

Prevalence

**An alternative way of thinking about statistics in
neuroimaging experiments**

Robin Ince, CuttingEEG Dundee Garden, October 2023



University
of Glasgow



Bayesian inference of population prevalence

Robin AA Ince^{1*}, Angus T Paton¹, Jim W Kay², Philippe G Schyns^{1,3}


<https://elifesciences.org/articles/62461>

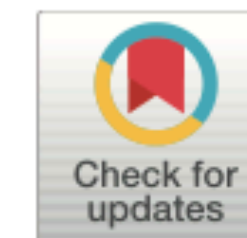
<https://github.com/robince/bayesian-prevalence>

<https://estimate.prevalence.online/>

Forum

Within-participant statistics for cognitive science

Robin A.A. Ince ^{1,*}
Jim W. Kay,² and
Philippe G. Schyns¹



<https://doi.org/10.1016/j.tics.2022.05.008>

Background

Statistics for experiments in neuroscience and neuroimaging

- What is the goal of our statistical analyses?
- Often: to infer a causal relationship between some aspect of the external world or behaviour and some aspect of measured neural activity



- Avoid being fooled by randomness
- **Generalise** from our sample of participants to the wider population

Background

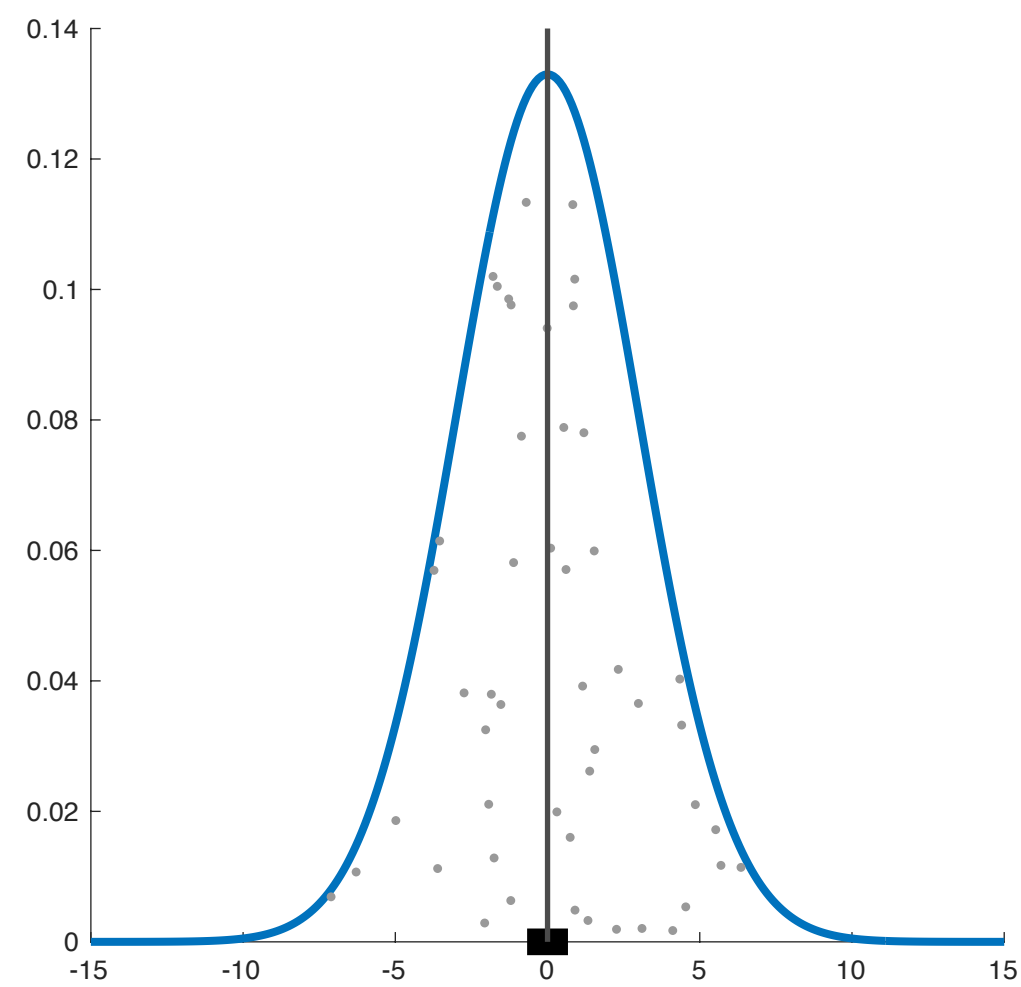
Population Inference

- **Generalise:** Population inference vs case study
- **To make an inference about the population requires a model of the population**
- Common approach: models population with a normal distribution. Participants as random effects in a linear model (Holmes & Friston, 1998)
- Usually focus on population mean
(i.e. reject null hypothesis the population mean effect is 0)

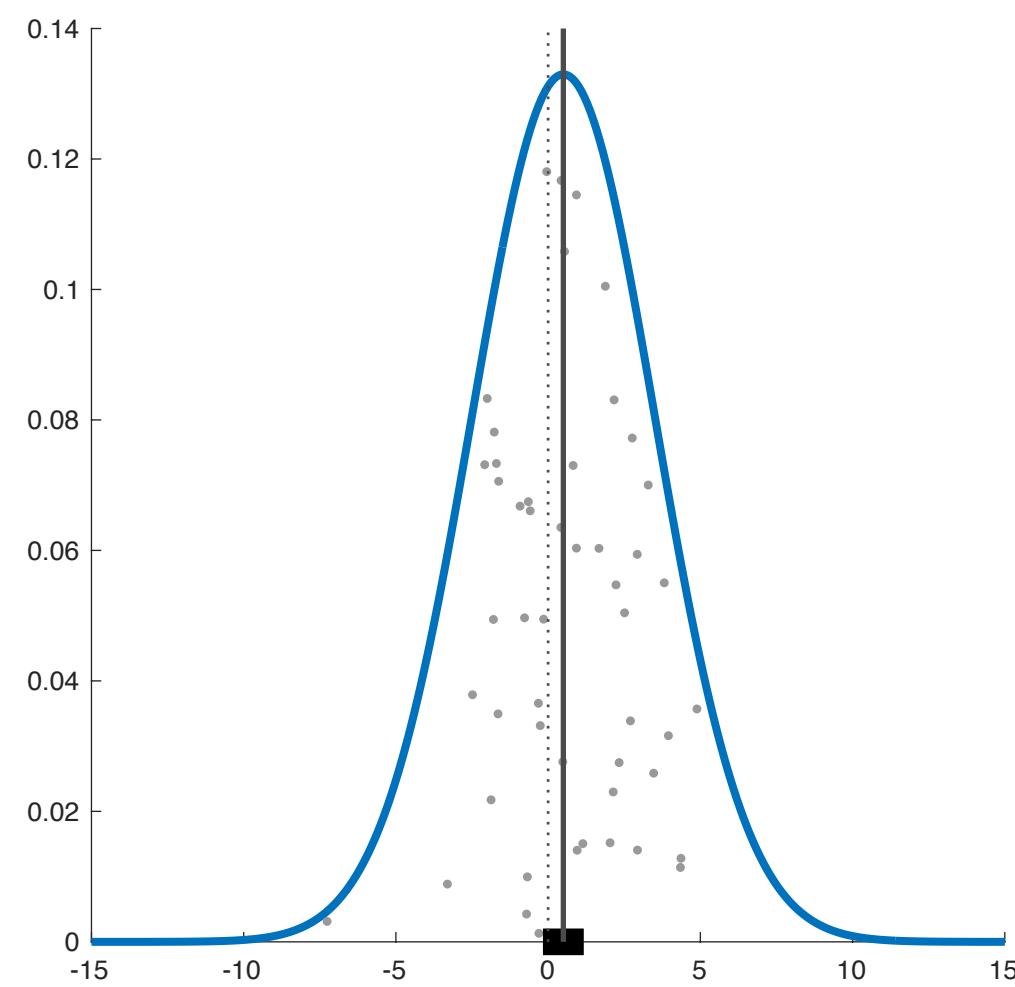
Background

Random Effects Population Inference

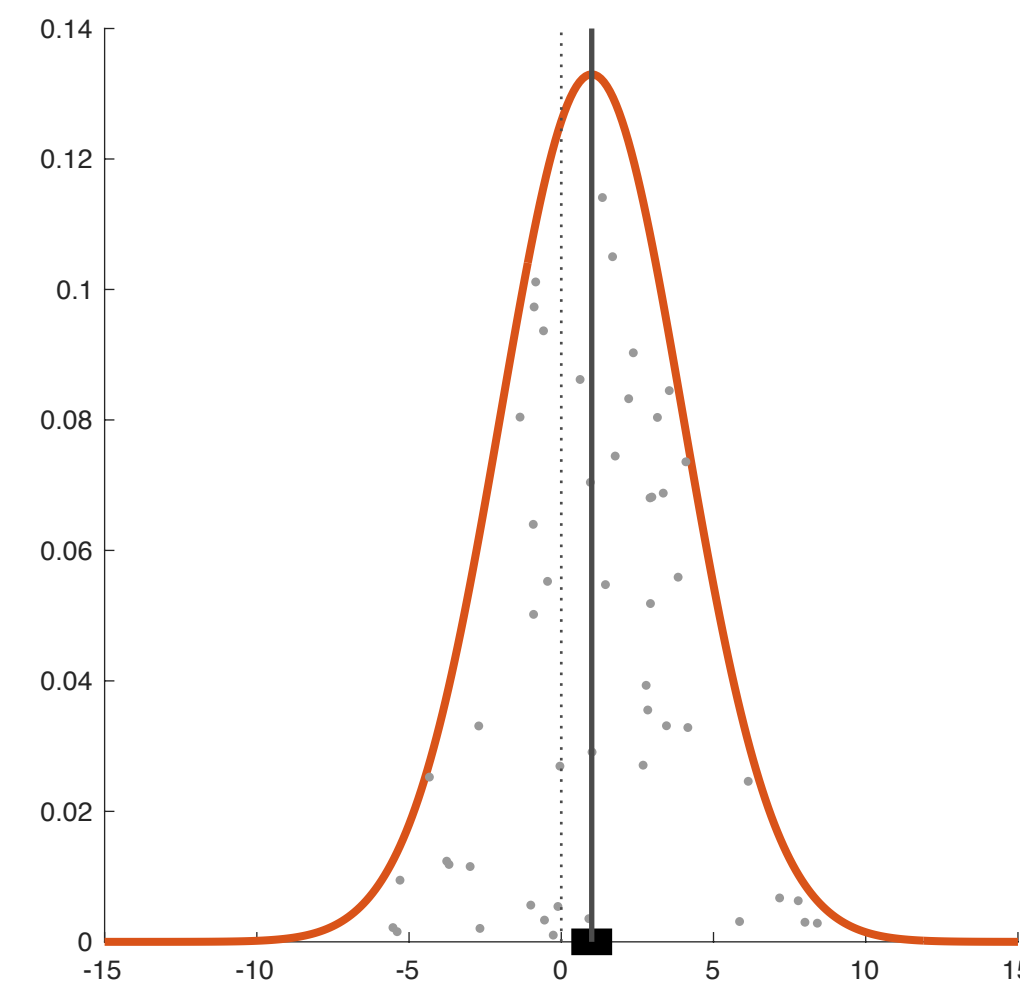
- Second level t-test. Model population with normal distribution. Compare mean across subjects to variance across subjects



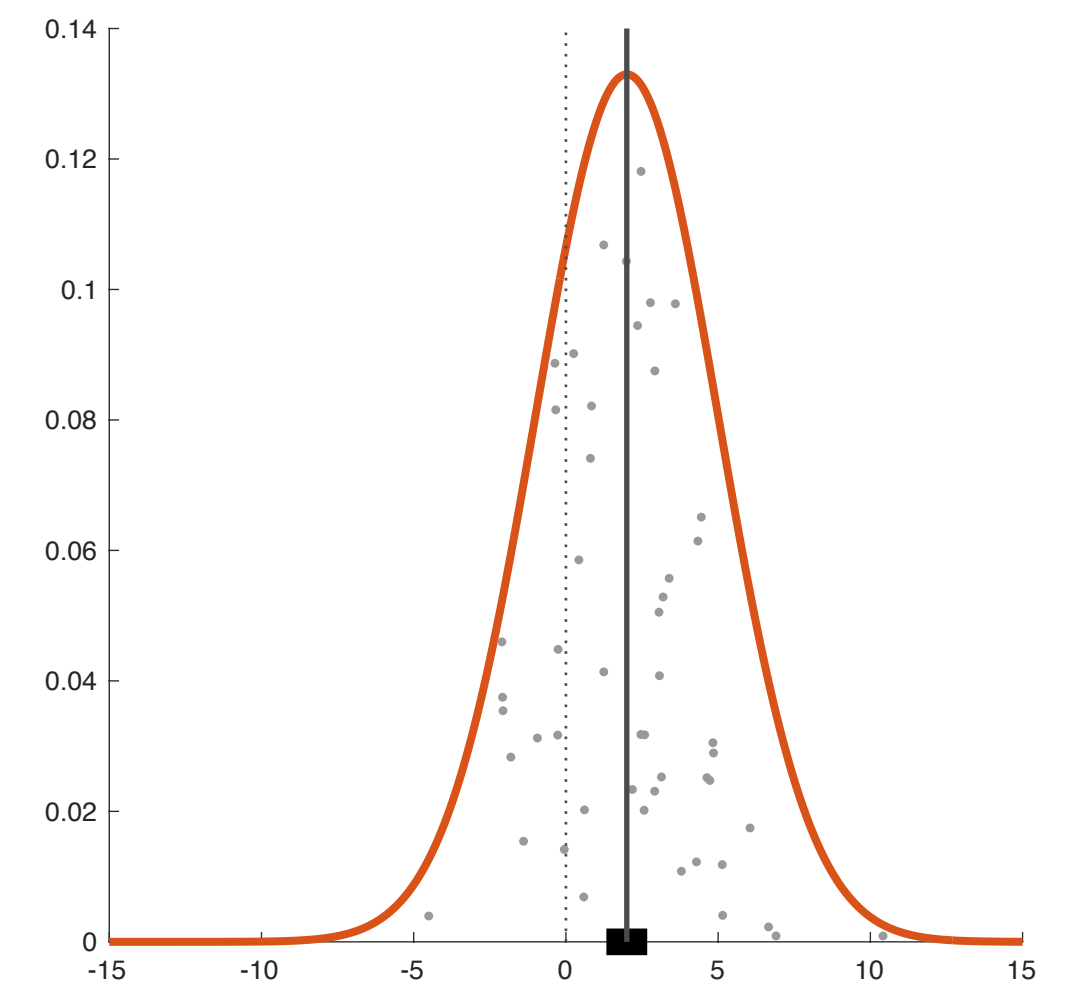
$$\mu_{pop} = 0$$



$$\mu_{pop} = 0.5$$



$$\mu_{pop} = 1$$



$$\mu_{pop} = 2$$

Second level t-test

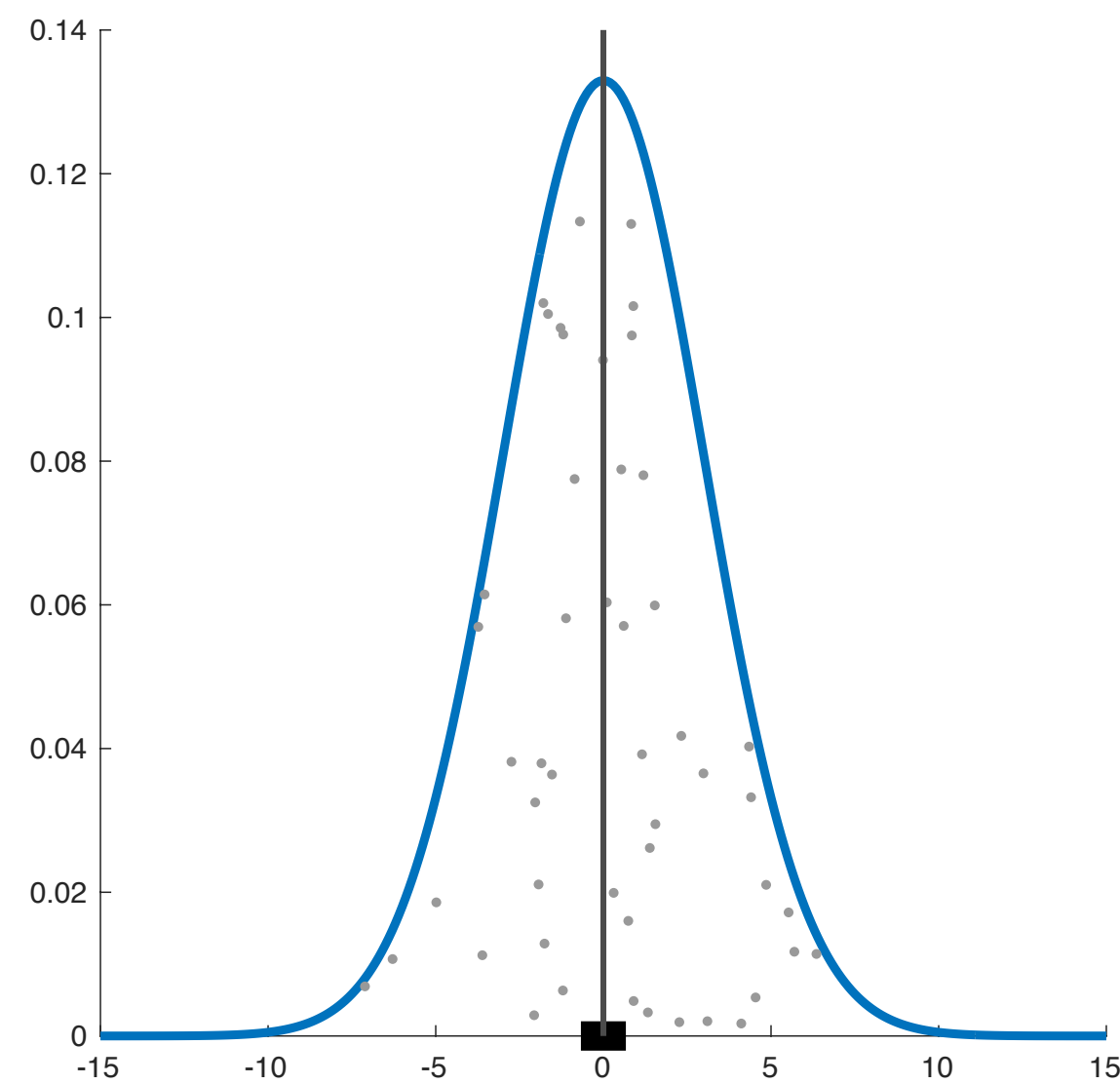
No effect in the population?

$$N = 50$$

$$\sigma_w = 10$$

$$\sigma_b = 2$$

$$\mu_{pop} = 0$$



$$T = 20$$

Second level t-test

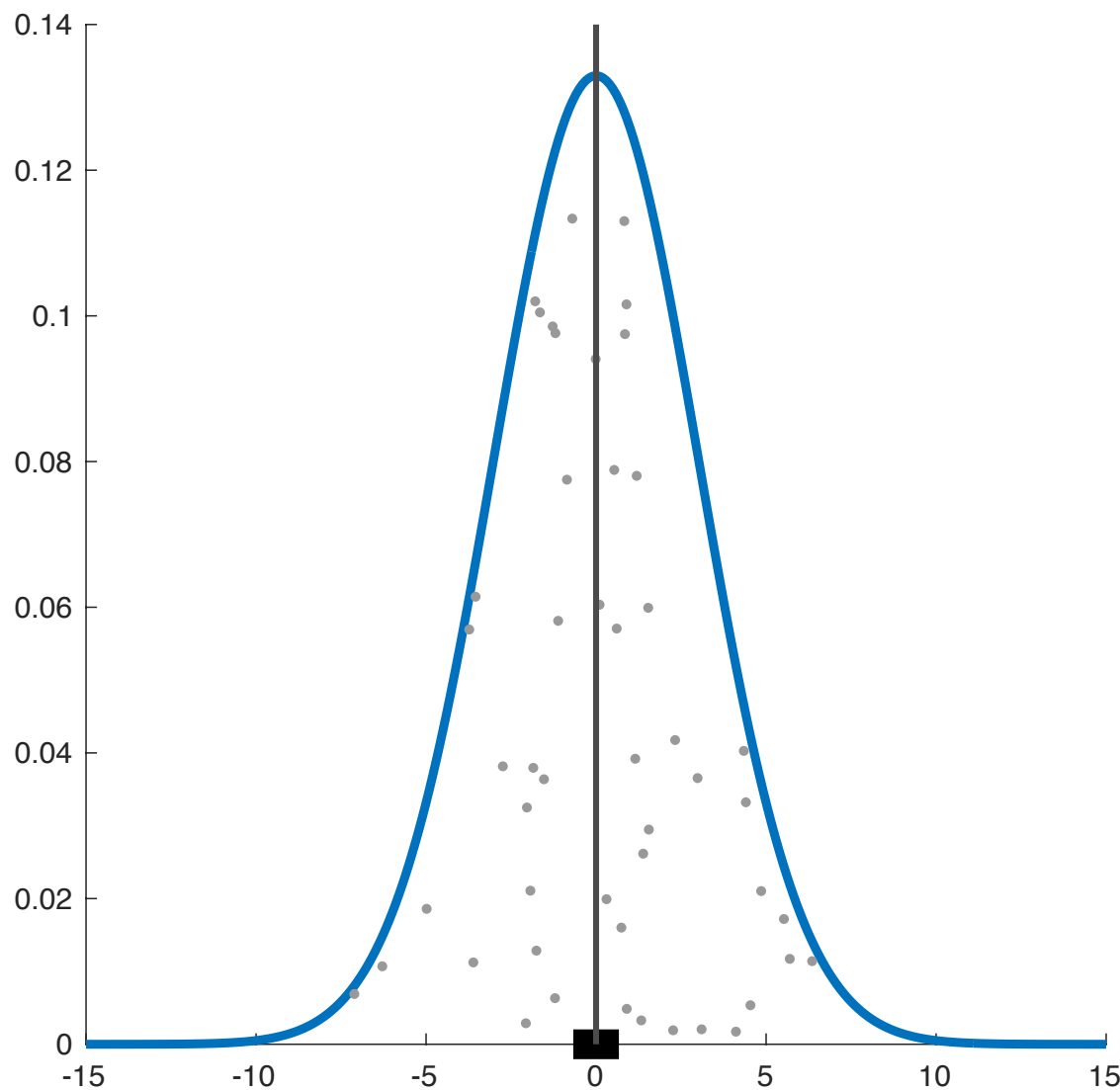
No effect in the population?

$N = 50$

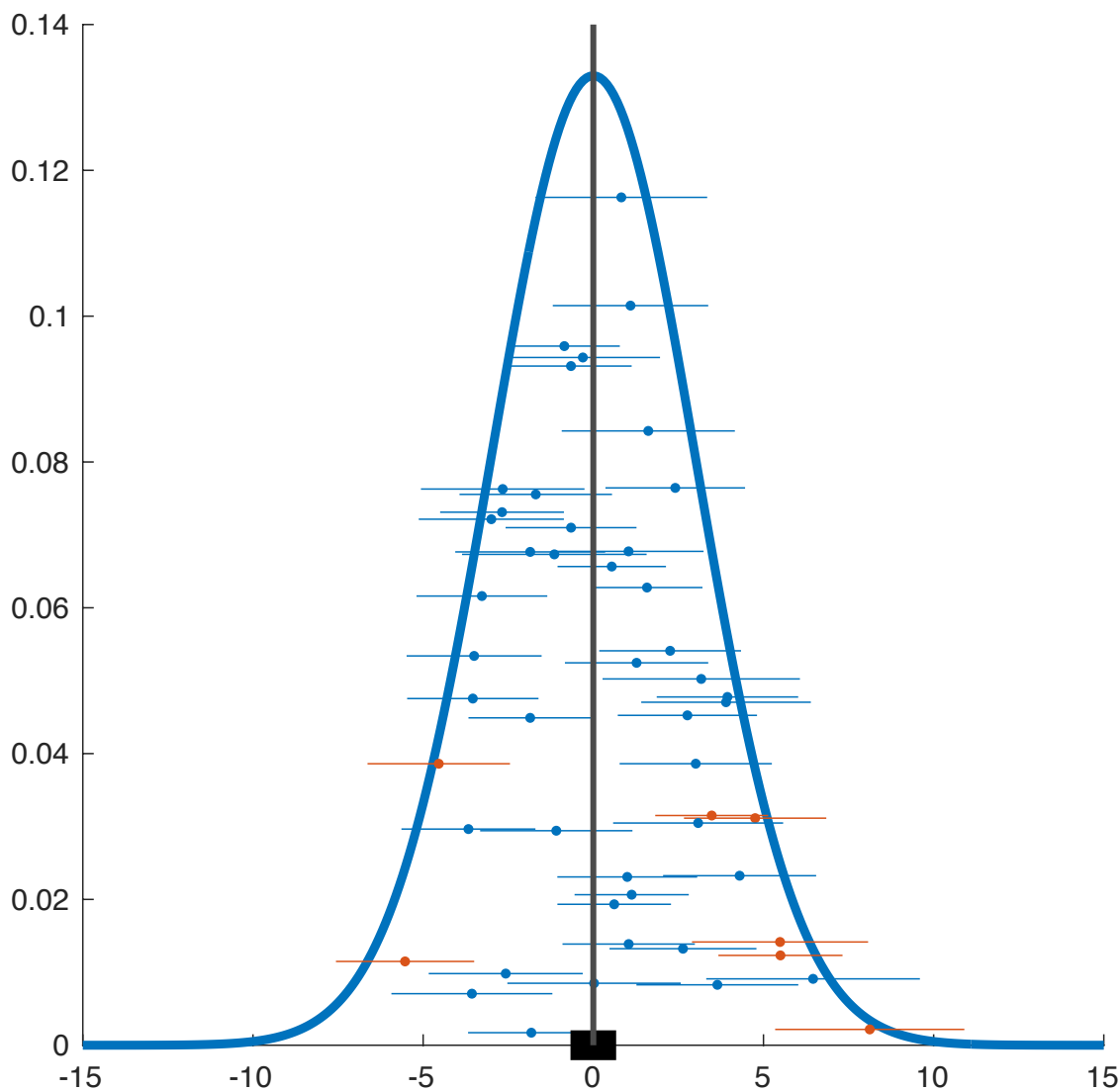
$\sigma_w = 10$

$\sigma_b = 2$

$\mu_{pop} = 0$



$T = 20$

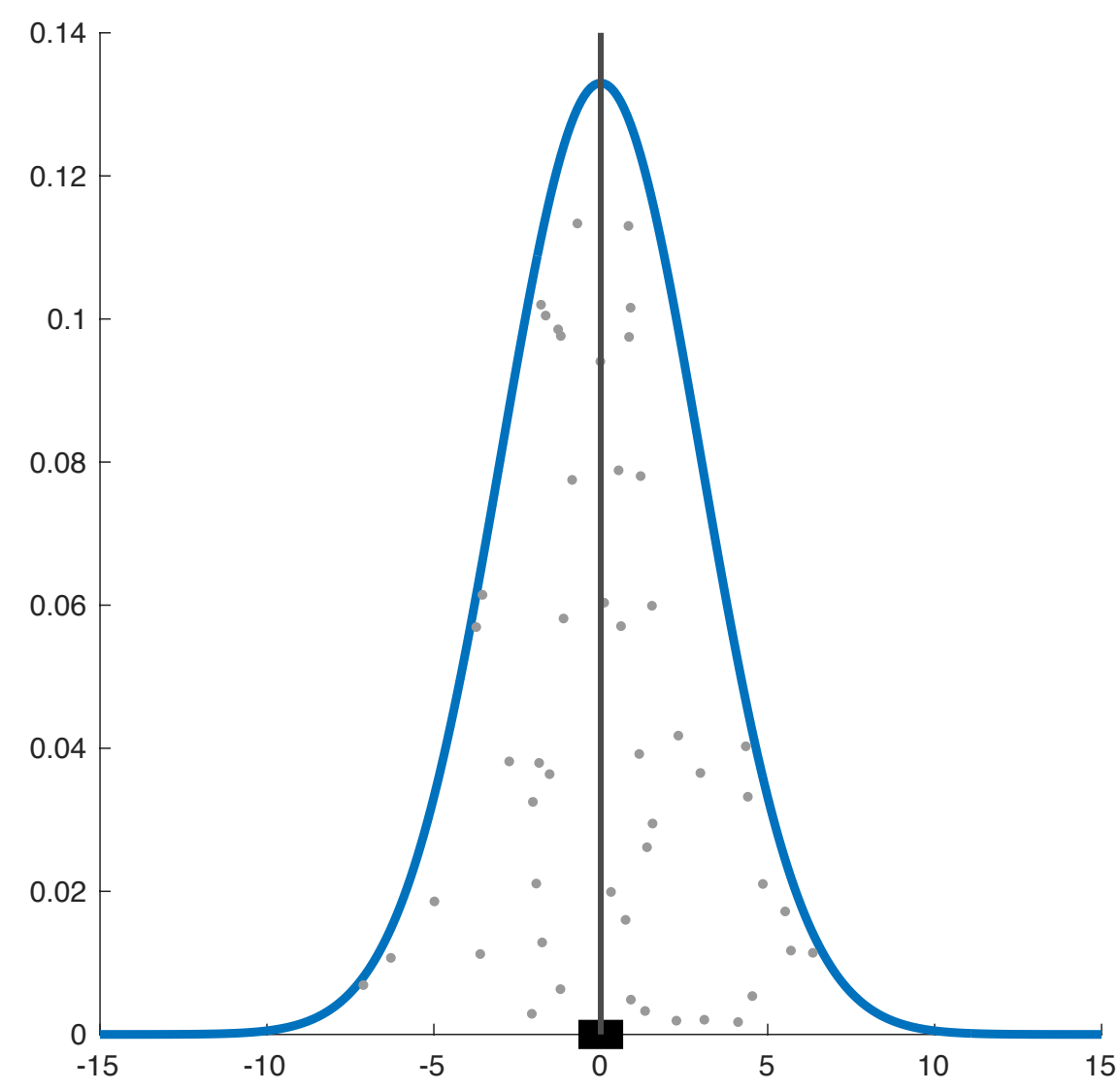


$T = 20$

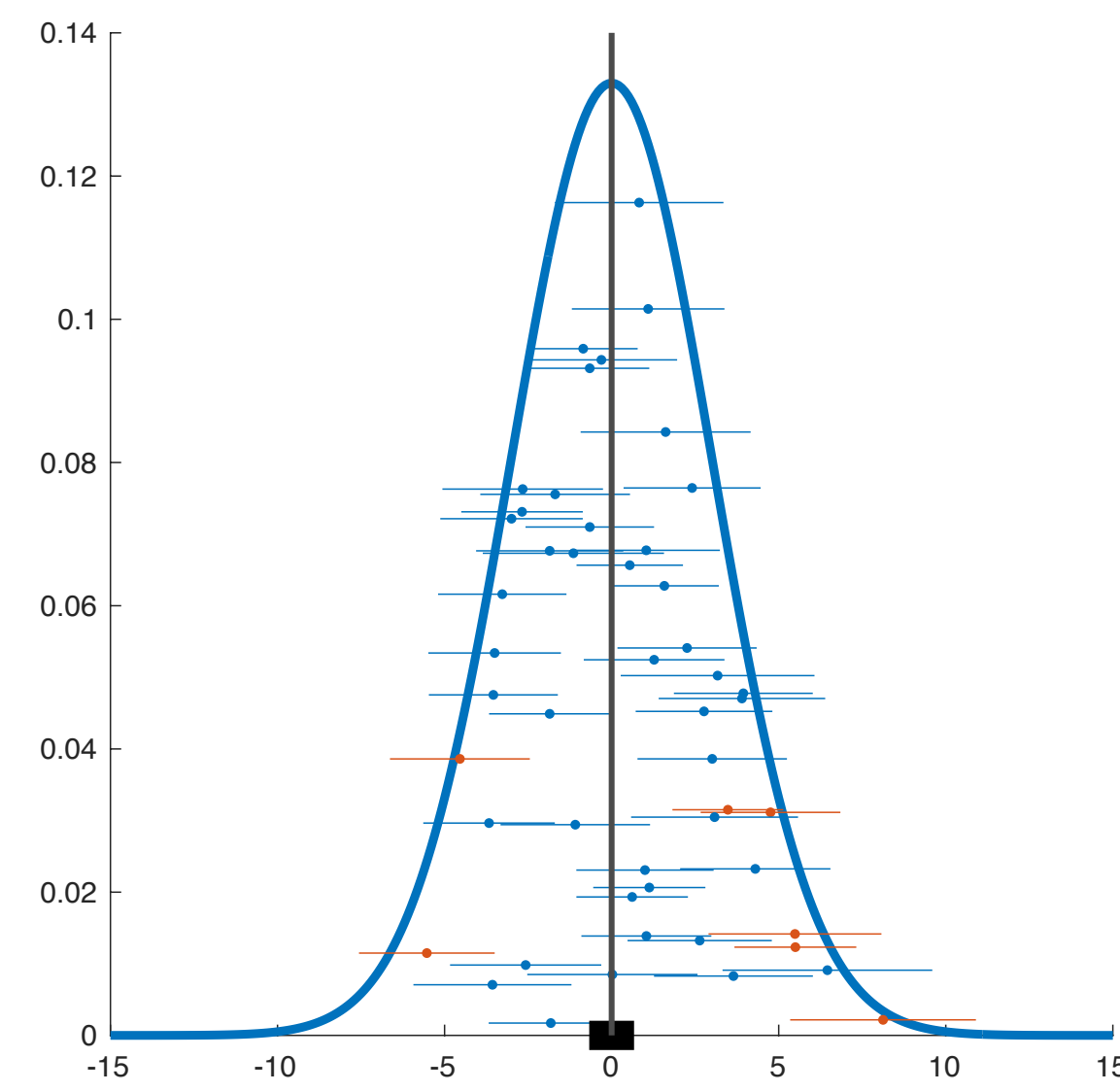
Second level t-test

No effect in the population?

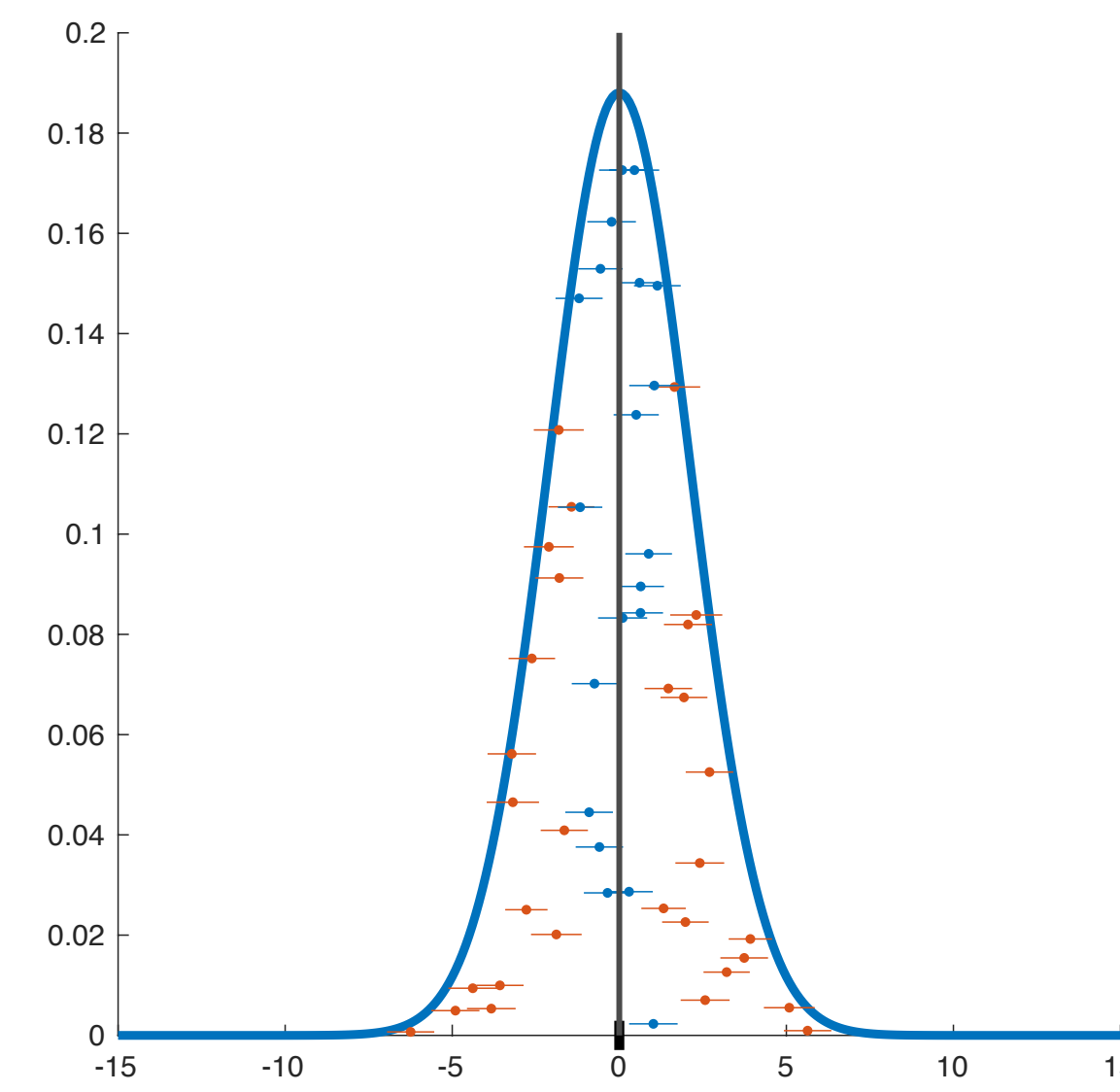
$$N = 50$$
$$\sigma_w = 10$$
$$\sigma_b = 2$$
$$\mu_{pop} = 0$$



$T = 20$



$T = 20$



$T = 200$

Second level t-test

No effect in the population?

$$N = 50$$

$$\sigma_w = 10$$

$$\sigma_b = 2$$

$$\mu_{pop} = 0$$

- Under the null model that is used in almost every study, it is possible to have highly reliable but heterogenous effects across participants
- We know neuroimaging results are very heterogenous across participants
- We might be missing some effects by focussing on the population mean

Second level t-test

No effect in the population?

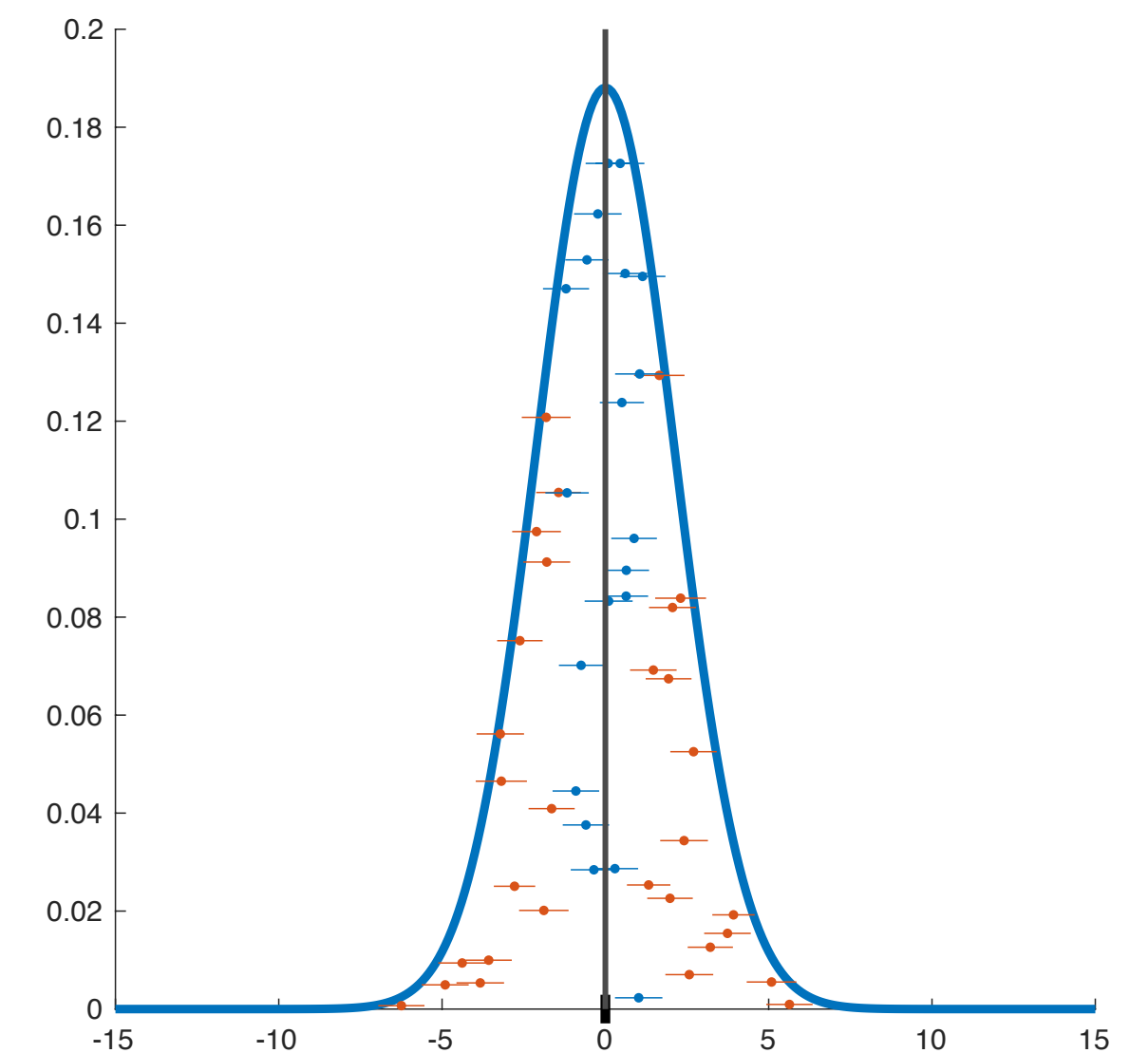
- Under the null model that is used in almost every study, it is possible to have highly reliable but heterogenous effects across participants
- We know neuroimaging results are very heterogenous across participants
- We might be missing some effects by focussing on the population mean

$$N = 50$$

$$\sigma_w = 10$$

$$\sigma_b = 2$$

$$\mu_{pop} = 0$$



$$T = 200$$

Second level t-test

No effect in the population?

$$N = 50$$

$$\sigma_w = 10$$

$$\sigma_b = 2$$

$$\mu_{pop} = 0$$

- Under the null model that is used in almost every study, it is possible to have highly reliable but heterogenous effects across participants
- We know neuroimaging results are very heterogenous across participants
- **We might be missing some effects by focussing on the population mean**

Second level t-test

No effect in the population?

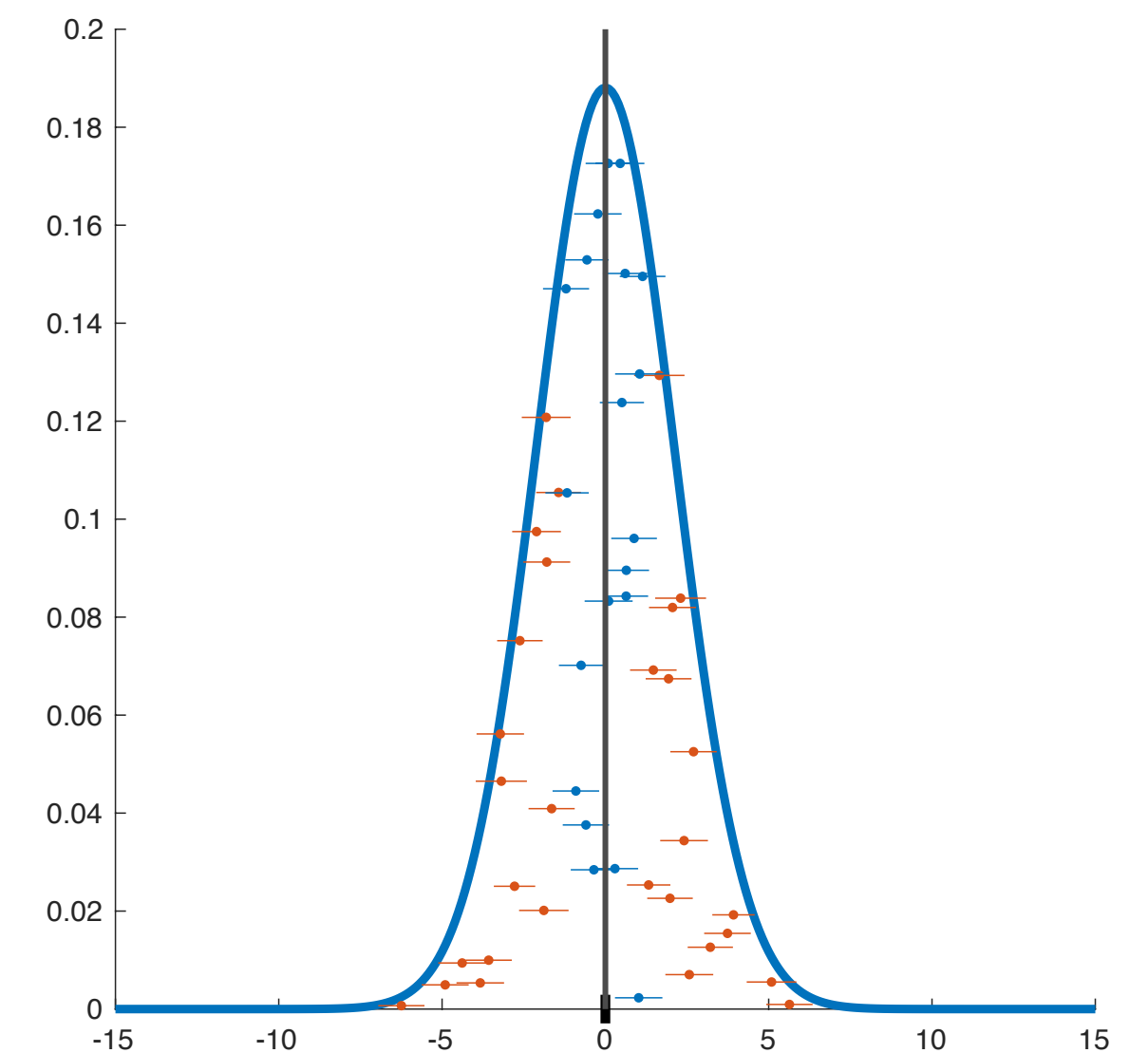
- Under the null model that is used in almost every study, it is possible to have highly reliable but heterogenous effects across participants
- We know neuroimaging results are very heterogenous across participants
- **We might be missing some effects by focussing on the population mean**

$$N = 50$$

$$\sigma_w = 10$$

$$\sigma_b = 2$$

$$\mu_{pop} = 0$$



$$T = 200$$

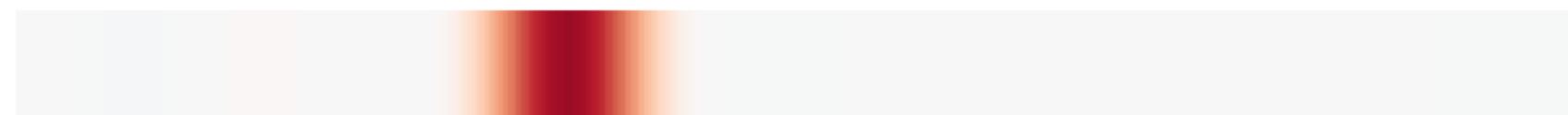
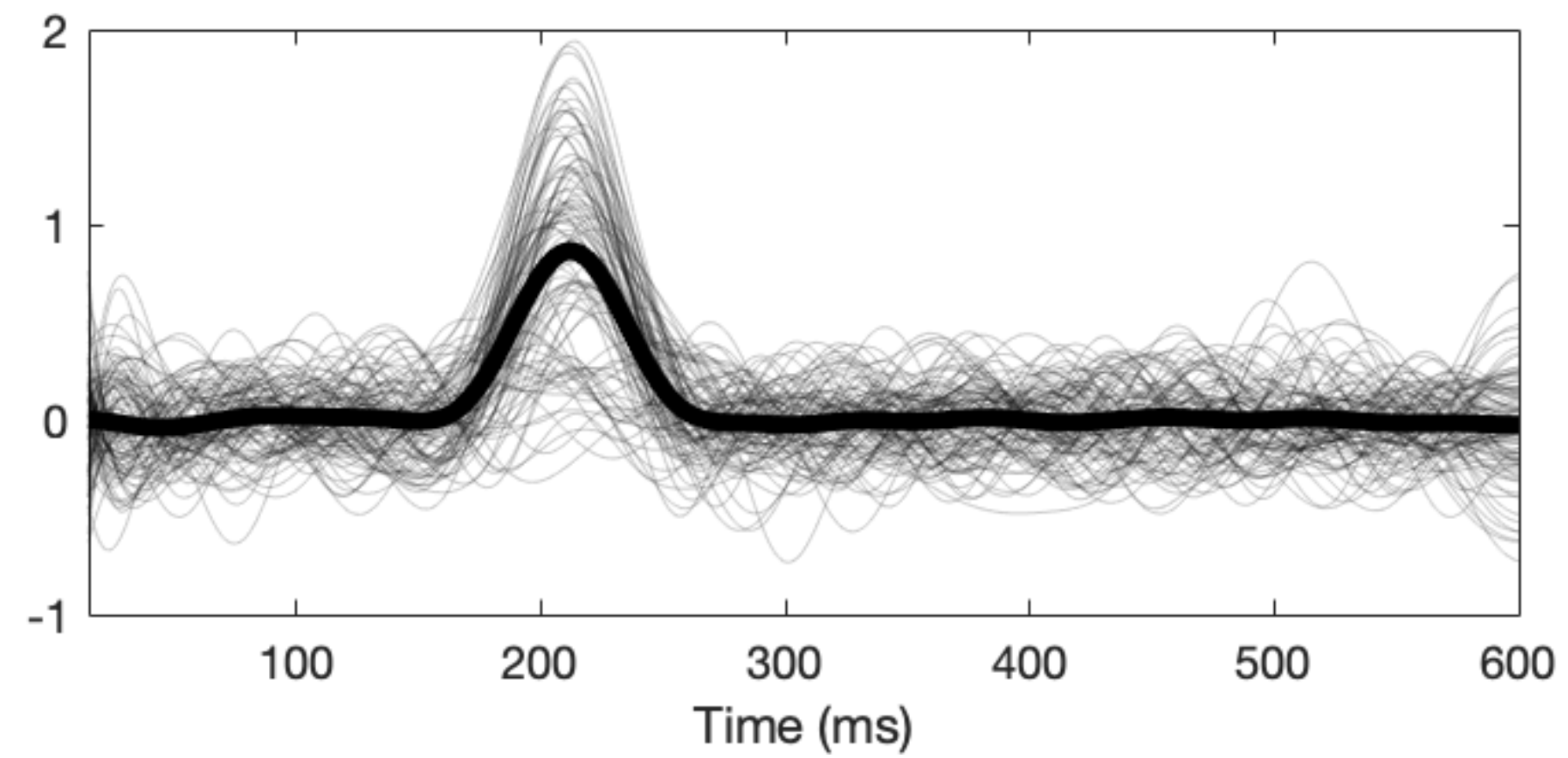
Tutorial

- We might be missing some effects by focussing on the population mean
- Investigate this issue using simulation
- Goal: understand the difference between statements about population prevalence and population mean
- Goal: understand through simulation situations where prevalence and population mean results might diverge
- Stretch goal: apply this to some of your own data

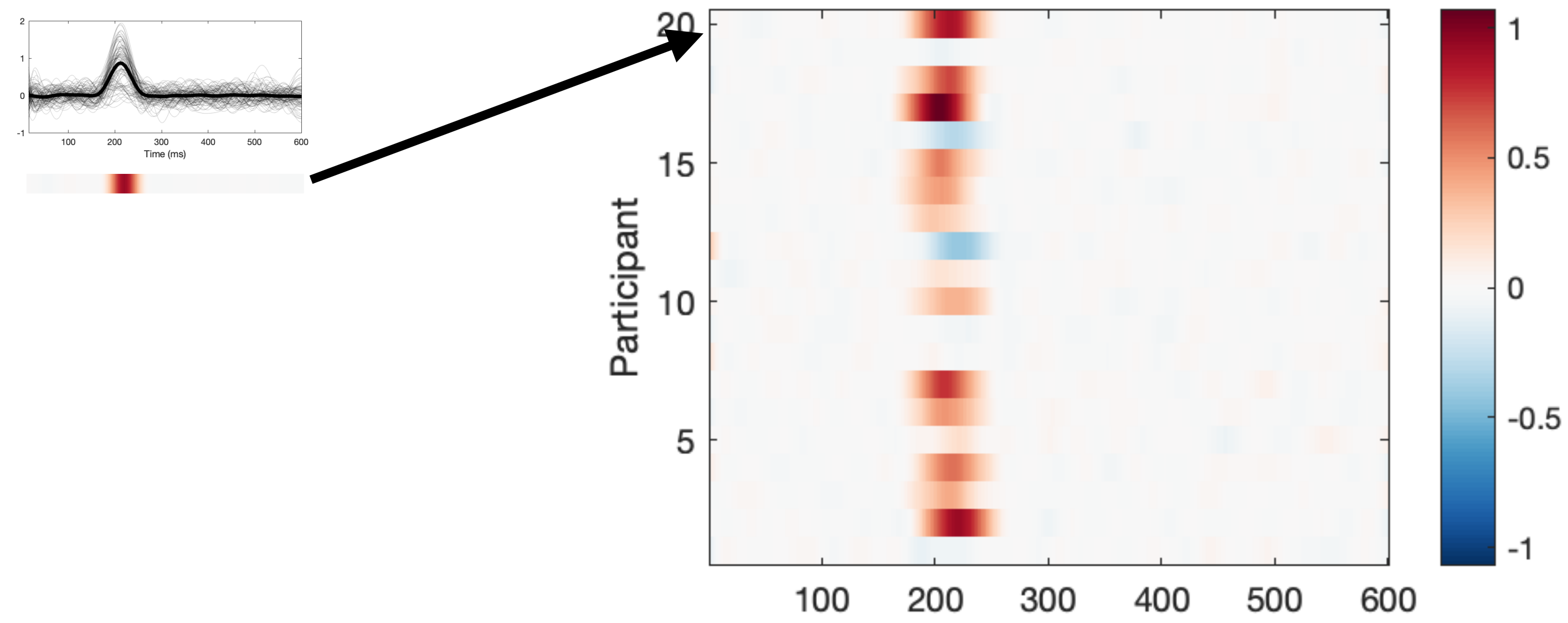
Tutorial

- Load up Prevalence_Tutorial_1 and follow along

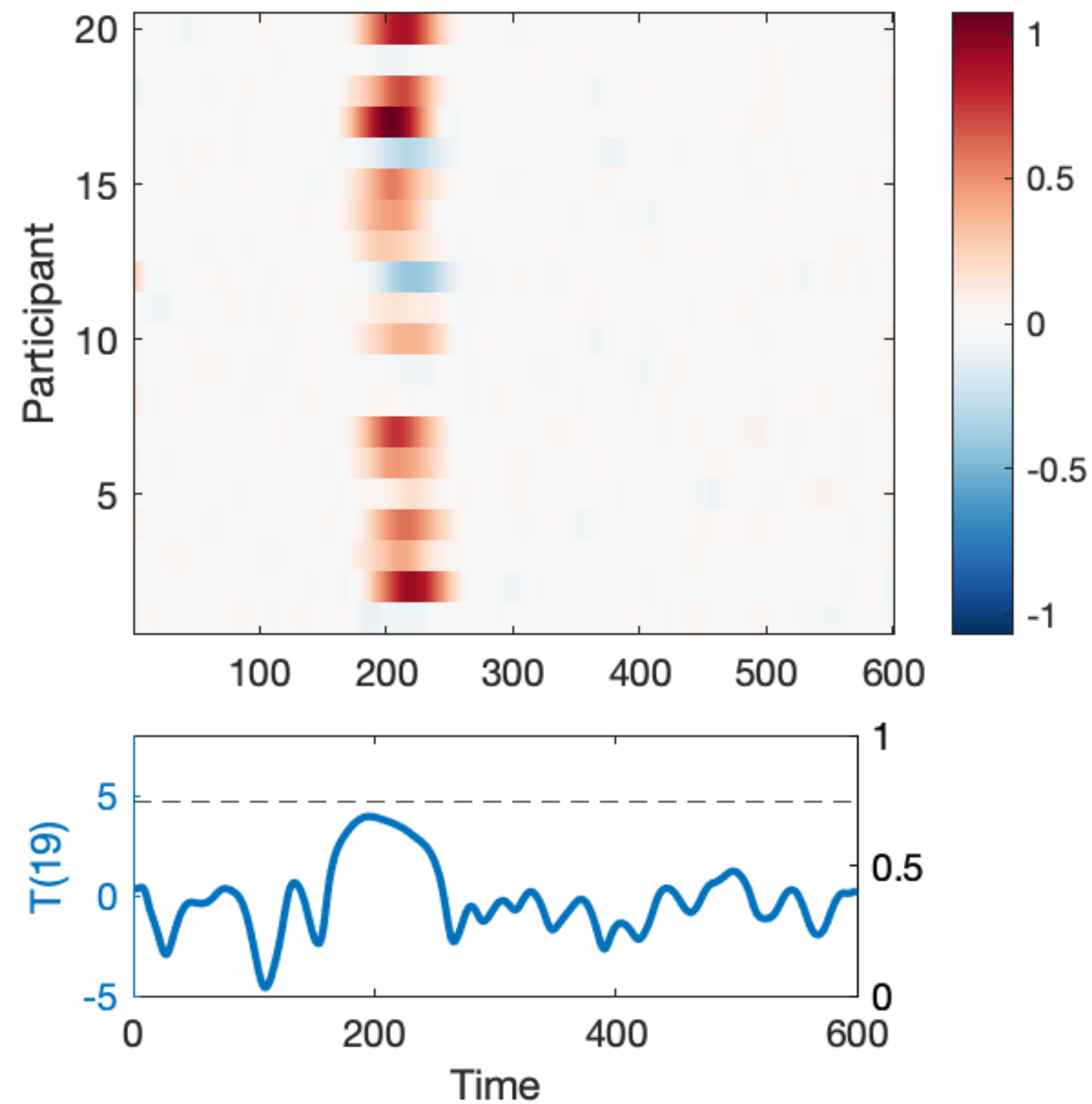
Example: Simulated EEG



Example: Simulated EEG

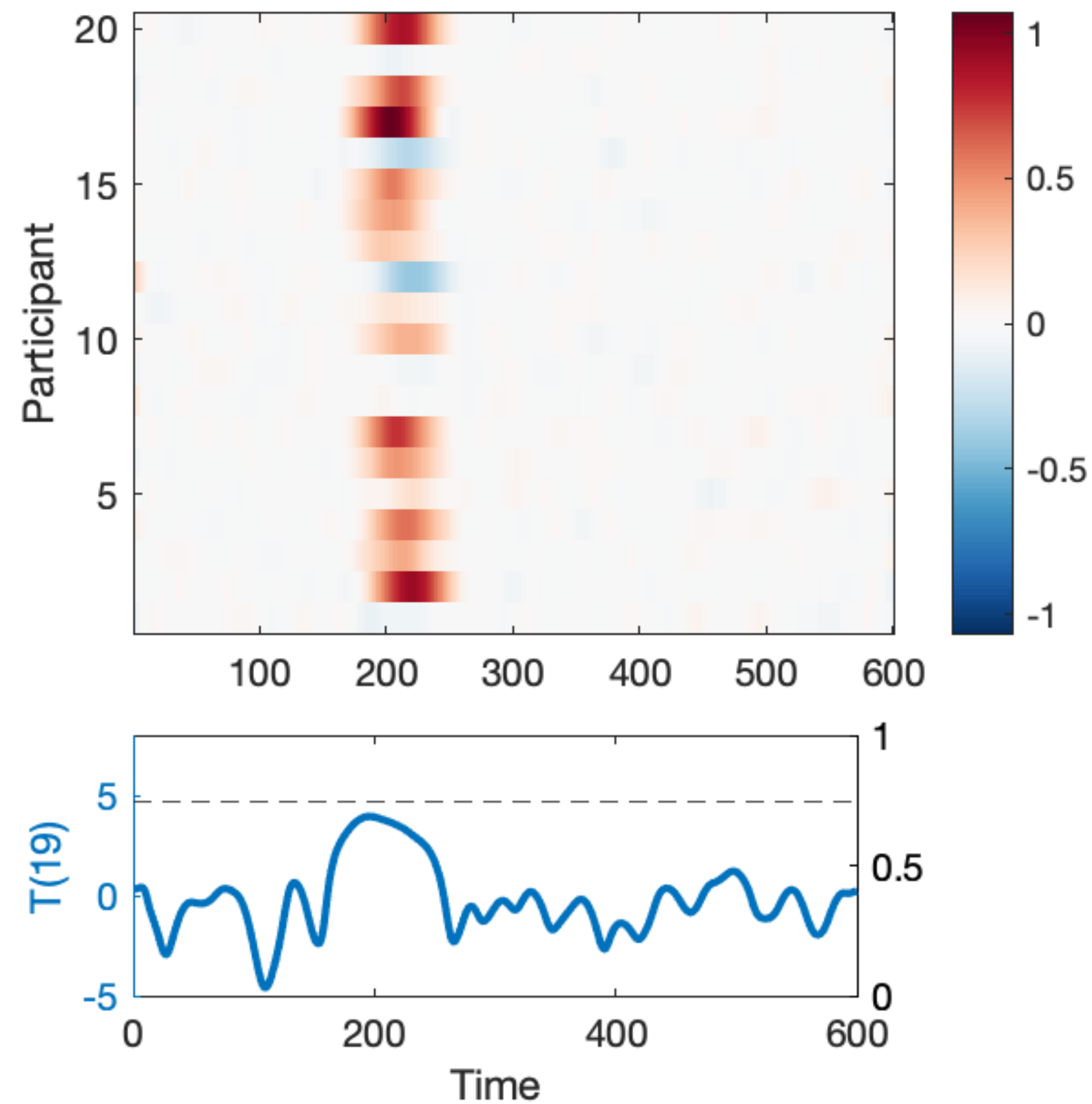


Example: Simulated EEG



- T-test at each time point
- Bonferroni correction for multiple comparisons
- $p = 0.05 / 600$ time points

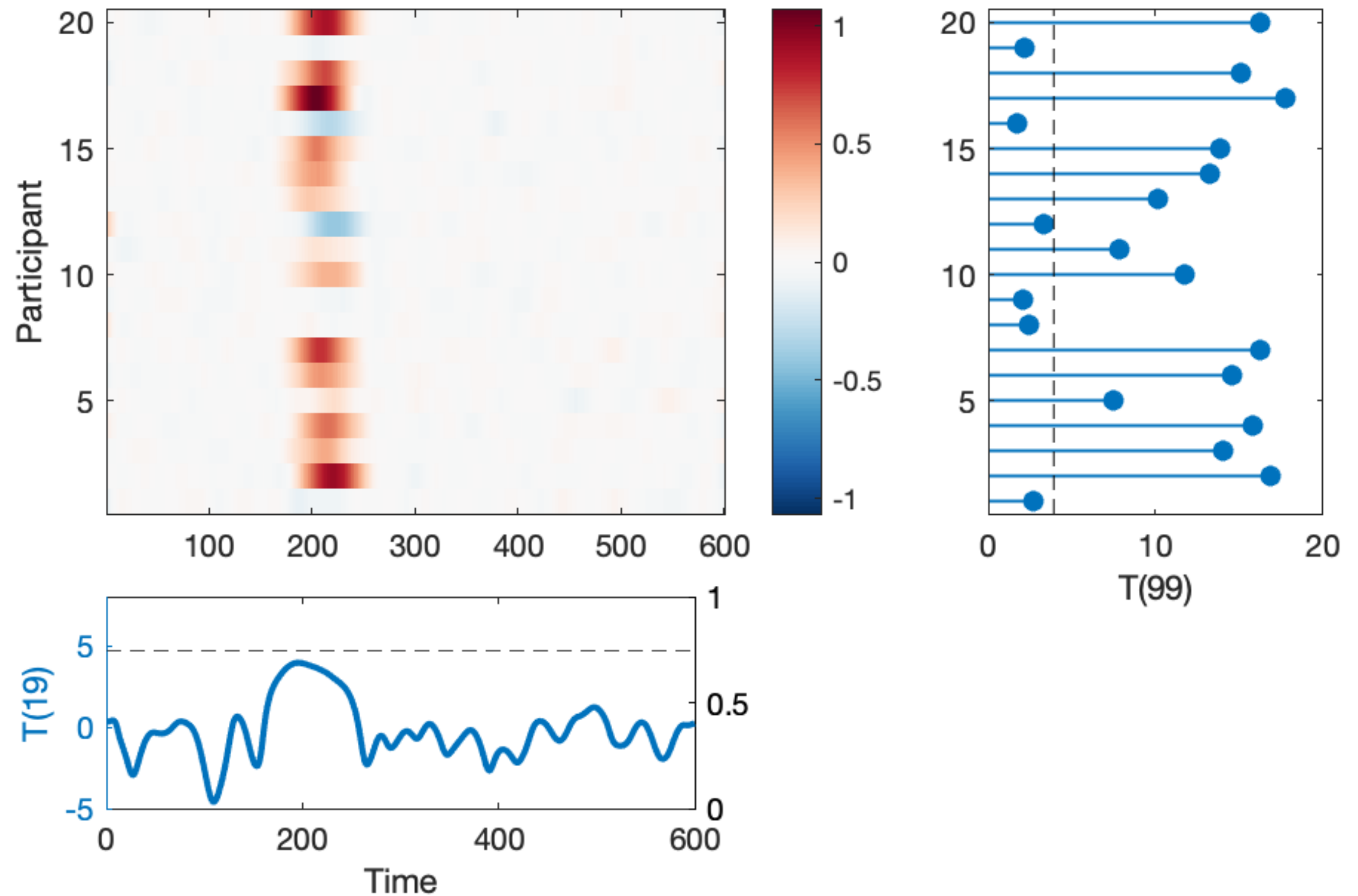
Example: Simulated EEG



- Variance across participants is too high for us to reject the null hypothesis at any time point
- Is there anything else we can learn about the population from this data?

Example: Simulated EEG

- Test the hypothesis in each participant separately
- Same Bonferroni correction
- Reject null hypothesis 14 / 20
- Does this tell us anything about the population?



Does this tell us anything about the population?

Yes!

- If the null hypothesis was true in every member of the population, then the probability of a positive result in our test is $p=0.05$
- Probability of 14 (or more) out of 20 participants testing positive under the *global null* (null true for every member of the population): $1 - \text{binocdf}(13,20,0.05) : 1.7 \times 10^{-14}$
- This is the p-value for a null hypothesis that the proportion of participants ***in the population*** who would show a true positive effect in this experiment ***is zero***
- $p = 1.7 \times 10^{-14}$: as surprising as 45 heads in a row (quite surprising, quite strong evidence by standards usually required for publication, $p=0.001$: ~10 heads in a row)

Population Prevalence

- The mean is not the only parameter of the population we can estimate quantitatively!
- Can think about the proportion of participants that show the effect, or the ***prevalence***

Prevalence Model

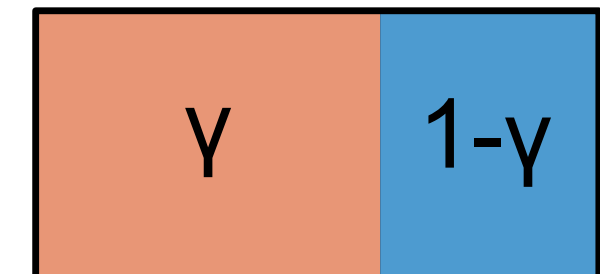
Need a model of the population

- Proportion γ of the population have a property, the remainder don't.
- If we could measure the property directly, this would be a simple binomial distribution
- But we measure with an error prone within-participant statistical test with false positive rate (significance level) α and sensitivity (1 - false negative rate) β
- So probability of a participant randomly sampled from the population testing positive is given by:

$$\theta = (1 - \gamma)\alpha + \gamma\beta$$

- And then our experimental sample is modelled with a binomial distribution:

$$P(X = k|\theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$



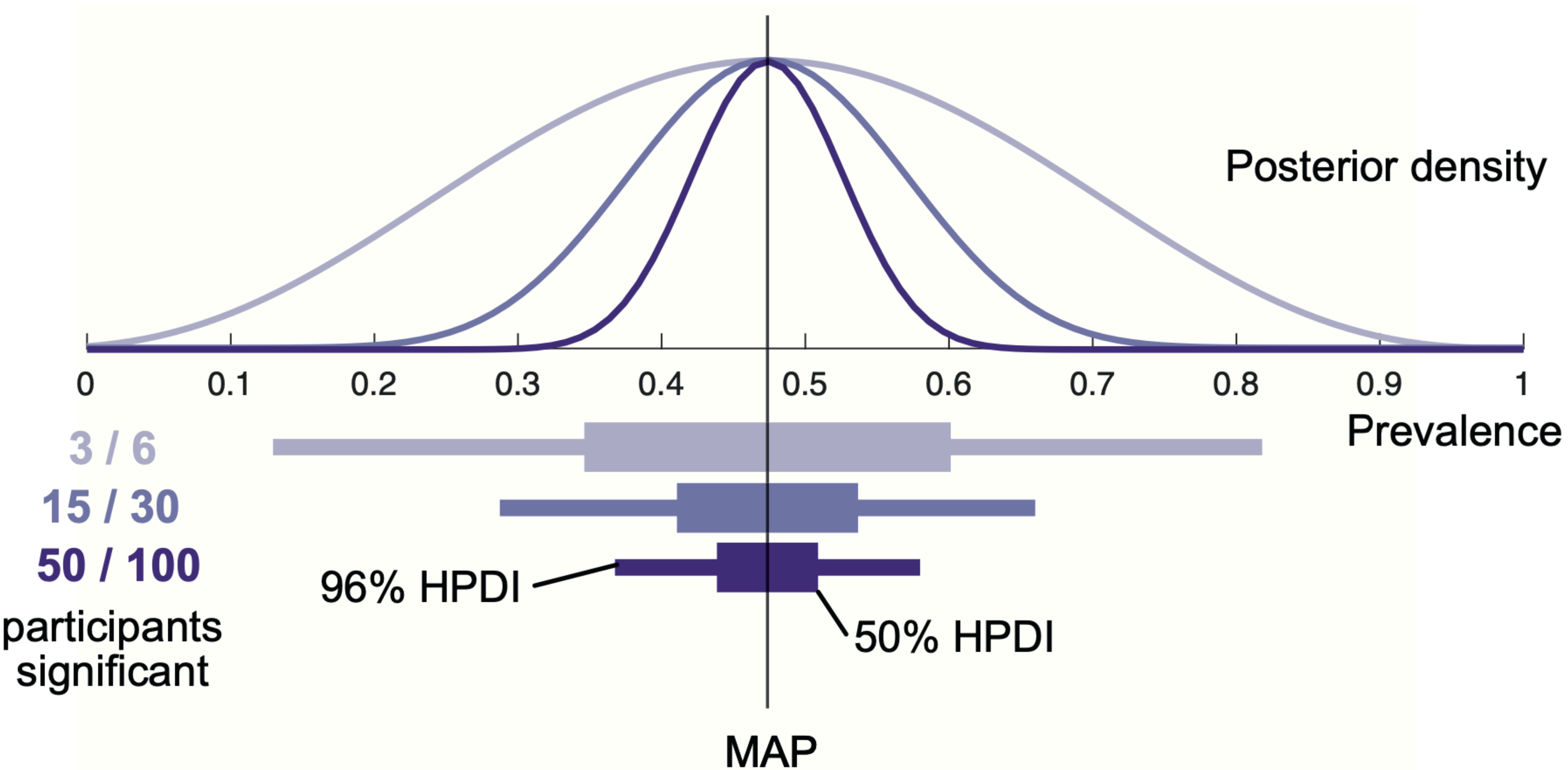
Bayesian Prevalence

<https://github.com/robince/bayesian-prevalence>

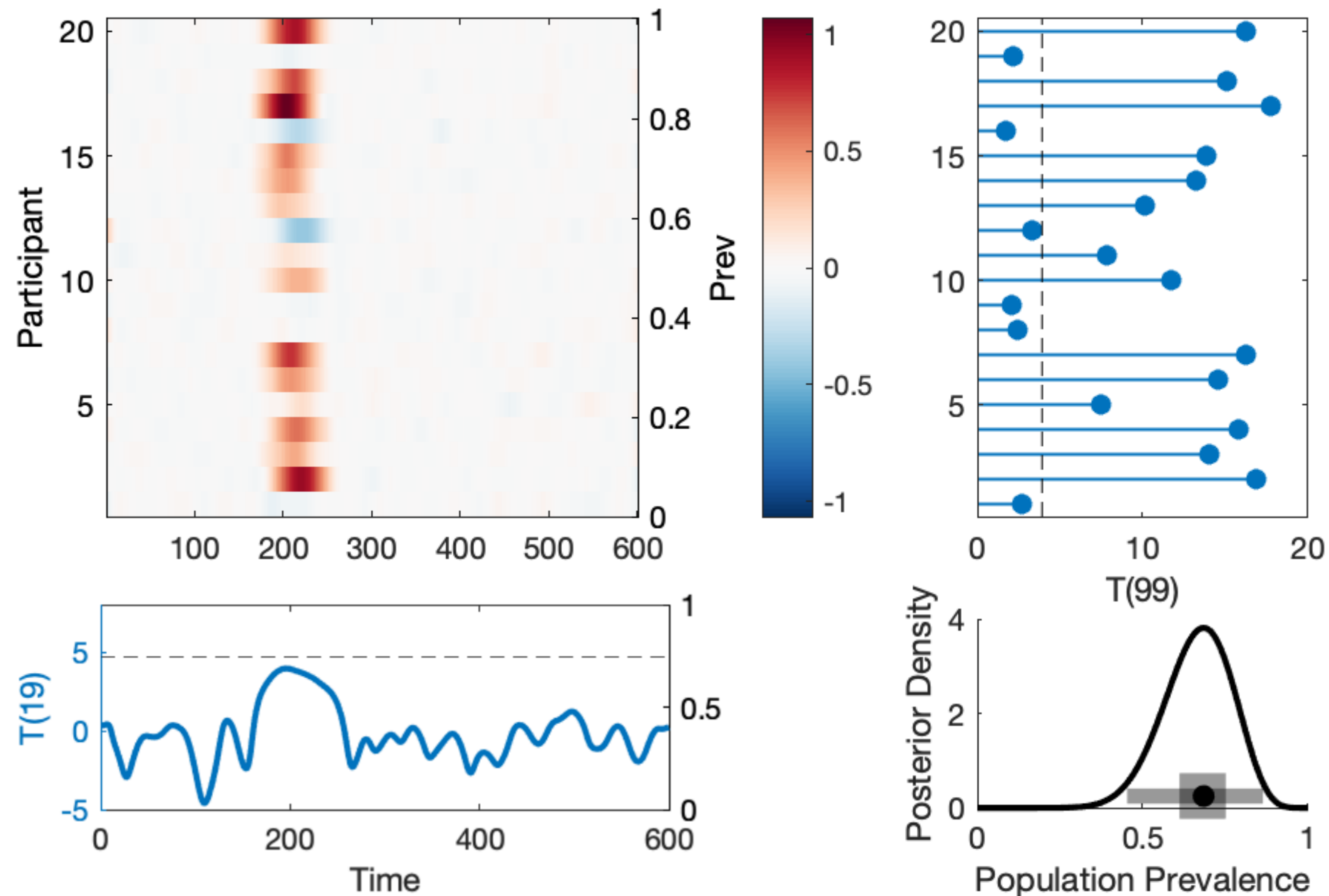
- Simple Bayesian estimation methods applied to the NHST prevalence model.
- Uniform prior on γ
- Get a full posterior distribution, reflecting belief in the possible population values, given the observed experiment
- Maximum A Posteriori (MAP) : most likely posterior value
- Highest Posterior Density Intervals (HPDI) : credible intervals in which the population value falls with specified probability

Bayesian Prevalence

Posterior, MAP and HPDI



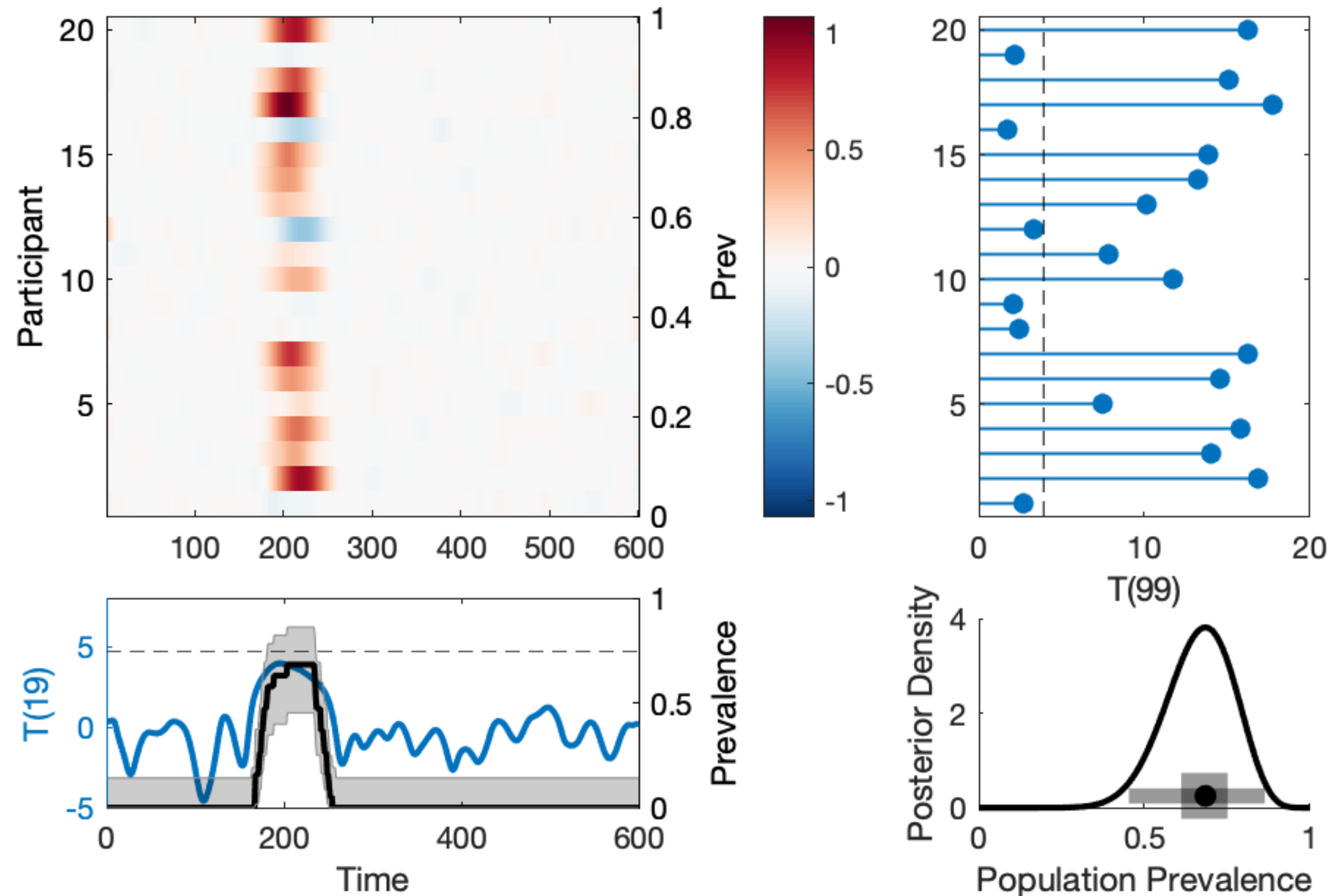
Example: Simulated EEG



- 14 / 20 at $p=0.05$
- MAP = 0.68
- 50% HPDI = [0.61 0.75]
- 96% HPDI = [0.45 0.86]

Example: Simulated EEG

Bayesian Prevalence at each time point



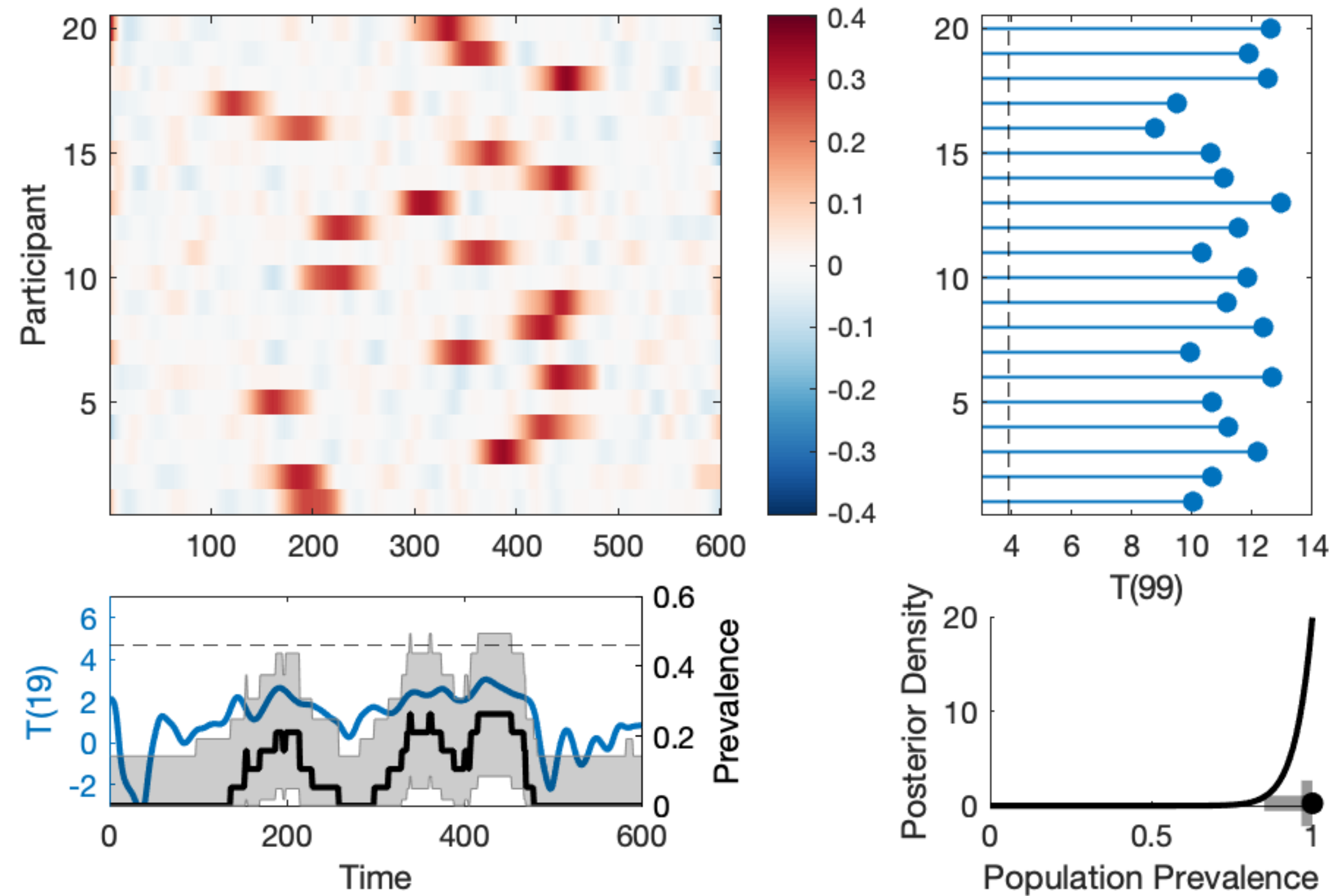
- Can localise the effect in time

Bayesian Prevalence

- Quantitative estimation of a population parameter: the prevalence
- Probability that a new randomly chosen participant would present a true positive result in a given experiment
- Within-participant replication probability!
- Quantified at the population level, with explicit uncertainty (not reducing to a brittle binary inference)

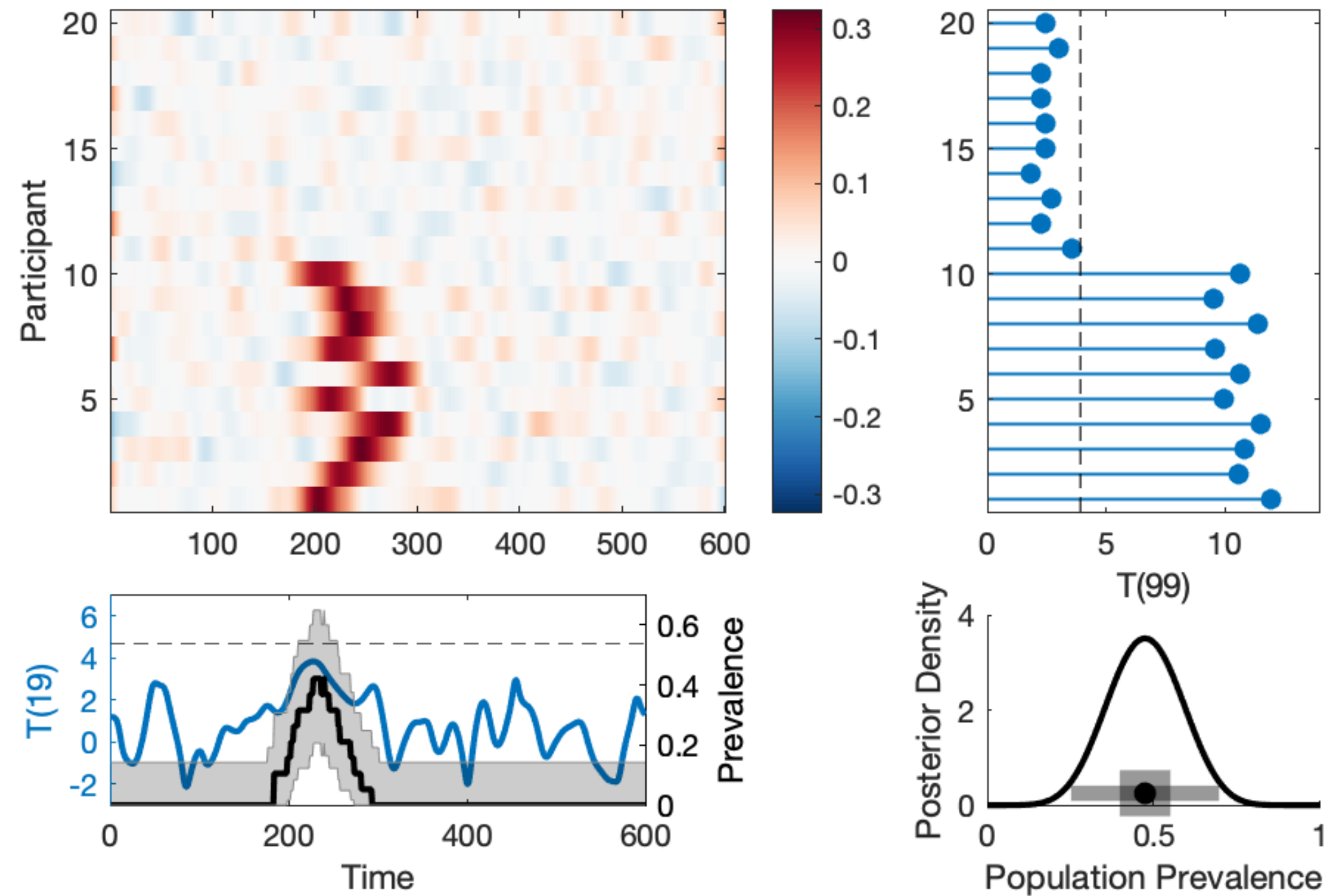
More motivating examples

Variable alignment



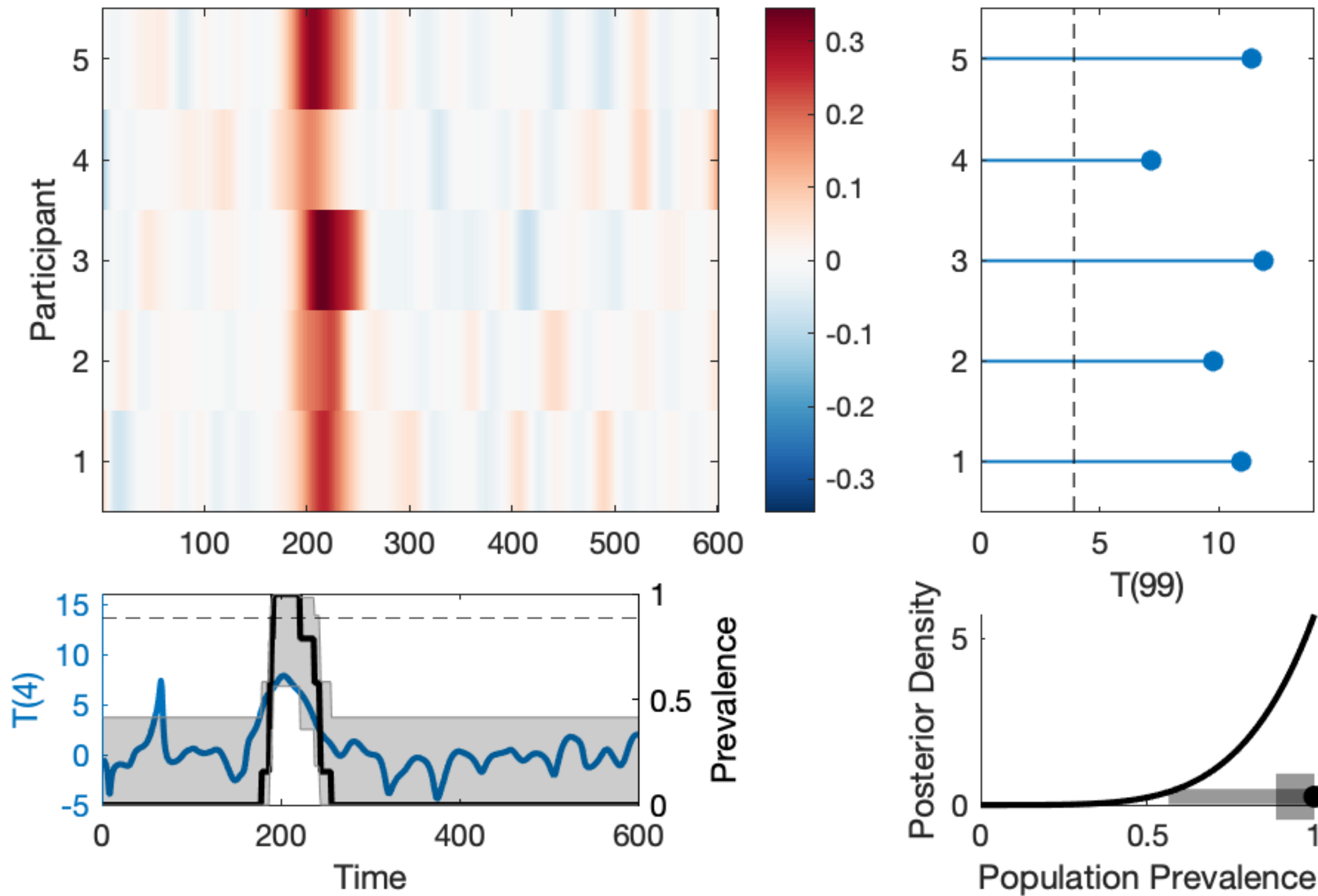
More motivating examples

Subgroups



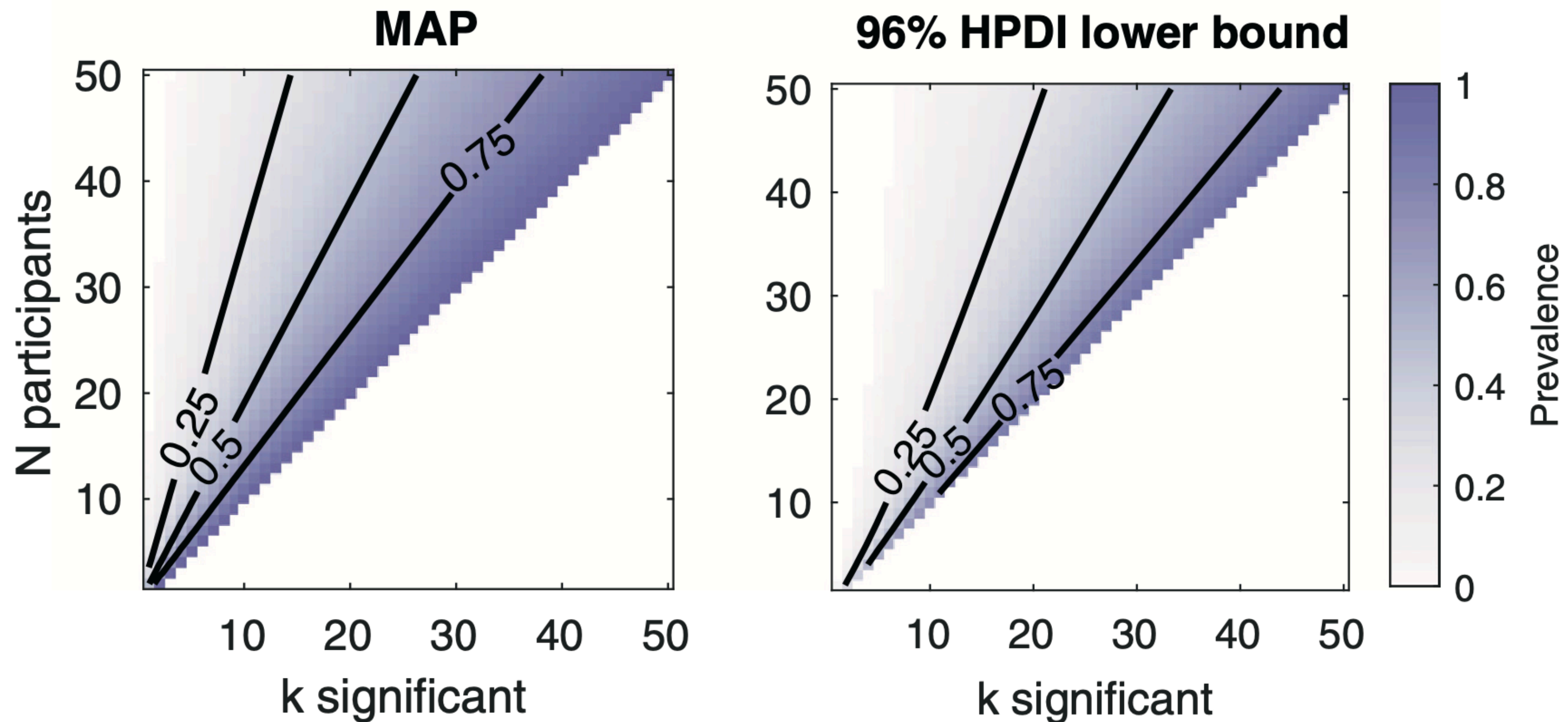
More motivating examples

Small N



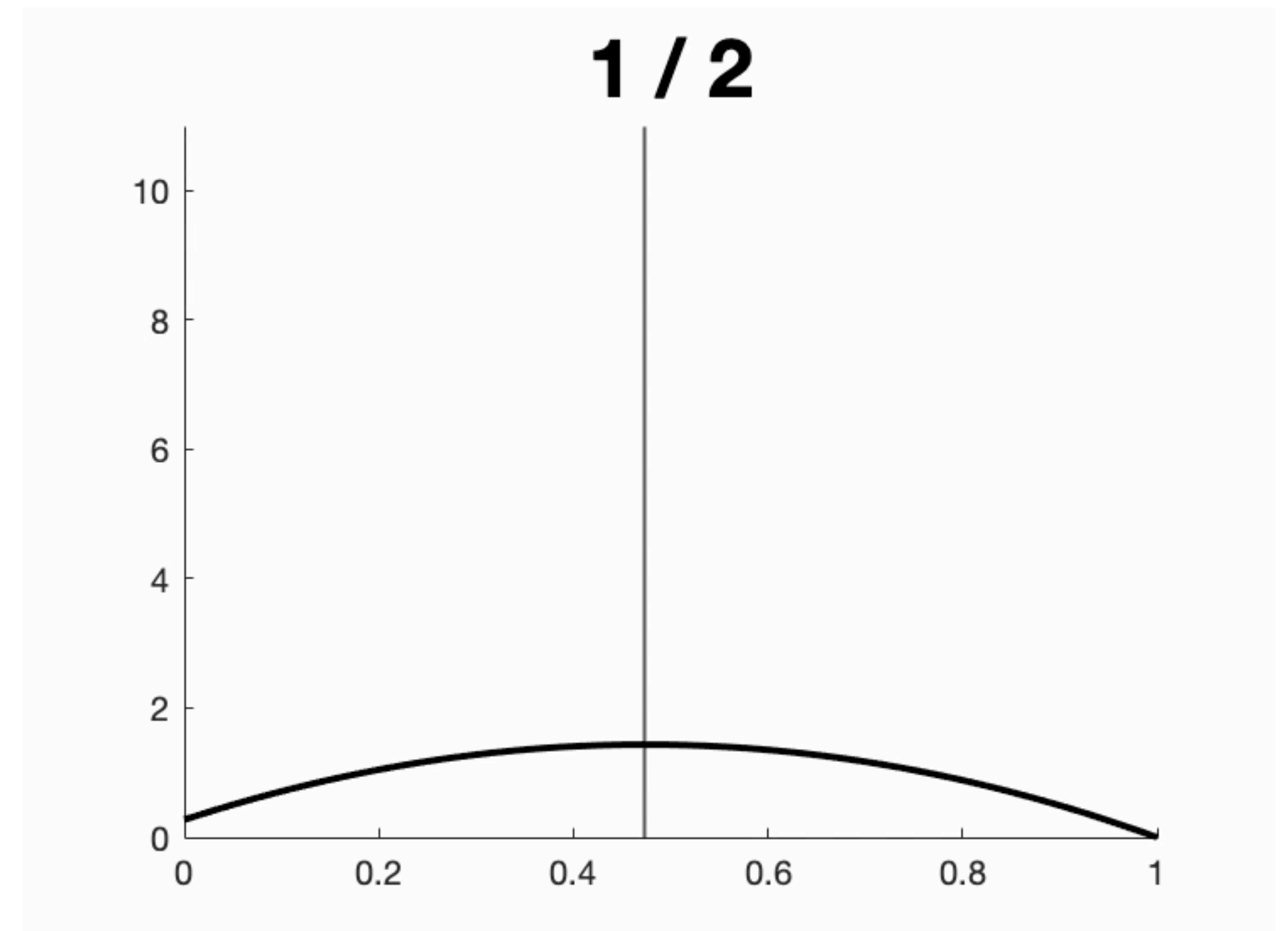
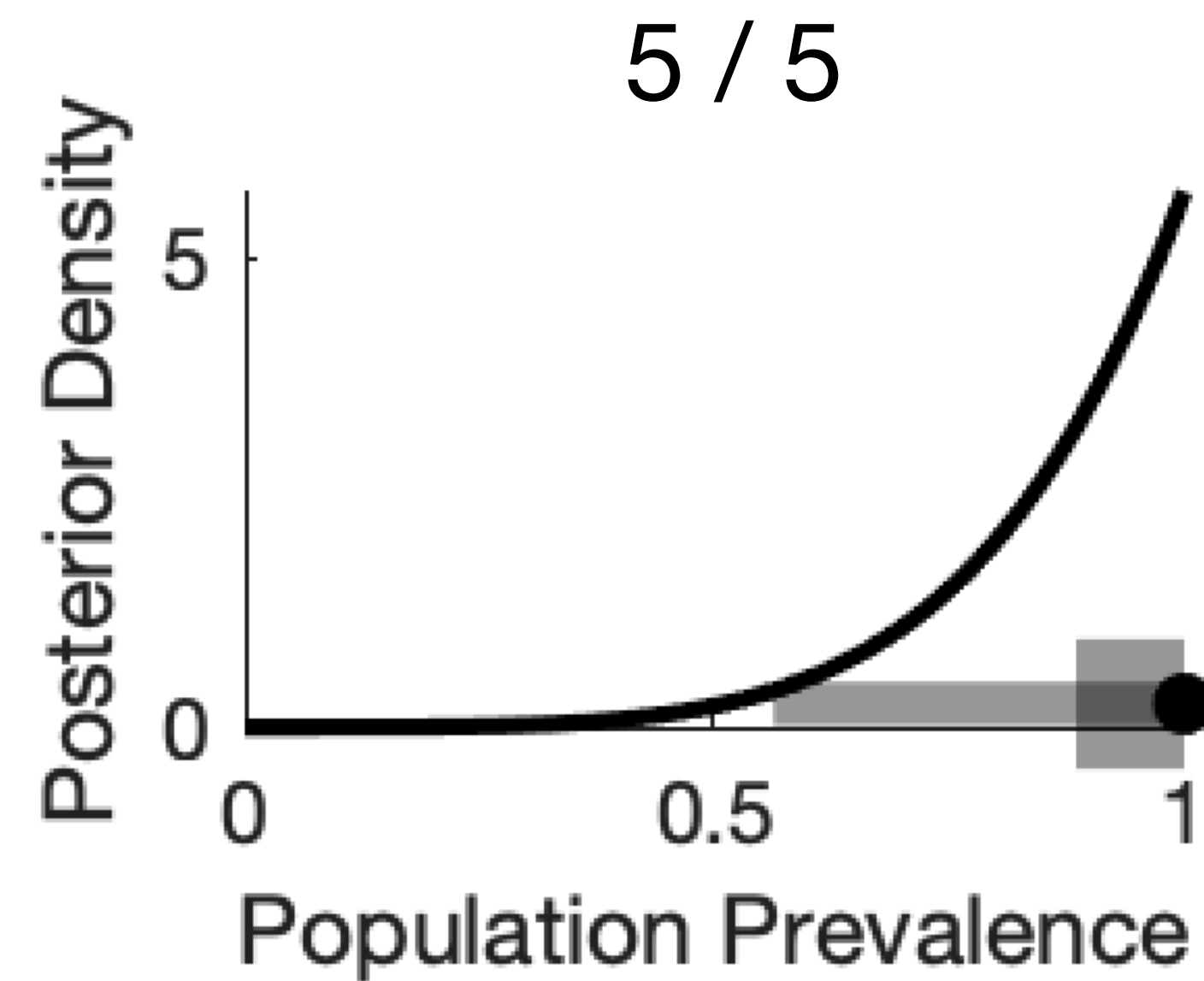
Scaling with Sample Size

- You can investigate this scaling in the tutorial.



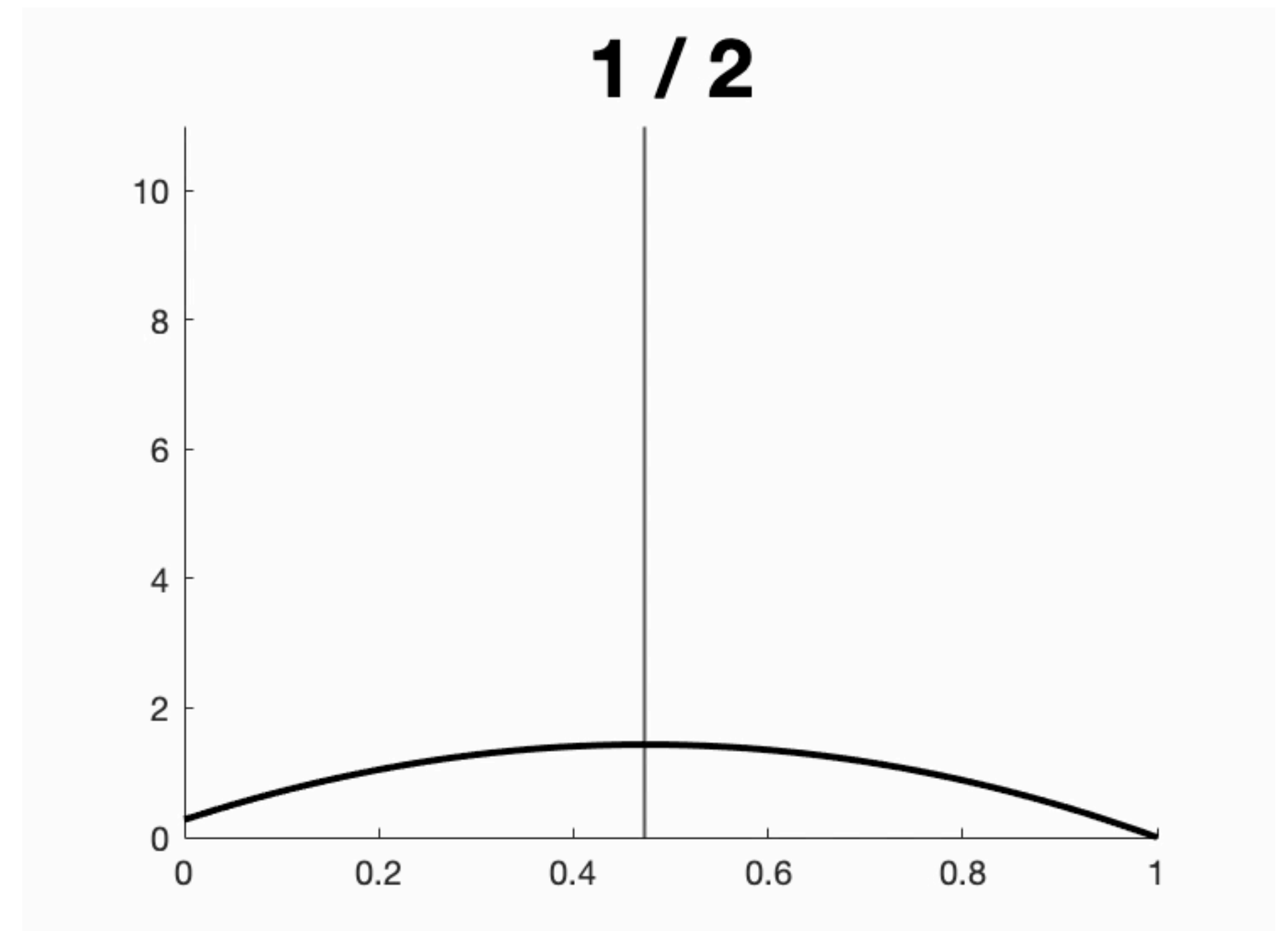
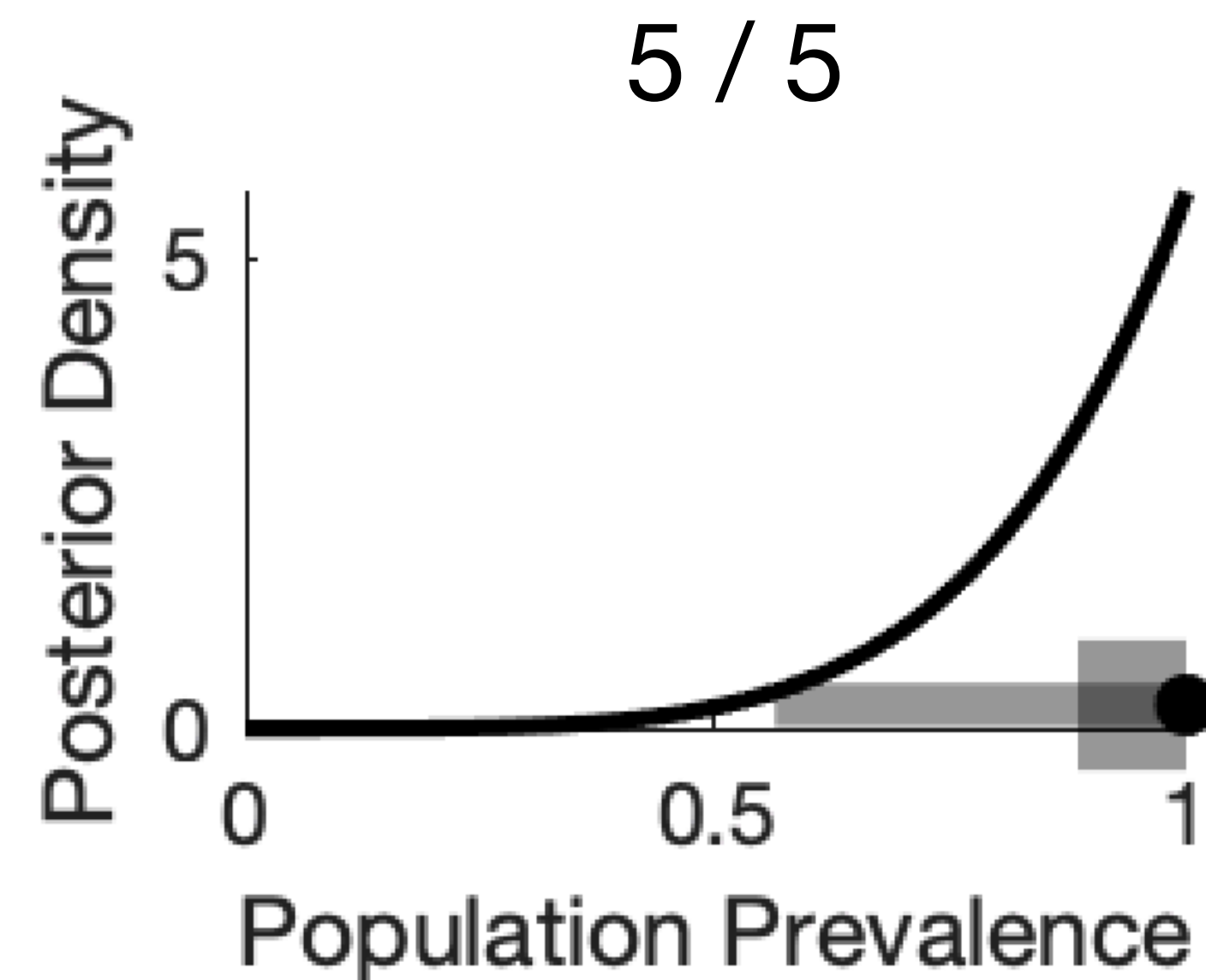
YOU CAN'T LEARN ANYTHING ABOUT THE POPULATION FROM 5 SUBJECTS ! RAAARRR!

But you can though



YOU CAN'T LEARN ANYTHING ABOUT THE POPULATION FROM 5 SUBJECTS ! RAAARRR!

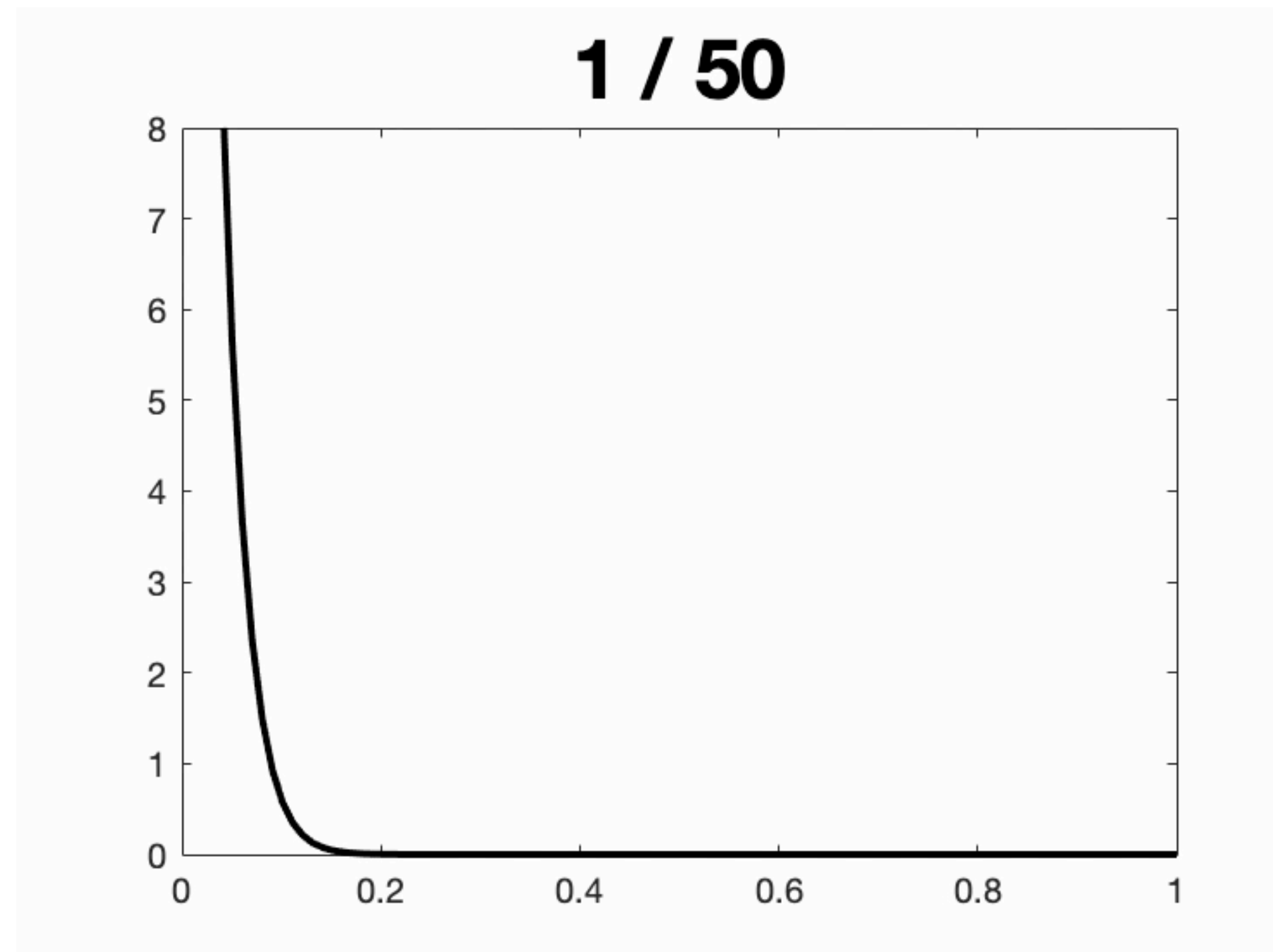
But you can though



Bayesian Prevalence Inference

Scaling Properties

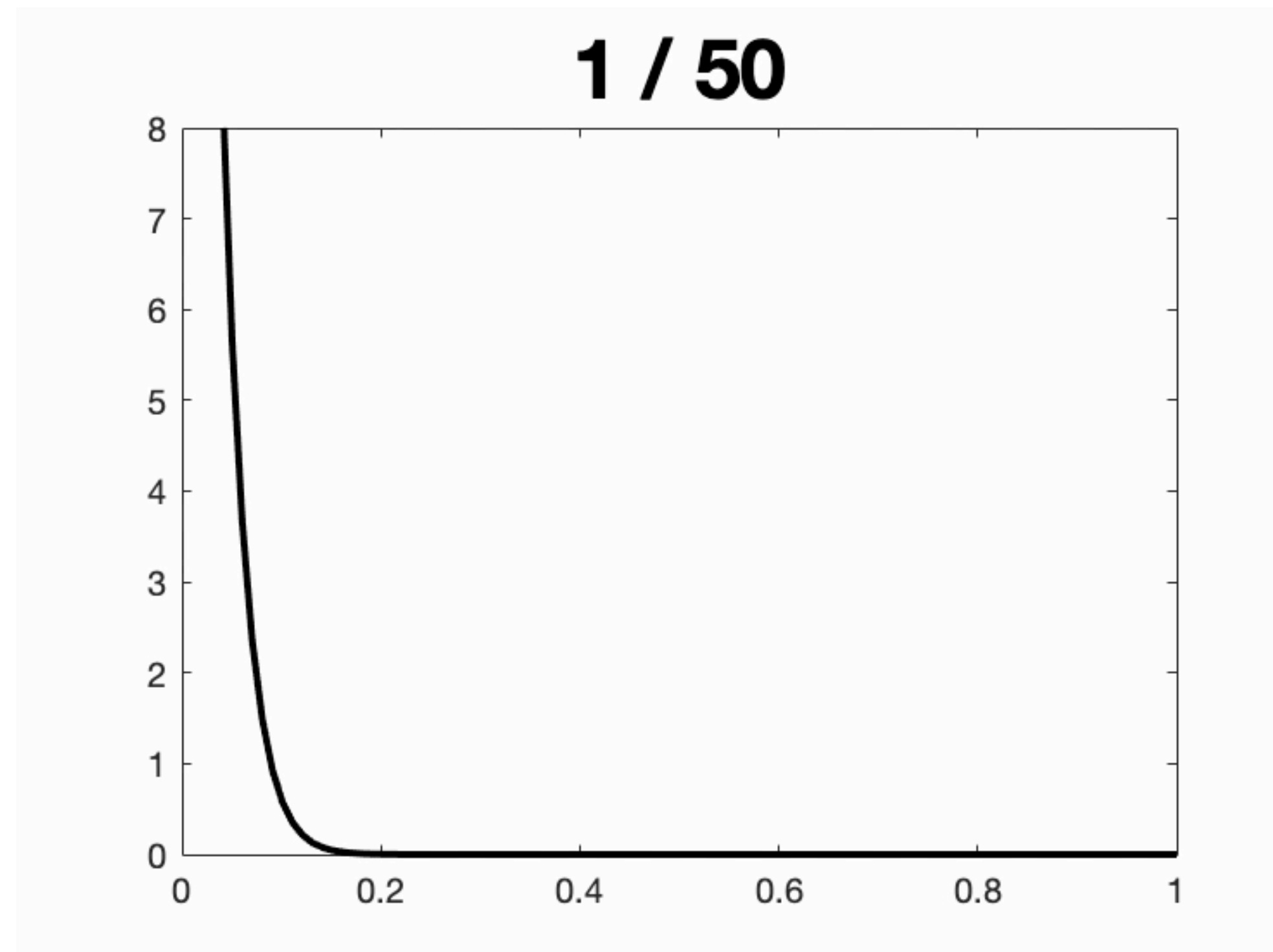
- Bayesian posterior for different numbers of positive test subjects



Bayesian Prevalence Inference

Scaling Properties

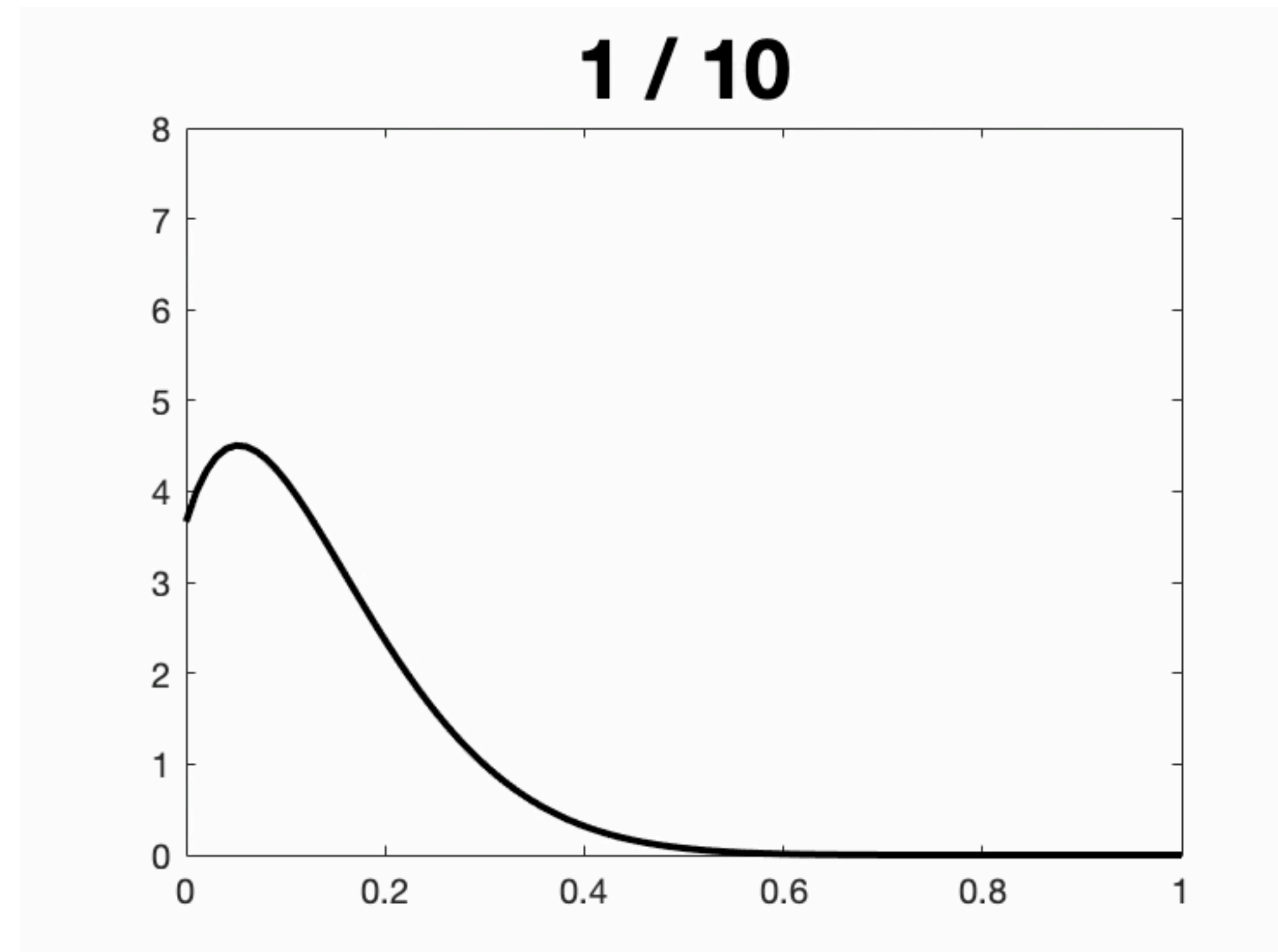
- Bayesian posterior for different numbers of positive test subjects



Bayesian Prevalence Inference

Scaling Properties

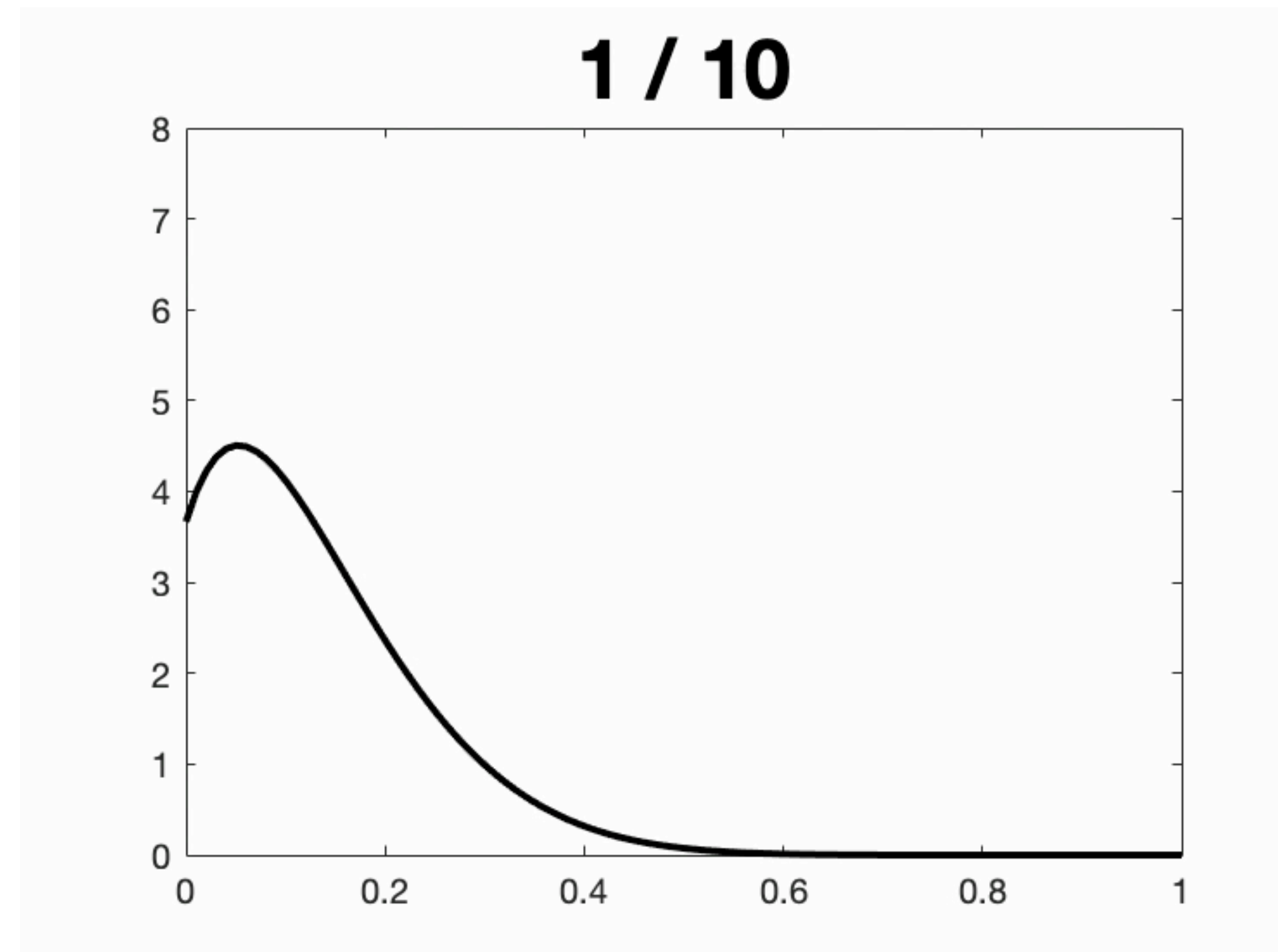
- Bayesian posterior for different numbers of positive test subjects



Bayesian Prevalence Inference

Scaling Properties

- Bayesian posterior for different numbers of positive test subjects



Interim Summary

Bayesian Prevalence

- Alternative perspective to population mean inference
- Provides a bridge from within-participant testing to quantitative statements about the population.
- Can apply to any within-participant test or model (power contrasts, encoding/decoding models, RSA, behavioural models)
- Output is a quantitative Bayesian estimate with associated uncertainty (not a binary significance result)

```
map = bayesprev_map(k, Nsub, a);  
pp = bayesprev_posterior(linspace(0,1,100), k ,Nsub, a);  
h = bayesprev_hpdi(0.96, k, Nsub, a);
```

Other functions

- Difference in prevalence for two tests applied to the same sample
- Difference in prevalence for the same test applied to samples from two different populations
- Prevalence as a function of effect size (don't have to do $p=0.05$ NHST within-participant)

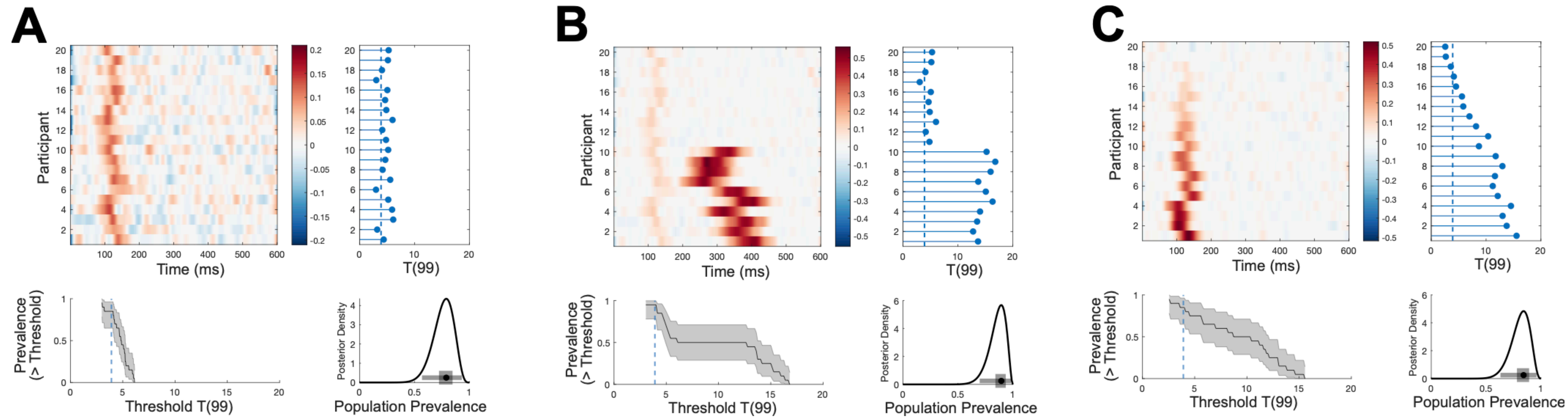
Other functions

- Difference in prevalence for two tests applied to the same sample
- Difference in prevalence for the same test applied to samples from two different populations
- **Prevalence as a function of effect size (don't have to do $p=0.05$ NHST within-participant)**

Prevalence as a function of effect size

- $p=0.05$ is an arbitrary choice of threshold.
- What if we chose a different threshold?
- We just need to know the false positive rate of exceeding the threshold “by chance” (i.e. under a suitable null of no within-participant effect)
- Calculations in the tutorial

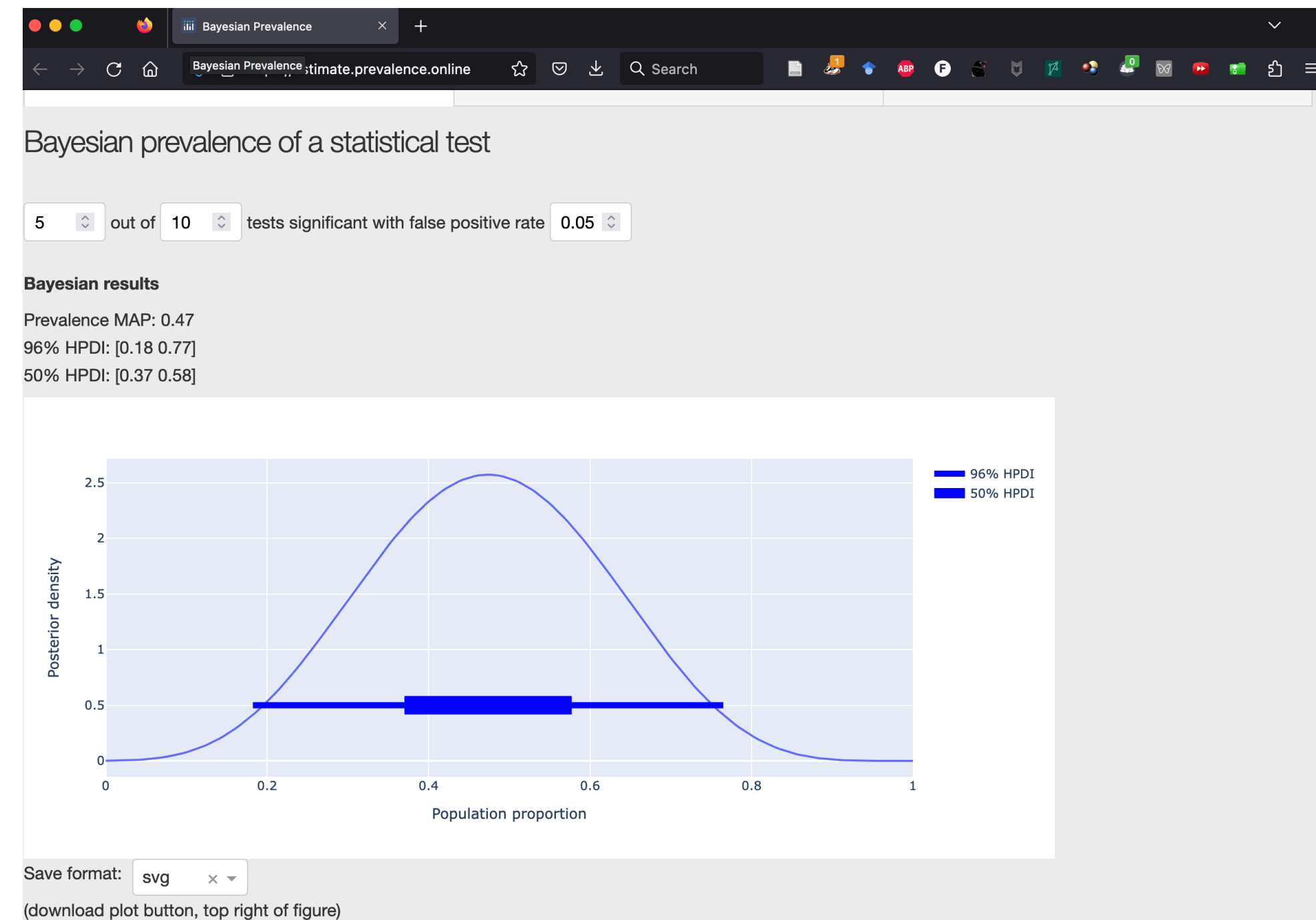
Prevalence as a function of effect size



Code and WebApp

- Code in Matlab / Python / R (simple functions):
<https://github.com/robince/bayesian-prevalence>

- Webapp
<https://estimate.prevalence.online/>



If you want to estimate prevalence online just go to:

<https://estimate.prevalence.online/>



Summary

- What we learn about the world is shaped by the scientific methods we use to study it: neuroimaging, almost exclusively uses ***population mean random effects***
- Prevalence is alternative way of thinking about what we learn about the population from an experiment, with two key advantages
- **Robustness:** replicating an effect in several participants is a stronger and more robust result than current standards of evidence (c.f. noninvasive brain stimulation)
- **Increased sensitivity to heterogenous effects:** we may be completely missing effects that are too heterogenous to have significant population mean - these neural effects which are more variable across individuals may be more useful as biomarkers (c.f. simulated EEG examples in the talk)

Acknowledgements

- Philippe Schyns
- Jim Kay

Discussion Time!

Discussion

Timeliness

Editorial

Consideration of Sample Size in Neuroscience Studies

parameters. While many of these practices typically rely on large sample sizes, some areas of neuroscience make statistical inferences on individual subjects, implementing a sort of exploration-then-estimation procedure across successive subjects (e.g., patients or nonhuman animal models in electrophysiology; machine-learning explorations of fMRI data; psychophysics and human brain lesion studies). These small-N approaches focus their statistical power on individual-level characterization of an effect; a finding is deemed present when all or a majority of a small pool of subjects show an effect, usually based on a large sample of trial-level observations (Smith and Little, 2018). It should be acknowledged that this approach only allows for statements that pertain to the existence and magnitude of effects in those subjects, rather than in the populations those subjects are drawn from. Many of the most robust findings in psychophysics have come from a small-N approach (Smith and Little, 2018), and it could be preferred ethically when animal welfare or vulnerable individuals are involved.

Consideration of Sample Size in Neuroscience Studies

Discussion

Timeliness




parameters. While many of these practices typically rely on large sample sizes, some areas of neuroscience make statistical inferences on individual subjects, implementing a sort of exploration-then-estimation procedure across successive subjects (e.g., patients or nonhuman animal models in electrophysiology; machine-learning explorations of fMRI data; psychophysics and human brain lesion studies). These small-N approaches focus their statistical power on individual-level characterization of an effect; a finding is deemed present when all or a majority of a small pool of subjects show an effect, usually based on a large sample of trial-level observations (Smith and Little, 2018). It should be acknowledged that this approach only allows for statements that pertain to the existence and magnitude of effects in those subjects, rather than in the populations those subjects are drawn from. Many of the most robust findings in psychophysics have come from a small-N approach (Smith and Little, 2018), and it could be preferred ethically when animal welfare or vulnerable individuals are involved.

Individual Participant

Discussion points

What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis

Maxwell L. Elliott¹ , Annchen R. Knodt¹, David Ireland², Meriwether L. Morris¹, Richie Poulton², Sandhya Ramrakha², Maria L. Sison¹, Terrie E. Moffitt^{1,3,4,5}, Avshalom Caspi^{1,3,4,5} , and Ahmad R. Hariri¹ 

- In neuroimaging publication pressure has selected for effects with low between participant variance, therefore not suited for individual difference studies (Elliot et al. 2020) **or use as biomarkers (?)**
- Samples of patients may have particularly heterogenous neural or behavioural effects, and be difficult to recruit. In this regime prevalence might be able to statistically identify candidate biomarker effects that are invisible to the group mean.

Individual Participant

Discussion points

- Example: brain stimulation. What is more relevant, average effect size if the stimulation was applied to everyone, or the proportion of people who have a change that is measurable in a 30 minute task after 20 minutes of stimulation?
- Any practical application (of neuroimaging, neurostimulation, behavioural interventions) will ultimately depend on the degree to which effects are reliable **within individuals**

Discussion Points

Within-participant statistics

- Solves the **replication crisis**! (or at least reduces it a lot)
- The **individual participant** is the most relevant replication unit for cognitive science (Smith & Little, 2018; Thiebaut de Schotten et al., 2017)

Psychonomic Bulletin and Review (2018) 25:2083–2101
<https://doi.org/10.3758/s13423-018-1451-8>

THEORETICAL REVIEW

Small is beautiful: In defense of the small-*N* design

Philip L. Smith¹ · Daniel R. Little¹

Special issue: Editorial

Identical, similar or different? Is a single brain model sufficient?

Michel Thiebaut de Schotten^{a,b,c,*} and Tim Shallice^{d,e,*}

- Interpretation is grounded to the **specific experiment**, reducing temptation to over-generalise (Yarkoni, 2020, *The Generalizability Crisis*, BBS; Broers 2021)

<https://www.discovermagazine.com/the-sciences/are-effect-sizes-in-psychology-meaningless>

Replication Crisis

Discussion points

- Graded quantitative output not binary NHST
- Replication is built in
- Protection against researcher degrees of freedom
- More severe test of hypothesis (Mayo, 2018, *Statistical inference as severe testing*)
- Don't need to abandon null hypothesis testing or redefine scientific frameworks
- Some effects might not be strong enough to detect in individuals, but there might be others that are missed in population mean because of being too variable

Replication Crisis

Discussion points

- Graded quantitative output not binary NHST
- Replication is built in
- Protection against researcher degrees of freedom
- More severe test of hypothesis (Mayo, 2018, *Statistical inference as severe testing*)
- Don't need to abandon null hypothesis testing or redefine scientific frameworks
- Some effects might not be strong enough to detect in individuals, but there might be others that are missed in population mean because of being too variable



Jack Gallant
@gallantlab

Think of it this way. If you can show an effect in a single subject, then each additional single-subject result is a replication of your experiment! The only drawback of small N studies is generalization ambiguity, but IMHO that is a secondary consideration.

9:54 PM · Mar 27, 2018 · [Twitter Web Client](#)

Individual Participant

Discussion points

- “if psychology is to be a mature quantitative science, its primary theoretical aim should be to investigate systematic, functional relationships as they are manifested at the individual participant level and that, wherever possible, it should use methods that are optimized to identify relationships of this kind” (Smith and Little, 2018)
- “It is more useful to study one animal for 1000 hours than to study 1000 animals for one hour” — B. F. Skinner (quoted in Smith and Little, 2018)



much of what makes good science. To us, it is a source of irony that, in the current climate of uncertainty and methodological re-evaluation, studies that embody what we believe are characteristics of good science can be rejected by journal editors as a priori “unreliable.” We therefore wish to challenge the reductive view that the only route to reliable psychological knowledge is via large samples of participants. Parenthetically, we note that while



Individual Participant

Discussion points

- Alignment (ECoG, iEEG, sEEG, 7T fMRI)
- Small samples: dense sampling, precision imaging, deep imaging; clinical studies
- Discovery led (new recording modalities).
 - No power analysis for multivariate neuroimaging.
 - Researcher degrees of freedom.
- Normal distribution with large variance (common) implies existence of participants with strong effects (in both directions if mean close to 0)
- Think about your population model and its implications (and look at variance as well as mean in LMEM)

Individual Participant

Alignment

- Provided FWER is controlled within a region per-participant, can combine results for population inference without having precise alignment.
- For example, defining anatomical ROI in each participant, can model an fMRI contrast voxelwise controlling FWER over the region, and then report at the population level the prevalence of an activation in that region, without requiring cross-participant overlap at the voxel level (i.e. without requiring smoothing, functional alignment or averaging signal within the region).
- Or think about a parametric effect of decision confidence in post-stimulus EEG alpha power. As long as within participant inference has FWER controlled over time points, can report the prevalence of subjects having an effect between 500-1000ms post-stimulus, without requiring all subjects to overlap
- Better way to deal with high between participant variability? (e.g. laminar high field fMRI)

Interpretation grounded to specific experiment

Discussion points

- Direct and easily interpretable population estimate of replication probability, with uncertainty.
- Avoids temptation to misinterpret NHST results (Greenland et al, 2016) or overgeneralise (Yarkoni, 2020)
- Effect sizes automatically relevant (constrained by the experiment). So even large online studies no issue of significant but not meaningful effects.
- Consider experimental time instead of effect size. Prevalence of effect in 1h experiment vs 10mins vs 10hours

Interpretation grounded to specific experiment

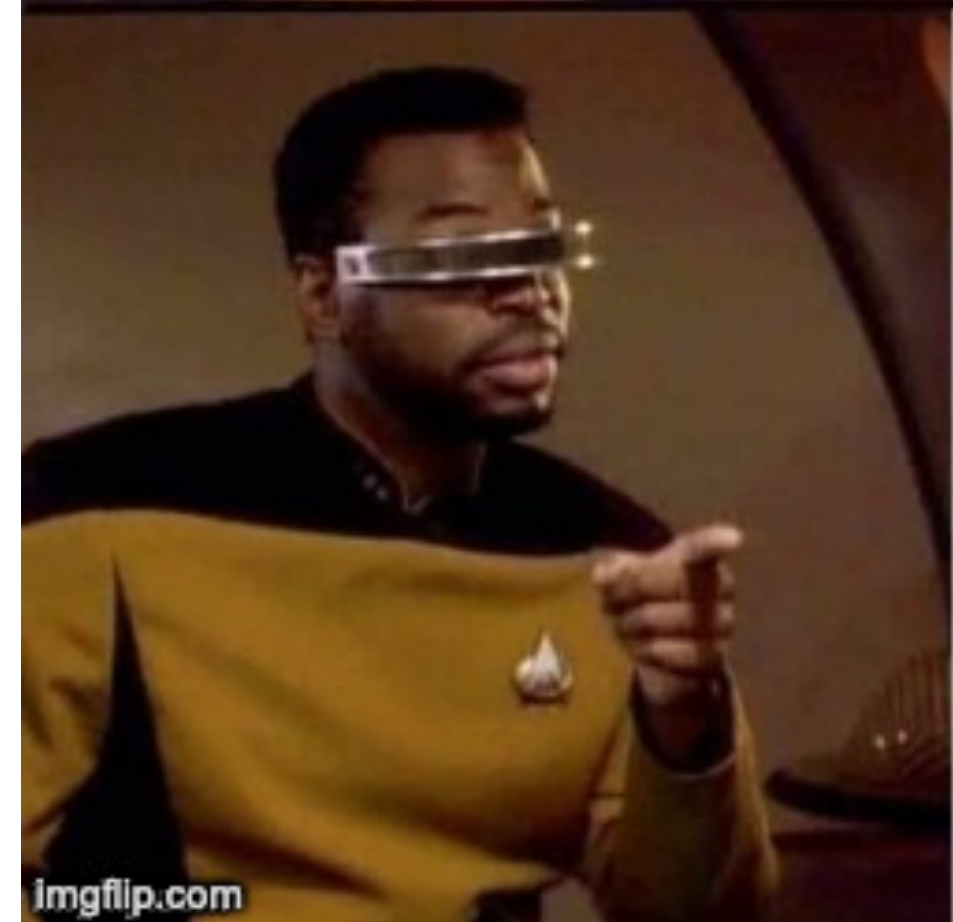
- “From a generalizability standpoint, then, the key question is how closely the verbal and quantitative expressions of one’s hypothesis align with each other.” Yarkoni (2020) BBS
- Reject population mean null hypothesis: “there is an effect” (prone to overgeneralisation)
- Prevalence: XX% of the population would show an effect greater than Y in this particular experiment. (not prone to overgeneralisation - directly linked to the experimental design and effect size)

Interpretation grounded to specific experiment

Discussion points



**SMALLEST
EFFECT SIZE
OF INTEREST**



**SHORTEST
EXPERIMENT
OF INTEREST**

When the Numbers Do Not Add Up: The Practical Limits of Stochastologicals for Soft Psychology

Perspectives on Psychological Science
2021, Vol. 16(4) 698–706
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1745691620970557
www.psychologicalscience.org/PPS


Nick J. Broers 

Department of Methodology and Statistics, Faculty of Psychology and Neuroscience, Maastricht University

- “effect sizes do not really exist independently of the adopted research design that led to their manifestation”
- “These statistics help to establish an observed effect size, but that effect size, it is crucial to realize, did not have an independent existence in nature before the theorist constructed this research design as a means for eliciting a predicted (ordinal) effect”
- “The arbitrariness of the outcome scales reflects the indifference of the researchers toward whatever quantitative outcome the study might yield. The only purpose of quantification was to enable [significance testing] to underwrite the ordinal theoretical prediction. The conclusion must then be that observed effect sizes have no meaning outside the research design in which they were established.”

Take home message

- Think about your population model
- Think about what you want to learn about the population
- Population average? Or effects within individuals?

Questions?

