326	3.090	

[a] *Discussed in Section 19.3.*
**95% level of confidence.*
***99% level of confidence.*

Resampling-Based Statistics:

Two Popular Resampling Methods



RANDOMIZATION, BOOTSTRAP AND MONTE CARLO METHODS IN BIOLOGY

Second Edition

Bryan F. J. Manly

Texts in Statistical Science

CHAPMAN & HALL/CRC

1. Permutation Tests (also called “Randomization Tests”)

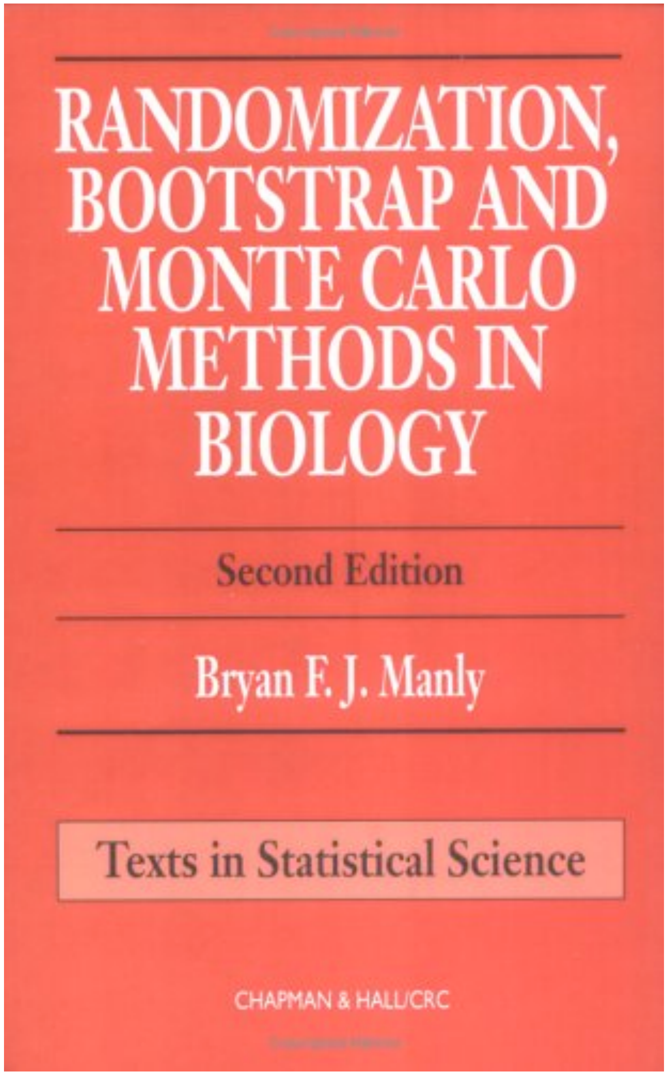
2. Bootstrap Statistics

Advantages of Permutation Tests & Bootstrap Statistics

1. Non-parametric (i.e., make minimal assumptions about population distributions)
2. Can be used in situations for which there is no analytic solution
3. Simple to use and easily provide confidence intervals
4. Useful for multiple comparison correction

Resampling-Based Statistics:

Two Popular Resampling Methods



RANDOMIZATION, BOOTSTRAP AND MONTE CARLO METHODS IN BIOLOGY

Second Edition

Bryan F. J. Manly

Texts in Statistical Science

CHAPMAN & HALL/CRC

1. Permutation Tests (also called “Randomization Tests”)

2. Bootstrap Statistics

Permutation Tests

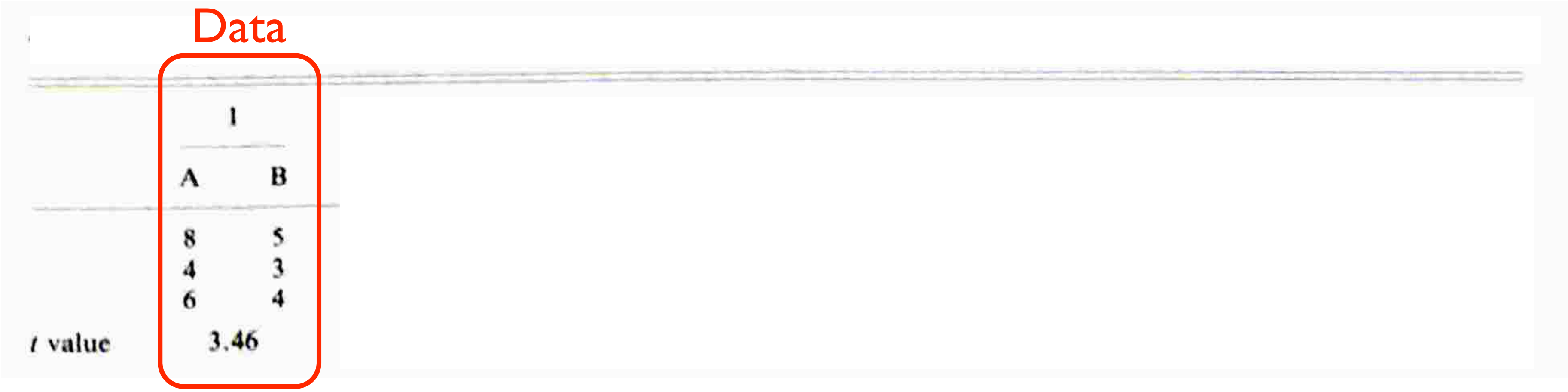
1. Old idea (Neyman, 1923; Fisher, 1935) but too computationally intensive to be widely used until relatively recently

2. Test the null hypothesis that the observations in multiple groups of data are exchangeable (i.e., they were just as likely to occur in one condition/group as any other)

Hypothetical Experiment #1

- Two conditions: A & B
- Within-subject design
- Three subjects

Observed Data

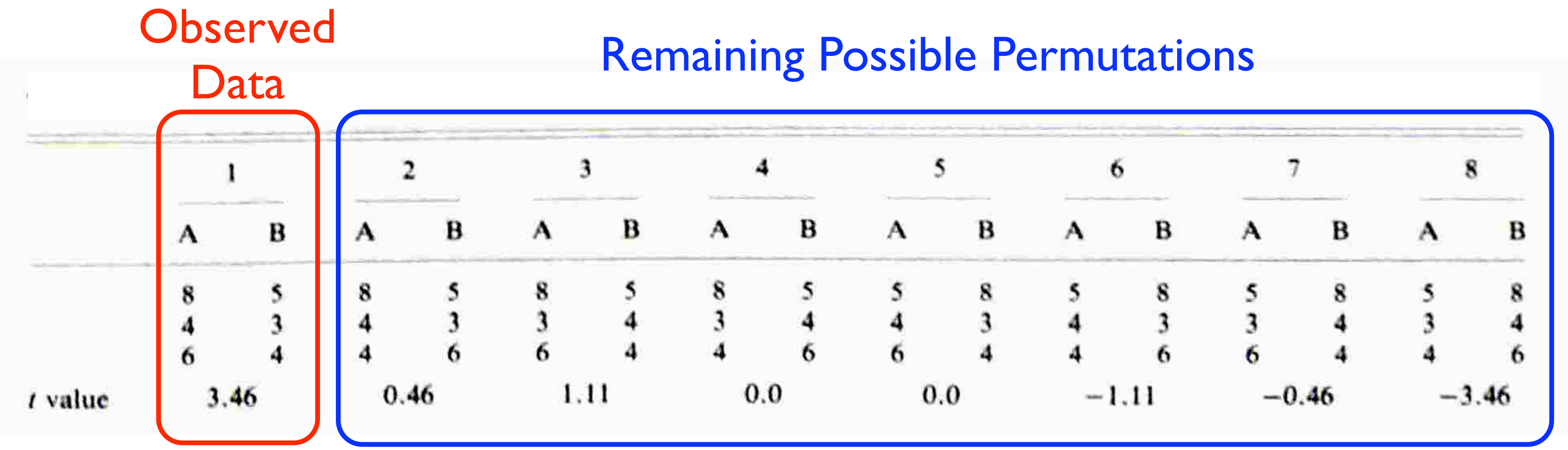


	1	
	A	B
	8	5
	4	3
	6	4
t value	3.46	

from: Blair & Karniski (1993) *Psychophysiology*

Null Hypothesis

- Observations in Condition A could have just as likely come from Condition B (and vice-versa)
- Each possible permutation of observations equally likely



Observed Data

Remaining Possible Permutations

	1		2		3		4		5		6		7		8	
	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
	8	5	8	5	8	5	8	5	5	8	5	8	5	8	5	8
	4	3	4	3	3	4	3	4	4	3	4	3	3	4	3	4
	6	4	4	6	6	4	4	6	6	4	4	6	6	4	4	6
t value	3.46		0.46		1.11		0.0		0.0		−1.11		−0.46		−3.46	

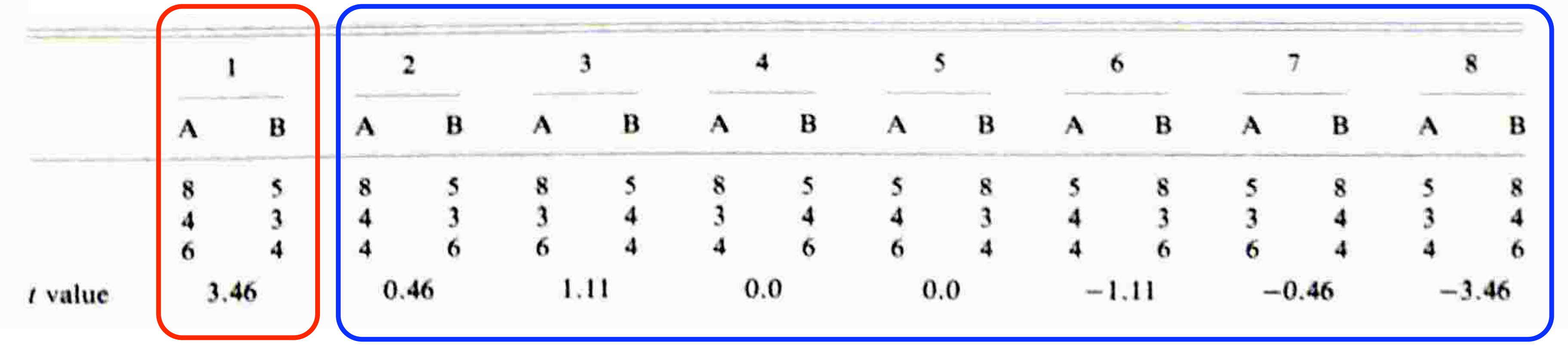
Sub1	orig	orig	orig	orig	flip	flip	flip	flip
Sub2	orig	orig	flip	flip	orig	orig	flip	flip
Sub3	orig	flip	orig	flip	orig	flip	orig	flip

2^n possible permutations

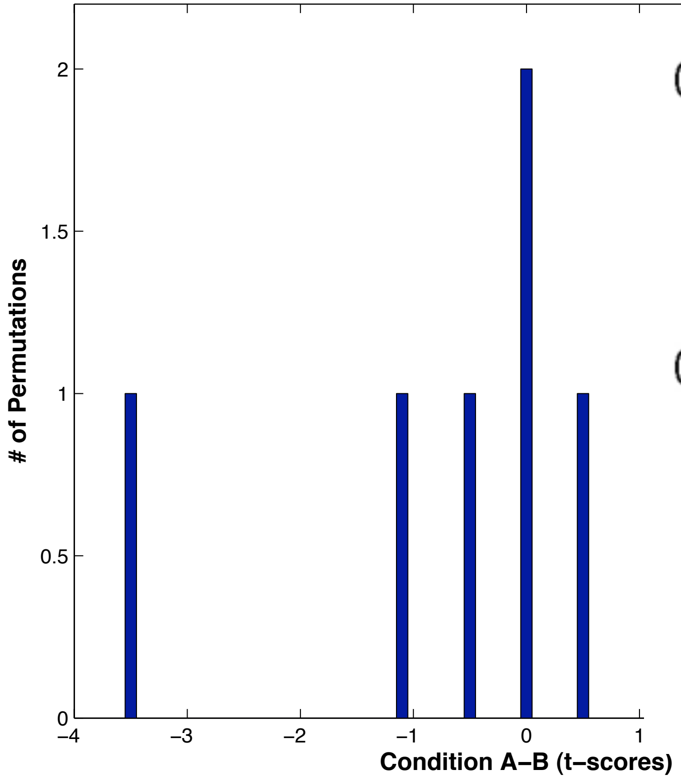
Null Hypothesis

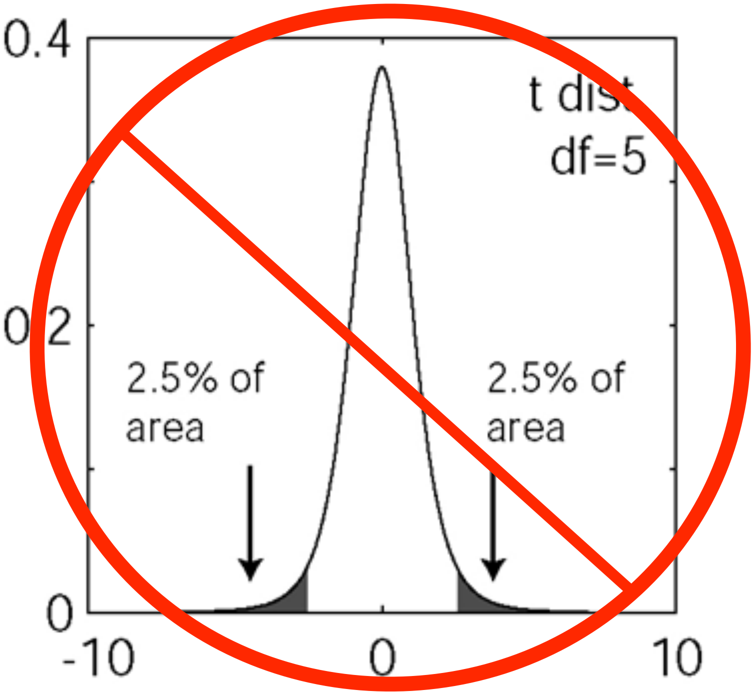
Observed Data

Remaining Possible Permutations



	1		2		3		4		5		6		7		8	
	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
	8	5	8	5	8	5	8	5	5	8	5	8	5	8	5	8
	4	3	4	3	3	4	3	4	4	3	4	3	3	4	3	4
	6	4	4	6	6	4	4	6	6	4	4	6	6	4	4	6
t value	3.46		0.46		1.11		0.0		0.0		−1.11		−0.46		−3.46	

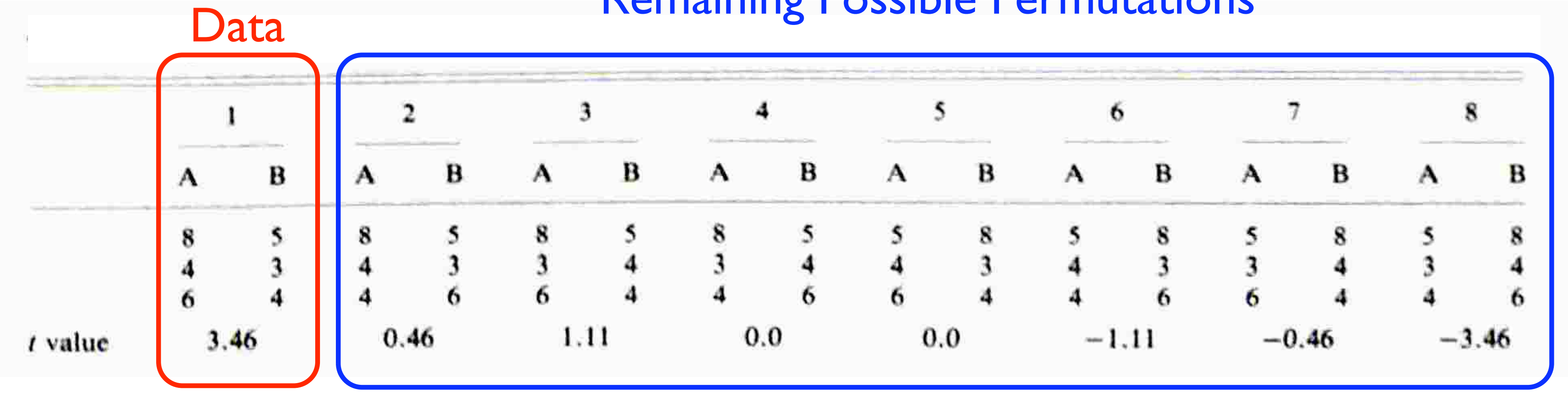




Null Hypothesis

Observed Data

Remaining Possible Permutations



	1		2		3		4		5		6		7		8	
	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
	8	5	8	5	8	5	8	5	5	8	5	8	5	8	5	8
	4	3	4	3	3	4	3	4	4	3	4	3	3	4	3	4
	6	4	4	6	6	4	4	6	6	4	4	6	6	4	4	6
t value	3.46		0.46		1.11		0.0		0.0		−1.11		−0.46		−3.46	

Permutation	1	2	3	4	5	6	7	8
t	-3.46	-1.11	-0.46	0	0	0.46	1.11	3.46

Null Hypothesis

Permutation	1	2	3	4	5	6	7	8
t	-3.46	-1.11	-0.46	0	0	0.46	1.11	3.46

Observed Difference

p=0.125

Decision Rule: If observed difference is the most positive permutation, reject null hypothesis (upper tailed test).

α=1/8=0.125

Null Hypothesis

Permutation	1	2	3	4	5	6	7	8
t	-3.46	-1.11	-0.46	0	0	0.46	1.11	3.46

Observed Difference

p=0.25

Decision Rule: If observed difference is the most positive or negative, reject null hypothesis (two tailed test).

α=2/8=0.25

Hypothetical Experiment #2

- Two conditions: A & B
- Within-subject design
- 25 subjects

2^{25} (i.e., 33,554,432) permutations

Approximate distribution of null hypothesis with thousands of random permutations.

Hypothetical Experiment #3

- Two groups: A & B
- Between-subject design
- 3 "A" subjects, 2 "B" subjects

Group	Observed Data
A	5
A	18
A	-23
B	9
B	3

Null Hypothesis

- Observations in Group A could have just as likely come from Group B (and vice-versa)
- Each possible permutation of observations equally likely

Group	Observed Data	Perm 2	Perm 3	Perm 4
A	5	5	5	5
A	18	9	18	18
A	-23	3	9	3
B	9	-23	-23	-23
B	3	18	3	9

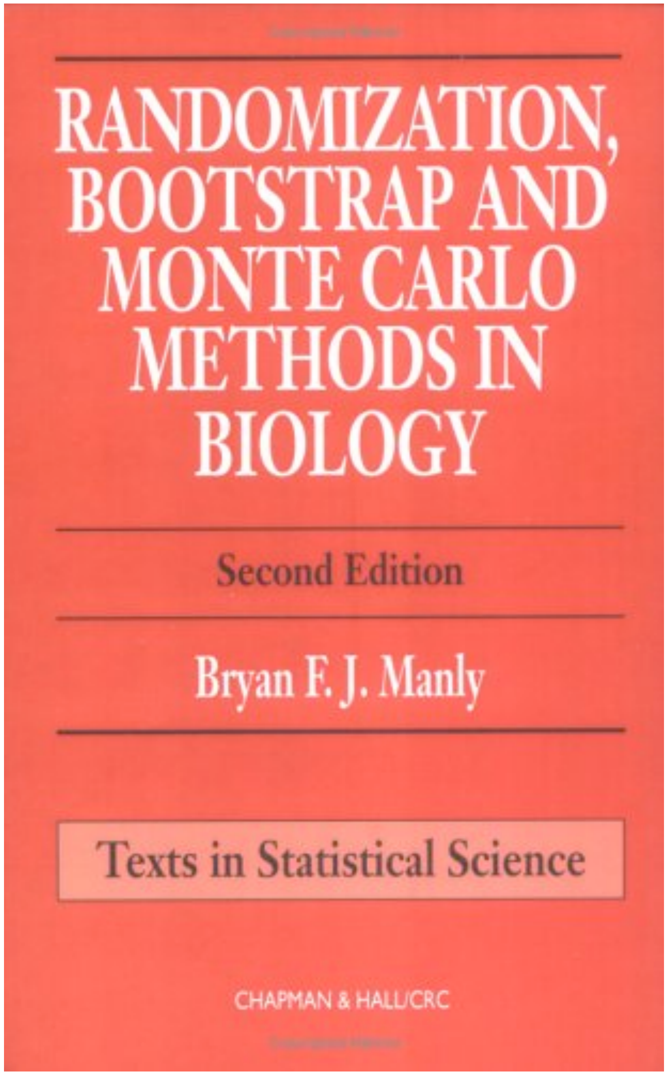
etc... →

Possible Permutations:

$$\binom{5}{3} = \frac{5!}{3!(5-2)!} = 10$$

Resampling-Based Statistics:

Two Popular Resampling Methods



RANDOMIZATION, BOOTSTRAP AND MONTE CARLO METHODS IN BIOLOGY

Second Edition

Bryan F. J. Manly

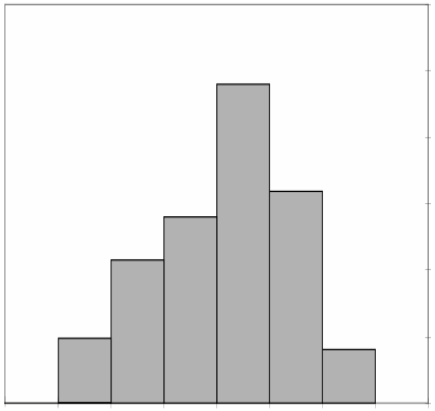
Texts in Statistical Science

CHAPMAN & HALL/CRC

1. Permutation Tests (also called “Randomization Tests”)
2. Bootstrap Statistics

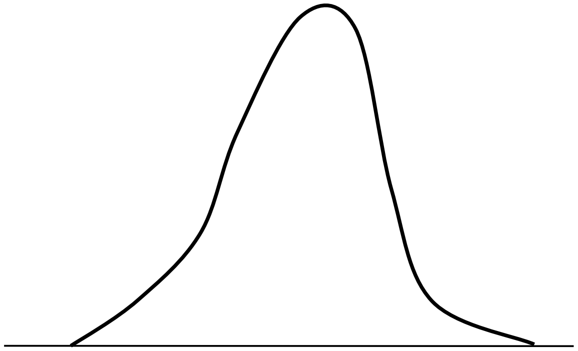
Sample and Population

What we observed



Sample

What we sampled from



Population

Bootstrap Statistics: Treat the sample as if it is the population

Hypothetical Experiment #4

- Two conditions: A & B
- Within-subject design
- Three subjects

Observed Data

A	B
8	5
4	3
6	4

Observed Difference

A-B
3
1
2

Mean Difference: **2**

Hypothetical Experiment #4

- Two conditions: A & B
- Within-subject design
- Three subjects

Bootstrap Sample

Observed Difference

Make a "bootstrap" sample by randomly selecting one of the difference values three times

	A-B	A-B*
	3	
	1	
	2	
Mean Difference:	2	

Hypothetical Experiment #4

- Two conditions: A & B
- Within-subject design
- Three subjects

Observed Difference

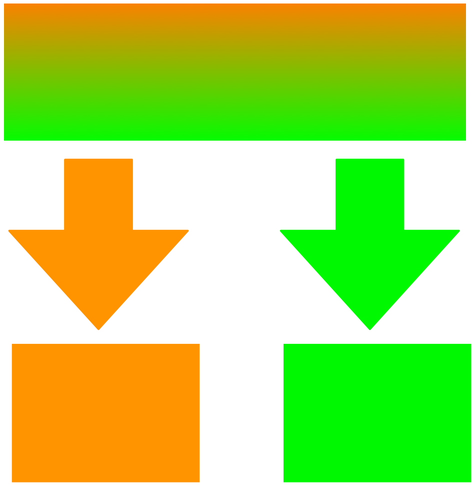
Bootstrap Sample

Make a "bootstrap" sample by randomly selecting one of the difference values three times

	A-B	A-B*
	3	2
	1	3
	2	3
Mean Difference:	**2**	**2.7**

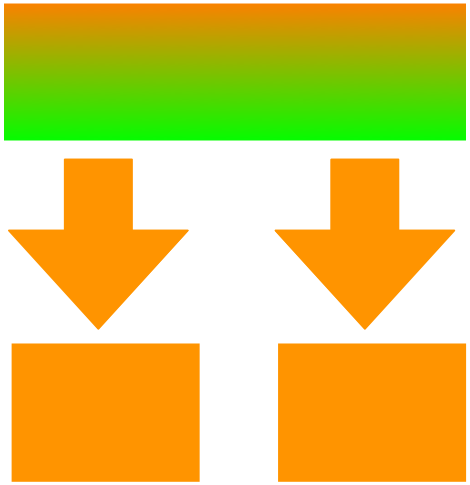
Bootstrap versus Permutation

Permutation



Each data point gets picked exactly once

Bootstrap



Each data point can be picked zero, one, or multiple times

Hypothetical Experiment #4

- Two conditions: A & B
- Within-subject design
- Three subjects

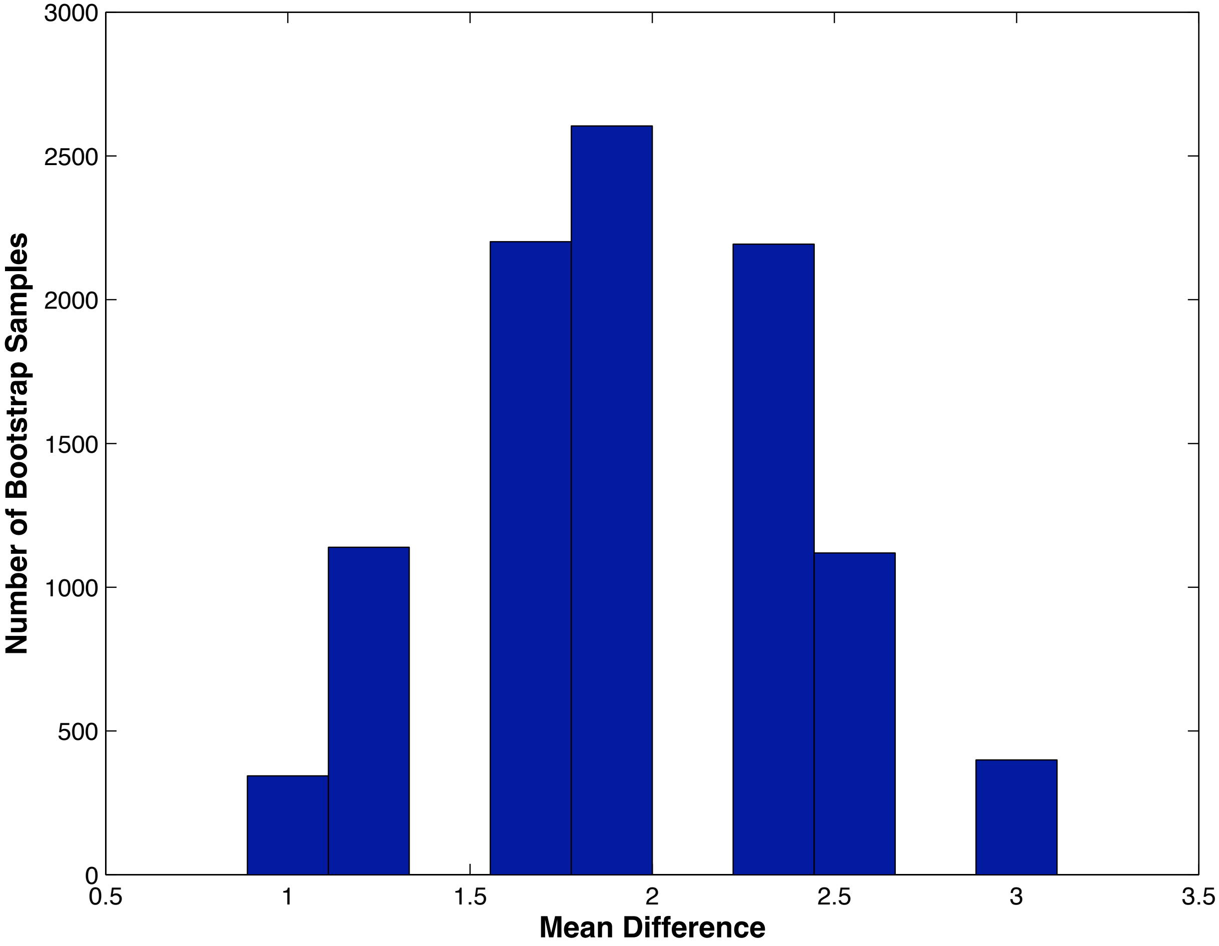
Observed Difference

Bootstrap Samples

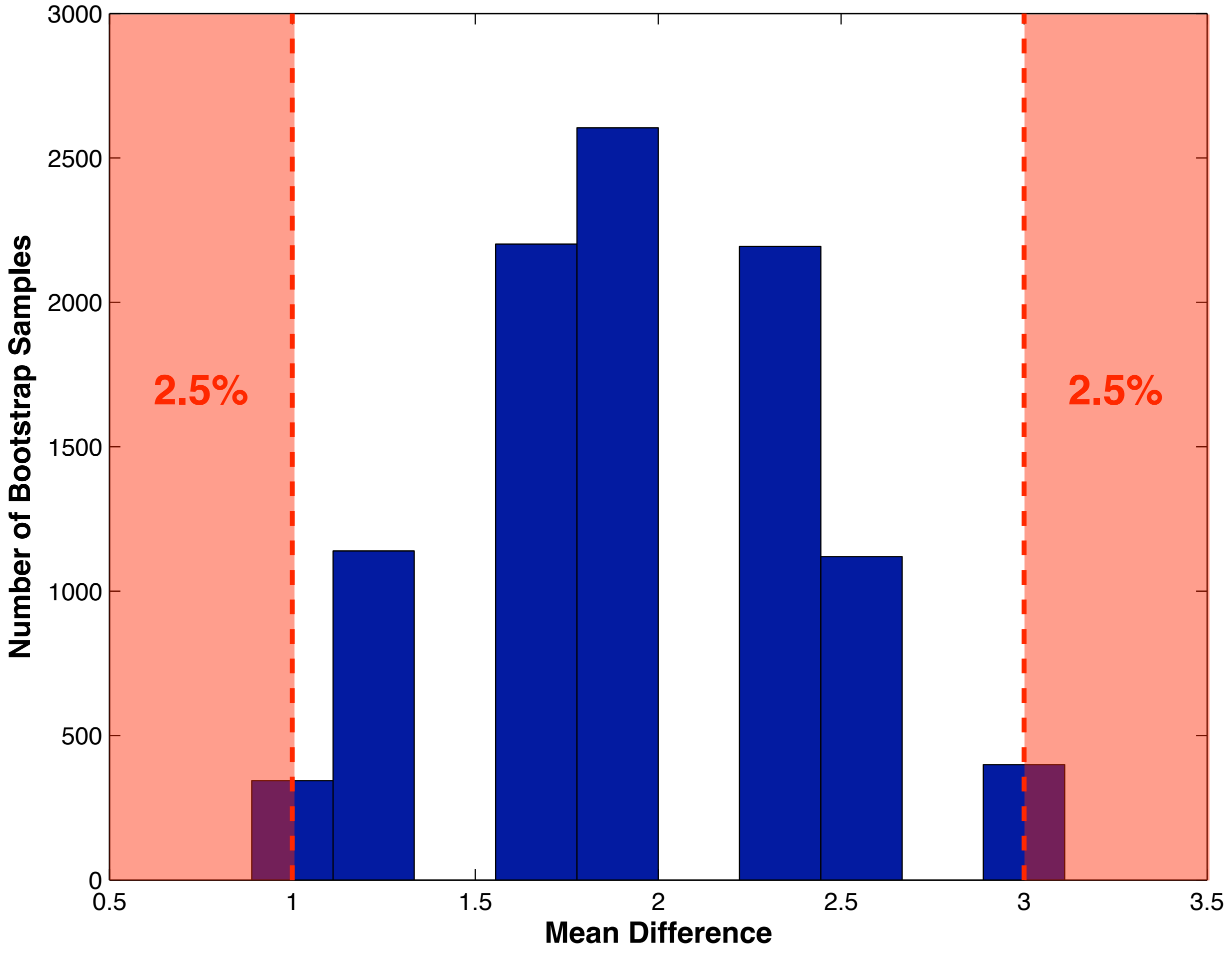
Make lots (thousands) of bootstrap samples

	A-B	A-B*	A-B*	A-B*	etc...
	3	2	2	3	
	1	3	2	2	
	2	3	1	2	
Mean Difference:	**2**	**2.7**	**1.7**	**2.3**	

Distribution of Mean of 10,000 Bootstrap Samples



Distribution of Mean of 10,000 Bootstrap Samples



“Percentile Bootstrap” Confidence Intervals

Presentation Outline

Summary:

- **“Classic** **Statistics**
  - Parametri
- **Resampling-Based Inferential Statistics**
  - Randomization/permutation tests
  - Bootstrap statistics
- **Correcting for Multiple Comparisons**
  - Permutation test based control of family-wise error
  - Benjamini methods for control of false discovery rate
  - Evaluating multiple comparison correction on simulated ERP data

Advantages of Permutation Tests & Bootstrap Statistics

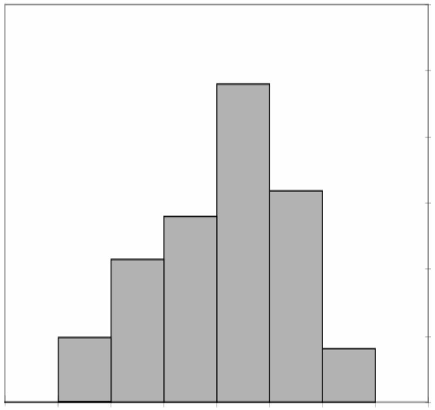
1. Non-parametric (i.e., make minimal assumptions about population distributions)
2. Can be used in situations for which there is no analytic solution
3. Simple to use and easily provide confidence intervals
4. Useful for multiple comparison correction

Coming up next!

Disadvantages of Permutation Tests & Bootstrap Statistics

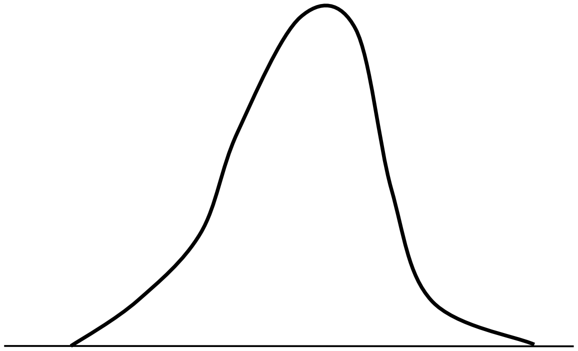
1. Poor performance with small sample sizes
   - Might be inaccurate

What we observed



Sample

What we sampled from



Population

	Permutation	Bootstrap
Simple Analyses (e.g., *t*-tests, correlation)	✓ Always Accurate	Asymptotically Accurate
Complex Analyses (e.g., multifactor ANOVAS)	Asymptotically Accurate or Not Applicable	✓ Asymptotically Accurate

Disadvantages of Permutation Tests & Bootstrap Statistics

1. Poor performance with small sample sizes

- Might be inaccurate
- Limited set of possible *p*-values

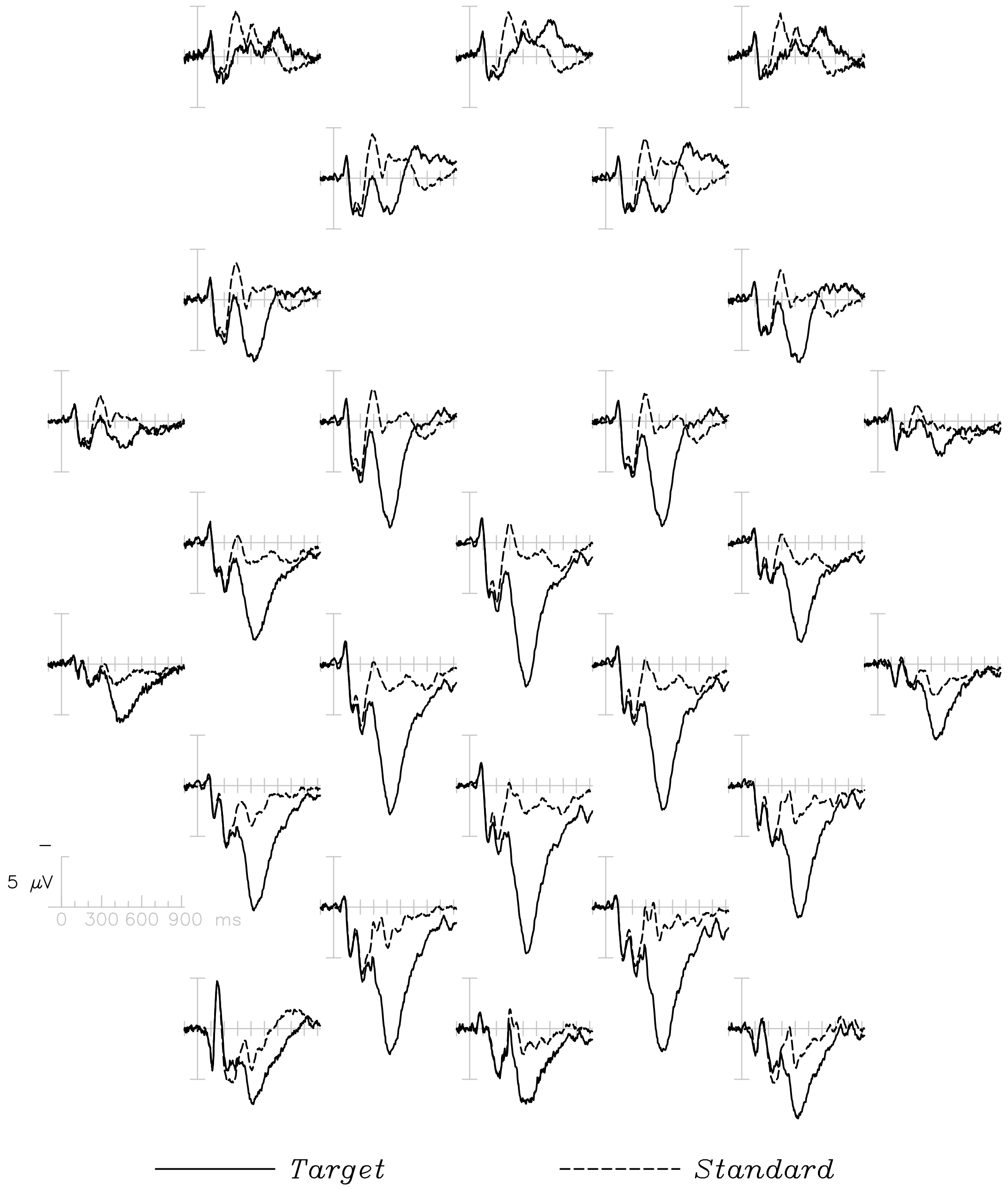
Disadvantages of Permutation Tests & Bootstrap Statistics

1. Poor performance with small sample sizes
   - Might be inaccurate
   - Limited set of possible *p*-values
2. Not practical for computationally intensive analyses (e.g., non-linear regression via gradient descent)

Presentation Outline

- **"Classic" Analytical Inferential Statistics**
  - Parametric & non-parametric
- **Resampling-Based Inferential Statistics**
  - Randomization/permutation tests
  - Bootstrap statistics
- **Correcting for Multiple Comparisons**
  - Permutation test based control of family-wise error
  - Benjamini methods for control of false discovery rate
  - Evaluating multiple comparison correction on simulated ERP data

Potentially Lots of Possible Statistical Tests



Conventional ERP study:

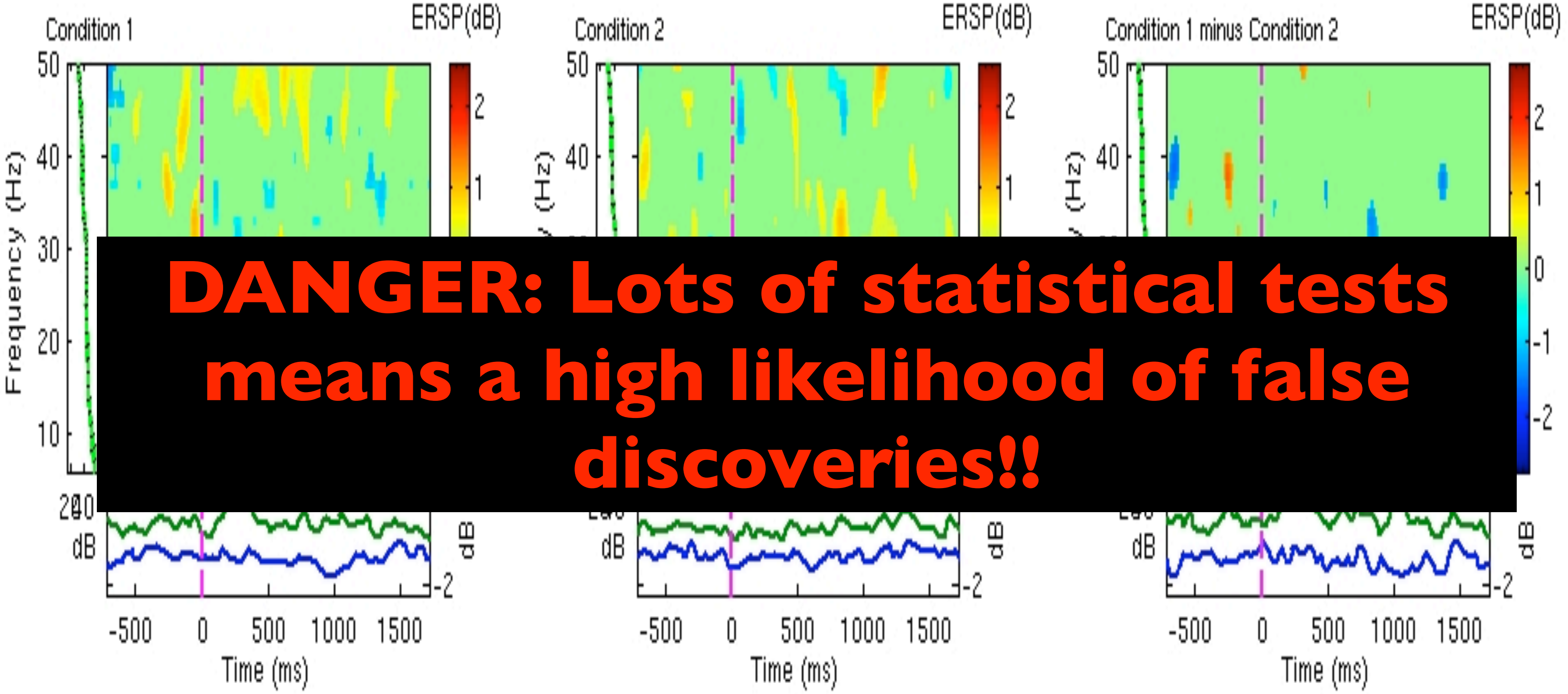
-2 conditions

-26 electrodes

-218 time points (50-920 ms)

-5,668 dependent variables

Potentially Lots of Possible Statistical Tests



DANGER: Lots of statistical tests means a high likelihood of false discoveries!!

Even more dependent variables with time-frequency analyses!!

Hypothetical Experiment #4

- Two conditions: A & B
- Within-subject design
- Three subjects
- Two dependent variables: X & Y

X

	A	B	A-B
Sub1	-4	28	-32
Sub2	3	-13	16
Sub3	36	30	6

t_x=-0.23

Y

A	B	A-B
141	-121	262
142	72	70
67	163	-96

t_y=0.76

Control of Family-Wise Error Rate (FWER)

$FWER = P(R_F > 0) = \alpha_{fam}$

$R_F =$ number of false discoveries in the family of tests

Control of Family-Wise Error Rate (FWER)

$$FWER = P(R_F > 0) = \alpha_{fam}$$

R_F = number of false discoveries in the family of tests

This "family" consists of two tests:

X

Su			
Su			
Sub3	36	30	6

Y

67	163	-96

FWER control provides same degree of certainty as a priori tests!!

t_x=-0.23

t_y=0.76

Control of Family-Wise Error Rate (FWER)

$$FWER = P(R_F > 0) = \alpha_{fam}$$

$$R_F = \text{number of false discoveries in the family of tests}$$

Bonferroni Correction:

$$\text{Desired "family - wise alpha"} = \text{Desired } \alpha_{fam} = 0.05$$

$$\text{Bonferroni "test - wise alpha"} = \alpha_{test} = \frac{\text{Desired } \alpha_{\text{fam}}}{\#\ \text{of comparisons}} = \frac{0.05}{2} = 0.025$$

$$\text{True } \alpha_{fam} \leq \text{Desired } \alpha_{\text{fam}}$$

Control of Family-Wise Error Rate (FWER)

$$FWER = P(R_F > 0) = \alpha_{fam}$$

$$R_F = \text{number of false discoveries in the family of tests}$$

Bonferroni Correction:

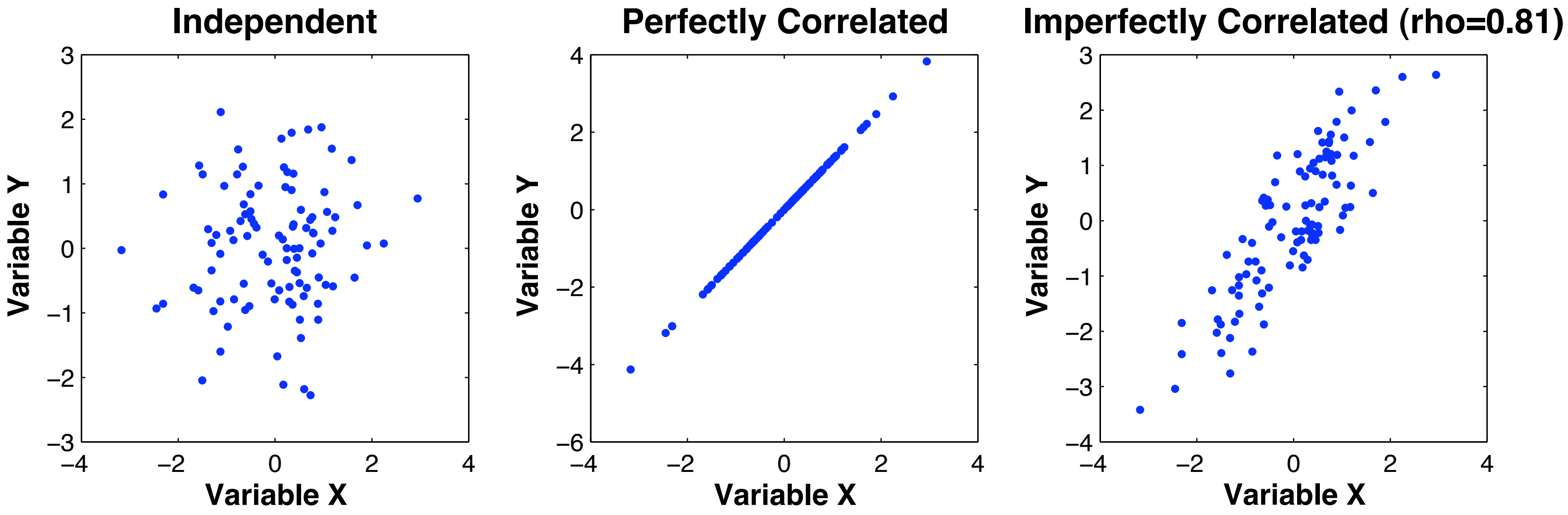
$$\text{Desired "family - wise alpha"} = \text{Desired } \alpha_{fam} = 0.05$$

$$\text{Bonferroni "test - wise alpha"} = \alpha_{test} = \frac{\text{Desired } \alpha_{\text{fam}}}{\#\ \text{of comparisons}} = \frac{0.05}{2} = 0.025$$

$\text{True } \alpha_{fam} \leq \text{Desired } \alpha_{\text{fam}}$ ⟵ Might be overly conservative

Bonferroni Correction

- Desired α_{fam}: 5%
- Bonferroni α_{test}: 2.5%



4.9(±0.3)%

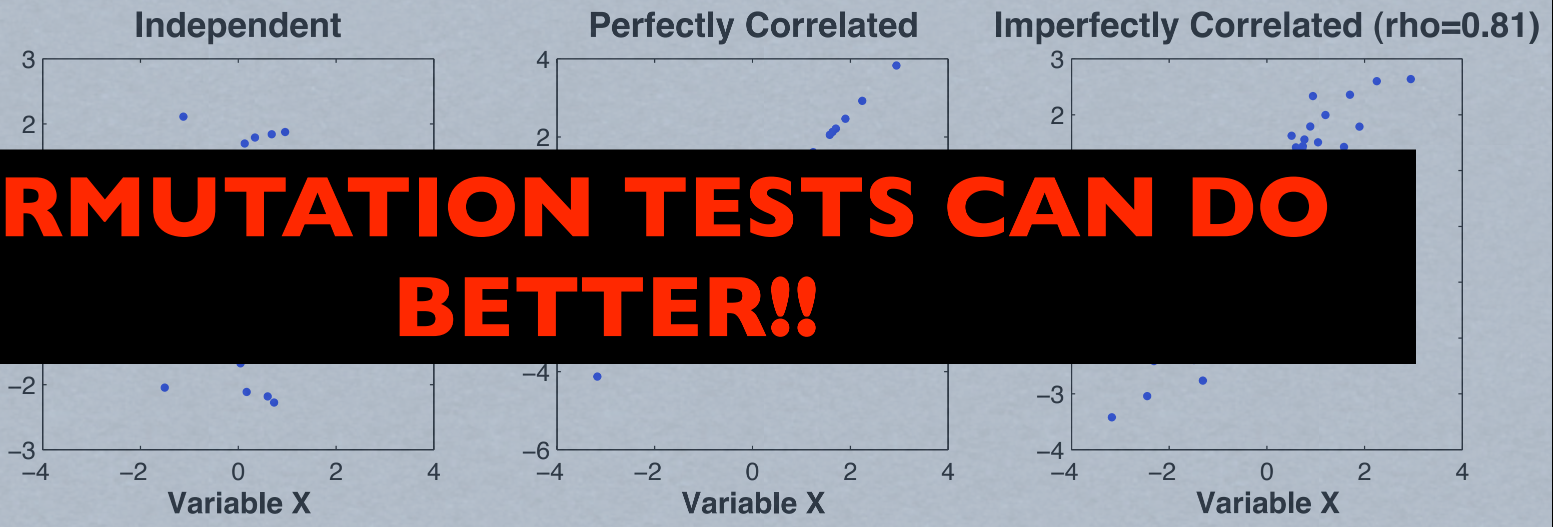
2.3(±0.3)%

4.1(±0.3)%

Estimated true family-wise α level (95% Confidence Intervals)

Bonferroni Correction

- Desired α_{fam}: 5%
- Bonferroni α_{test}: 2.5%



PERMUTATION TESTS CAN DO BETTER!!

4.9(±0.3)%

2.3(±0.3)%

4.1(±0.3)%

Estimated true family-wise α level (95% Confidence Intervals)

Permutation Test

Observed Values (Permutation #1)

X

	A	B	A-B
Sub1	-4	28	-32
Sub2	3	-13	16
Sub3	36	30	6

t_x=-0.23

Y

A	B	A-B
141	-121	262
142	72	70
67	163	-96

t_y=0.76

t_{max}=most extreme t-score=0.76

Permutation Test

Permutation #2

X

	A	B	A-B
Sub1	28	-4	32
Sub2	3	-13	16
Sub3	36	30	6

t_x=2.38

Y

A	B	A-B
-121	141	-262
142	72	70
67	163	-96

t_y=-1.00

t_{max}=most extreme *t*-score=2.38

Null Hypothesis

Permutation	1	2	3	4	5	6	7	8
t_{max}	-2.377	-2.372	-1.27	-0.76	0.76	1.27	2.372	2.377

Decision Rule: If observed difference is most positive or negative, reject null hypothesis (two tailed test).

Critical t=±2.377

Null Hypothesis

Permutation	1	2	3	4	5	6	7	8
t_{max}	-2.377	-2.372	-1.27	-0.76	0.76	1.27	2.372	2.377

Decision Rule: If observed difference is most positive or negative, reject null hypothesis (two tailed test).

Critical t=±2.377

$\alpha_{fam}=2/8=0.25$

Permutation Test

Observed Values (Permutation #1)

X

	A	B	A-B
Sub1	-4	28	-32
Sub2	3	-13	16
Sub3	36	30	6

t_x=-0.23

Y

A	B	A-B
141	-121	262
142	72	70
67	163	-96

t_y=0.76

Perm Test Critical t=±2.377

Permutation Test

Observed Values (Permutation #1)

X

	A	B	A-B
Sub1	-4	28	-32
Sub2	3	-13	16
Sub3	36	30	6

t_x=-0.23

Y

A	B	A-B
141	-121	262
142	72	70
67	163	-96

t_y=0.76

Perm Test Critical t=±2.377

Retain null hypothesis

(i.e., neither X nor Y significantly differ across A & B)

Corrects for Multiple Comparisons by Raising Critical t

X

	A	B	A-B
Sub1	-4	28	-32
Sub2	3	-13	16
Sub3	36	30	6

t_x=-0.23

Y

A	B	A-B
141	-121	262
142	72	70
67	163	-96

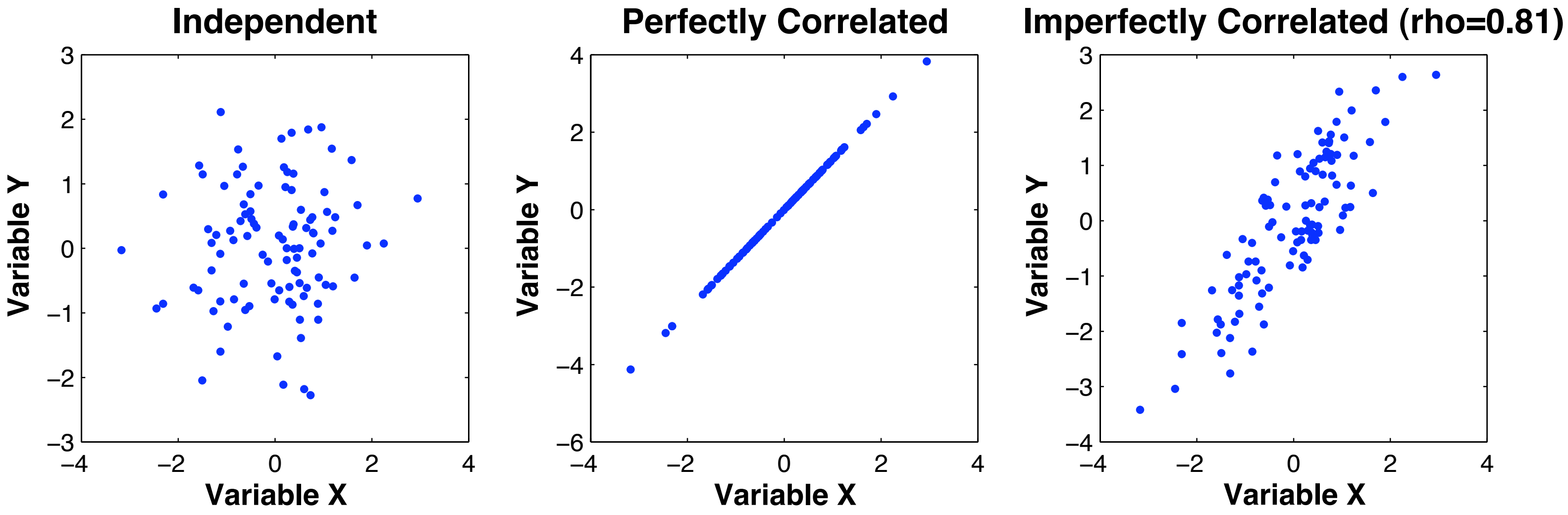
t_y=0.76

Perm Test Critical t=±2.377

Repeated Measures t-test Critical t
(no correction for two comparisons)=±2.353

t_{max} Permutation Test

- Desired α_{fam}: 5%



4.9(±0.3)%

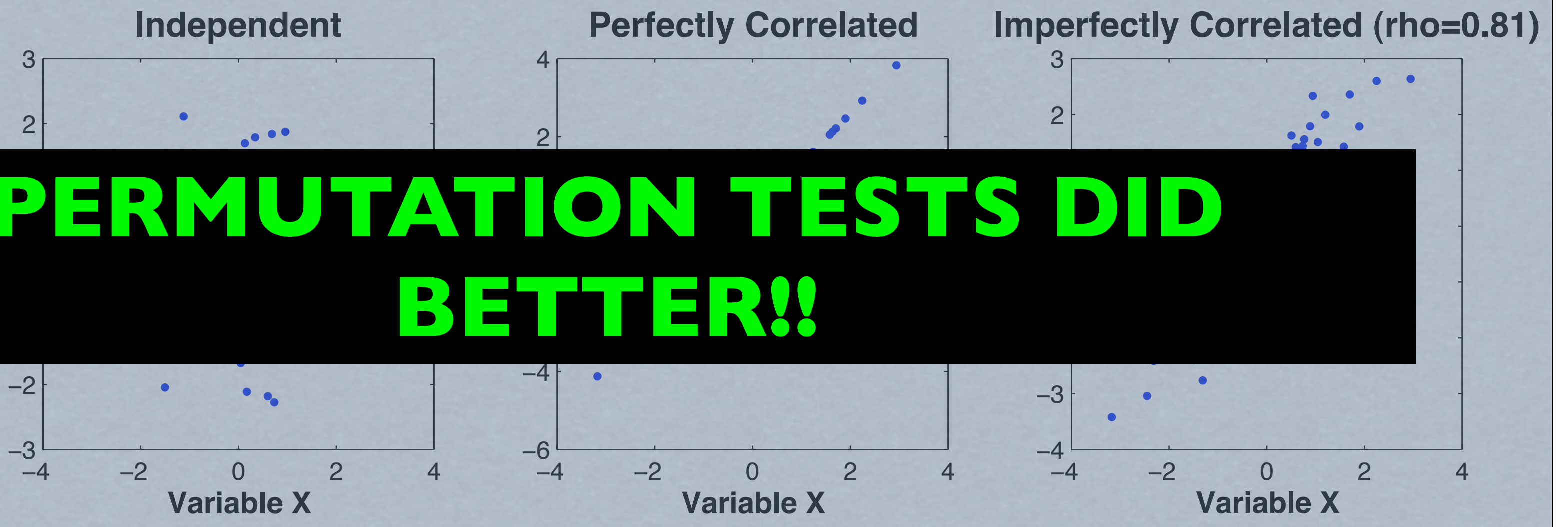
4.8(±0.3)%

5.1(±0.3)%

Estimated true family-wise α level (95% Confidence Intervals)

t_{max} Permutation Test

- Desired α_{fam}: 5%



PERMUTATION TESTS DID BETTER!!

4.9(±0.3)%

4.8(±0.3)%

5.1(±0.3)%

Estimated true family-wise α level (95% Confidence Intervals)

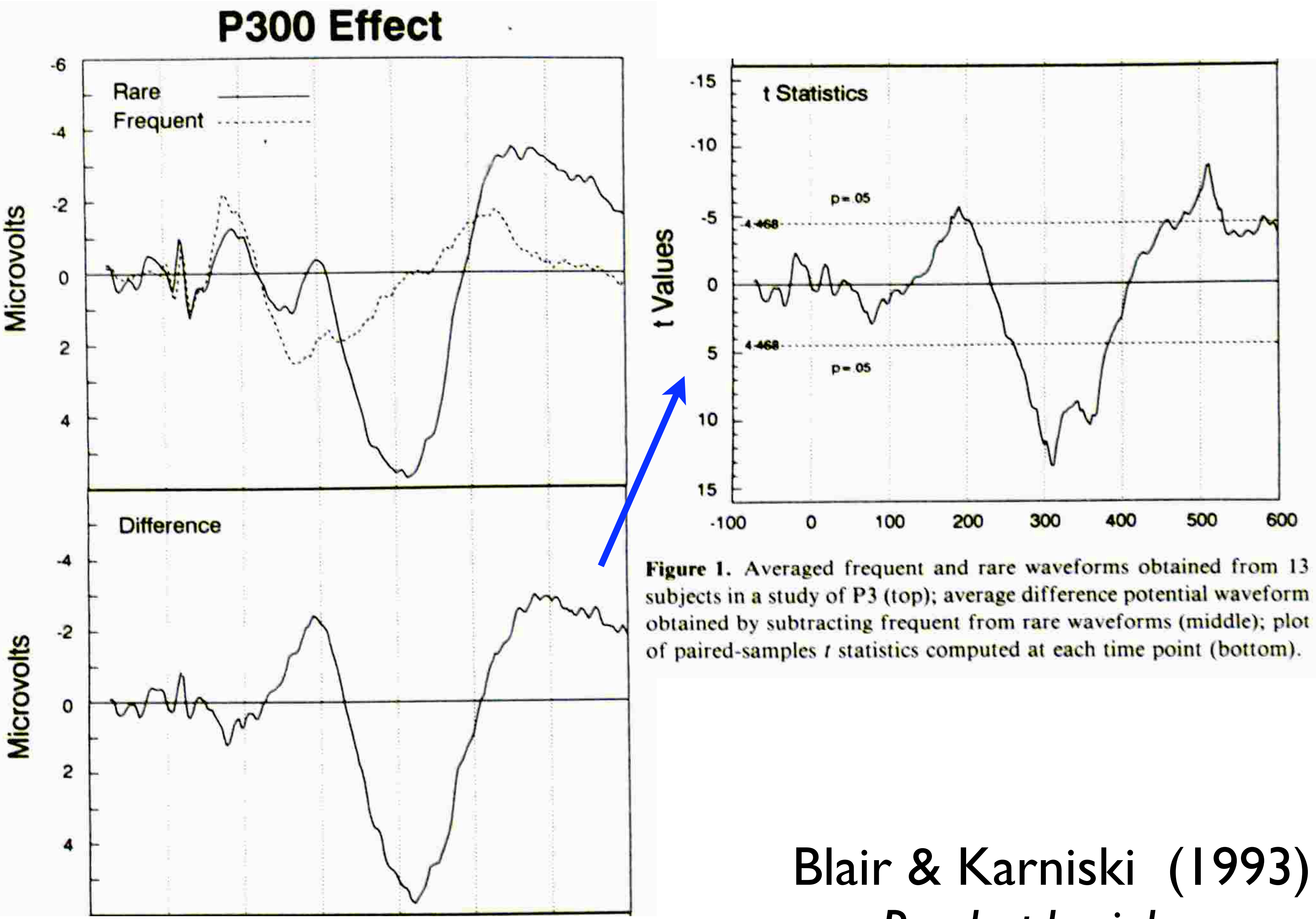
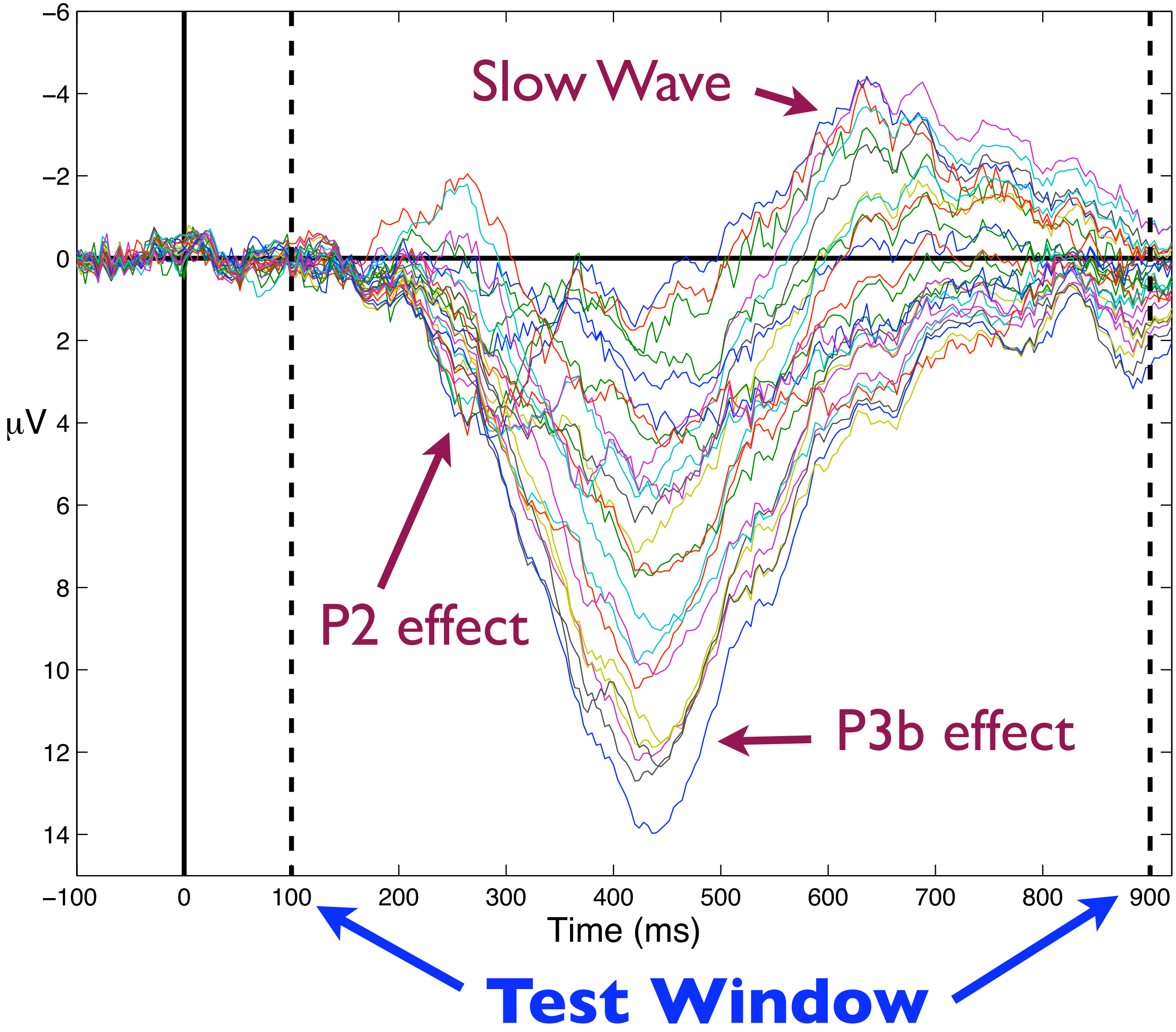


Figure 1. Averaged frequent and rare waveforms obtained from 13 subjects in a study of P3 (top); average difference potential waveform obtained by subtracting frequent from rare waveforms (middle); plot of paired-samples *t* statistics computed at each time point (bottom).

Blair & Karniski (1993)
Psychophysiology

Target-Standard Difference Wave (26 electrodes)



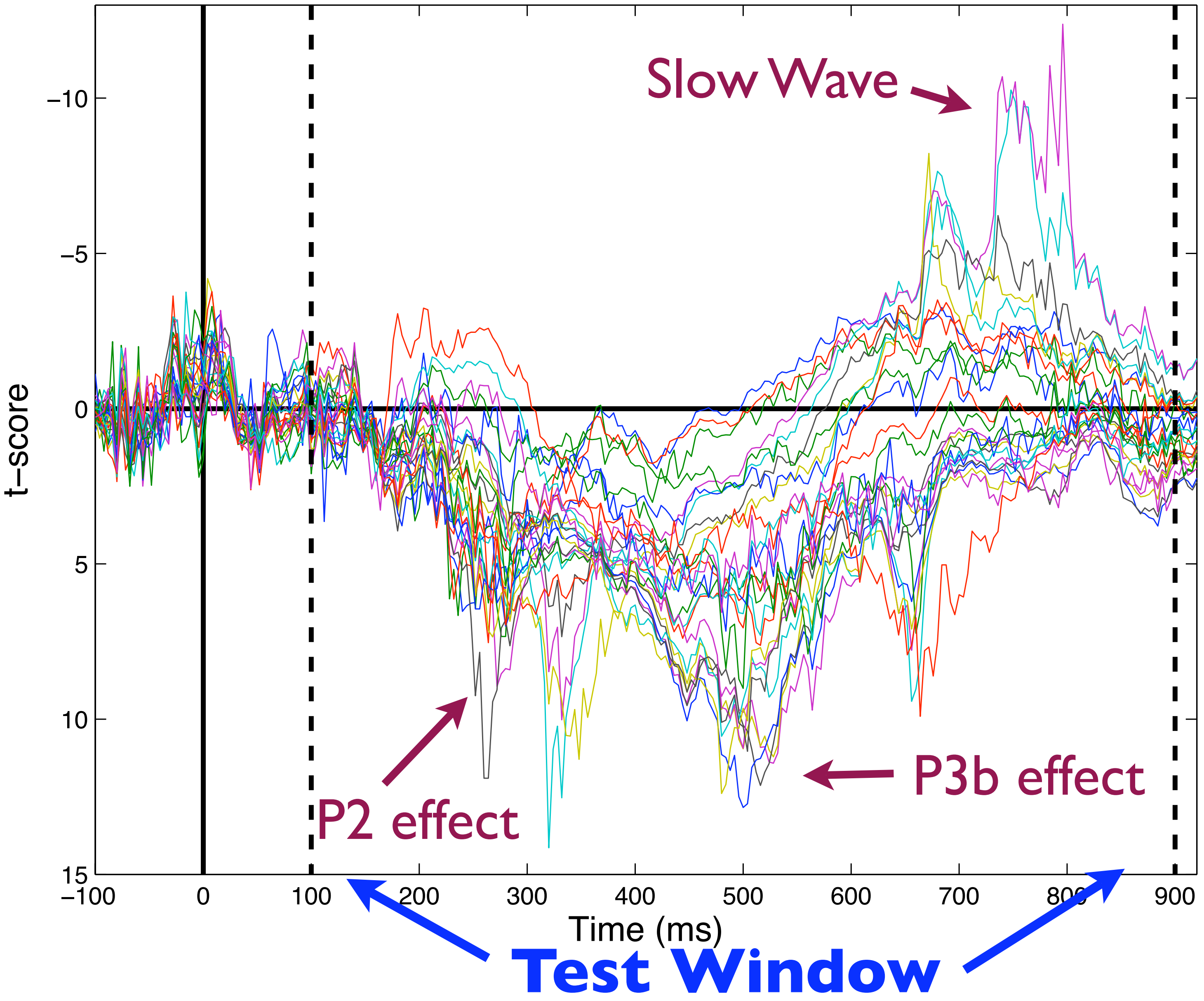
Slow Wave

P2 effect

P3b effect

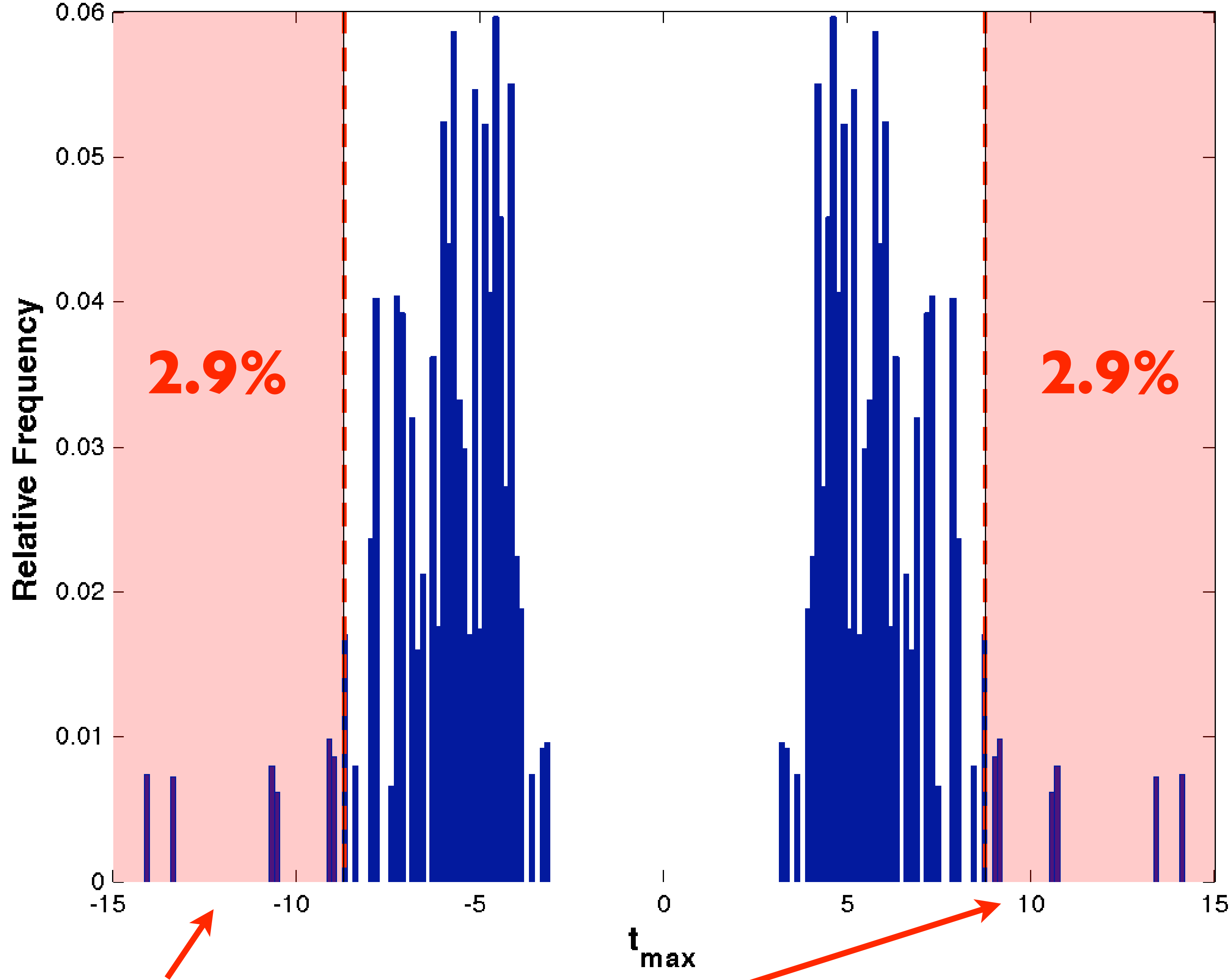
Test Window

Target-Standard Difference Wave (26 electrodes)



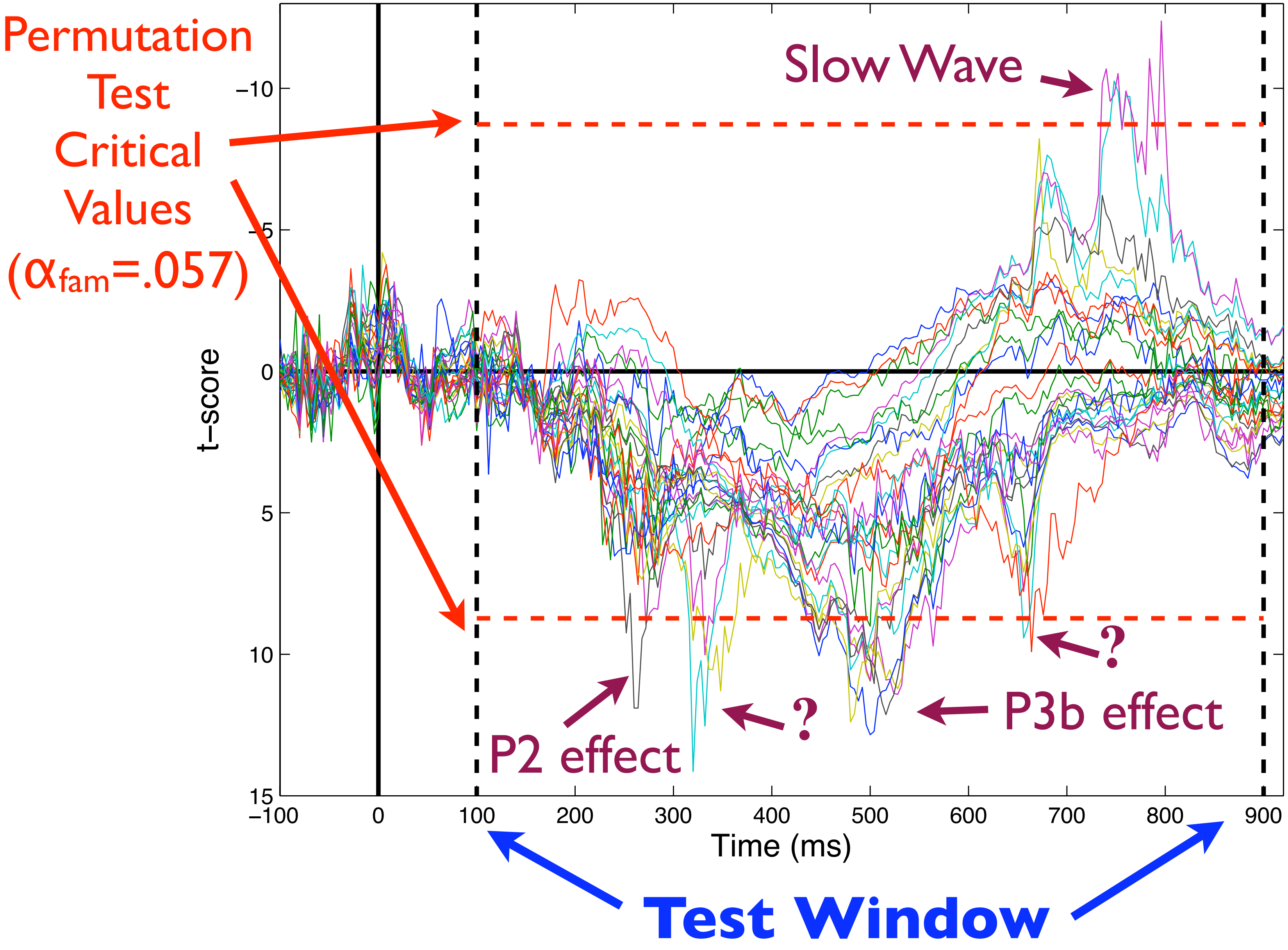
Test Window

t_{max} Distribution from 5000 Permutations



Critical Regions

Target-Standard Difference Wave (26 electrodes)



Permutation Test Critical Values (α_{fam}=.057)

Slow Wave

P2 effect

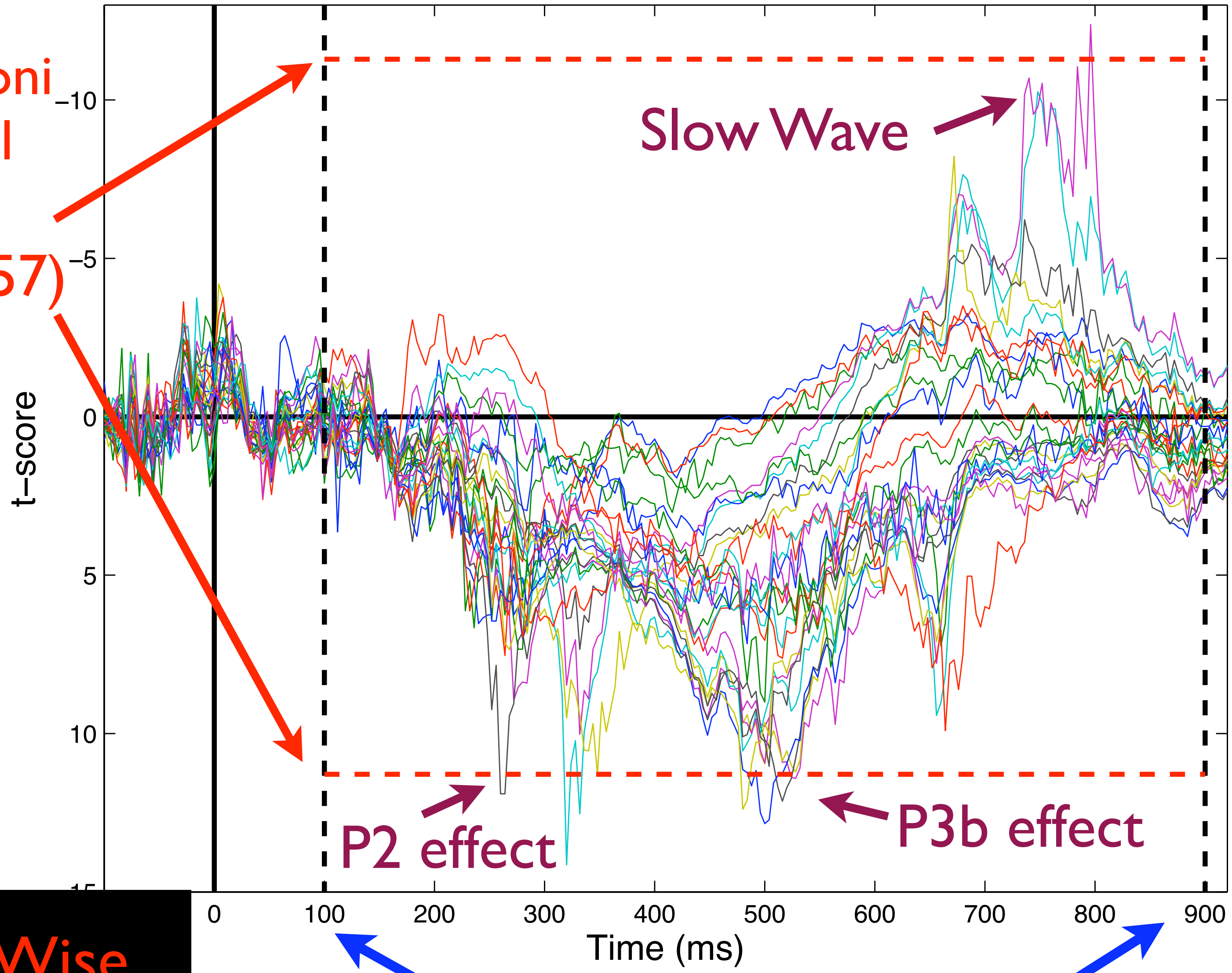
?

P3b effect

?

Test Window

Target-Standard Difference Wave (26 electrodes)



Bonferroni Critical Values (α_{fam}=.057)

Slow Wave

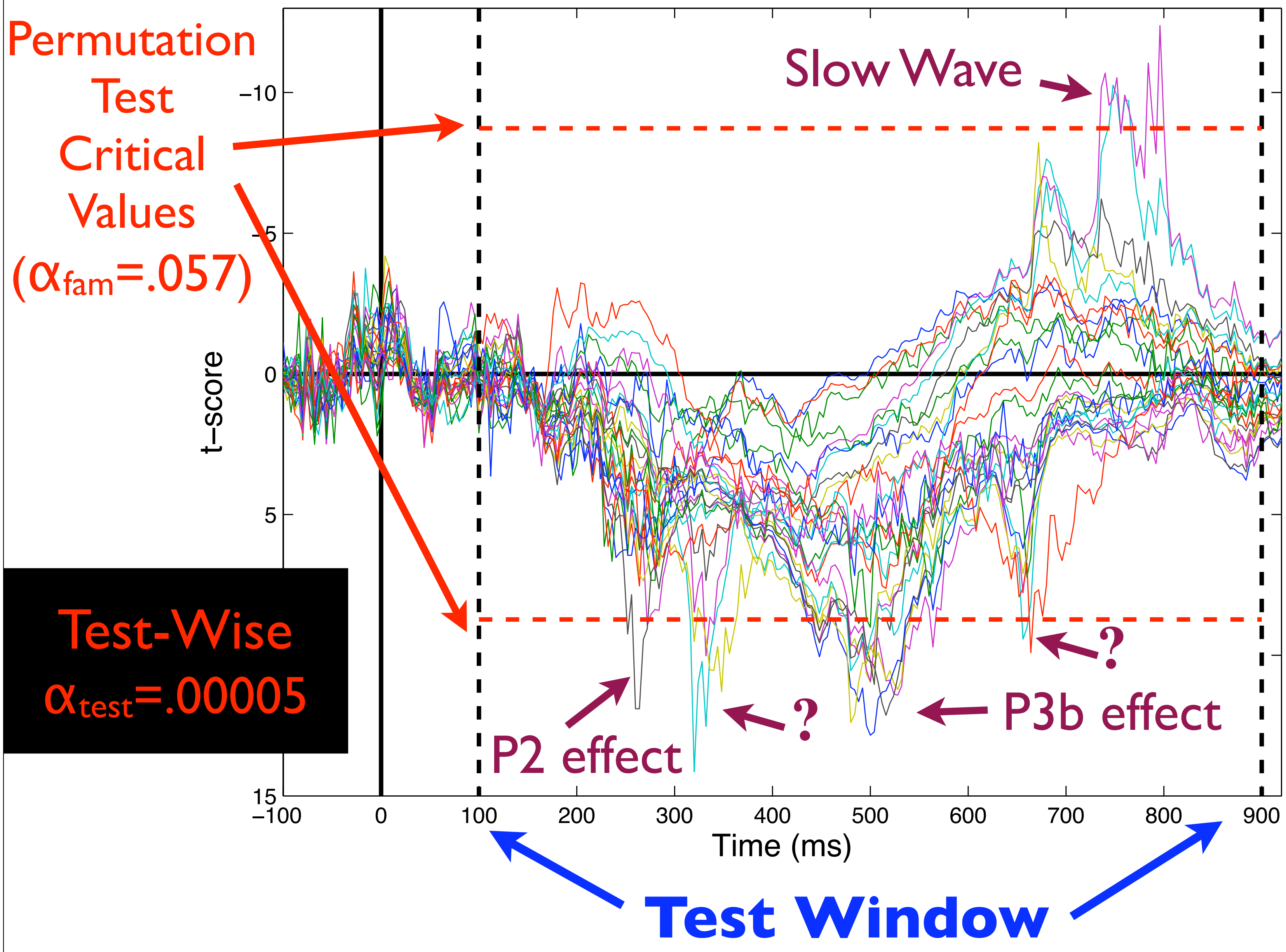
P2 effect

P3b effect

Test Window

Test-Wise α_{test}=.00001

Target-Standard Difference Wave (26 electrodes)



Permutation Tests: Some Pros

1. FWER control provides the same degree of certainty as more selective a priori tests
2. Guaranteed accuracy for simple tests (e.g., *t*-tests, correlation)
3. Relatively powerful when dependent variables are highly correlated (like EEG)

Permutation Tests: Some Cons

1. For more complicated tests (e.g., two factor ANOVAs) the results are only "asymptotically exact" (like bootstrapping).
2. Power can still be rather weak with a larger number of comparisons

Presentation Outline

- **“Classic” Analytical Inferential Statistics**
    - Parametric & non-parametric
- **Resampling-Based Inferential Statistics**
    - Randomization/permutation tests
    - Bootstrap statistics
- **Correcting for Multiple Comparisons**
    - Permutation test based control of family-wise error
    - ⇨ Benjamini methods for control of false discovery rate
    - Evaluating multiple comparison correction on simulated ERP data

Control of Family-Wise Error Rate (FWER)

$$FWER = P(R_F > 0) = \alpha$$

$$R_F = \text{number of false discoveries in the family of tests}$$

If FWER=5%, you have a 5% chance that one or more of your significant *p*-values is a mistake.

Control of Family-Wise Error (FWER)

$$FWER = P(R_F > 0) = \alpha$$

$$R_F = \text{number of false discoveries in the family of tests}$$

If FWER=5%, you have a 5% chance that one or more of your significant *p*-values is a mistake.

Control of False Discovery Rate (FDR)

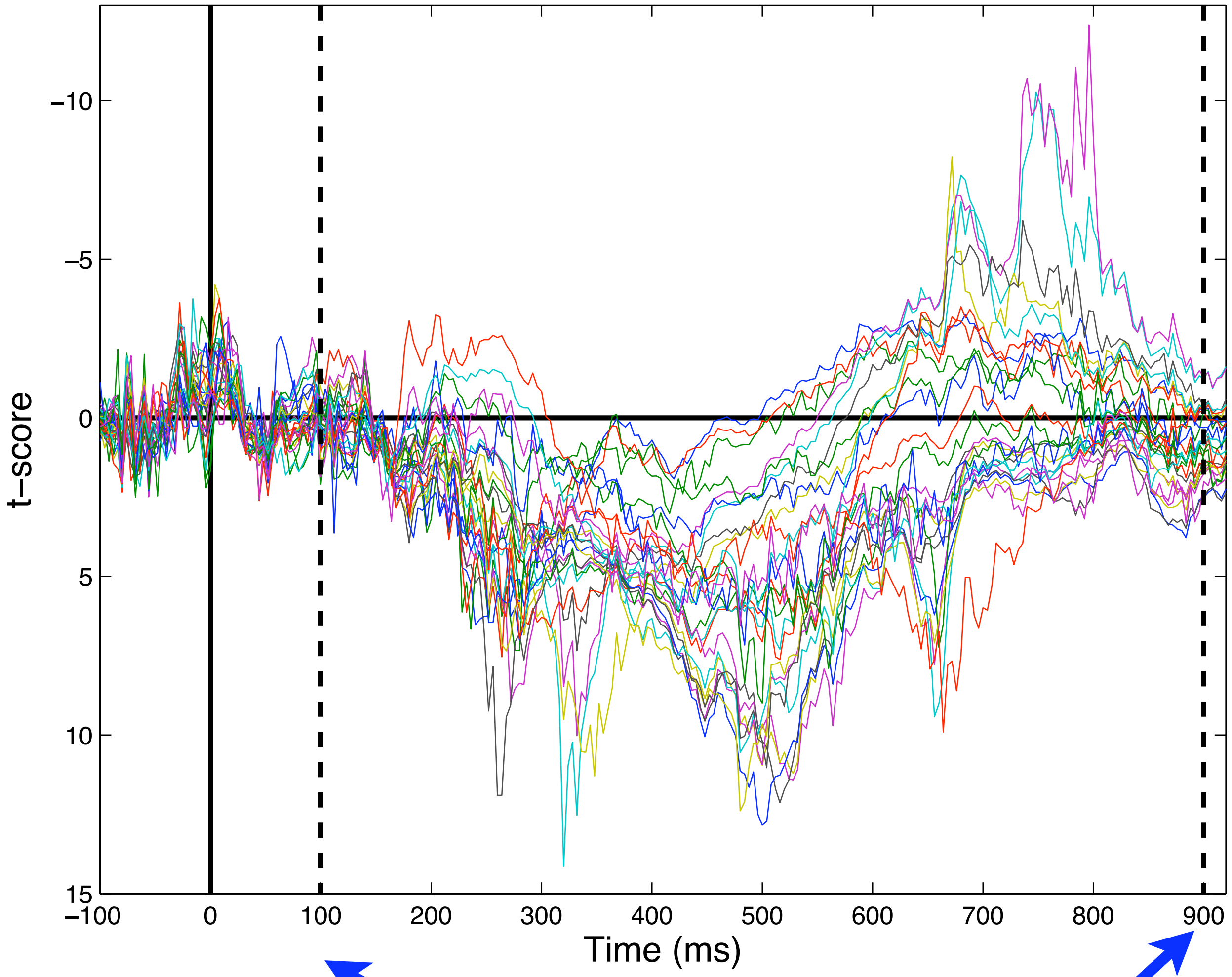
$$\text{False Discovery Proportion} = FDP = \begin{cases} \dfrac{R_F}{R} \text{ if } R > 0 \\ 0 \text{ if } R = 0 \end{cases}$$

$$R = \text{number of rejected null hypotheses}$$

$$FDR = E(FDP) = \alpha$$

If FDR=5%, on average, 5% of your significant *p*-values are mistakes.

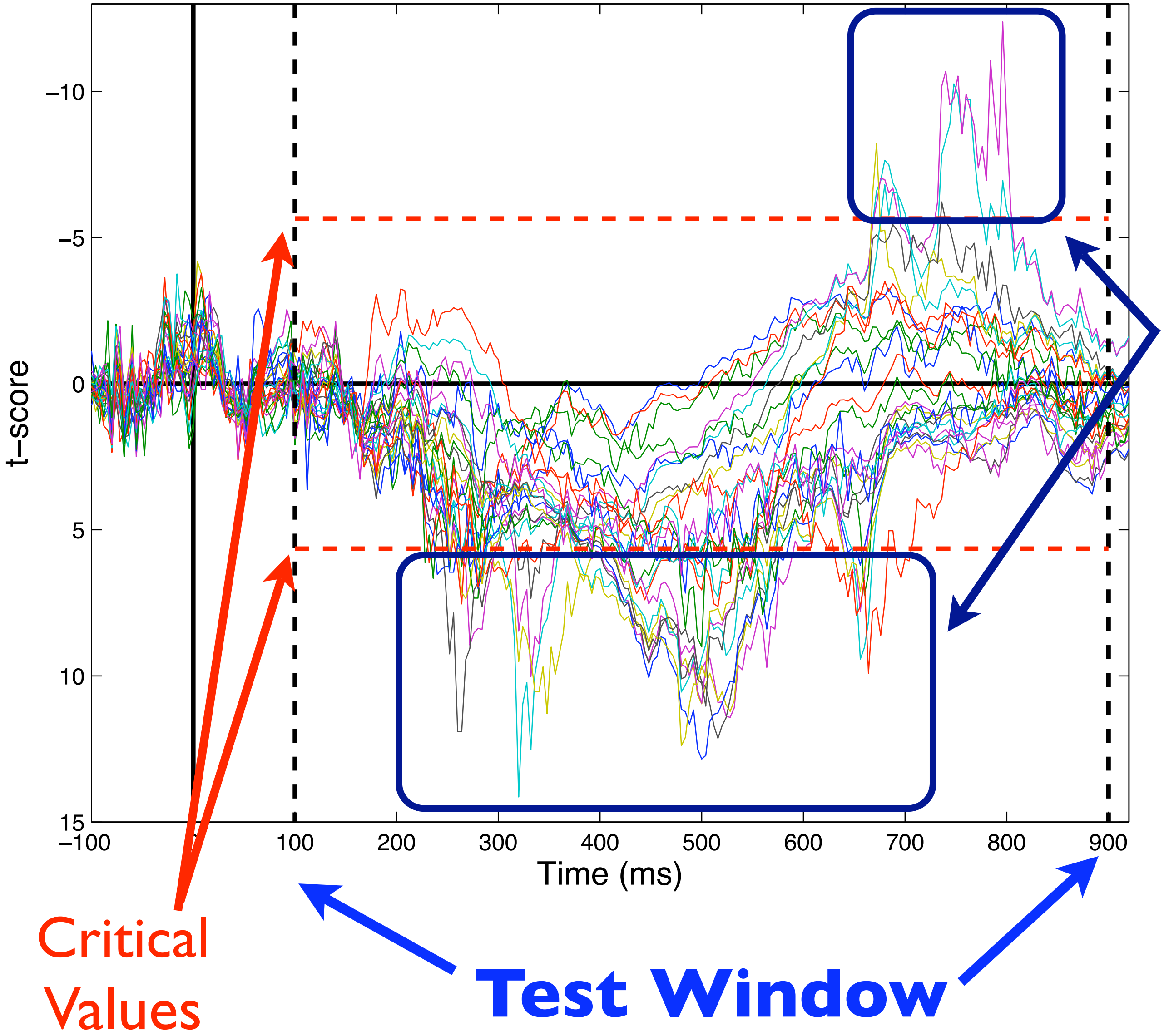
Imagine you replicate an experiment thousands of times



- FDR≤5%

Test Window

Imagine you replicate an experiment thousands of times

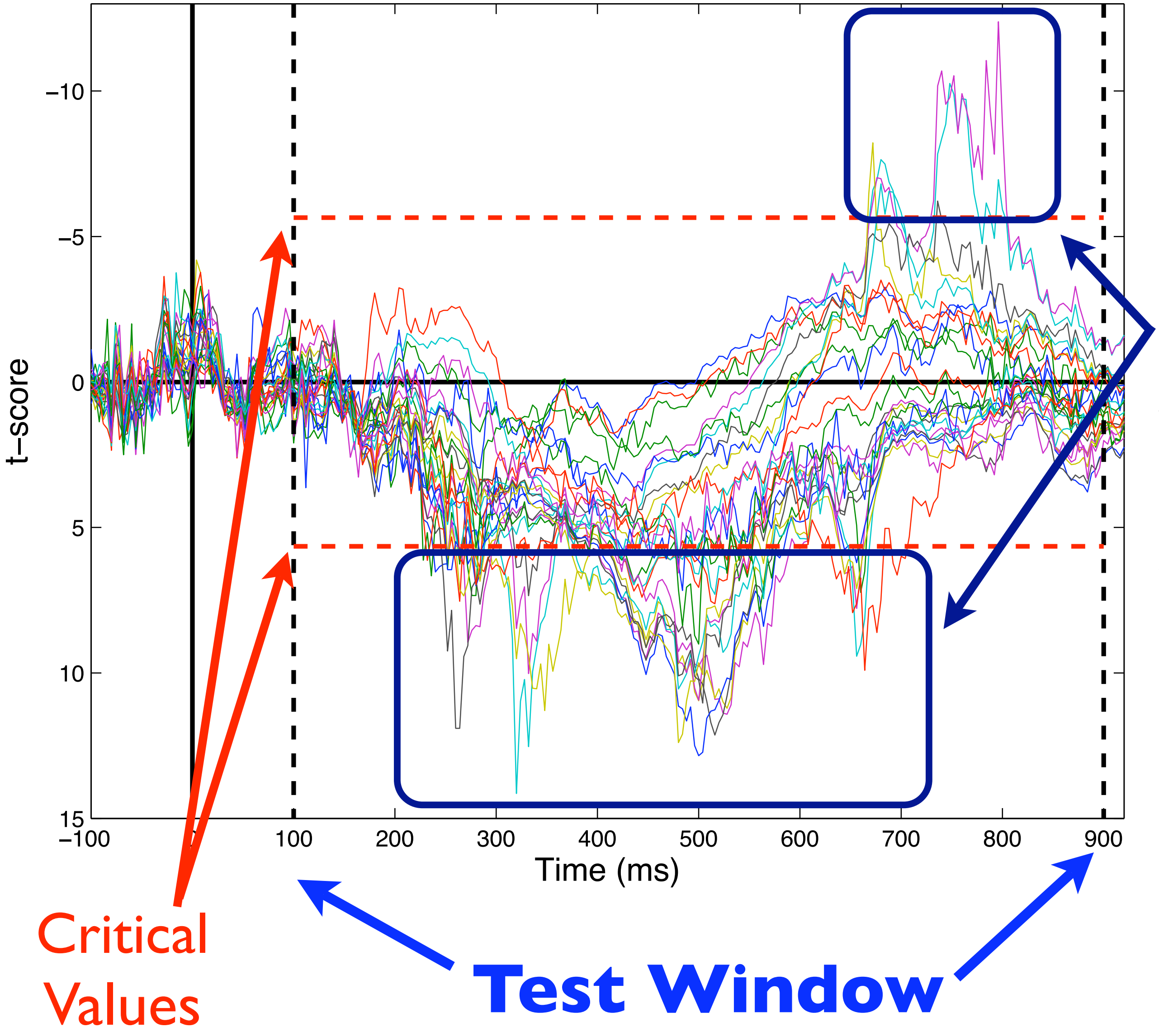


Critical Values

Test Window

- FDR≤5%
- 100 significant *t*-tests, on average

Imagine you replicate an experiment thousands of times

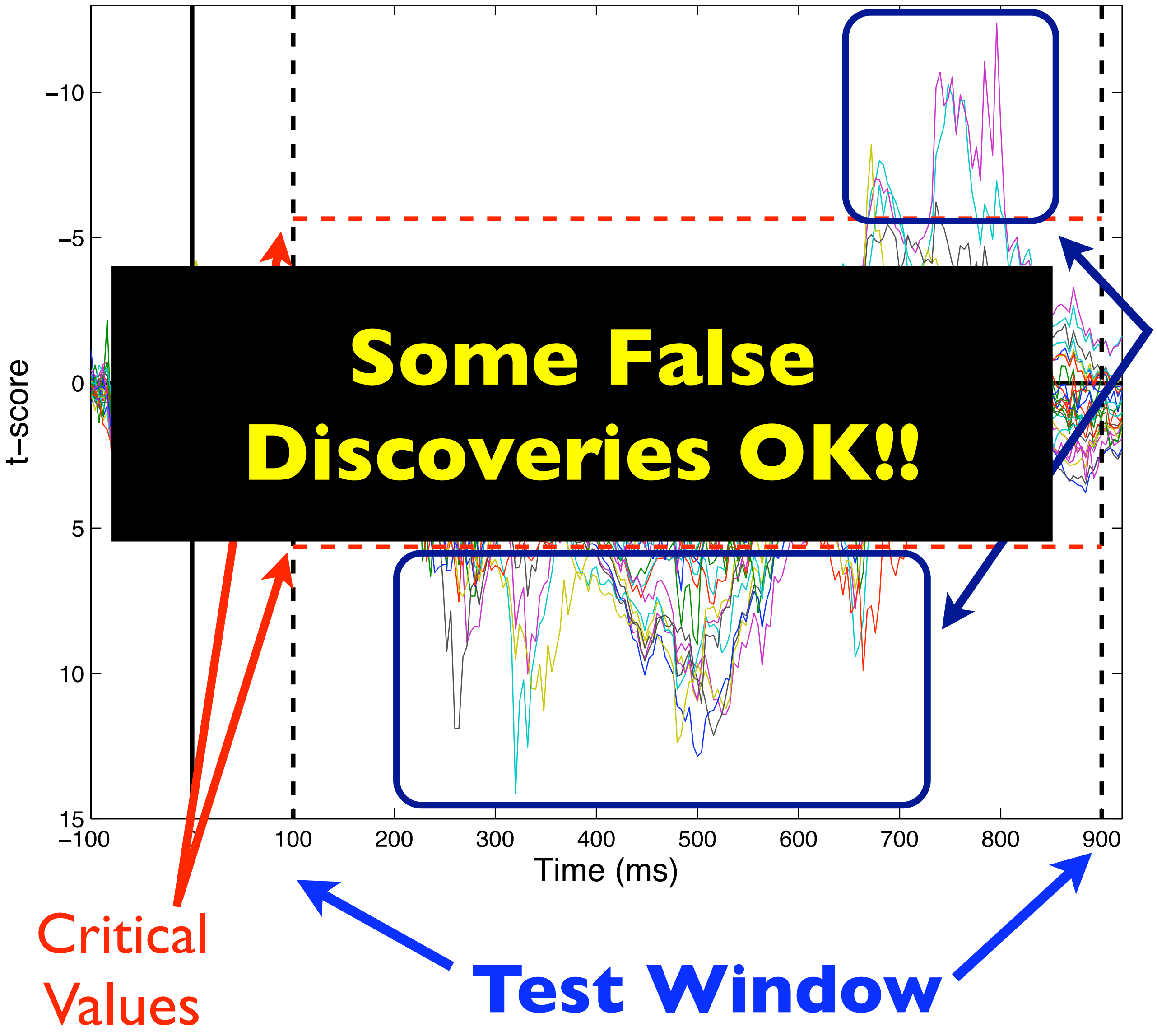


Critical Values

Test Window

- •FDR≤5%
- •100 significant *t*-tests, on average
- •5 *t*-tests or less will be false discoveries, on average

Imagine you replicate an experiment thousands of times



Some False Discoveries OK!!

Critical Values

Test Window

- •FDR≤5%
- •100 significant *t*-tests, on average
- •5 *t*-tests or less will be false discoveries, on average

Most Popular FDR Control Algorithm

Benjamini & Hochberg (1995)

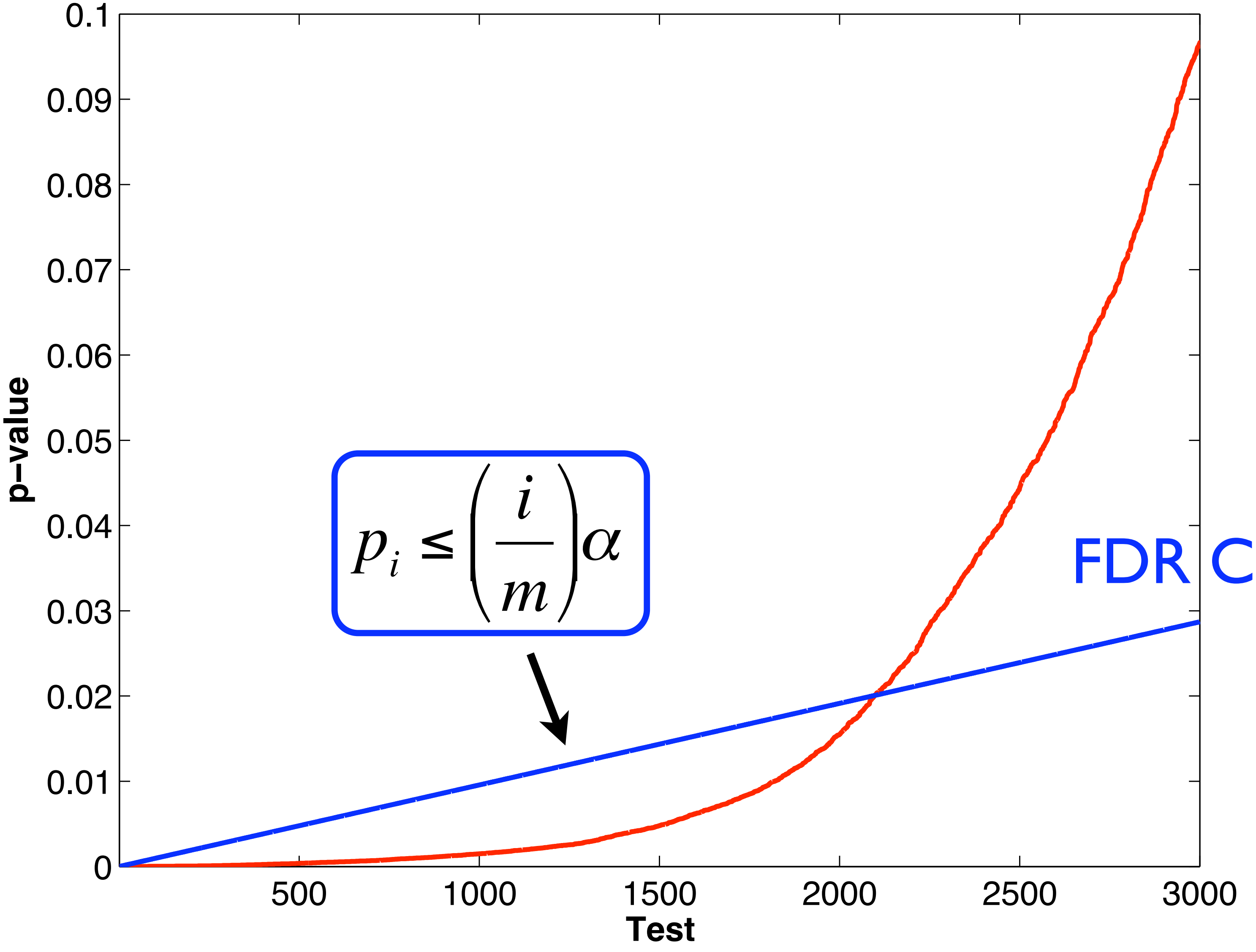
1. Sort the *p*-values from the entire family of *m* tests (i.e., *m* is the total number of hypothesis tests) in order of smallest to largest. p_i refers to the *i*th largest *p*-value.

2. Define *k*, as the largest value of *i* for which the following is true:

$$p_i \leq \left(\frac{i}{m}\right)\alpha$$

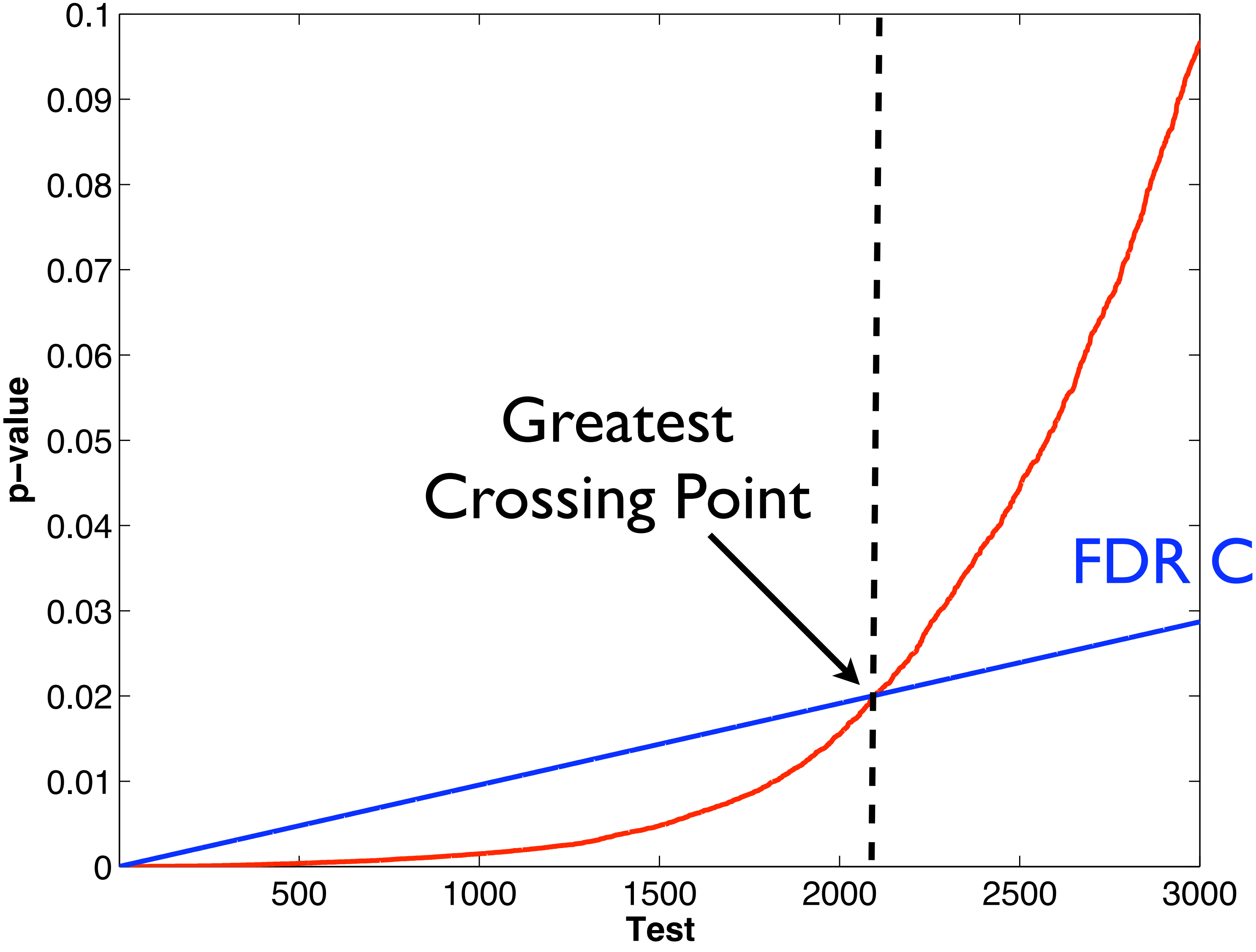
3. If at least one value of *i* satisfies that relationship, then hypotheses 1 though *k* are rejected. If not, no hypotheses are rejected.

Sorted *p*-Values



$$p_i \le \left(\frac{i}{m}\right)\alpha$$

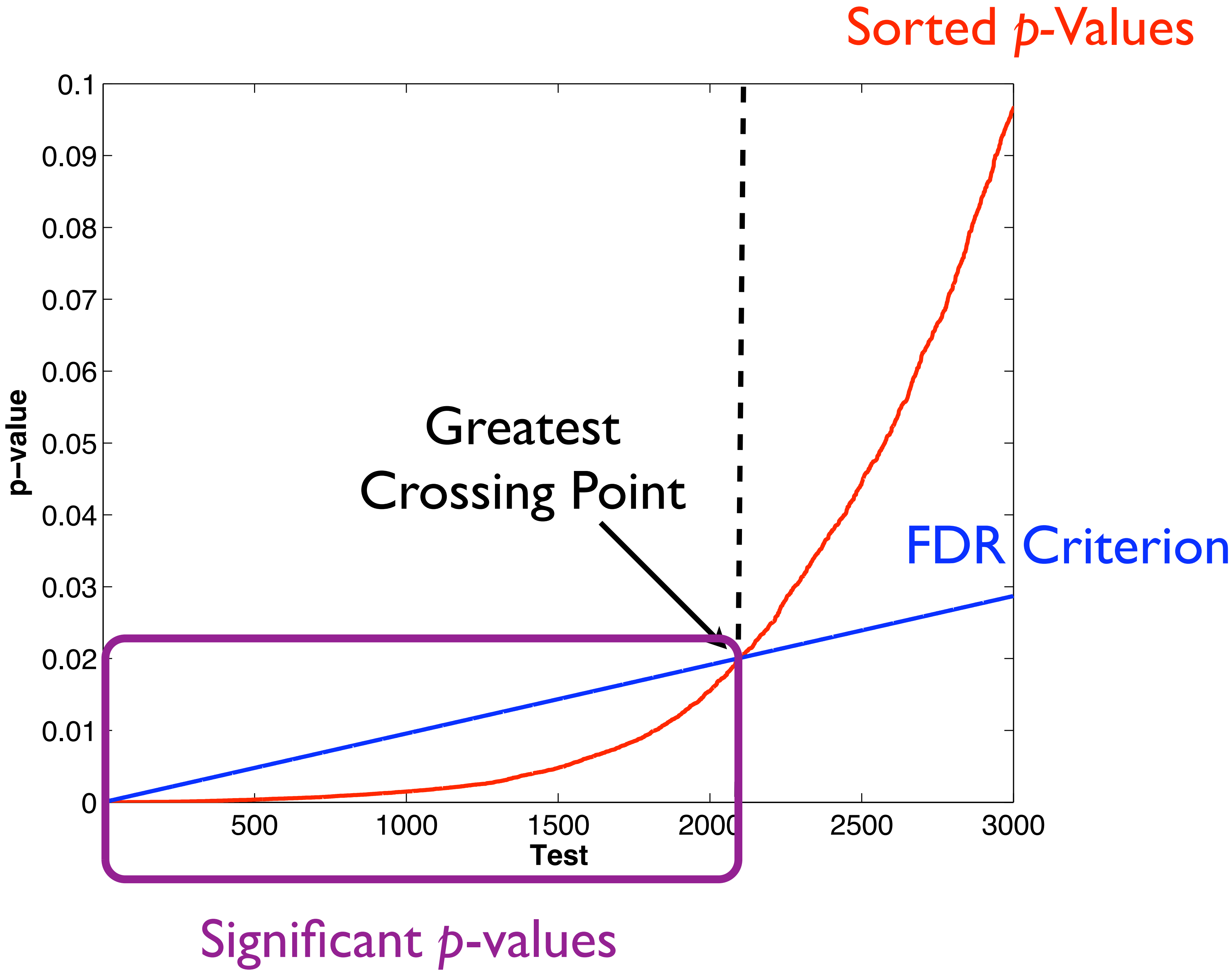
FDR Criterion



Sorted *p*-Values

Greatest Crossing Point

FDR Criterion



Sorted *p*-Values

Greatest Crossing Point

FDR Criterion

Significant *p*-values

Most Popular FDR Control Algorithm

Benjamini & Hochberg (1995)

1. If the dependent variables are independent or exhibit positive regression dependency, the BH algorithm guarantees:

$$FDR \leq \left(\frac{m_0}{m}\right)\alpha$$

where m_0 equals the number of null hypotheses that are true and m equals the total number of null hypotheses.

2. If the dependent variables are Gaussian, then positive regression dependency means that none of the variables are negatively correlated.

Benjamini & Yekutieli (2001) *The Annals of Statistics*

Most Popular FDR Control Algorithm

Benjamini & Hochberg (1995)

Problem

1 If the dependent variables are independent or exhibit positive regression dependency, the BH algorithm guarantees:

$$FDR \leq \left(\frac{m_0}{m}\right)\alpha$$

where m_0 equals the number of null hypotheses that are true and m equals the total number of null hypotheses.

2. If the dependent variables are Gaussian, then positive regression dependency means that none of the variables are negatively correlated.

Benjamini & Yekutieli (2001) *The Annals of Statistics*

More General Variant of BH FDR Control Algorithm

Benjamini & Yekutieli (2001)

1. Sort the *p*-values from the entire family of *m* tests (i.e., *m* is the total number of hypothesis tests) in order of smallest to largest. p_i refers to the *i*th largest *p*-value.

2. Define *k*, as the largest value of *i* for which the following is true:

New BY Criterion → $p_i \leq \left(\dfrac{i}{m \sum_{j=1}^{m} \frac{1}{j}} \right) \alpha$

Original BH Criterion → $p_i \leq \left(\dfrac{i}{m} \right) \alpha$

3. If at least one value of *i* satisfies that relationship, then hypotheses 1 though *k* are rejected. If not, no hypotheses are rejected.

More General Variant of BH FDR Control Algorithm

Benjamini & Yekutieli (2001)

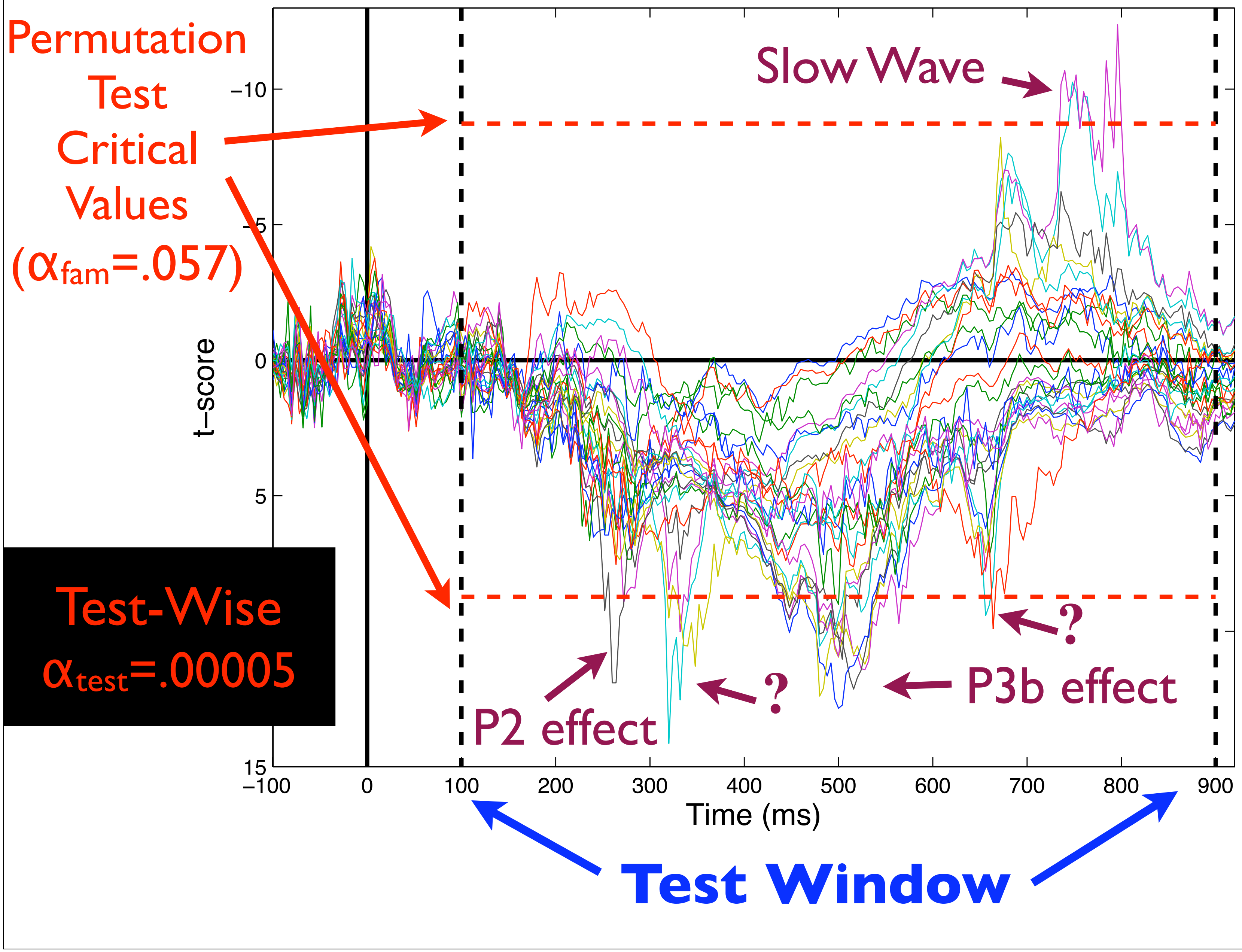
1. Regardless of dependent variable dependency structure, BY algorithm guarantees:

$$FDR \leq \left(\frac{m_0}{m}\right)\alpha$$

where m_0 equals the number of null hypotheses that are true and m equals the total number of null hypotheses.

Benjamini & Yekutieli (2001) *The Annals of Statistics*

Target-Standard Difference Wave (26 electrodes)



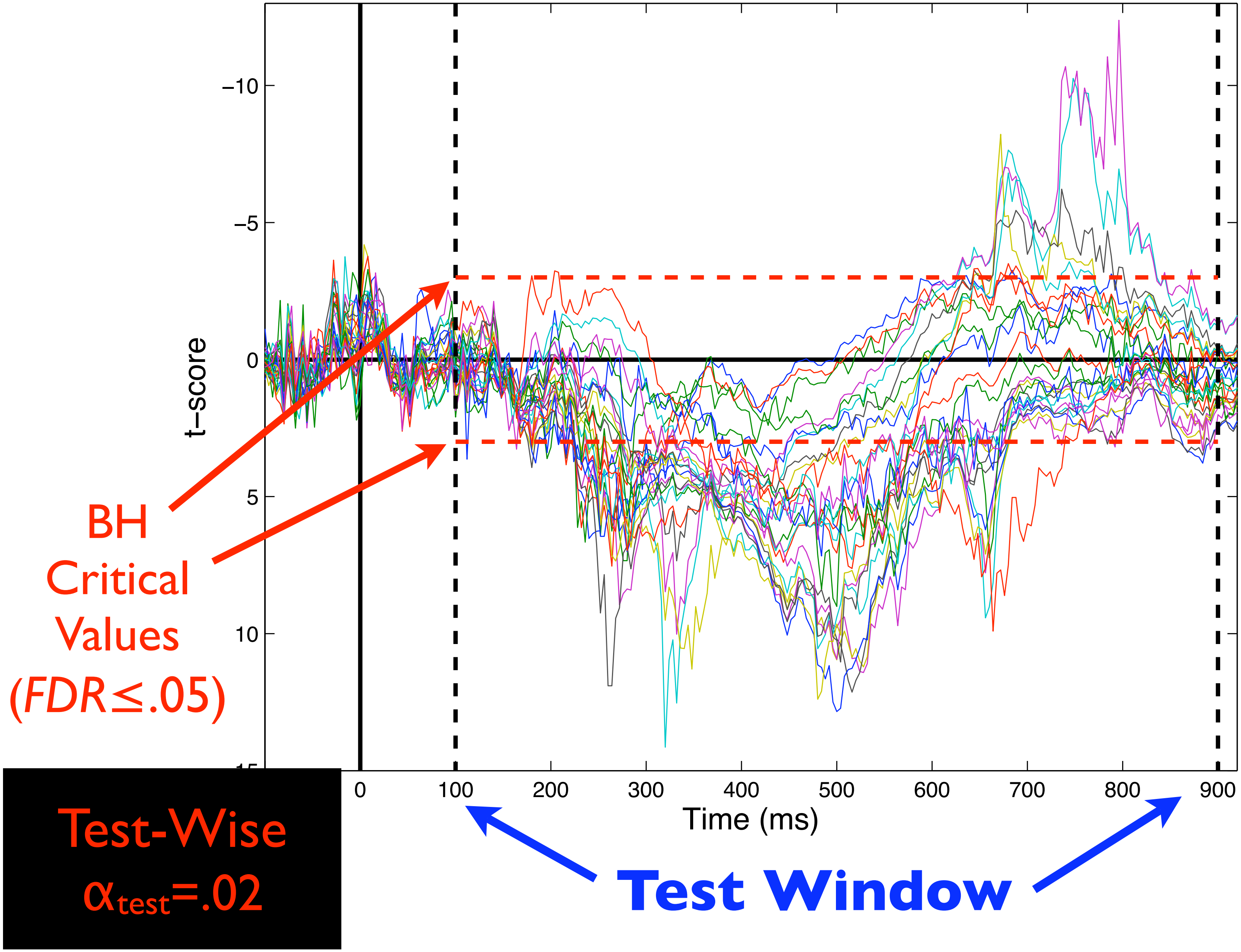
Permutation Test Critical Values (α_{fam}=.057)

Test-Wise

α_{test}=.00005

Test Window

Target-Standard Difference Wave (26 electrodes)

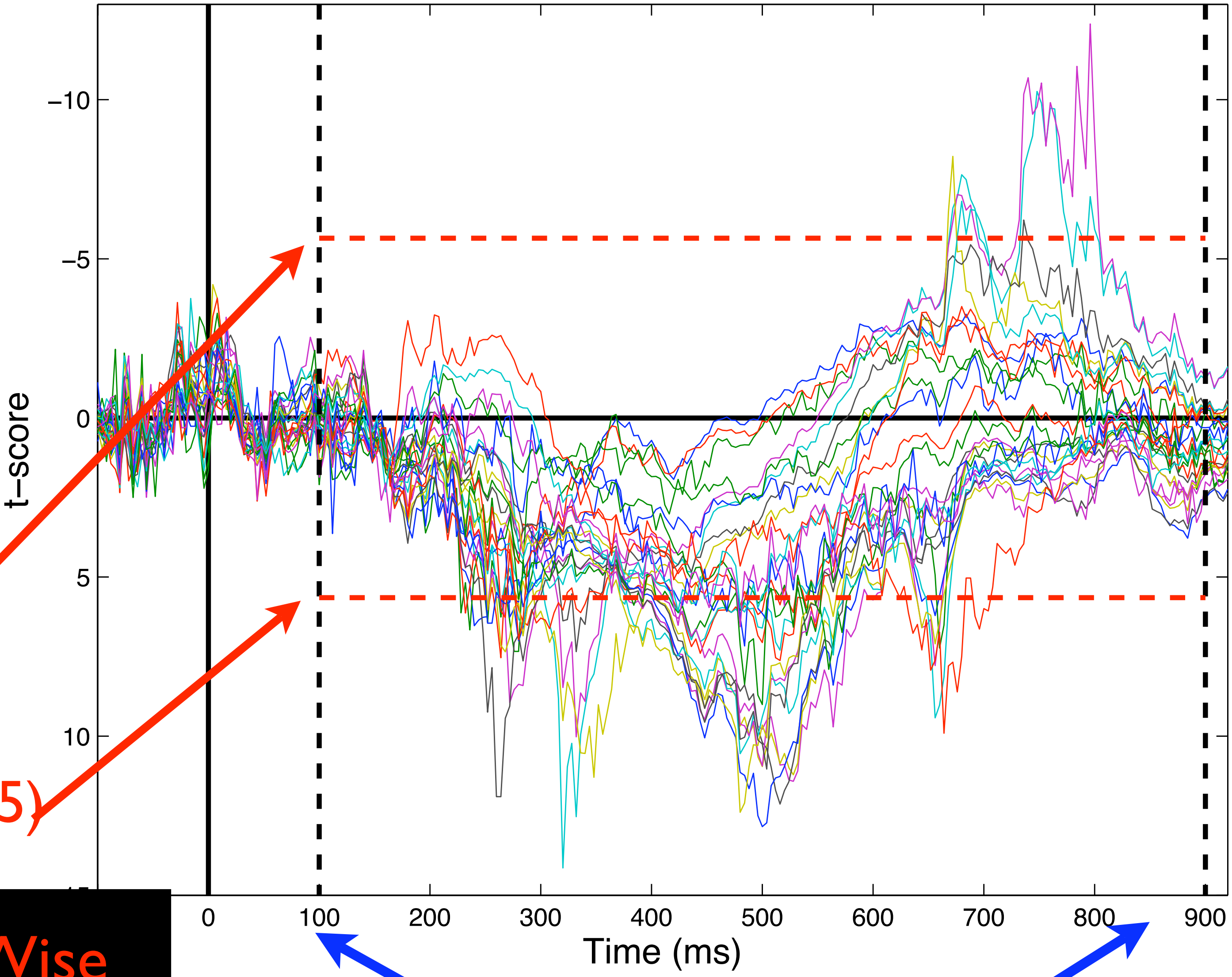


BH Critical Values (*FDR*≤.05)

Test-Wise α_{test}=.02

Test Window

Target-Standard Difference Wave (26 electrodes)

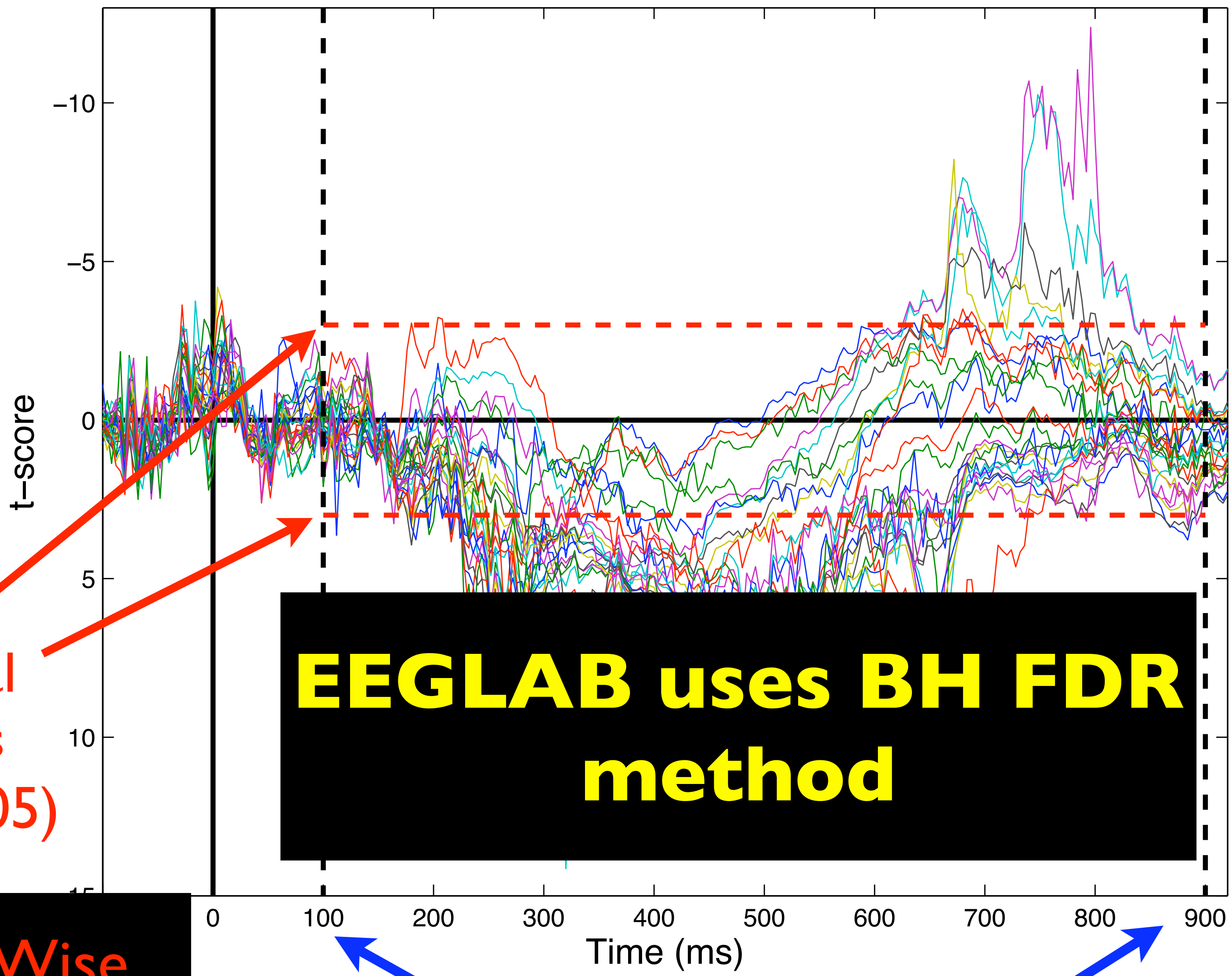


BY Critical Values (*FDR*≤.05)

Test-Wise α_{test}=.0008

Test Window

Target-Standard Difference Wave (26 electrodes)



BH Critical Values (*FDR*≤.05)

EEGLAB uses BH FDR method

Test-Wise α_{test}=.02

Test Window

FDR Control: Pros

1. With a large number of comparisons, FDR is generally more powerful than FWER control (especially if an appreciable proportion of null hypotheses are false).

2. If all null hypotheses are true, FDR control=FWER control. Thus, if you find effects with FDR control you can be 1-α confident that some effect is present.

3. Benjamini procedures can be used with any hypothesis test (simply requires test *p*-values).

FDR Control: Cons

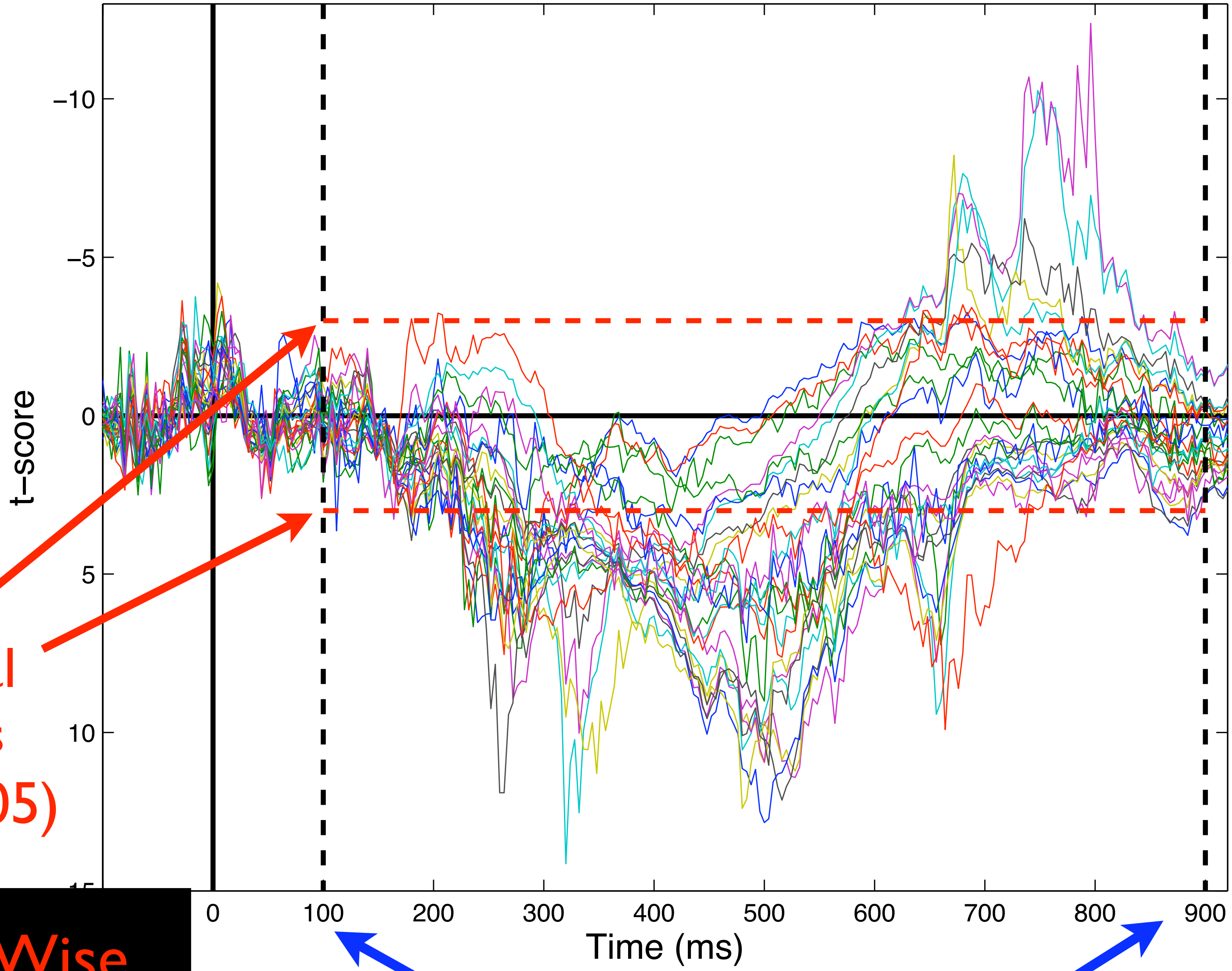
1. FDR control may lead to a high proportion of false positives with some frequency

When applied to simulated data and an α-level of 10%, Korn et al. (2004) found that the BH algorithm produces 29% or more false discoveries 10% of the time.

FDR Control: Cons

1. FDR control may lead to a high proportion of false positives with some frequency
2. FDR can be difficult to interpret as effects may disappear when analyses become more selective

Target-Standard Difference Wave (26 electrodes)

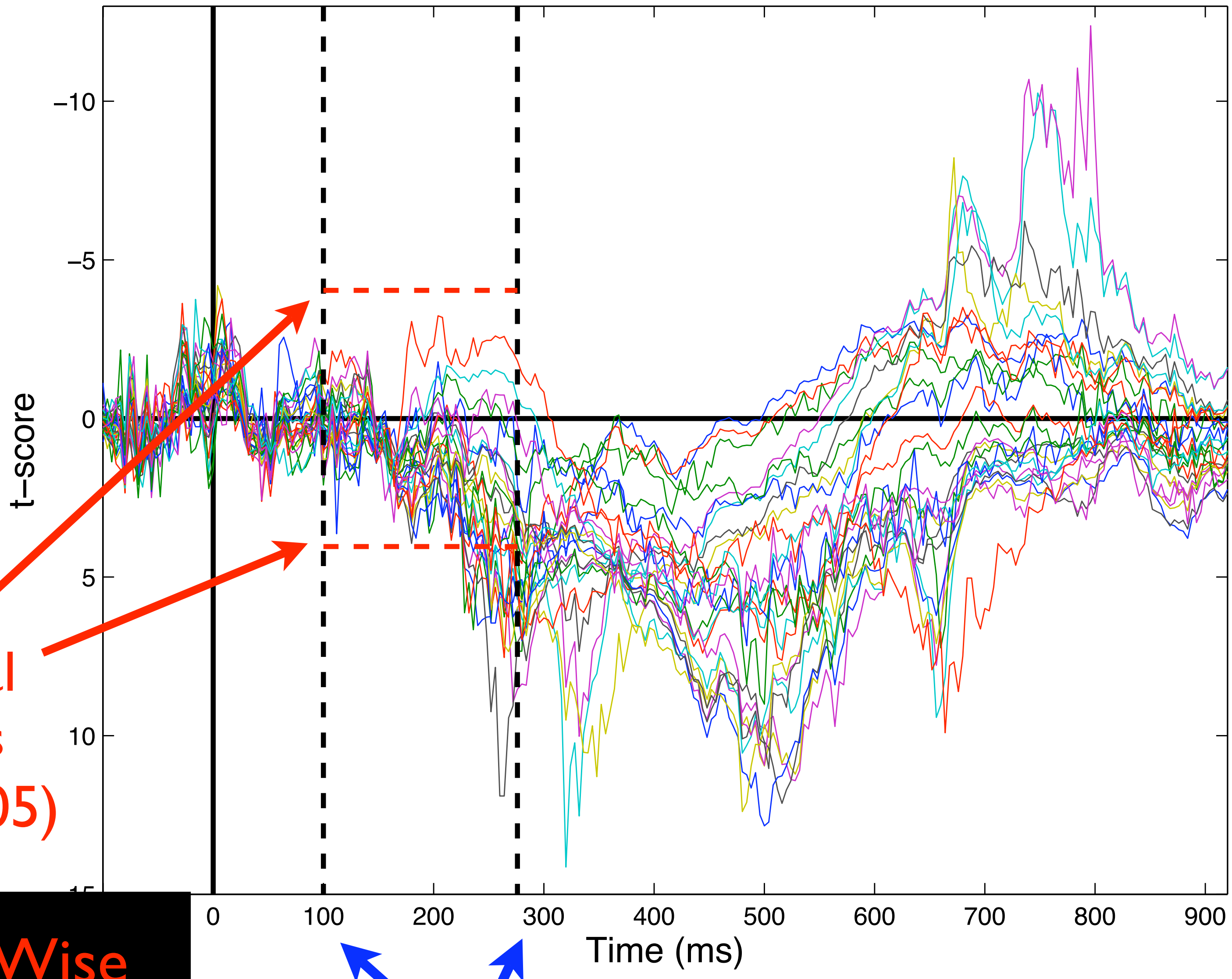


BH Critical Values (*FDR*≤.05)

Test-Wise α_{test}=.02

Test Window

Target-Standard Difference Wave (26 electrodes)



BH Critical Values (*FDR*≤.05)

Test-Wise α_{test}=.005

Test Window

FDR Control: Cons

1. FDR control may lead to a high proportion of false positives with some frequency
2. FDR can be difficult to interpret as effects may disappear when analyses become more selective
3. More powerful and popular FDR control algorithm (BH) is not guaranteed to work for data with negatively correlated variables

FDR Control: Cons

1. FDR control may lead to a high proportion of false positives with some frequency

2. FDR can be difficult to interpret as effects may disappear when analyses become more selective

3. More powerful and popular FDR control algorithm (BH) is not guaranteed to work for data with negatively correlated variables

   - However, recent work by Clarke & Hall (2009) shows that for light tailed data (e.g., Gaussian) multiple comparison correction procedures will behave as if the data were independent if the number of variables is large enough

Presentation Outline

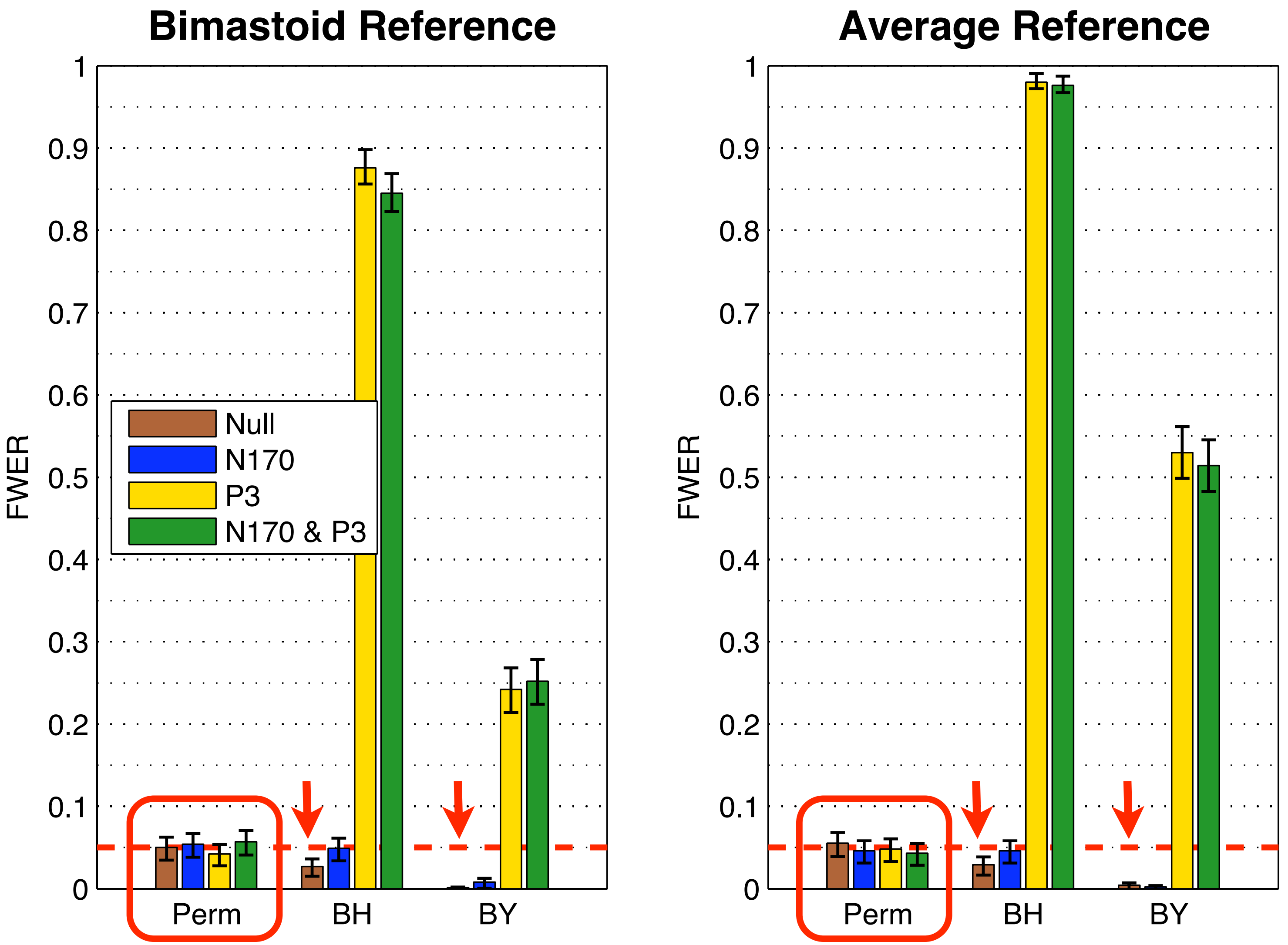
- **"Classic" Analytical Inferential Statistics**
  - Parametric & non-parametric
- **Resampling-Based Inferential Statistics**
  - Randomization/permutation tests
  - Bootstrap statistics
- **Correcting for Multiple Comparisons**
  - Permutation test based control of family-wise error
  - Benjamini methods for control of false discovery rate
  - Evaluating multiple comparison correction on simulated ERP data

ERP Simulations

- **Simulation Parameters**
  - Simulated ERP noise estimated from ERP noise in a real ERP study
  - 26 electrodes, 201 time points (100-900 ms)
  - Average & bimastoid reference
  - Negatively correlated dependent variables ranged from 13-51%
- **ERP Effects**
  1. *Null effect:* 0% of comparisons differ from 0
  2. *Focal effect ("N170"):* 0.2% of comparisons differ from 0
  3. *Broad effect ("P300"):* 18.9% of comparisons differ from 0
  4. *Combined focal & broad effect:* 19.1% of comparisons differ from 0

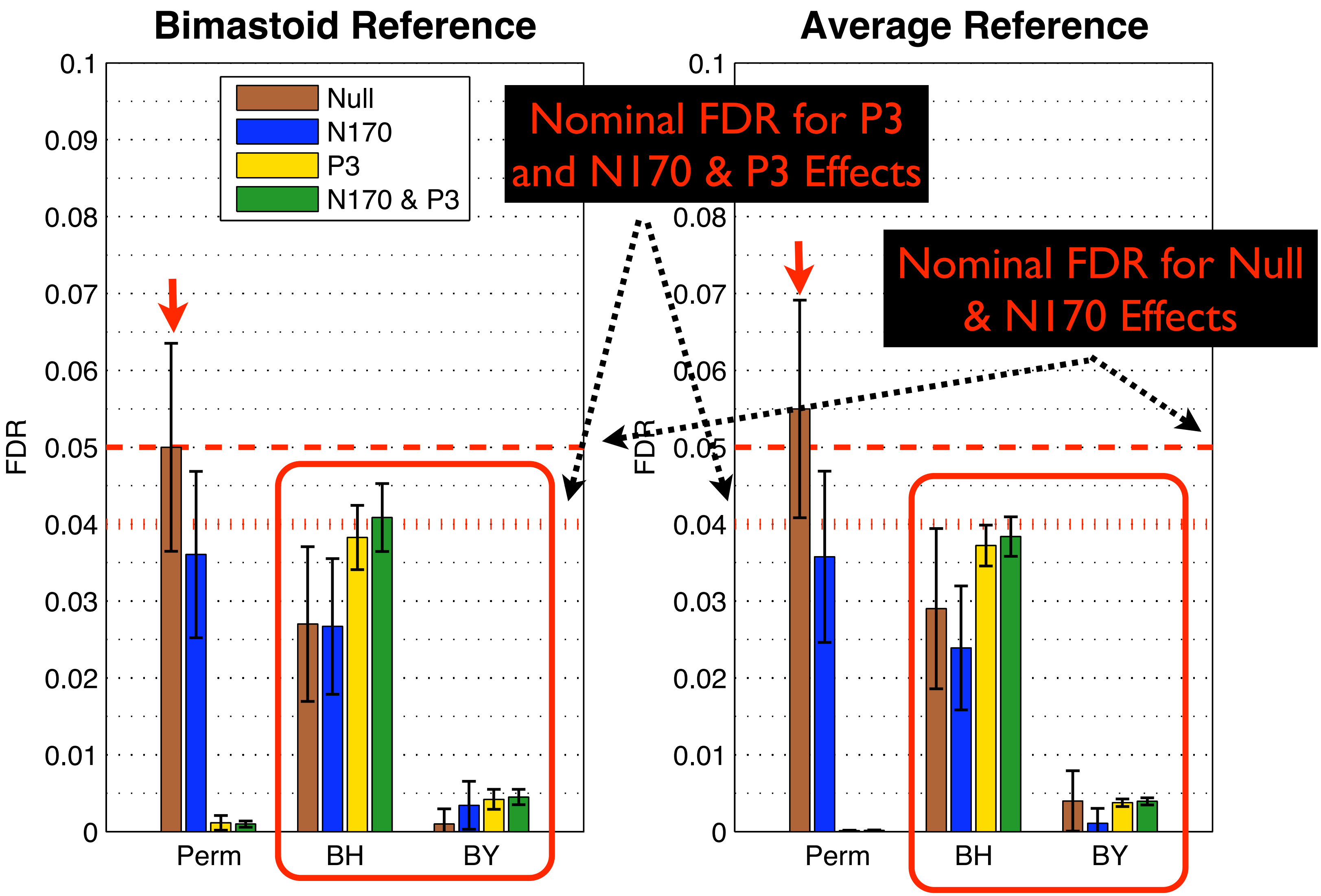
Groppe, Urbach, & Kutas (*in prep*)

Family Wise Error Rate



Perm=t_{max} permutation test FWER control; *BH*=Benjaminin & Hochberg FDR control; *BY*=Benjamini & Yekutieli FDR control

False Discovery Rate



Bimastoid Reference

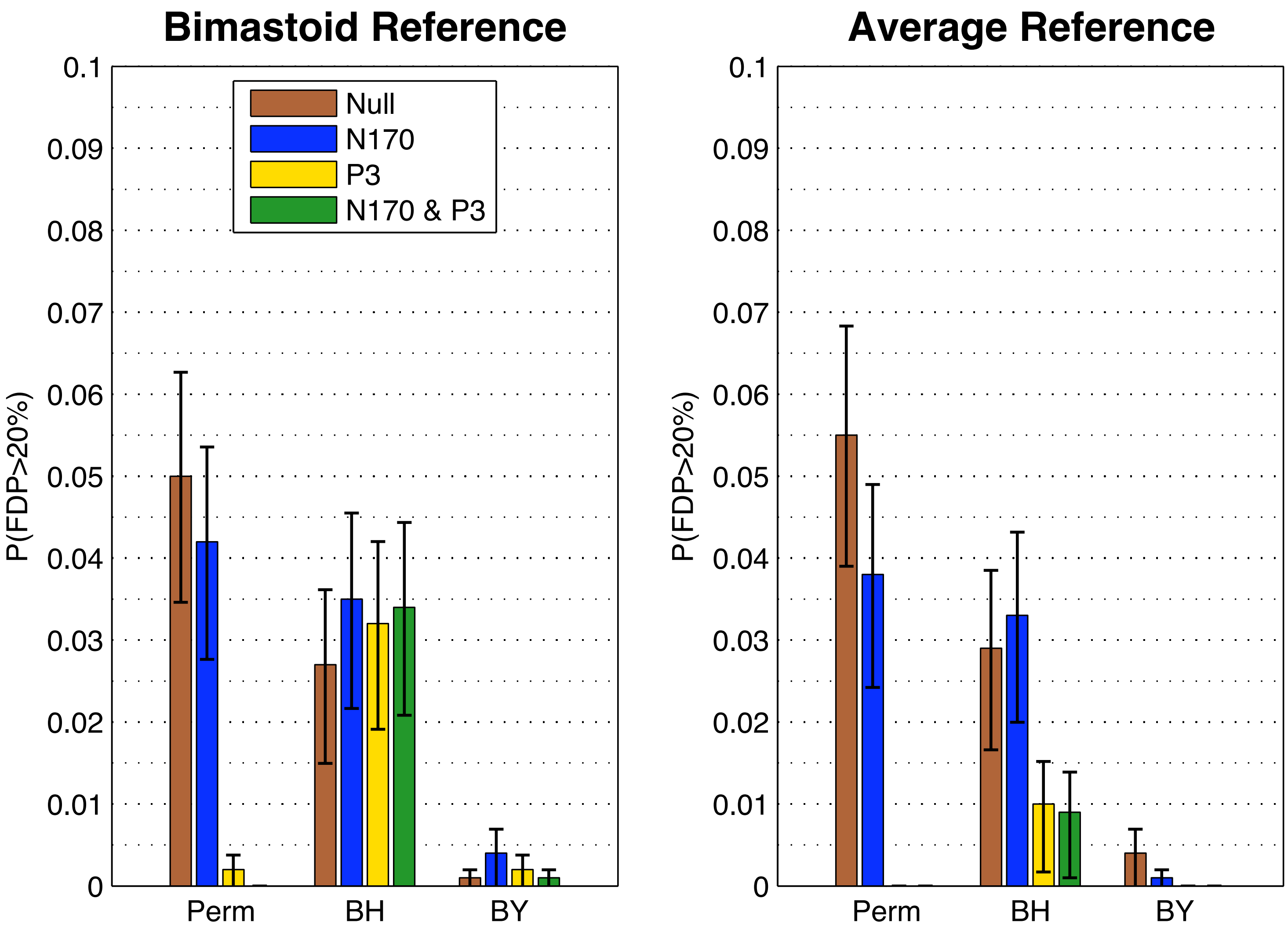
Average Reference

Nominal FDR for P3 and N170 & P3 Effects

Nominal FDR for Null & N170 Effects

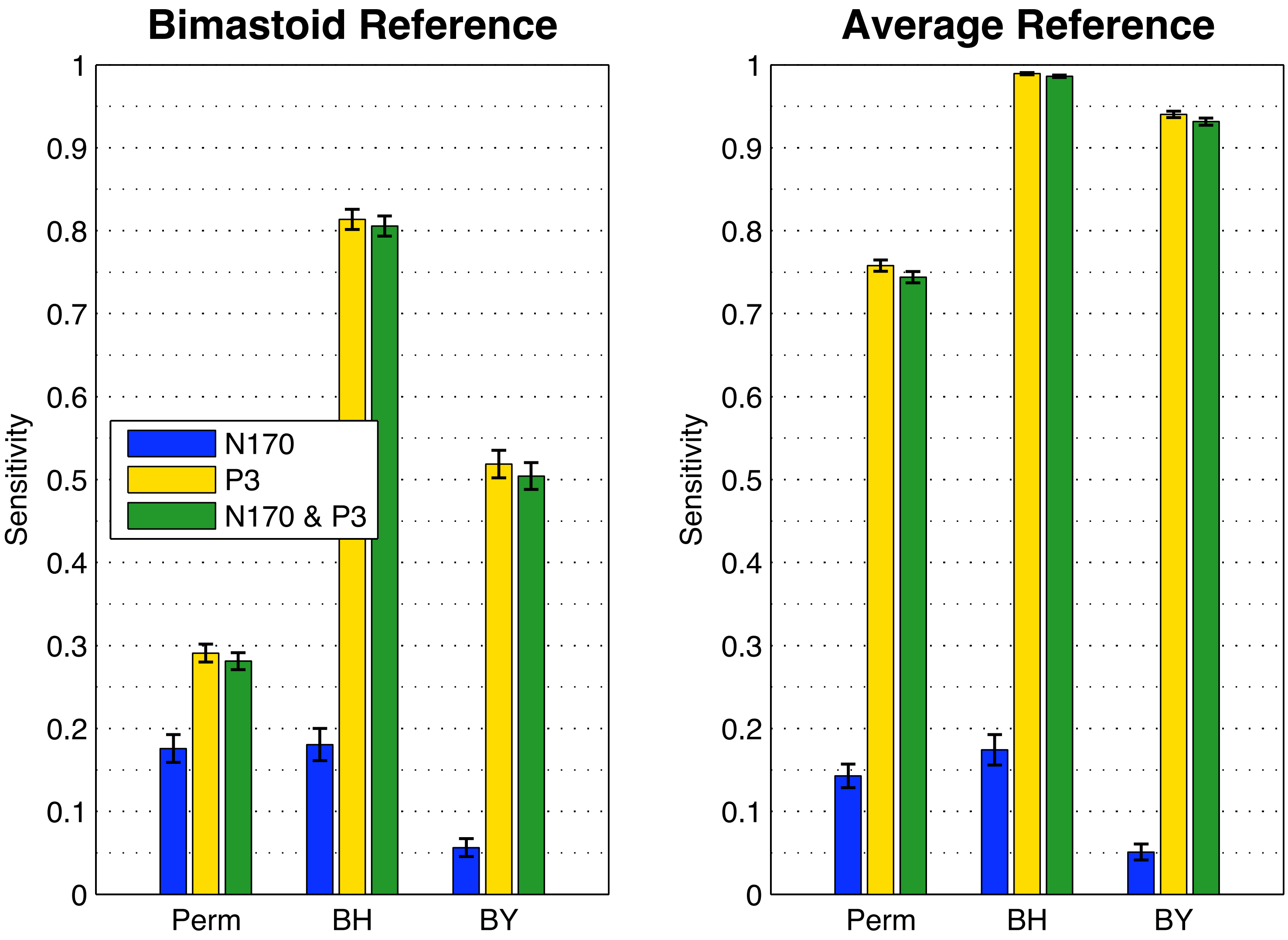
Perm=$t_{\max}$ permutation test FWER control; *BH*=Benjaminin & Hochberg FDR control; *BY*=Benjamini & Yekutieli FDR control

Probability of 20% or More False Discovery Proportion



Perm=$t_{\max}$ permutation test FWER control; *BH*=Benjaminin & Hochberg FDR control; *BY*=Benjamini & Yekutieli FDR control

Mean Proportion of Effects Detected



Perm=$t_{\max}$ permutation test FWER control; *BH*=Benjaminin & Hochberg FDR control; *BY*=Benjamini & Yekutieli FDR control

Presentation Outline

- **"Classic" Analytical Inferential Statistics**
  - Parametric & non-parametric
- **Resampling-Based Inferential Statistics**
  - Randomiz
  - Bootstrap

Summary:

- **Correcting for Multiple Comparisons**
  - Permutation test based control of family-wise error
  - Benjamini methods for control of false discovery rate
  - Evaluating multiple comparison correction on simulated ERP data

Summary

1. **FWER control via permutation tests:**

- **Pros:**
  - Relatively powerful because EEG is highly correlated
  - Same degree of error control as a priori analyses
- **Cons:**
  - May sacrifice considerable power when applied to large numbers of comparisons
  - Only guaranteed to work for simple analyses

Summary

2. **FDR control via BH & BY procedures:**

- **Pros:**

- Relatively powerful because of less conservative error measure
- More general than permutation test procedures and often more powerful

- **Cons:**

- Can be difficult to interpret due to invalid statistical assumptions, potentially high proportions of false discoveries, and interactions between variables
- Simulations found **no** evidence that these FDR procedures are prone to the former two problems when applied to ERPs

Yet More Multiple Comparison Correction Procedures

1. Control of False Discovery Exceedance (FDX)

(also called control of FDP)

$$FDX = P(FDP > c)$$

$$FDP = \begin{cases} \dfrac{R_F}{R} \text{ if } R > 0 \\ 0 \text{ if } R = 0 \end{cases}$$

2. Control of Generalized Family-Wise Error Rate (GFWER)

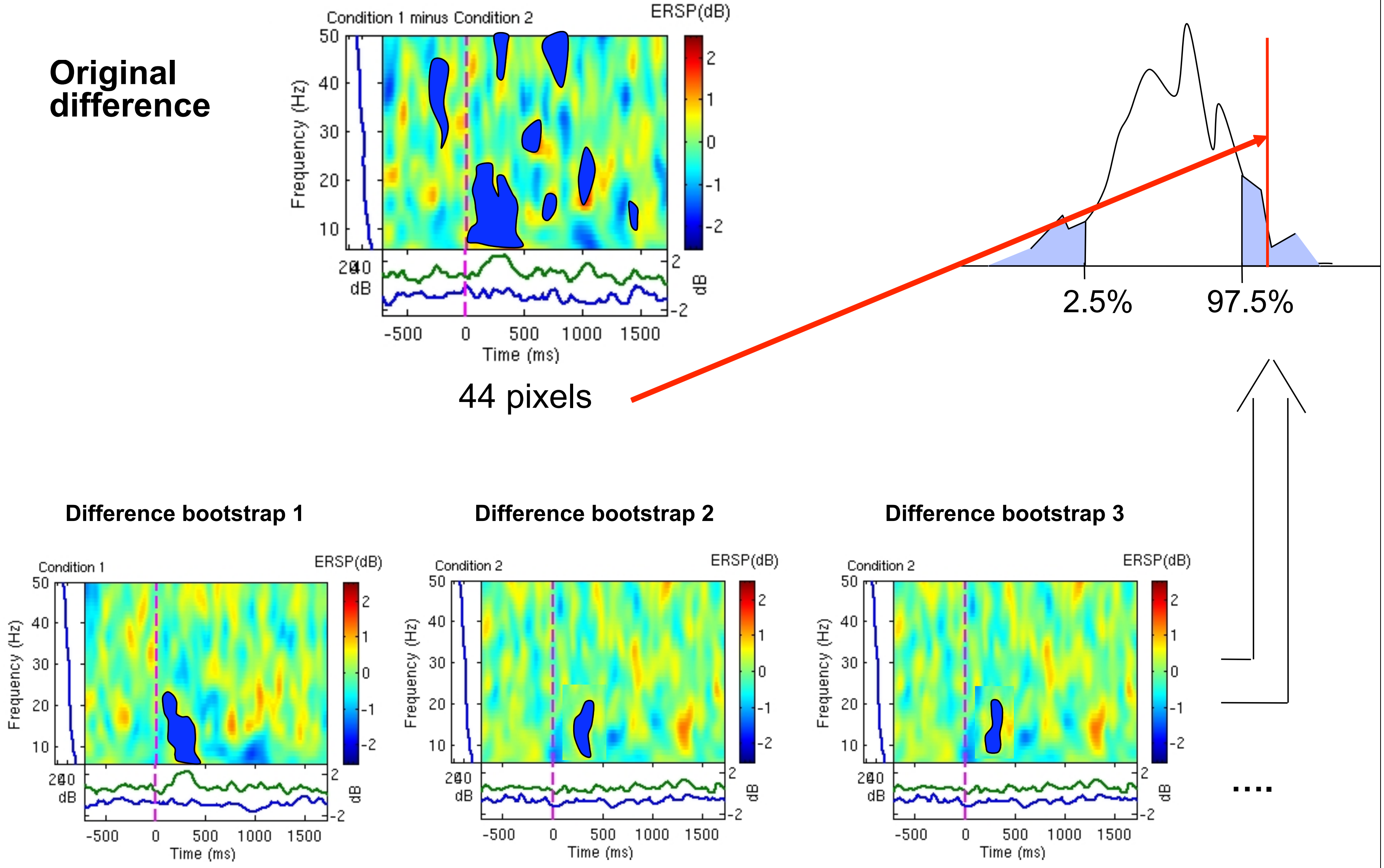
$$GFWER = P(R_F > u)$$

$u =$ an acceptable number of false discoveries

3. Control of Local False Discovery Rate:

Bootstrap based control of FDR (Efron, 2004)

Cluster correction for multiple comparisons



Original difference

44 pixels

2.5%

97.5%

Difference bootstrap 1

Difference bootstrap 2

Difference bootstrap 3

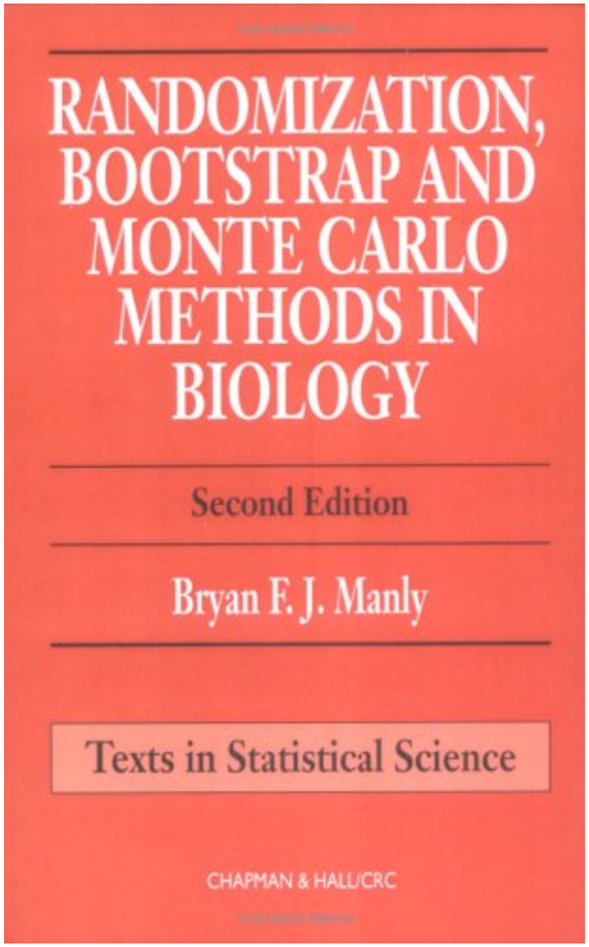
....

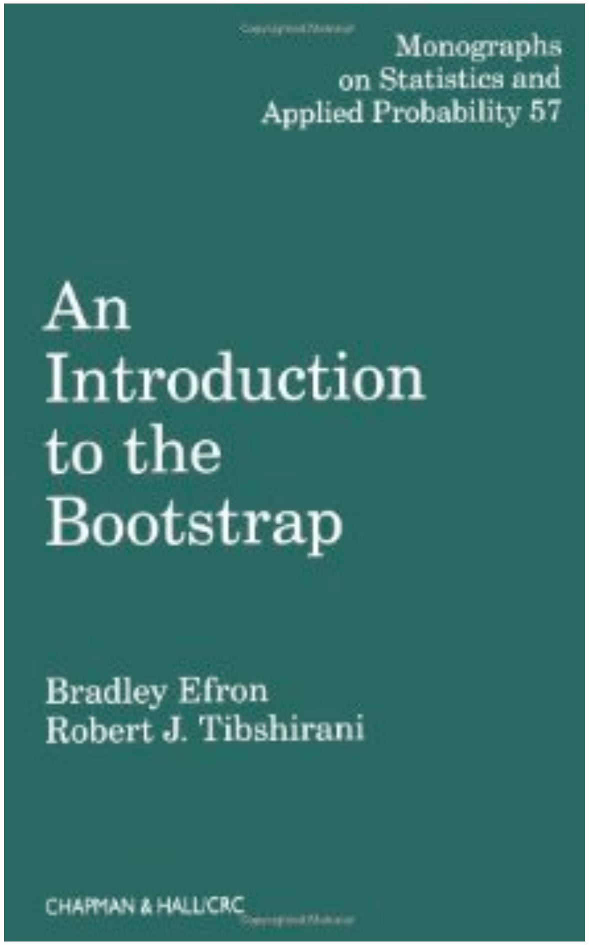
Maris & Oostenveld (2007) *Jnl of Neuro Methods*

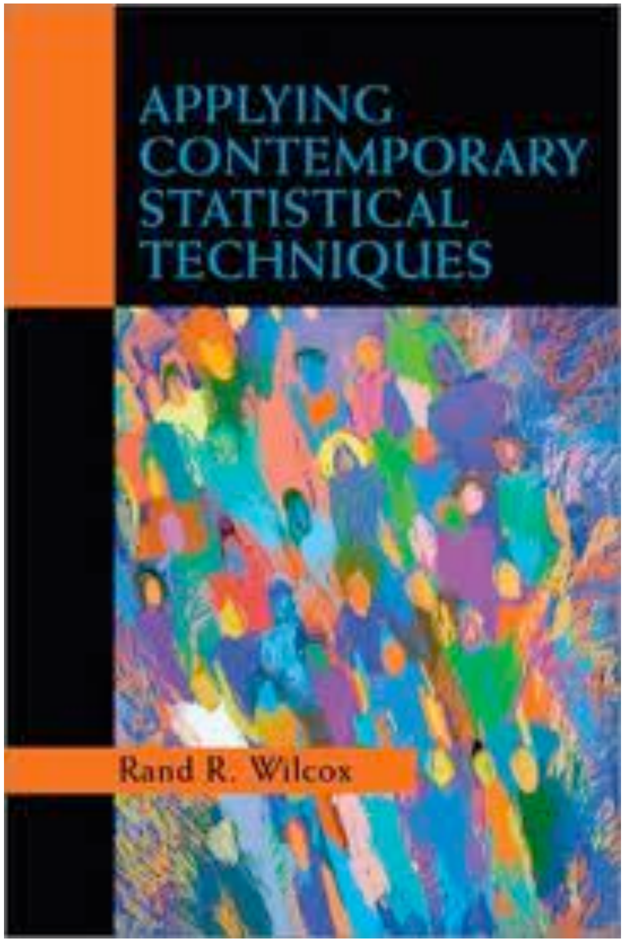
Presentation Outline

- **“Classic” Analytical Inferential Statistics**
  - Parametric & non-parametric
- **Resampling-Based Inferential Statistics**
  - Randomization/permutation tests
  - Bootstrap statistics
- **Correcting for Multiple Comparisons**
  - Permutation test based control of family-wise error
  - Benjamini methods for control of false discovery rate
  - Evaluating multiple comparison correction on simulated ERP data

Recommended Textbooks


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



Recommended Papers

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

Groppe, D.M., Urbach, T.P., Kutas, M. (in prep) Mass univariate analysis of event-related potentials.

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

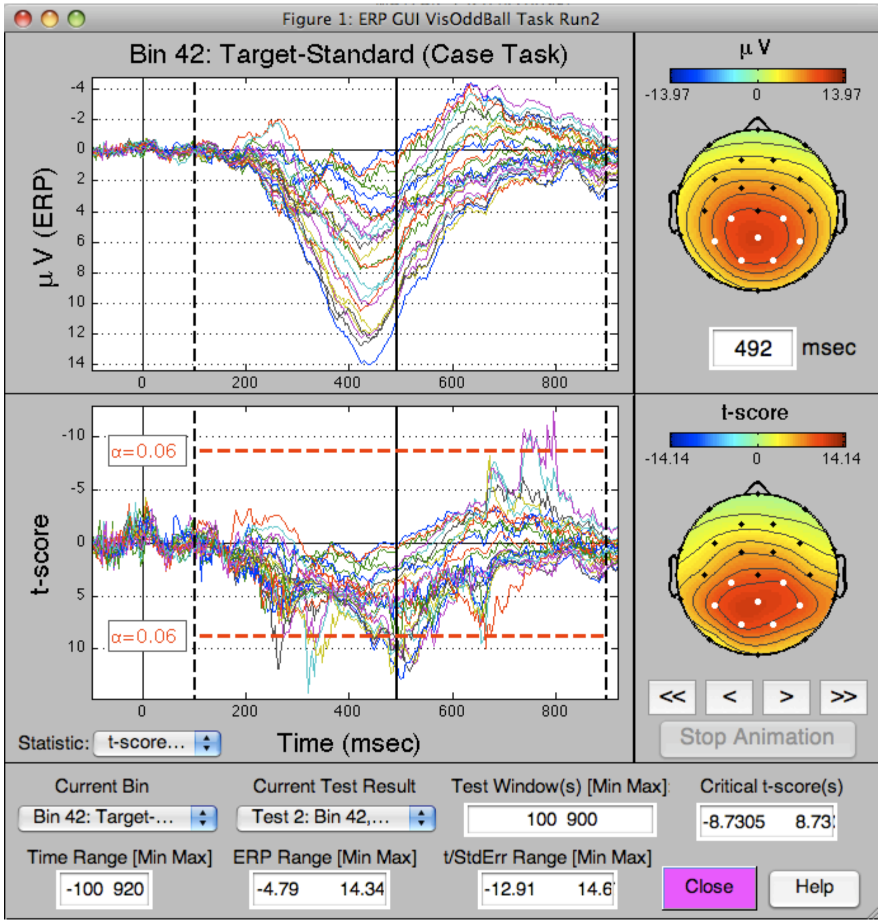
Maris, 2004. Randomization tests for ERP topographies and whole spatiotemporal data matrices. *Psychophysiology*, 41: 142-151

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

Thanks to G. Rousselet

Thanks!

EEGLAB Compatible Software
for ERP Analysis



492 msec

Statistic: t-score...

Current Bin: Bin 42: Target-...

Current Test Result: Test 2: Bin 42,...

Test Window(s) [Min Max]: 100 900

Critical t-score(s): -8.7305 8.73

Time Range [Min Max]: -100 920

ERP Range [Min Max]: -4.79 14.34

t/StdErr Range [Min Max]: -12.91 14.6

Questions:

dgroppe@cogsci.ucsd.edu

http://openwetware.org/wiki/Mass_Univariate_ERP_Toolbox

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

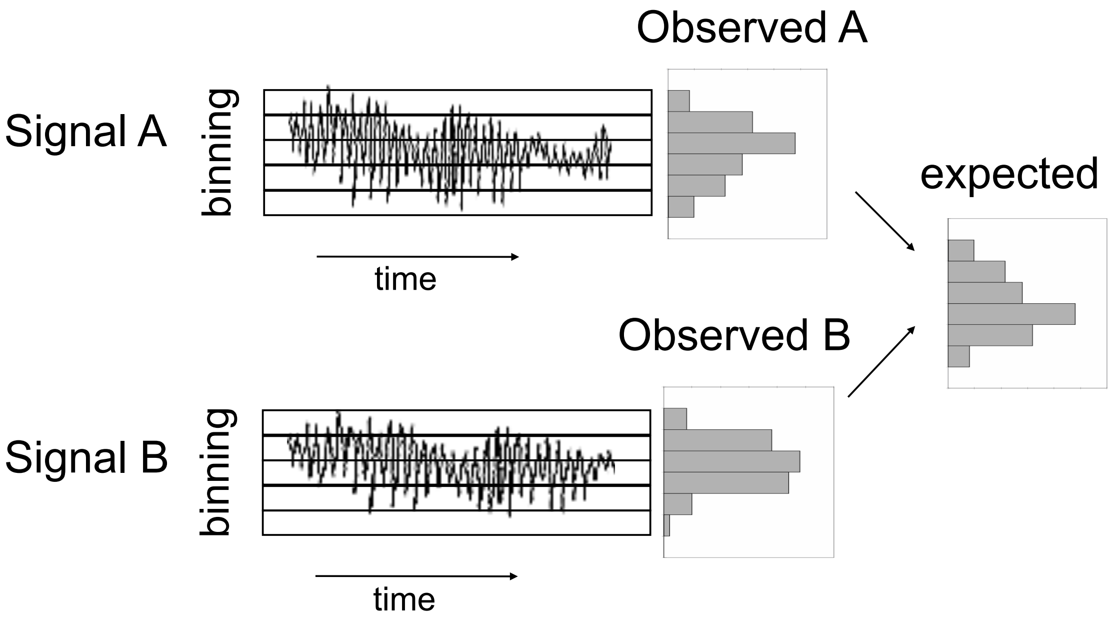
```
a = { rand(3,4,10) rand(3,4,10) rand(3,4,10); ...
      rand(3,4,10) rand(3,4,10) rand(3,4,10)+0.5 };

% 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
```

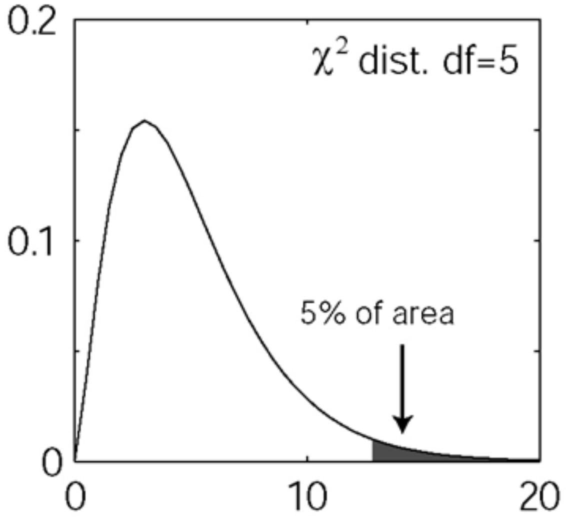
Non-parametric statistics

Do not assume a distribution for the data

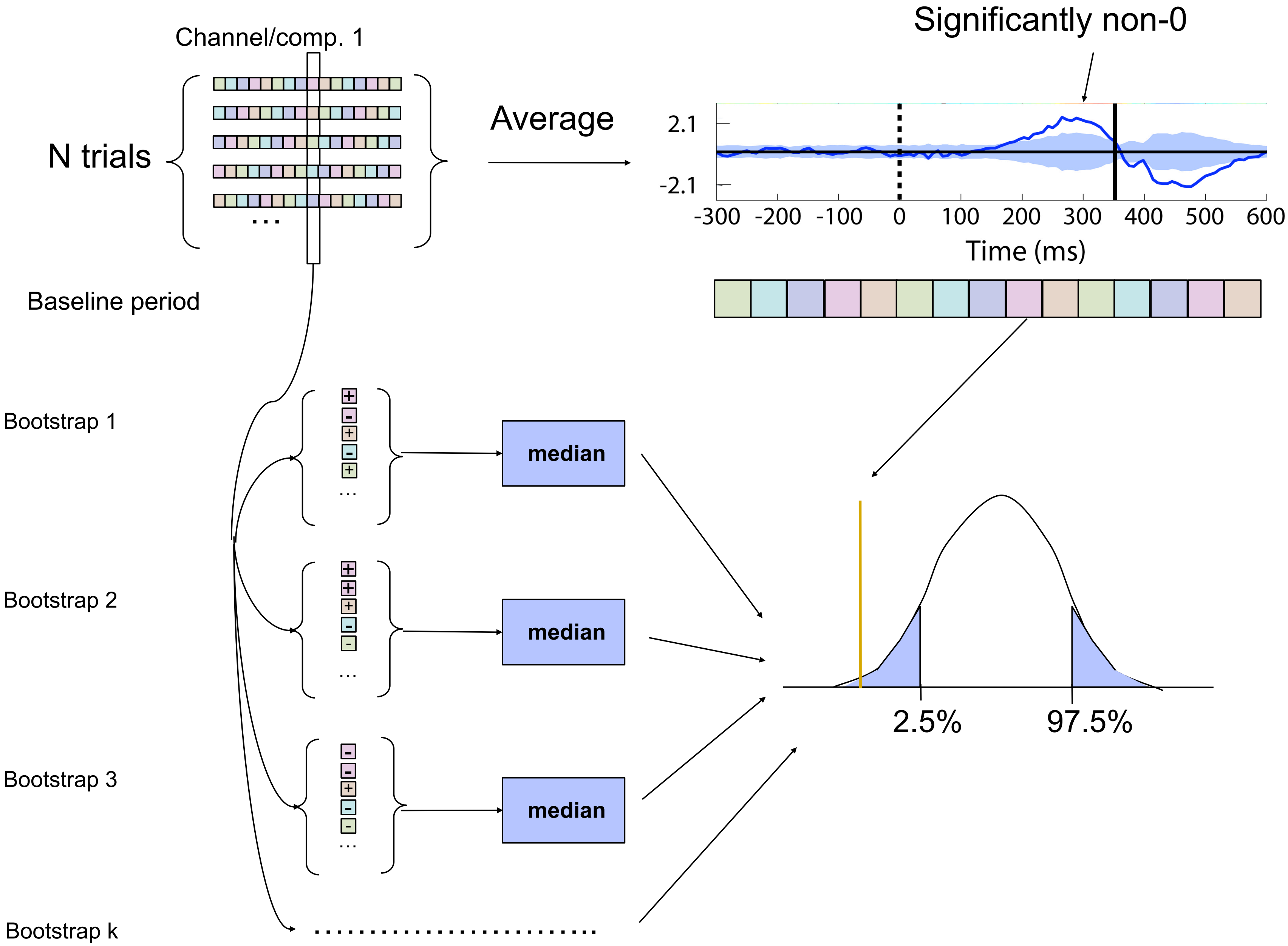
χ2 is used to compare 2 or more unpaired samples

$$\chi^2 = \sum_{i,j} (Observed_{i,j} - expected_{i,j})^2 / expected_{i,j}$$





Bootstrap for ERPs and time-frequency



Channel/comp. 1

N trials

Average

Significantly non-0

Baseline period

Bootstrap 1

Bootstrap 2

Bootstrap 3

Bootstrap k

median

median

median