Program for this afternoon

1. Lecture on RSA (30 mins)
2. Hands-on RSA exercise (45 mins)
3. RSA wrap-up, Q&A (15 mins)
4. Break (20 mins)
5. **Lecture on statistics (30 mins)**
6. Hands-on statistics exercise (45 mins)
7. Statistics wrap-up, Q&A (15 mins)

Statistical analysis of MEG

Featuring:
MNE-Python and the cluster-based permutation test

Aalto-yliopisto
Aalto-universitetet
Aalto University

CM
HC

Marijn van Vliet

STATISTICAL ERRORS

P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

BY REGINA NUZZO

Cognition and Neurosciences

Hail the impossible: *p*-values, evidence, and likelihood

TOBIAS JOHANSSON

Kristianstad University, Sweden

Johansson, T. (2011). Hail the impossible: *p*-values, evidence, and likelihood. *Scandinavian Journal of Psychology 52*, 113–125.

Significance testing based on *p*-values is standard in psychological research and teaching. Typically, research articles and textbooks present and use *p* as a measure of statistical evidence against the null hypothesis (the Fisherian interpretation), although using concepts and tools based on a completely different usage of *p* as a tool for controlling long-term decision errors (the Neyman–Pearson interpretation). There are four major problems with using *p* as a measure of evidence and these problems are often overlooked in the domain of psychology. First, *p* is uniformly distributed under the null hypothesis and can therefore never indicate evidence for the null. Second, *p* is conditioned solely on the null hypothesis and is therefore unsuited to quantify evidence, because evidence is always relative in the sense of being evidence for or against a hypothesis relative to another hypothesis. Third, *p* designates probability of obtaining evidence (given the null), rather than strength of evidence. Fourth, *p* depends on unobserved data and subjective intentions and therefore implies, given the evidential interpretation, that the evidential strength of observed data depends on things that did not happen and subjective intentions. In sum, using *p* in the Fisherian sense as a measure of statistical evidence is deeply problematic, both statistically and conceptually, while the Neyman–Pearson interpretation is not about evidence at all. In contrast, the likelihood ratio escapes the above problems and is recommended as a tool for psychologists to represent the statistical evidence conveyed by obtained data relative to two hypotheses.

Using *p* in the Fisherian sense as a measure of statistical evidence is deeply problematic, both statistically and conceptually.

Aalto-yliopisto
Aalto-universitetet
Aalto University

CM
HC

comment

Redefine statistical significance

We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson

Insight from the Fieldtrip manual

If you are doing **hypothesis-driven** research, you should not guide your statistical analysis by a visual inspection of the data; you should state your hypothesis up-front and avoid data dredging or p-hacking.

On the other hand: if you are doing **exploratory** research, you should not compute p-values. Effect sizes are interesting and relevant to report for both exploratory and hypothesis-driven research.

— Robert Oostenveld (probably)

https://www.fieldtriptoolbox.org/example/effectsize/

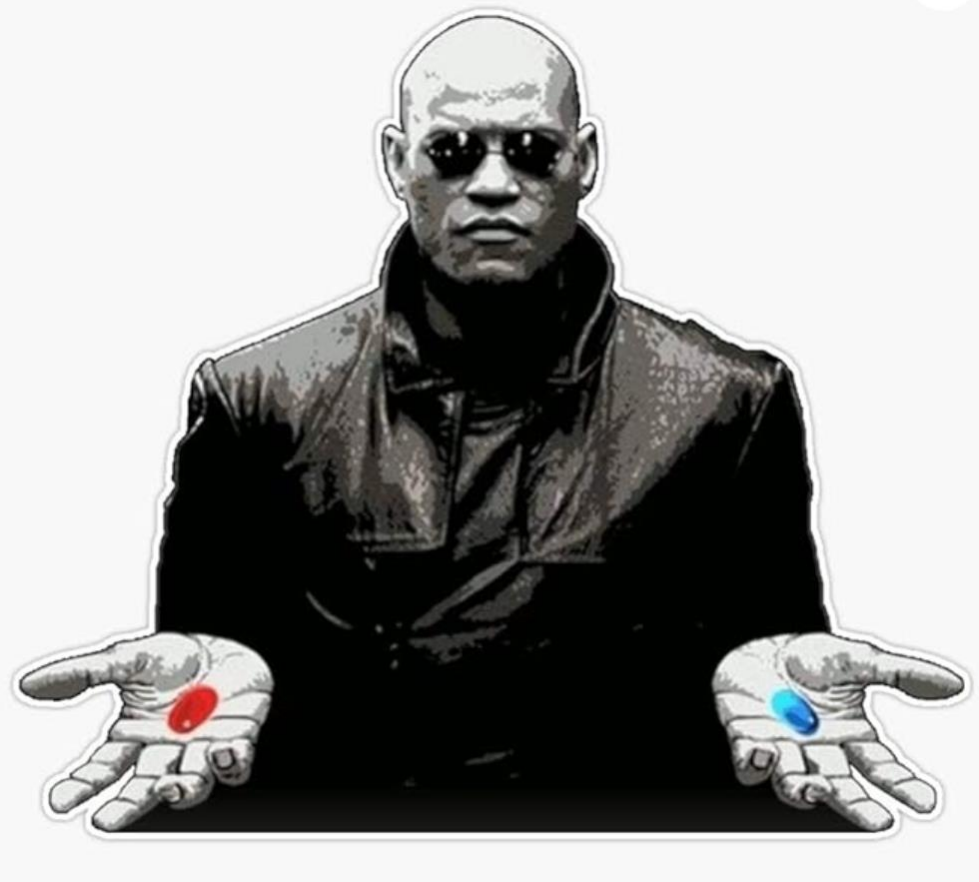
Marijn's advice

effect-size: is there an effect in the data?
where in the data is this effect?
do we possibly care about this effect?

p-value: is this effect due to inherent randomness
or is it systematic in the data? (~SNR)

replication: is this effect a true effect?

Marijn van Vliet

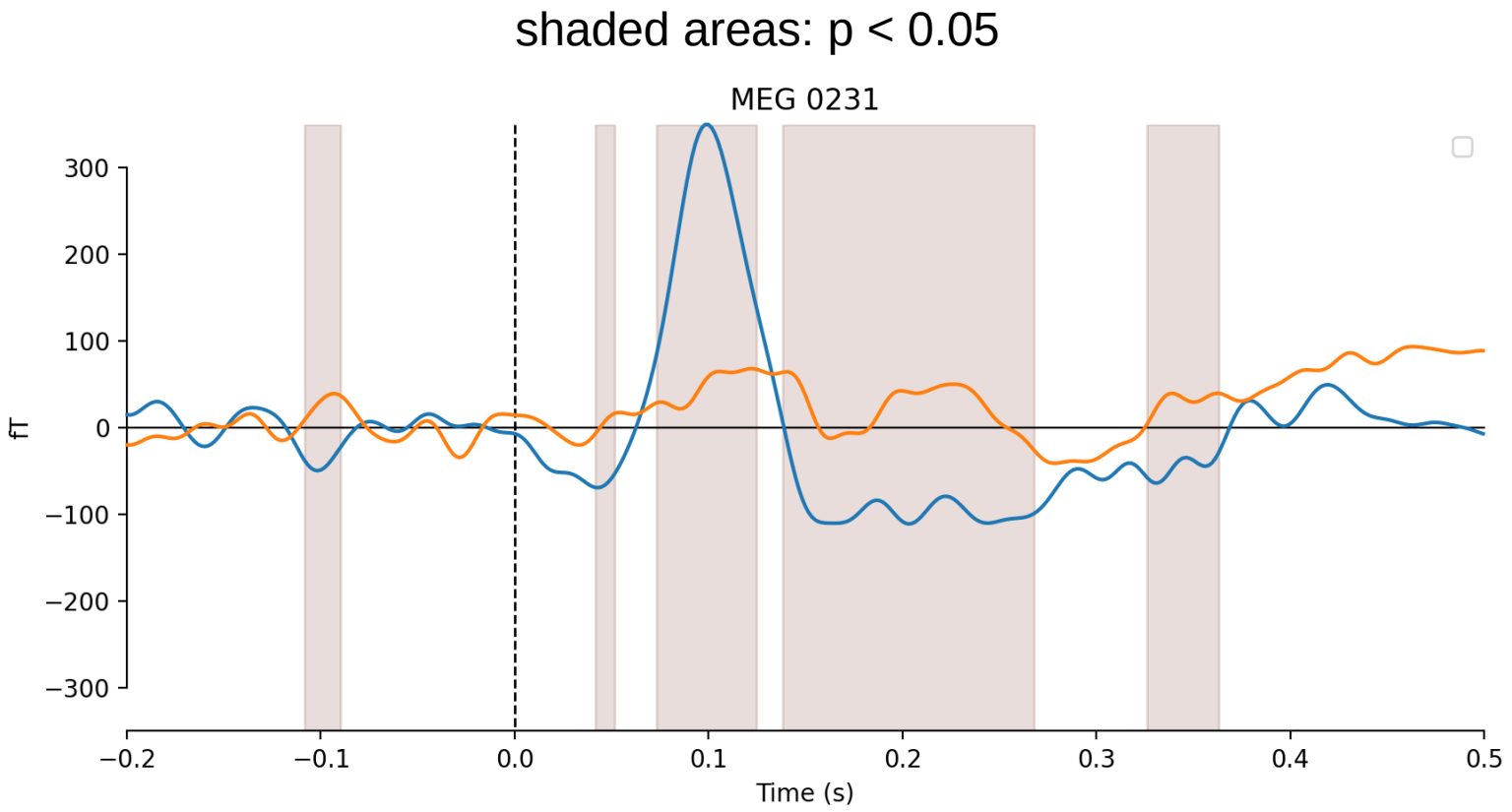


exploration

hypothesis testing

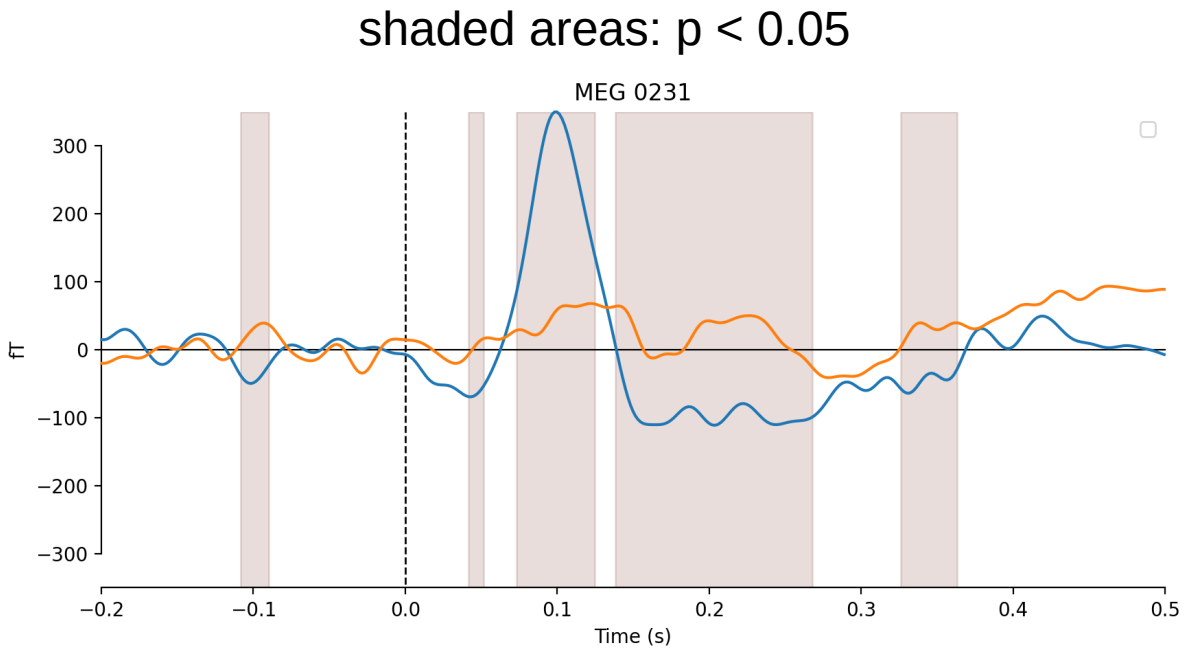
Multiple comparisons “problem”

MNE-Python sample data, channel #20, left auditory vs left visual



Presence of “significant” p-value (<0.05) before stimulus onset?

Multiple comparisons “problem”

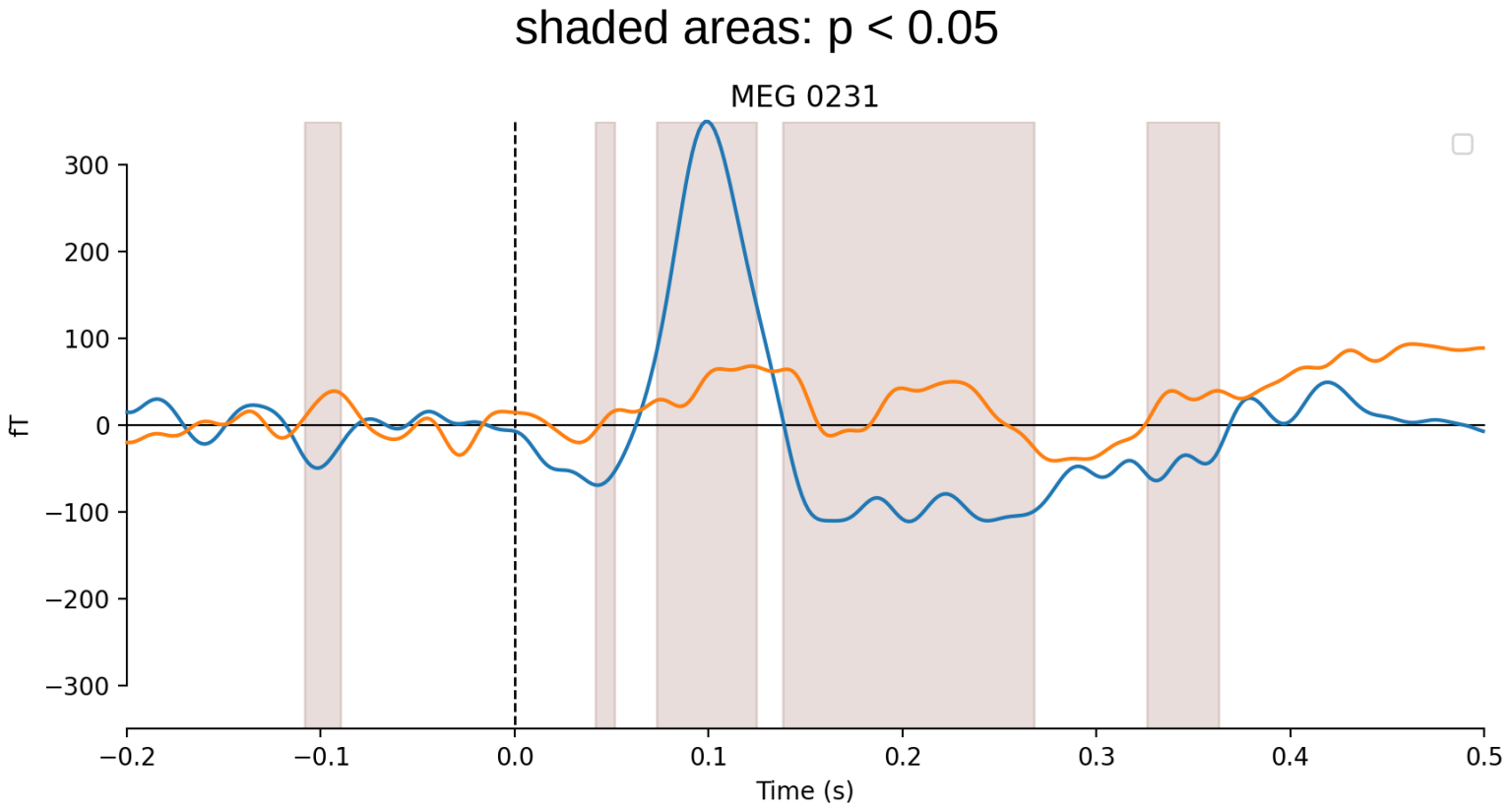


421 t-tests performed.
Around 0.05 × 421 = 21 of them should be “significant”

Solution to multiple comparisons

Stop relying on them.

Multiple comparisons “problem”



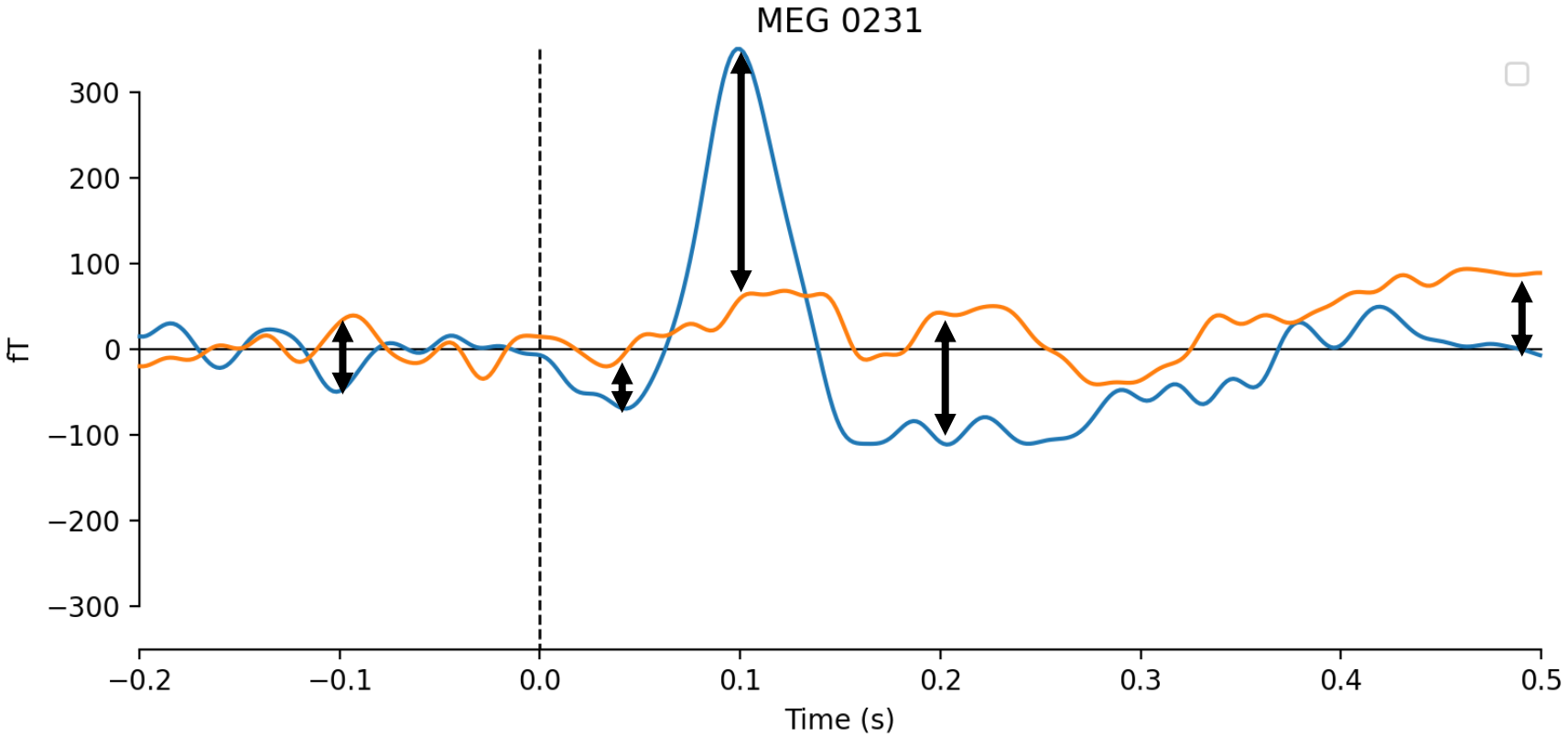
We are **exploring**. During exploration, our question is often:
“at what time(s) is there a difference between the two conditions?”
p-values are the wrong tool to answer this question.

!

Statistical tests are uninformative about location and extent of an effect!

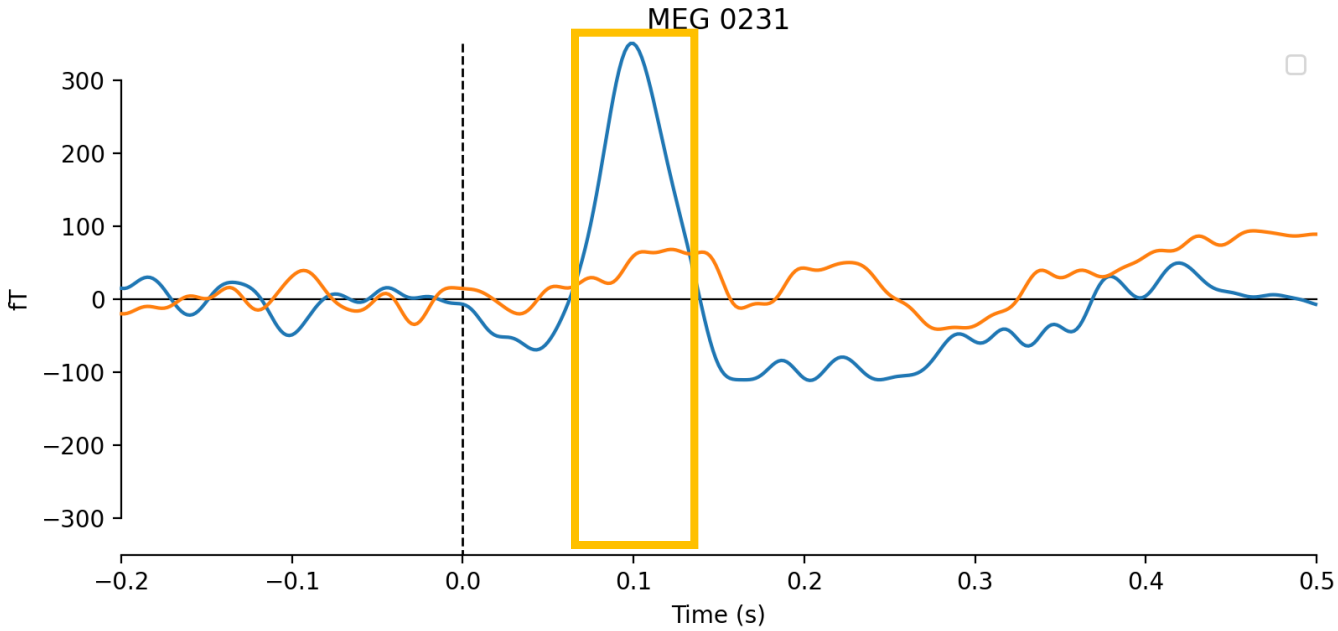
use effect size for this purpose.

In exploration effect-size is king



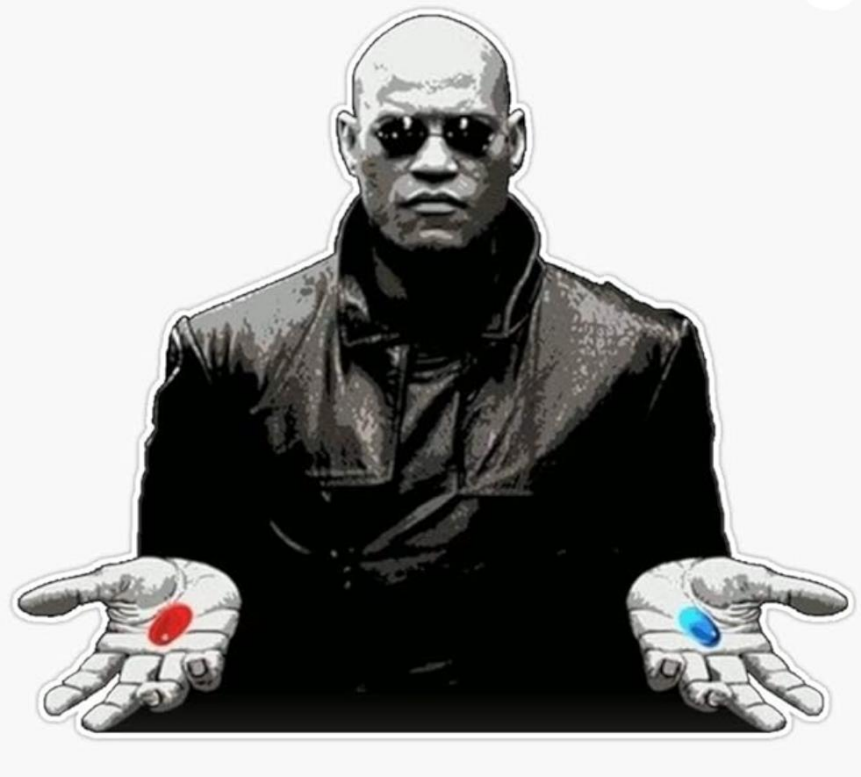
Does anything stand out to us?

In exploration effect-size is king



- Define region of interest (ROI) based on where data diverges/converges.
- **Report extent of ROI and effect-size within ROI**
- Report single p-value to give a sense of SNR.

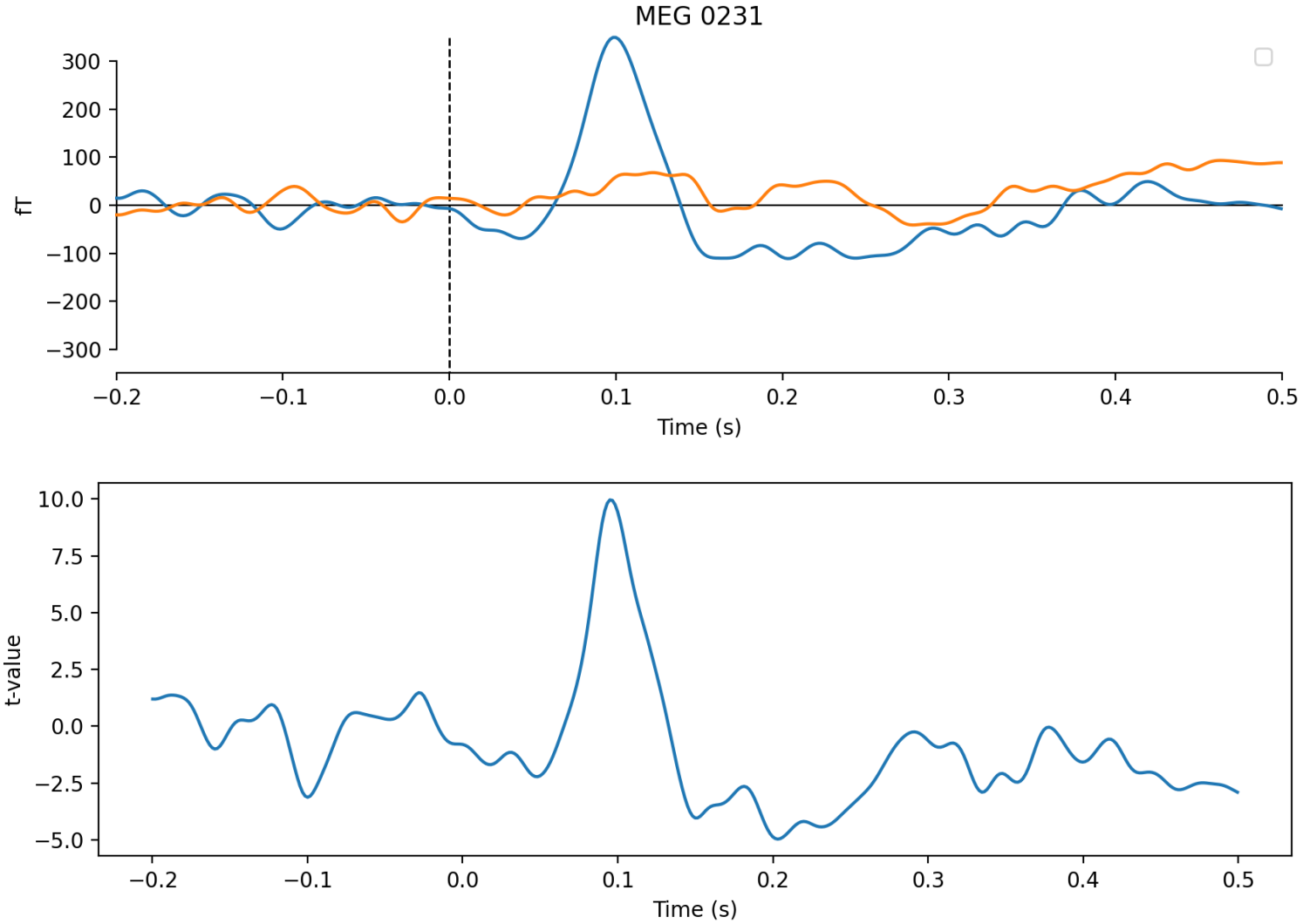
How to obtain single p-value?



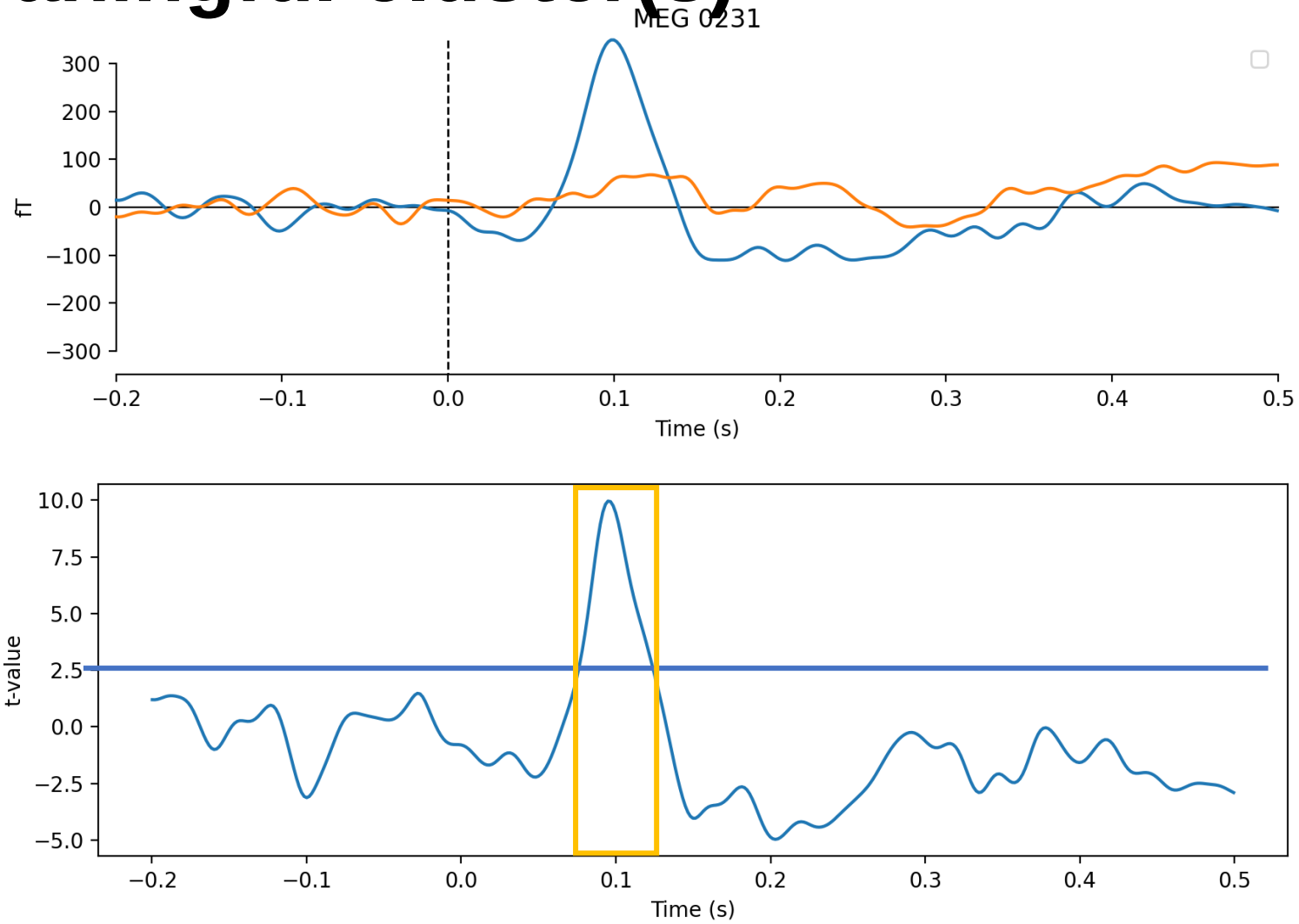
mean across ROI
single statistical test

cluster-based permutation test

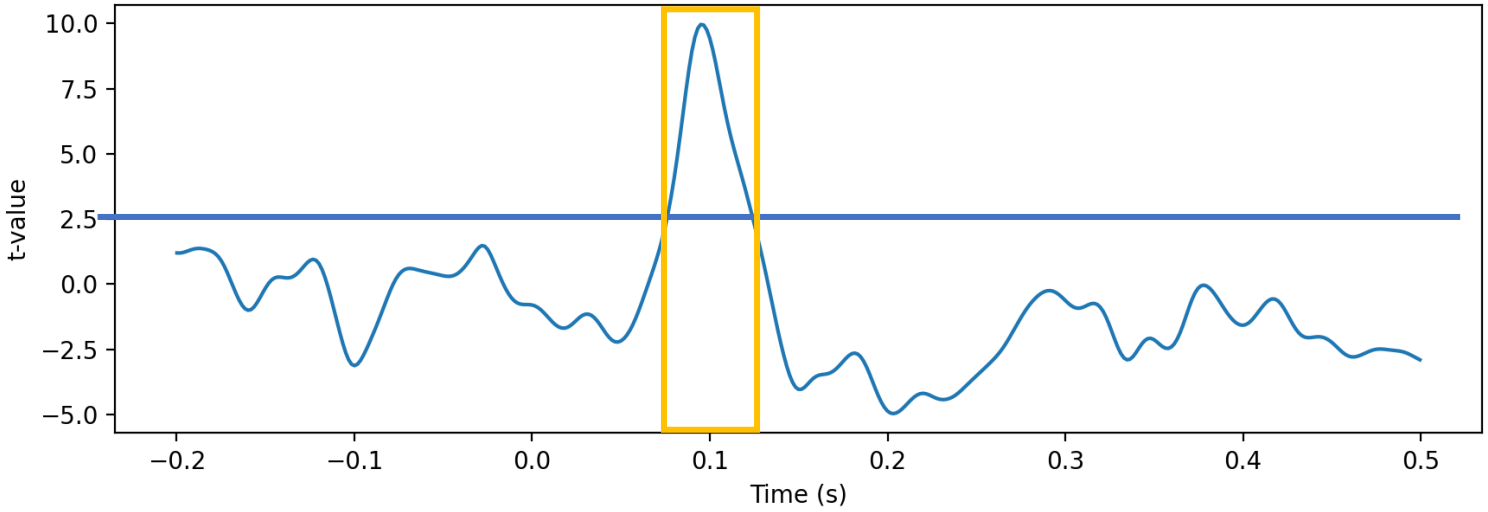
Step 1: make statistical map



Step 2: define threshold that isolates meaningful cluster(s)

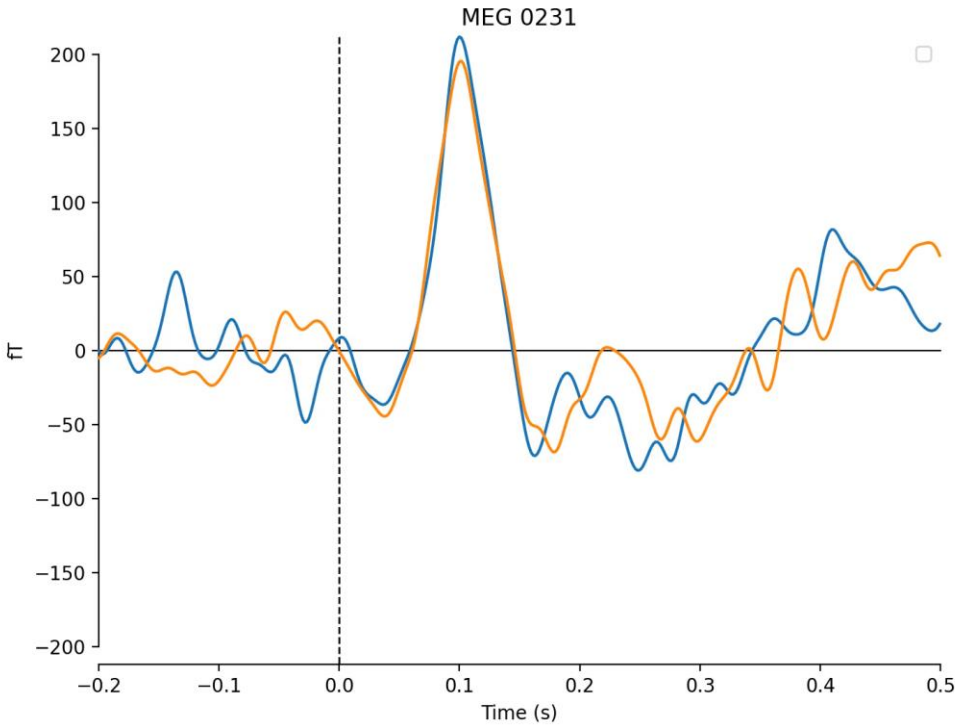


Step 3: compute sum statistic across cluster



Sum of t-values in cluster: 189.98

Step 4: permute the data



1. Shuffle condition labels
2. Form clusters using the same statistic and threshold as before
3. Compute **sum** of statistic in each permuted cluster

Step 3: compute p-value

1. For each original cluster, find the percentage of permutations that resulted in a cluster with a higher sum-statistic.

2. This yields for each cluster, a p-value.

3. **If you have defined multiple clusters, take the smallest p-value**

!

In cluster-based permutation tests, a single cluster is never “significant” or “not-significant”

?

“But each cluster has a p-value!”

!

The cluster's p-value is computed against the entire dataset, not just that one cluster!

That's why we take the smallest p-value as final outcome of the test.

PSYCHOPHYSIOLOGY

ORIGINAL ARTICLE

Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location

Jona Sassenhagen, Dejan Draschkow

Nevertheless, researchers consistently employ cluster-based permutation test results to support claims not only about the existence of a significant difference, but also about the (spatial or temporal) extent or location of effects. This procedure, while common, is inapplicable.

Misuses of the method are ubiquitous in the literature; while we will abstain from identifying individual "offenders," the authors themselves must admit to having inappropriately utilized the method in the way here described, and further examples can be abundantly found in the published literature.

Aalto-yliopisto
Aalto-universitetet
Aalto University

CM
HC

Sassenhagen and Draschkow, 2019

!

If you want a p-value for a single ROI, you must first restrict the data to that ROI before doing the cluster-based permutation test

Reporting

1. Statistical test and threshold used to form clusters.
2. “We found a significant difference between the conditions (p-value)”
3. “This result was mostly driven by the following clusters”

Do **not** say that a **cluster** was significant or not.

Here's how to do a cluster-based permutation test in MNE-Python

Find the answer in the repository:

`rsa/statistics.ipynb`