Bayesian inference of population prevalence

Robin A. A. Ince[1], Jim W. Kay[2], Philippe G. Schyns[1,3]

[1] Institute of Neuroscience and Psychology, University of Glasgow
[2] Department of Statistics, University of Glasgow
[3] School of Psychology, University of Glasgow

Abstract

Within neuroscience, psychology and neuroimaging, the most frequently used statistical approach is null-hypothesis significance testing (NHST) of the population mean. An interesting alternative is to perform NHST within individual participants and then infer, from the proportion of participants showing an effect, the prevalence of that effect in the population. We propose a novel Bayesian method to estimate such population prevalence which has several advantages over population mean NHST. First, it provides a population level inference currently missing for designs with small numbers of participants such as traditional psychophysics, animal electrophysiology and precision imaging. Second, it delivers a quantitative estimate with associated uncertainty instead of reducing an experiment to a binary inference on a population mean. Bayesian prevalence is widely applicable to a broad range of studies in neuroscience, psychology and neuroimaging. Its emphasis on detecting effects within individual participants could help address replicability issues. To facilitate the applicability of Bayesian prevalence, we provide code in Matlab, Python and R.

Introduction

Within neuroscience, psychology and neuroimaging the common experimental paradigm is to run an experiment on a sample of participants and then infer and quantify any effect of the experimental manipulation in the population from which the sample was drawn. For example, in a psychology experiment a particular set of stimuli (e.g. visual and/or auditory stimuli) might be presented to a sample of human participants who are asked to categorize the stimuli or perform some other task. Each participant repeats the procedure a number of times with different stimuli (experimental trials) and their responses and reaction times are recorded. In a neuroimaging experiment, the same procedure is employed while neuroimaging signals are recorded in addition to behavioural responses. The researcher analyses these responses in order to infer something about the population from which the sample of participants were drawn. To simplify terminology and fix ideas, in the remainder of this article we focus on a broad class of experiments in psychology and neuroimaging which feature human participants and non-invasive recording modalities. This experimental paradigm encompasses a large range of experimental setups including psychophysics, categorisation and perception, with ongoing dynamic stimulation, multi-modal neuroimaging and complex behavioural tasks. However, we emphasise that our arguments are general and apply equally to other experimental model organisms or sampling units (e.g. a sample of single unit neural recordings from a population of neurons within a certain brain region).

In this standard experimental paradigm, the implicit goal is usually to determine the presence of a causal relationship between the experimental manipulation and the response of interest. For example, between a stimulus property, and the neural activity in a particular brain region, as reflected through non-invasive neuroimaging signals, or between neural activity and behaviour. A properly controlled experimental design in which other extraneous factors are held constant (i.e. a randomised control trial) enables a causal interpretation of a correlational relationship (Angrist and Pischke, 2014; Pearl, 2009). We use tools from statistics to evaluate the measured effect, and to ensure we are not being fooled by randomness—i.e. that what we observe does not result from random fluctuations that we might expect to see by chance even if there was no effect. This is often formalised as Null Hypothesis Significance Testing (NHST). We *reject* the *null hypothesis* when the probability, if the null hypothesis was true, of observing an effect as large as that which we do observe, is less than some prespecified value (often 0.05). Simply stated, it would be unlikely to obtain the observed effect due to chance, if the null hypothesis was true.

It has long been noted that an experimenter usually wishes to infer something about the population from which the experimental participants are selected (Holmes and Friston, 1998), rather than something about the specific sample of participants that were examined (i.e. a case study). Importantly, any statistical inference from a sample to the population requires a model of the effect in the population. The ubiquitous approach in psychology and neuroimaging is to model the effect in the population with a normal distribution and perform inference on the mean of this model: the population mean (see Methods). For example, the null hypothesis is often that the population mean is zero, and the probability of the observed sample data under this null is computed, taking the variance across the sample as an estimate of the variance across the population. However, an alternative and equally valid question is to ask how typical is an effect in the population (Friston et al., 1999a). That is, we

can infer an effect in each individual of the sample, and from that infer the *prevalence* of the effect in the population—i.e. the proportion of the population that would show the effect, if tested (Allefeld et al., 2016; Donhauser et al., 2018; Friston et al., 1999b; Rosenblatt et al., 2014). The results of these two approaches, considering population mean versus population prevalence, can differ, particularly when effects are heterogenous across participants.

Here, we argue that in many experimental applications in psychology and neuroscience, the individual participant is the natural replication unit of interest (Little and Smith, 2018; Nachev et al., 2019; Smith and Little, 2018; Thiebaut de Schotten and Shallice, 2017). This is because many aspects of cognition, and the neural mechanisms underlying them, are likely to be heterogenous across individuals. Therefore, we should seek to quantify effects within individual participants, and ensure that our results can be reliably distinguished from chance within individual participants. We argue that with this shift in perspective towards experimental assessment within individual participants, our statistical focus at the population level could also shift from NHST of the population mean to estimating the population prevalence: the proportion of individuals in the population who would show an above chance effect in a specific experiment.

We present here a simple but novel Bayesian method to estimate population prevalence based on the results of within-participant NHST, including prevalence differences between groups of participants or between tests on the same participants. This can also be applied without explicit within-participant NHST to provide prevalence of different effect sizes, giving a new view on what can be learned about the population from an experimental sample. We suggest that applying Bayesian population prevalence estimation in studies which are sufficiently powered within individual participants could address many of the recent issues raised regarding replicability in psychology and neuroimaging (Benjamin et al., 2018; Ioannidis, 2005). This approach provides a population prevalence estimate with associated uncertainty and therefore avoids reducing an entire experiment to a binary NHST inference on a population mean (McShane et al., 2019).

Results

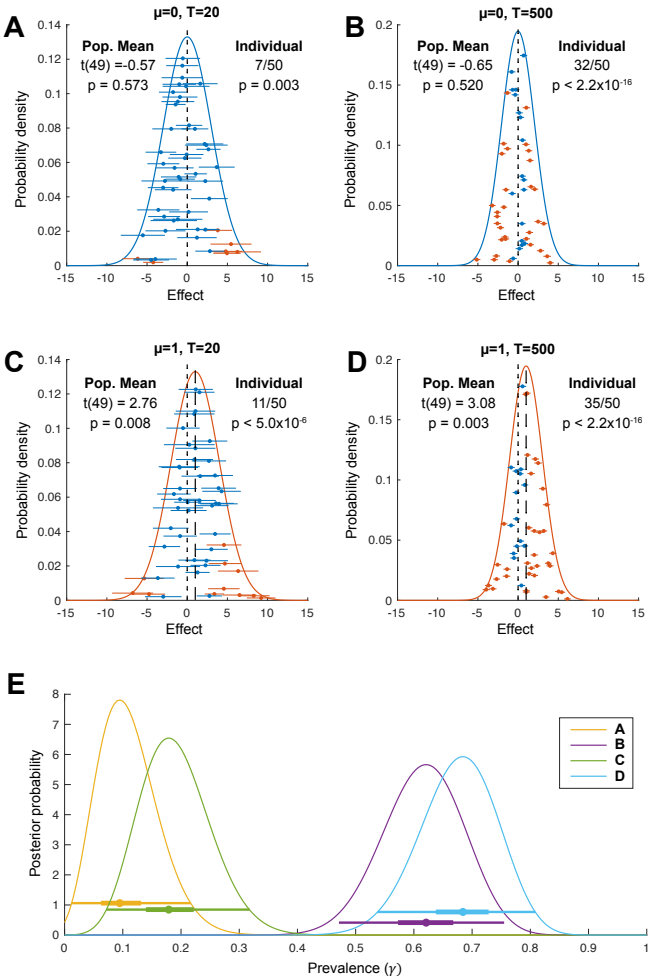

Figure 1: *Population vs individual inference*. For each simulation we sample $N = 50$ individual participant mean effects from a normal distribution with population mean μ (**A,B**: $\mu = 0$; **C,D**: $\mu = 1$) and between-participant standard deviation $\sigma_b = 2$. Within each participant, T trials (**A,C**: $T = 20$; **B,D**: $T = 500$) are drawn from a normal distribution with the participant-specific mean and a common within-participant standard deviation $\sigma_w = 10$ (Baker et al. 2019). Orange and blue indicate, respectively, exceeding or not exceeding a $p = 0.05$ threshold for a t-test at the population level (on the within-participant means, population normal density curves) or at the individual participant level (individual sample means +/- s.e.m.). **E**: Bayesian posterior distributions of population prevalence for the 4 simulated data sets. Points show Bayesian maximum a posteriori estimates. Thick and thin horizontal lines indicate 50% and 96% highest posterior density intervals respectively.

Population vs within-participant statistical tests

To illustrate our key point, we simulate data from the standard hierarchical population model that underlies inference of a population mean effect based on the normal distribution (Friston, 2007; Holmes and Friston, 1998; Penny and Holmes, 2007) (see Methods). Figure 1 illustrates the results of four simulations that differ in the true population mean, μ (A, B: $\mu = 0$; C, D, $\mu = 1$) and in the number of trials per participant (A, C: 20 trials, B, D: 500 trials).

For each simulation we performed inference based on a standard two-sided one-sample t-test against zero at two levels. First, we applied the standard summary statistic approach: we took the mean across trials for each participant and performed the second-level population t-test on these per-participant mean values. This provides an inference on the population mean, taking into account between-participant variation. This is equivalent to inference on

the full hierarchical model in a design where participants are separable (Holmes and Friston, 1998) (see Methods). The modelled population distribution is plotted as a solid curve, coloured according to the result of the population-mean inference (orange for significant population mean, blue for non-significant population mean). Second, we performed inference within each participant, applying the t-test on within-participant samples, separately for each participant. The sample mean +/- s.e.m. is plotted for each participant (orange for significant participants, blue for non-significant participants).

The population mean inference correctly fails to reject the null hypothesis for panels A and B (ground truth $\mu = 0$), and correctly rejects the null in panels C and D (ground truth $\mu = 1$). But consider carefully panels B and C, which illustrate our main point. With 500 trials in panel B, 32/50 participants (orange markers) show a significant within-participant result. The probability of this happening by chance, if there was no effect in any members of the population, can be calculated from the cumulative density function of the binomial distribution. In this case it is tiny—for a test with false positive rate $\alpha = 0.05$, and no effect in any individual, $p < 2.2 \times 10^{-16}$ (below 64-bit floating-point precision). Compare that to $p = 0.008$ for the population t-test in panel C. Thus, the panel B results provide very much stronger statistical evidence of a non-zero effect *at the population level* – the observed results are very unlikely if the proportion of individuals in the population who show the effect is zero. This would be ignored in analyses based only on the population mean. Considering inference at the individual level, panel C results (11/50 significant) have $p = 4.9 \times 10^{-6}$ if the proportion of the population with the effect was zero, i.e. there was no effect within any individuals. Thus, even panel C, simulating an experiment with only 20 trials per participant, provides stronger evidence for a population effect from the within-participant perspective than from the population mean perspective.

Obviously these two different p-values are calculated from two different null hypotheses. The population t-test tests the null hypothesis that the population mean is zero:

$$H_0: \mu_{pop} = 0$$

while the p-value for the number of significant within-participant tests comes from the null hypothesis that there is no effect in any individual in the population, termed the *global null* (Allefeld et al., 2016; Donhauser et al., 2018; Nichols et al., 2005):

$$H_0: \mu_i = 0 \text{ for all } i$$

These are testing different questions, but importantly both are framed at the population level and both provide a population level inference. We agree that it is important to align "the meaning of the quantitative inference with the meaning of the qualitative hypothesis we're interested in evaluating" (Yarkoni, 2019). We suggest that often, if the goal of the analysis is to infer the presence of a causal relationship within individuals in the population, the within-participant perspective may be more appropriate. Clearly the global null itself is quite a blunt tool, as it would be untrue mathematically if even one in a million participants showed an effect. The goal of the prevalence methods we present here is to quantify within-participant effects at the population level in a more meaningful and graded way.

We emphasise that this is the simplest possible illustrative example, intended to demonstrate the different perspective of within-participant vs. population mean inference. Although we have only considered the global null, the simulations show that the within-participant perspective can give a very different view of the evidence for a population level effect

provided by a specific data set. It is also important to emphasise that the simulated data values within individuals are considered to represent the *effect* of an experimental manipulation (e.g. the difference in a neuroimaging response between stimuli of different classes). Therefore, the results illustrated in Figure 1B represent strong evidence of a non-zero effect within many participants, albeit one that seems to be approximately balanced across participants between positive and negative directions.

The point we make here is applicable to any within-participant statistical test of an experimental manipulation. For example, we could consider a within-participant d-prime, a coefficient or contrast from a linear model (e.g. a General Linear Model of fMRI data), a cross-validated out-of-sample predictive correlation from a high-dimensional stimulus encoding model (e.g. a model predicting auditory cortex MEG responses to continuous speech stimuli), a rank correlation of dissimilarity matrices in a Representational Similarity Analysis, or parameters of computational models of decision making (e.g. the Diffusion Drift Model). In all of these cases, the distinction between performing inference on the population mean, vs. within individual participants still holds. We argue that performing NHST at the individual participant level is preferable for both conceptual reasons in psychology and neuroimaging, but also for practical reasons related to the replicability crisis (see Discussion).

Estimating population prevalence

The *p*-values under the global null are obtained from the cumulative density function of the binomial distribution, based on a within-participant false positive rate $\alpha = 0.05$. However, we can also model the number of above-threshold individuals in a sample when the true prevalence, the proportion of the population that show the effect, is γ. Consider a within-participant test with a false positive rate α and sensitivity β. In this case, the distribution of the number of significant individuals follows a binomial distribution with success probability $\theta = (1-\gamma)\alpha + \gamma\beta$. Here, we present a Bayesian approach to estimate population prevalence proportion γ, from this binomial model of within-participant testing, but note that alternative frequentist inference approaches can be used (Allefeld et al., 2016; Donhauser et al., 2018; Friston et al., 1999b) (see Methods).

The Bayesian approach provides a full posterior distribution for γ, from which we can obtain the *maximum a posteriori* (MAP) estimate, together with measures of uncertainty—e.g. highest posterior density intervals (HPDIs) or lower bound quantiles. Figure 1E shows the Bayesian posteriors, MAPs and HPDIs for the 4 simulated data sets in Figure 1A-D. Even though there is no population mean effect in Figure 1B, the MAP estimate of the prevalence is 0.62 (96% HPDI: [0.47 0.76]). Given the data, the probability that the population prevalence is greater than 47% is higher than 0.96. Therefore, we would consider it highly likely that more than 47% of the population would show an effect in an experiment with 500 trials.

Figure 2 illustrates how Bayesian prevalence inference scales with numbers of participants and trials. Figure 2A-C suggests that for the Bayesian prevalence metrics, there are benefits to having larger numbers of participants (decrease in variance of obtained MAP and HPDI width, increase in prevalence lower bound), but beyond around 50 participants these benefits become less pronounced. Figures 2E shows that inferred prevalence is mostly sensitive to the number of trials per participant (horizontal contours), and invariant to the

number of participants (although variance increases as in Figure 2A,C,F), whereas t-test power (Figure 2D) is mostly sensitive to number of participants (vertical contours) and largely invariant to number of trials beyond around 100 trials per participant (Baker et al., 2019). In sum, compared to the population mean t-test, prevalence exhibits greater sensitivity to the number of trials obtained per participant, and less sensitivity to the number of participants.

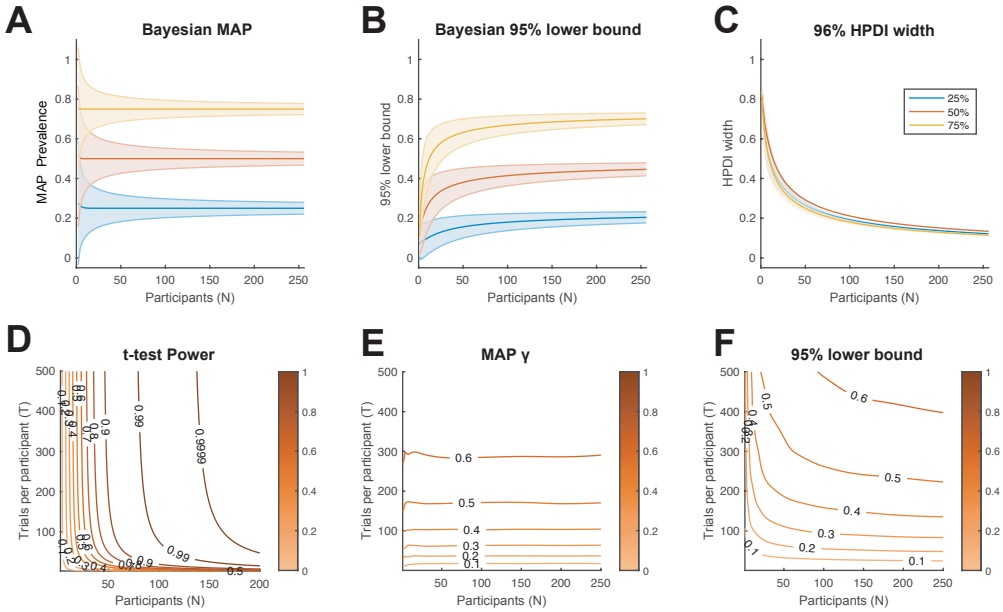

Figure 2: *Characterisation of Bayesian prevalence inference*. **A,B,C**: We consider the binomial model of within-participant testing for three ground truth population proportions: 25%, 50% and 75% (blue, orange, yellow, respectively). We show how **A**: the Bayesian MAP estimate, **B**: 95% Bayesian lower bound and **C**: 96% HPDI width, scale with the number of participants. Lines show theoretical expectation, shaded region shows +/- 1 s.d.. **D,E,F**: We consider the population model from Figure 1C,D ($\mu = 1$). **D**: Power contours for the population inference using a t-test (Baker et al. 2019). **E**: Contours of average Bayesian MAP estimate for γ. **F**: Contours of average 95% Bayesian lower bound for γ. From the prevalence perspective, the number of trials obtained per participant has a larger effect on the resulting population inference than does the number of participants.

Estimating differences in prevalence

Often the scientific question of interest might involve a contrast between groups of participants or experimental conditions. We therefore provide additional functions to directly estimate the difference in prevalence between two different groups of participants who undergo the same test, or two different tests which are applied to the same participants (see Methods).

For the difference in prevalence of two independent groups, the data required for the Bayesian prevalence estimation is the count of significant participants and the total number of participants in each group. We illustrate this with a simulated result. We specify the prevalence in the two populations as $\gamma_1 = 0.75$ and $\gamma_2 = 0.25$ respectively, and draw a random sample based on the respective binomial distribution for the parameters θ_i (see Methods). We simulate $N_1 = 60$ participants in the first group and $N_2 = 40$ participants in the second group. The results of one such draw gives $k_1 = 45$, $k_2 = 11$ positive tests in each group respectively. In this case, the MAP [96% HPDI] prevalence difference $\gamma_1 - \gamma_2$, calculated from these four values (k_1, k_2, N_1, N_2), is 0.49 [0.29 0.67], which closely matches

the ground truth. Figure 3A-B shows how the between group posterior prevalence difference estimates scale with the number of participants for three different simulations.

For the within-group difference, the input parameters are the number of participants significant in both tests, the number significant only in each of the two tests and the total number of participants. We simulate two tests applied to a group of $N = 50$ participants. Each test detects a certain property with false positive rate $\alpha = 0.05$. The ground truth prevalence's for the two properties are $\gamma_1 = 0.5$ and $\gamma_2 = 0.25$ respectively and correlation between the existence of each effect is $\rho_{12} = 0.2$ (i.e. participants who possess one property are more likely to have the other property). The results of one random draw from this model gives (see Methods) $k_{11} = 8$ participants with a significant result in both tests, $k_{10} = 19$ participants with a significant result in the first test but not the second and $k_{01} = 5$ participants with a significant result in the second but not the first. In this case, the MAP [96% HPDI] prevalence difference $\gamma_1 - \gamma_2$, calculated from these four values $(k_{11}, k_{10}, k_{01}, N)$, is 0.28 [0.08 0.46], which again matches the ground truth. Figure 3C-D shows the how the within group posterior prevalence difference estimates scale with the number of participants for three different ground truth situations, given as $[\gamma_1\ \gamma_2]\ \rho_{12}$.

Both these approaches are implemented using Monte Carlo methods, and the functions return posterior samples (Gelman, 2014). These posterior samples can be used to calculate other quantities, such as the posterior probability that one test or group has a higher prevalence than the other. The posterior log odds in favour of this hypothesis can be computed from by applying the logit function to the proportion of posterior samples in favour of a hypothesis. In the between group example above, the hypothesis $\gamma_1 > \gamma_2$ has a proportion 0.9999987 of posterior samples in favour, corresponding to log posterior odds of 13.5. In the above within group comparison the hypothesis $\gamma_1 > \gamma_2$ has a proportion 0.9979451 of posterior samples in favour, corresponding to log posterior odds of 6.2 (each computed from 10 million posterior samples).

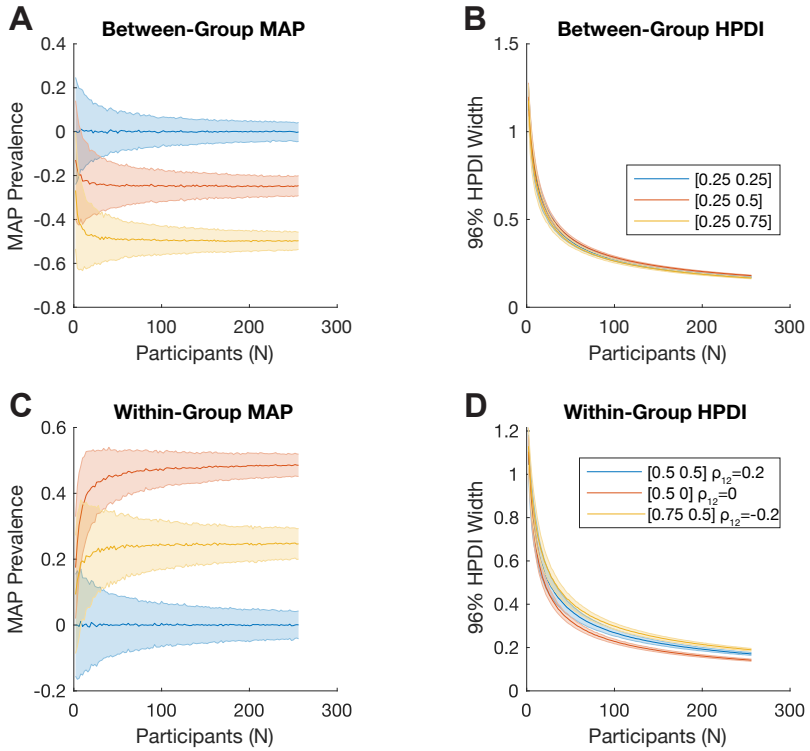

Figure 3: *Characterisation of Bayesian inference of difference of prevalence*. **A,B**: We consider two independent groups of participants with respective ground truth population prevalences $[\gamma_1, \gamma_2]$ of [25% 25%] (blue), [25% 50%] (red) and [25% 75%] (yellow). We show how **A**: the Bayesian MAP estimate of and **B**: 96% HPDI width, of the estimated between-group prevalence difference $\gamma_1 - \gamma_2$ scale with the number of participants. **C,D**: We consider two tests applied to the same sample of participants. Here each simulation is parameterised by the ground truth population prevalence of the two tested effects, $[\gamma_1, \gamma_2]$, as well as ρ_{12}, the correlation between the (binary) presence of the two effects across the population. We show this for [50% 50%] with $\rho_{12} = 0.2$ (blue), [50% 0%] with $\rho_{12} = 0$ (red), and [75% 50%] with $\rho_{12} = -0.2$ (yellow). We show how **C**: the Bayesian MAP estimate and **D**: 96% HPDI width, of the estimated within-group prevalence difference $\gamma_1 - \gamma_2$ scale with the number of participants.

Prevalence as a function of effect size

In the above, we have focussed on performing explicit statistical inference within each participant. A possible criticism of this approach is that the within-participant binarization of a continuous effect size can lose information. If the null distribution is the same for each participant, then the within-participant inference involves comparing each participant effect size, E_p, to a common statistical threshold $\hat{E}$. The prevalence estimation described above can therefore be interpreted as estimating the population prevalence of participants for which $E_p > \hat{E}$. In the NHST case, $\hat{E}$is chosen so that $P(E > \hat{E}) = \alpha$, (usually $\alpha = 0.05$) but we can in general consider the prevalence of participants with effects exceeding any value $\hat{E}$. We therefore estimate the prevalence of $E_p > \hat{E}$ as a function of $\hat{E}$. This can reveal if a data set provides evidence for population prevalence of a sub-threshold within-participant effect, as well as showing how population prevalence decreases for larger effects. If desired, a frequentist approach to control the error rate can be applied by using the method of maximum statistics to correct over the multiple prevalence inferences. Figure 4 demonstrates this approach for the simulated systems of Figure 1, showing results for both right-sided prevalence of $E_p > \hat{E}$ and left-sided $E_p < \hat{E}$ separately. Note that this approach requires the null distribution to be the same for each participant, and requires calculation of the false positive rate α for each effect size considered. It reveals everything we can learn

about the population prevalence of different effect sizes from our data set, exploring beyond the binarization of the within-participant inference.

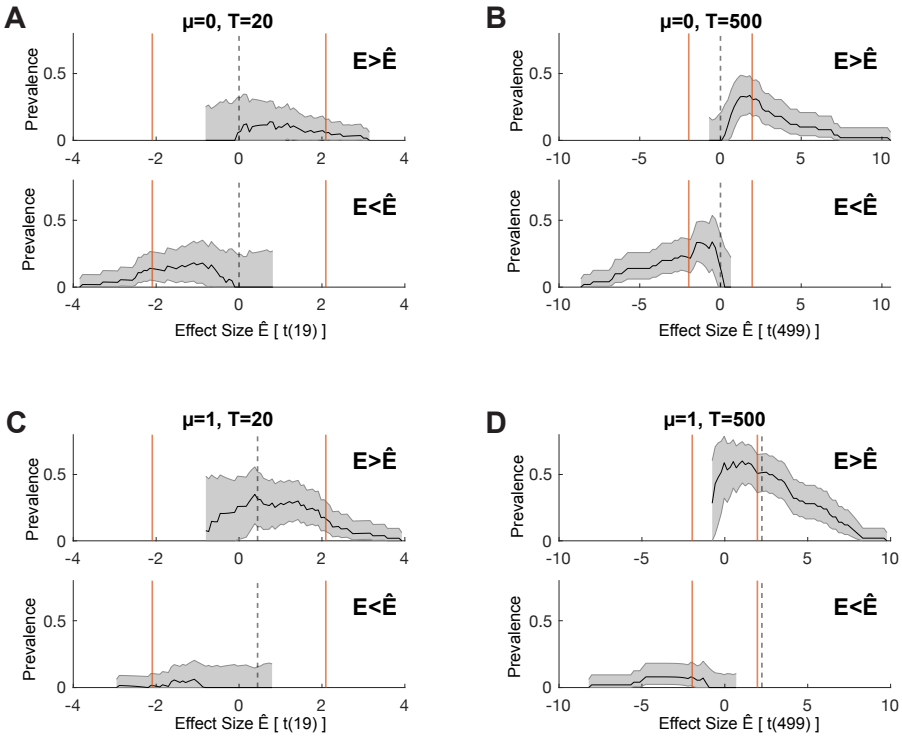

Figure 4: One-sided prevalence as a function of effect size. We consider the same simulated systems shown in Figure 1, showing both right-tailed ($E_p > \hat{E}$) and left-tailed ($E_p < \hat{E}$) prevalence as a function of effect size. Orange lines show the effect size corresponding to the two-sided $\alpha = 0.05$ within-participant test as used in Figure 1. Dashed lines show the effect size corresponding to the ground truth of the simulation. A,B: $\mu_{pop} = 0$, C,D: $\mu_{pop} = 1$. A,C: $k = 20$ trials, B,D: $k = 500$ trials. Black line shows MAP, shaded region shows 96% HPDI.

How to apply Bayesian prevalence in practice

As in the simulation of Figure 1, a typical population mean inference is often framed as a two-level summary statistic procedure. At the first level, the effect is quantified within each participant (e.g. a difference in mean response between two conditions). At the second level, the population mean is inferred under the assumption that the effect is normally distributed in the population (i.e. based on the mean and standard deviation of the measured effect across participants). Bayesian prevalence is similarly framed as a two-level procedure. At the first level, a statistical test is applied within each participant. The result of this test can be binarized via a within-participant NHST (e.g. using a parametric test as in our simulation, or alternatively using non-parametric permutation methods, but independently for each participant), or via an arbitrary effect size threshold $\tilde{E}$. At the second level, the binary results from the first level (i.e. the counts of significant participants) are the input to the Bayesian population prevalence computation. To accompany this paper, we provide code[1] in Matlab, Python and R to visualise the full posterior distribution of the population prevalence, as well

[1] To accompany this paper, we provide functions in Matlab, Python and R to calculate the Bayesian prevalence posterior density (e.g. to plot the full posterior distribution), the MAP estimate of the population prevalence, HPDI intervals of the posterior and lower bound quantiles of the posterior, as well as prevalence differences between samples or between tests within a sample. We also provide example scripts which produce posterior plots as in Figure 1E. See https://github.com/robince/bayesian-prevalence

as extract properties such as the maximum a posteriori (MAP) point estimate and highest posterior density intervals (HPDI). We also provide functions to provide Bayesian estimates of the difference in prevalence between two mutually exclusive participant groups to the same test (between-group prevalence difference) as well as the difference in prevalence between two different tests applied to a single sample of participants (within-group prevalence difference). We suggest reporting population prevalence inference results as the MAP estimate together with one or more HPDI's (e.g. with probability 0.5 or 0.96, see Methods).

It is important to stress that the second-level prevalence inference does not impose any requirements on the first level within-participant tests, other than that each test should have the same false positive rate of α (see Methods). It is not required, for example, that each participant have the same number of trials or degrees of freedom. The within-participant test can be parametric (e.g. a t-test) or non-parametric (e.g. based on permutation testing). It can be a single statistical test, or an inference over multiple tests (e.g. a neuroimaging effect within a certain brain region), provided the family-wise error rate is controlled at α (e.g. by using permutation tests with the method of maximum statistics).

Discussion

In this paper, we presented a conceptual point and a practical method. The conceptual point argues for a shift in perspective from population means to prevalence of effects detected within individuals. We agree that it is important to align "the meaning of the quantitative inference with the meaning of the qualitative hypothesis we're interested in evaluating" (Yarkoni, 2019). We argue that within-participant inference combined with estimation of population prevalence may in some cases better match the qualitative hypothesis researchers are interested in—i.e. in contrast to a binary inferential result that the population mean differs from zero. However, the fields of psychology and neuroimaging are currently dominated by the latter. The practical method to address the conceptual point is a simple but novel Bayesian approach to estimate population prevalence and associated uncertainty. Our method can easily be applied to almost any statistical evaluation of any experiment, provided a NHST can be performed at the individual participant level. The simulations presented here can also be used for simple power analyses when designing studies from this perspective.

Together, this conceptual point and practical method support an alternative perspective for statistics in which the individual participant is the most relevant experimental unit to consider for replication (Nachev et al., 2019; Smith and Little, 2018). From this perspective, power should be calculated for effects within individual participants. This gives a very different view of the strength of evidence provided by a data set and of the importance of sample size (for both participants and trials) compared to the more common population mean perspective (Baker et al., 2019). For example, the simulation of 50 participants with 20 trials in Figure 1C has $p = 0.008$ for a group mean different from zero, a result that is as surprising, under the null hypothesis, as observing 7 heads in a row from tosses of a fair coin (Obleser, 2019). This is weaker evidence than just 2 out of 5 participants showing an effect at $\alpha = 0.05$ ($p = 0.0012$ under the global null, or about as surprising as 10 heads in a row). 3 out of 5 significant participants corresponds to $p = 0.00003$ under the global null (as surprising as 15 heads in a row); this is substantially stronger evidence for a population level effect than is provided by the population mean inference in Figure 1D (from 50 participants, even with 500 trials). However, in the current scientific climate, the weaker result obtained from the larger sample size would commonly be viewed as providing more satisfactory evidence by most readers and reviewers. We would like to highlight this pervasive misunderstanding – i.e. that larger participant numbers automatically imply better evidence at the population level. The crux of our argument is that most studies focus on the difference between panels A and C (a small difference in population mean, ignoring the implications of large between participant variance in studies that are underpowered for within-participant effects), whereas moving towards the situation shown in panels B and D (increased power within individual participants) would provide both improved replicability as well as greater insight into the size and prevalence of effects in the population.

Note that while similar points regarding within-participant inference have been made elsewhere (Nachev et al., 2019; Smith and Little, 2018), they are typically considered to form a case-study, without an inferential link allowing generalization to the population (Neuroscience, 2020). The methods presented here address this concern by providing an inferential bridge from within-participant statistics to the population level. We have focussed on human participants in typical psychology or neuroimaging experiments, but the proposed

methods can be applied to infer population prevalence of effects in other types of experimental units. For example, our functions could be directly applied to identified single-unit neurons in electrophysiological recordings. Typically, only a subset of the identified neurons respond to a specific stimulus or experimental manipulation. Bayesian prevalence could be applied to estimate the proportion of neurons that respond in a particular way in the population of neurons in a particular brain region. Our between group comparison method could be used to formally compare the population prevalence of a certain type of responsive neuron between different brain areas or between different species. Thus, although it is common in electrophysiology to have individual neurons as the replication unit and perform inference at that level, the inferential bridge to the population that Bayesian prevalence provides offers a new perspective on the results of such studies.

The prevalence approach, however, is not without certain limitations. One criticism is that requiring the demonstration of within-participant effects sets a much higher bar of evidence. It might be impractical to reach sufficient within-participant power in some experimental designs. However, many statisticians have explicitly argued that the replicability crisis results from standards of evidence for publication being too weak (Benjamin et al., 2018). If so, the within-participant approach should lead to increased replicability. Indeed, it is common in neuroimaging studies to have no consistent pattern discernible in the effect size maps of individual participants, and yet such studies report a significant group mean effect in a focal region. In our view, this is problematic if our ultimate goal is to relate neuroimaging results to cognitive functions within individuals (Smith and Little, 2018). By contrast, as our simulations show (Figure 1B), strong evidence for a modulation can be evident in the absence of a population mean effect when the effect is heterogeneous across the population.

It is natural to expect considerable heterogeneity to exist across populations for a wide range of experimental tasks. In fact, the normal distribution that underlies most inferences on the population mean implies such heterogeneity (Figure 1B). It has recently been suggested that researchers should define a smallest effect size of interest (SESOI) (Lakens, 2017) and consider this when calculating statistical power. We suggest that the SESOI should also be considered at the individual level, and explicitly related to the population variance obtained from the hierarchical mixed effects model. If the population variance is large, then this directly implies the existence of individuals within the population with effects larger than the individual SESOI, but this is almost always ignored in applications of such modelling which usually focus only on inference on the population mean. In clinical studies of rare diseases often only limited numbers of participants are available, and heterogeneity can be higher than in the healthy population. If an experiment is sufficiently powered within individual participants, then Bayesian prevalence provides a statistical statement about the population of patients with the disease, even from a small sample and without assuming a normal distribution for the effect in the patient population.

Furthermore, the common assumption that effect sizes in the population follow a normal distribution is strong, although seldom justified (Lakens et al., 2018; Nachev et al., 2019; Smith and Little, 2018). For example, information processing, decision making, risk taking, and other strategies vary within the healthy population and across clinical and sub-clinical groups. In neuroimaging studies, there are related issues of variability in anatomical and functional alignment. To address these issues, results are often smoothed within individuals

before performing population mean inference. However, many new experimental developments, such as high-resolution 7T fMRI to image cortical layers, or high-density intracranial recordings in humans pose increasing difficulties in terms of alignment of data across participants. A major advantage of the prevalence approach is that we can perform within-participant inference corrected for multiple comparisons, and then perform population prevalence inference without requiring the precise alignment of the effects across participants. For example, one might report that 24/30 participants showed an EEG alpha band power effect between 500 ms and 1000 ms post stimulus, which implies a population prevalence MAP of 0.79 (96% HPDI [0.61 0.91]), without requiring these individual effects to occur at precisely the same time point in those 24 participants. Similarly, if within-participant inference is corrected for multiple comparisons, one can infer prevalence of, say, an effect in a certain layer of V1, without requiring precise alignment of the cortical location of the effect between participants.

In fMRI in particular many concerns have recently been raised with common statistical approaches. We suggest many of these issues could be ameliorated by the prevalence perspective. (Botvinik-Nezer et al., 2020) show large variability in inferential results obtained when different groups test the same hypotheses on the same data set. However, they note broad agreement between the population effect size maps of different pipelines, suggesting it is the final binary population inference that is inconsistent. It would be interesting to compare the consistency of graded prevalence results based on within-participant inference between different analysis pipelines in a similar way. (Elliott et al., 2020) consider the within-participant test-retest reliability of common fMRI designs. They note that reliable population mean effects can arise from unreliable within-participant measurements, and that this is the case in many common experimental designs which focus on statistical power from the perspective of the population mean. They argue such designs are therefore problematic for individual-differences research. Focussing on within-participant power as we suggest here may increase test-retest reliability and open the avenue to more robust individual-difference applications.

There are of course many cases where the population mean is indeed the primary interest, and in such cases inference on the mean using hierarchical models is the most appropriate analysis. We argue here only that this is not always true. Indeed, for many complex computational techniques from modelling of learning behaviour and decision making, to neuroimaging analysis techniques such as Representational Similarity Analysis or high-dimensional encoding models (Haxby et al., 2014), it is not currently possibly to employ a direct multi-level modelling approach down to the trial level, due to the complexity of the non-linear analysis functions and models employed. It is also possible to interrogate linear mixed-effect models in different ways to investigate the question of prevalence, for example examining the variances of the random-slopes for each participant, or explicitly computing prevalence of different effect sizes from the normal distribution fit to the population. It is possible extend Bayesian hierarchical models to explicitly account for different sub-groups of participants (Bartlema et al., 2014; Haaf and Rouder, 2019). However, these approaches are not currently widely adopted, cannot easily be applied to non-linear or high-dimensional analysis methods common in neuroimaging, and they add both mathematical and computational complexity compared to the second-level Bayesian prevalence method we present here, which is straightforward to apply to any first-level within-participant analysis.

The goal of any statistical analysis should be explicitly stated and justified (Lakens et al., 2018; Yarkoni, 2019). In psychology and neuroscience, this goal is often to demonstrate a relationship between the experimental manipulation and the measured response. That is, to infer a causal influence on the behavioural or neural measure in the population from which the participants were sampled. This is usually framed at the level of the population mean, but we argue that in cognitive neuroscience and neuroimaging the question might be better framed at the level of individuals.

Ensuring that studies are sufficiently powered to obtain reliable effect size estimates within individual participants has two main advantages. First, that each participant serves as an independent replication protects from inference problems that arise when an entire experiment is reduced to a binary significance classification. Second, the within-participant effect sizes themselves provide a valuable description of the effect in the population. With studies that are sufficiently powered within individuals, comparisons between groups can look beyond means (Rousselet et al., 2017) to provide more scientific insight. For example, the empirical distribution of within-participant effect sizes might deviate from the implicitly modelled normality, possibly revealing subgroups. The practice of collecting enough data to perform within-participant inference is not a new idea – much of traditional psychophysics employs this approach (Smith and Little, 2018). We have employed this technique with EEG (Schyns et al., 2011; Smith et al., 2006) and MEG (Ince et al., 2015; Zhan et al., 2019) and in fMRI it is becoming more common to collect large quantities of data for fitting high-dimensional cross-validated machine learning models within individual participants (Huth et al., 2016; Stansbury et al., 2013). Recently, this practise has also been adopted in the resting state fMRI field where it is termed “dense sampling” or “precision imaging”(Braga and Buckner, 2017; Gordon et al., 2017; Laumann et al., 2015; Poldrack, 2017). The benefit of the prevalence perspective is to obtain a population level inference from such studies, even when the number of participants is small.

Fundamentally, many questions in cognitive neuroscience apply to behaviour at an individual level. They are therefore better answered with statistical techniques that are also framed at that level. We suggest that the combination of within-participant statistical tests with population prevalence inference provides a valuable alternative to current norm—i.e. the testing of population means assuming a normal distribution. However, the two approaches are not mutually exclusive and ideally one could report within-participant effect sizes and inferred population prevalence together with an inference on the population mean, assuming a normal distribution whenever appropriate.

While the prevalence framework still relies on a NHST dichotomisation within each participant, each participant becomes an independent replication. Population prevalence can be inferred in a graded way, by providing continuous valued estimates and bounds on prevalence, rather than reducing an entire experiment to a single binary population NHST result. This addresses many of the problems noted with the NHST framework (McShane et al., 2019) and also reduces the risk of questionable research practices, such as *p*-hacking (Forstmeier et al., 2017). Importantly, inferred prevalence should be taken as only one aspect that contributes to the strength of evidence provided by a data set, and should be assessed alongside the quality of the experimental design, within-participant power and the effect

sizes within individuals. We have also shown how to estimate the prevalence of difference effect size thresholds which avoids focussing on a single within-participant dichotomisation.

It is important to emphasize that Bayesian prevalence has a broad range of applicability – spanning dense sampling studies with high within-participant power, as well as more traditional sampling models (more participants, fewer trials, e.g. Figure 1C). It is applicable to any behavioural study, including detailed computational models of behaviour, provided that model comparison or inference on model parameters can be performed within individuals. In neuroimaging, Bayesian prevalence can be applied to any imaging modality (EEG, MEG, fMRI, intracranial EEG), individual neurons within a brain region (to infer the proportion of responsive neurons), and with any statistical approach, including non-linear and multivariate analysis methods. The crucial requirement is an adequately powered experiment to detect effects within individual participants (or other units of interest, e.g. neurons). We argue that ensuring experiments are adequately powered to detect effects within individuals would have a wide range of advantages increasing the robustness and replicability of results.

Conclusions

While the problems that underlie the replication crisis are being increasingly recognised, there is currently no consensus about alternative statistical approaches to address these problems. Here, we propose that shifting our focus to quantifying and inferring effects within individuals addresses many of the pressing concerns recently highlighted in psychology and neuroimaging (Amrhein et al., 2019; Benjamin et al., 2018; Forstmeier et al., 2017; Ioannidis, 2005; McShane et al., 2019). The prevalence approach is completely general and broadly applicable as it places no assumptions on the within-participant tests nor on the distribution of effects in the population. Further, prevalence does not require a Bayesian treatment; frequentist inference approaches can be used. The crucial point is to shift our perspective to first evaluate effects within individual participants, whom we believe represent the natural replication unit for psychology and neuroimaging.

Methods

Simulations from hierarchical population model

The data shown in Figure 1 were simulated from the standard hierarchical model:

$$y_{ij} \sim N(\mu_i, \sigma_w^2)$$
$$\mu_i \sim N(\mu_{pop}, \sigma_b^2)$$

Where y_{ij} denotes the measurement made on the j^{th} trial (out of t) of the i^{th} participant (out of n). μ_i represents the mean of each individual participant, σ_w represents a common within-participant standard deviation over trials, σ_b represents the standard deviation between participants and μ_{pop} represents the overall population mean. This can equivalently be written as

$$y_{ij} = \mu_{pop} + \eta_{ij} + \epsilon_i$$

where $\eta_{ij} \sim N(0, \sigma_w^2)$, and $\epsilon_i \sim N(0, \sigma_b^2)$. Note that under this model the distribution of the within participant means is $N(\mu_{pop}, \sigma_b^2 + \frac{1}{t}\sigma_w^2)$.

Binomial model of population prevalence when the true state of sampled units is unknown

We consider a population of experimental units (for example, human participants, or individual neurons) which are of two types: those which have a particular experimental effect, and those which don't. We are interested in estimating the population prevalence γ, which is the proportion of the population from which the sample was drawn that have the effect ($0 < \gamma < 1$). If the true status of each individual unit could be directly observed, then the sample could be modelled with a binomial distribution with probability parameter γ. However, we cannot directly observe the true status of each unit. Instead, we apply to each unit a statistical test following the NHST framework. This test has a false positive rate α, and sensitivity β. Thus, the probability that a randomly selected unit from the population which does not possess the defined effect will produce a positive test result is α, while the probability that randomly selected unit that does possess the defined effect will produce a positive test result is β. Under the assumption that the units are independent and α and β are constant across units, the number of positive tests k in a sample of size n can be modelled as a binomial distribution with parameter θ (Donhauser et al., 2018; Friston et al., 1999a, 1999b):

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

Frequentist estimation of and inference on population prevalence

Various frequentist approaches can be used with the above binomial model of statistical testing. First, the maximum likelihood of the population prevalence parameter can be obtained as:

$$\hat{\gamma} = \frac{k/n - \alpha}{\beta - \alpha}$$

Standard bootstrap techniques (Efron and Tibshirani, 1994) can give percentile bootstrap confidence intervals as an indication of uncertainty in this estimate. We can also explicitly test various null hypotheses at the population level. For example, we can test a compound null hypothesis $\gamma < 0.5$, termed the majority null, (Allefeld et al., 2016; Donhauser et al., 2018). This is chosen with the idea that a prevalence of >50% supports a claim the effect is *typical* in the population. Other explicit compound nulls of this form can also be tested (e.g. that $\gamma < 0.25$ or $\gamma < 0.75$). Alternatively, it is possible to infer a lower bound on the population prevalence, by finding the largest value γ_c, such that $p(X > k \mid \gamma < \gamma_c) < 0.05$ (Allefeld et al., 2016; Donhauser et al., 2018). This inferred lower bound provides a more graded output than a binary significance results of testing against a specific compound null (i.e. the continuous value γ_c).

Bayesian estimation of population prevalence

We apply standard Bayesian techniques to estimate the population prevalence parameter of this model (Gelman, 2014). Assuming a Beta prior distribution for θ with shape parameters r, s, together with a Binomial likelihood function, the posterior distribution for θ is given by a Beta distribution with parameters $(k + r, n - k + s)$, truncated to the interval $[\alpha, \beta]$, where k is the number of participants showing an above-threshold effect out of n tested. In the examples shown here we use a uniform prior (beta with shape parameters $r = s = 1$), as in the general case there is no prior information regarding θ. This implies a uniform prior also for γ, so, a priori, we consider any value of population prevalence equally likely. While we default to the uniform prior, the code supports any beta distribution as a prior. Alternative priors could be implemented via Markov chain Monte Carlo methods (Gelman, 2014) together with the models described here.

Under the uniform prior, the Bayesian maximum a posteriori (MAP) estimate for γ is available analytically and is equivalent to the maximum likelihood estimate:

$$\gamma_{map} = \frac{k/n - \alpha}{\beta - \alpha}$$

Following (McElreath, 2016) we present 96% Highest Posterior Density Intervals (HPDIs) here to emphasise the arbitrary nature of this value and reduce the temptation to interpret the interval as a frequentist p=0.05 inference.

One important caveat is that the sensitivity of the test, β, is not known a priori and will differ with effect size across participants. In general, a test with lower sensitivity allows inference of a higher prevalence for an observed k, because some of the observed negative results will be missed true-positive results. We therefore take $\beta = 1$ as a conservative approach, which leads to the smallest maximum likelihood, MAP, or lower bound population prevalence estimates (Allefeld et al., 2016; Donhauser et al., 2018; Friston et al., 1999b). Note that similar Bayesian approaches have been applied in the field of epidemiology, where sometimes multiple complementary diagnostic tests for a disease are applied with or without a gold standard diagnosis in a subset of the sampled units (Berkvens et al., 2006; Enøe et al., 2000; Joseph et al., 1995).

Bayesian estimation of the prevalence difference between two independent groups

We consider here a situation where the same test is applied to units sampled from two different populations and, in addition to the prevalence within each population, we wish to directly estimate the difference in prevalence between the two populations. For example, this could be human participants from two different cultures, a group of patients vs a group of healthy controls, or a population of neurons recorded in a transgenic animal model vs those recorded from a wildtype model. We denote the prevalence within each group as γ_1, γ_2, respectively, and similarly the number of significant within-participant tests in each group by k_1, k_2 out of n_1, n_2 total participants in each group. Assuming independent uniform priors on the prevalences, and associated θ_i variables as above with:

$$\theta_i = (1 - \gamma_i)\alpha + \gamma_i\beta$$

then the posterior distribution for (θ_1, θ_2) is given by the product of two truncated beta distributions, with parameters $(k_i + 1, n_i - k_i + 1)$ respectively, both truncated to the interval $[\alpha, \beta]$. The prevalence difference can be obtained as:

$$\gamma_1 - \gamma_2 = (\theta_1 - \theta_2)/(\beta - \alpha)$$

For non-truncated beta distributions, an analytic exact result is available (Pham-Gia and Turkkan, 1993). This result could be extended to provide an exact distribution for the prevalence difference, but the mathematical expressions involved are quite complex. It is simpler to employ Monte Carlo methods, which can provide as close an approximation to the exact answer as desired. Here we use Monte Carlo methods to draw samples from the posterior for (θ_1, θ_2), obtaining samples for the prevalence difference with the above expression. We use these samples to numerically compute the MAP and HPDIs.

Bayesian estimation of the prevalence difference of two different tests within the same sample of participants

In this situation we consider that two different test procedures are applied to a single sample of n units. We assume both tests have the same values of α and β, and define:

$$\theta_i = (1 - \gamma_i)\alpha + \gamma_i\beta = \alpha + (\beta - \alpha)\gamma_i$$

Here θ_i is the probability that a randomly selected unit from the population will show a positive result on the i^{th} test and γ_i is the population prevalence associated with the i^{th} test. Now, each unit provides one of four mutually exclusive results based on the combination of binary results from the two tests. k_{11} represents the number of units which have a positive result in both tests, k_{10}, k_{01} represent the number of units which have a positive result only in the first or second test respectively, and k_{00} is the number of units which do not show a positive result in either test. So $\sum_{i,j} k_{ij} = n$. We can define analogous variables $\boldsymbol{\theta} = \{\theta_{ij}\}$ representing the population proportion of units for each of the four combined test outcomes. Note that $\theta_{ij} > 0$ and $\sum_{i,j} \theta_{ij} = 1$. The marginal success probabilities θ_i can be expressed as:

$$\theta_1 = \theta_{11} + \theta_{10}, \quad \theta_2 = \theta_{11} + \theta_{01}$$

and so

$$\gamma_1 - \gamma_2 = (\theta_{10} - \theta_{01})/(\beta - \alpha)$$

The marginal probabilities θ_i are subject to the constraints

$$\alpha < \theta_i < b$$

and so

$$\alpha < \theta_{11} + \theta_{10} < \beta, \quad \alpha < \theta_{11} + \theta_{01} < \beta$$

Assuming a uniform prior, and a multinomial distribution for the k_{ij}, the posterior of $\boldsymbol{\theta}$ is a truncated Dirichlet distribution with parameters $m_{ij} = k_{ij} + 1$ subject to the constraints above (which are analogous to the truncation of the Beta posterior distribution in the case of a single test). We use a Monte Carlo approach to draw samples from this posterior following a modified stick-breaking process.

- Draw a sample θ_{11} from a $Beta(m_{11}, m_{10} + m_{01} + m_{00})$ distribution truncated to the interval $[0, b]$.
- Draw a sample z_{10} from a $Beta(m_{10}, m_{01} + m_{00})$ distribution truncated to the interval $\left[\max\left(\frac{a-\theta_{11}}{1-\theta_{11}}, 0\right), \frac{b-\theta_{11}}{1-\theta_{11}}\right]$. Set $\theta_{10} = (1 - \theta_{11})z_{10}$.
- Draw a sample z_{01} from a $Beta(m_{01}, m_{00})$ distribution truncated to the interval $\left[\max\left(\frac{a-\theta_{11}}{1-\theta_{11}-\theta_{10}}, 0\right), \min\left(\frac{b-\theta_{11}}{1-\theta_{11}-\theta_{10}}, 1\right)\right]$. Set $\theta_{01} = (1 - \theta_{11} - \theta_{10})z_{01}$.
- Set $\theta_{00} = 1 - \theta_{11} - \theta_{10} - \theta_{01}$.
- Then $(\theta_{11}, \theta_{01}, \theta_{10}, \theta_{00})$ is a draw from the required truncated Dirichlet distribution, and $(\theta_{10} - \theta_{01})/(\beta - \alpha)$ is a draw from the posterior distribution of the prevalence difference.

We use these samples to numerically compute properties like the MAP estimate and HPDIs.

To specify a ground truth to simulate data from two tests applied to the same participants (Figure 3) we require γ_1 and γ_2, the population prevalences of the two tested effects, together with ρ_{12}, the correlation between the presence of the two effects across the population. From this we can calculate γ_{11}, the proportion of individuals in the population that possess both effects as:

$$\gamma_{11} = \gamma_1\gamma_2 + \rho_{12}\sqrt{\gamma_1(1-\gamma_1)\gamma_2(1-\gamma_2)}$$

Similarly, we can define γ_{ij} representing the population proportions corresponding to the other test result configurations. Then we can generate multinomial samples using the parameters θ_{ij} computed as:

$$\begin{aligned}
\theta_{11} &= b^2\gamma_{11} + ab\gamma_{10} + ab\gamma_{01} + a^2\gamma_{00} \\
\theta_{10} &= a + (b-a)\gamma_1 - \theta_{11} \\
\theta_{01} &= a + (b-a)\gamma_2 - \theta_{11} \\
\theta_{00} &= 1 - \theta_{11} - \theta_{10} - \theta_{01}
\end{aligned}$$

Further mathematical details of the Bayesian estimation of prevalence and prevalence differences are given in supplemental notes available with the code at https://github.com/robince/bayesian-prevalence.

Prevalence as a function of effect size threshold

Estimating the prevalence of $E_p > \tilde{E}$ proceeds as for prevalence inference based on within-participant NHST. One additional step is that it is necessary to calculate α, the false positive rate under the null hypothesis of no effect, for each threshold value $\tilde{E}$. This is simply the probability of $E_p > \tilde{E}$ under the null hypothesis of no effect. In the examples shown here we

calculate this from the cumulative distribution function of the appropriate t-distribution, but for other tests this could also be estimated non-parametrically. A number of $\tilde{E}$ values are selected, linearly spaced over the observed range of the sample. For each of these values the the count of number of participants satisfying the inequality and the α value corresponding to the inequality are used to obtain the Bayesian posterior for prevalence. Note that this can be applied to either tail $E_p > \tilde{E}$, or $E_p < \tilde{E}$.

References

Allefeld, C., Görgen, K., and Haynes, J.-D. (2016). Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. NeuroImage *141*, 378–392.

Amrhein, V., Greenland, S., and McShane, B. (2019). Retire statistical significance. Nature *567*, 305–307.

Angrist, J.D., and Pischke, J.-S. (2014). Mastering 'Metrics: The Path from Cause to Effect (Princeton ; Oxford: Princeton University Press).

Baker, D.H., Vilidaite, G., Lygo, F.A., Smith, A.K., Flack, T.R., Gouws, A.D., and Andrews, T.J. (2019). Power contours: optimising sample size and precision in experimental psychology and human neuroscience. ArXiv:1902.06122 [q-Bio, Stat].

Bartlema, A., Lee, M., Wetzels, R., and Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. Journal of Mathematical Psychology *59*, 132–150.

Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.-J., Berk, R., Bollen, K.A., Brembs, B., Brown, L., Camerer, C., et al. (2018). Redefine statistical significance. Nature Human Behaviour *2*, 6.

Berkvens, D., Speybroeck, N., Praet, N., Adel, A., and Lesaffre, E. (2006). Estimating Disease Prevalence in a Bayesian Framework Using Probabilistic Constraints: Epidemiology *17*, 145–153.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C.F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J.A., Adcock, R.A., et al. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. Nature 1–7.

Braga, R.M., and Buckner, R.L. (2017). Parallel Interdigitated Distributed Networks within the Individual Estimated by Intrinsic Functional Connectivity. Neuron *95*, 457-471.e5.

Donhauser, P.W., Florin, E., and Baillet, S. (2018). Imaging of neural oscillations with embedded inferential and group prevalence statistics. PLOS Computational Biology *14*, e1005990.

Efron, B., and Tibshirani, R.J. (1994). An introduction to the bootstrap (CRC press).

Elliott, M.L., Knodt, A.R., Ireland, D., Morris, M.L., Poulton, R., Ramrakha, S., Sison, M.L., Moffitt, T.E., Caspi, A., and Hariri, A.R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis: Psychological Science.

Enøe, C., Georgiadis, M.P., and Johnson, W.O. (2000). Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Preventive Veterinary Medicine *45*, 61–81.

Forstmeier, W., Wagenmakers, E.-J., and Parker, T.H. (2017). Detecting and avoiding likely false-positive findings – a practical guide. Biological Reviews *92*, 1941–1968.

Friston, K.J. (2007). Statistical parametric mapping: the analysis of functional brain images (Academic Press).

Friston, K.J., Holmes, A.P., and Worsley, K.J. (1999a). How Many Subjects Constitute a Study? NeuroImage *10*, 1–5.

Friston, K.J., Holmes, A.P., Price, C.J., Büchel, C., and Worsley, K.J. (1999b). Multisubject fMRI Studies and Conjunction Analyses. NeuroImage *10*, 385–396.

Gelman, A. (2014). Bayesian data analysis (Boca Raton: CRC Press).

Gordon, E.M., Laumann, T.O., Gilmore, A.W., Newbold, D.J., Greene, D.J., Berg, J.J., Ortega, M., Hoyt-Drazen, C., Gratton, C., Sun, H., et al. (2017). Precision Functional Mapping of Individual Human Brains. Neuron *95*, 791-807.e7.

Haaf, J.M., and Rouder, J.N. (2019). Some do and some don't? Accounting for variability of individual difference structures. Psychon Bull Rev *26*, 772–789.

Haxby, J.V., Connolly, A.C., and Guntupalli, J.S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. Annual Review of Neuroscience *37*, 435–456.

Holmes, A., and Friston, K. (1998). Generalisability, random effects and population inference. Neuroimage *7*.

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., and Gallant, J.L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature *532*, 453–458.

Ince, R.A.A., van Rijsbergen, N., Thut, G., Rousselet, G.A., Gross, J., Panzeri, S., and Schyns, P.G. (2015). Tracing the Flow of Perceptual Features in an Algorithmic Brain Network. Scientific Reports *5*, 17681.

Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. PLOS Medicine *2*, e124.

Joseph, L., Gyorkos, T.W., and Coupal, L. (1995). Bayesian Estimation of Disease Prevalence and the Parameters of Diagnostic Tests in the Absence of a Gold Standard. Am J Epidemiol *141*, 263–272.

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. Social Psychological and Personality Science *8*, 355–362.

Lakens, D., Adolfi, F.G., Albers, C.J., Anvari, F., Apps, M.A.J., Argamon, S.E., Baguley, T., Becker, R.B., Benning, S.D., Bradford, D.E., et al. (2018). Justify your alpha. Nature Human Behaviour *2*, 168.

Laumann, T.O., Gordon, E.M., Adeyemo, B., Snyder, A.Z., Joo, S.J., Chen, M.-Y., Gilmore, A.W., McDermott, K.B., Nelson, S.M., Dosenbach, N.U.F., et al. (2015). Functional System and Areal Organization of a Highly Sampled Individual Human Brain. Neuron *87*, 657–670.

Little, D.R., and Smith, P.L. (2018). Replication is already mainstream: Lessons from small-N designs. Behavioral and Brain Sciences *41*.

McElreath, R. (2016). Statistical rethinking: a Bayesian course with examples in R and Stan (Boca Raton: CRC Press/Taylor & Francis Group).

McShane, B.B., Gal, D., Gelman, A., Robert, C., and Tackett, J.L. (2019). Abandon Statistical Significance. The American Statistician *73*, 235–245.

Nachev, P., Rees, G., and Frackowiak, R. (2019). Lost in translation. F1000Research *7*, 620.

Neuroscience, S. for (2020). Consideration of Sample Size in Neuroscience Studies. J. Neurosci. *40*, 4076–4077.

Nichols, T., Brett, M., Andersson, J., Wager, T., and Poline, J.-B. (2005). Valid conjunction inference with the minimum statistic. NeuroImage *25*, 653–660.

Obleser, J. (2019). Jonas Obleser on Twitter: “To have and to hold. #HowSurprisedShouldYouBe #Information #Surprise #RetireSignificance #ButDontBlameThePvalue https://t.co/fSBEUA5OZd” / Twitter.

Pearl, J. (2009). Causality (Cambridge: Cambridge University Press).

Penny, W., and Holmes, A. (2007). Random effects analysis. Statistical Parametric Mapping: The Analysis of Functional Brain Images 156–165.

Pham-Gia, T., and Turkkan, N. (1993). Bayesian analysis of the difference of two proportions. Communications in Statistics - Theory and Methods *22*, 1755–1771.

Poldrack, R.A. (2017). Precision Neuroscience: Dense Sampling of Individual Brains. Neuron *95*, 727–729.

Rosenblatt, J.D., Vink, M., and Benjamini, Y. (2014). Revisiting multi-subject random effects in fMRI: Advocating prevalence estimation. NeuroImage *84*, 113–121.

Rousselet, G.A., Pernet, C.R., and Wilcox, R.R. (2017). Beyond differences in means: robust graphical methods to compare two groups in neuroscience. European Journal of Neuroscience *46*, 1738–1748.

Schyns, P.G., Thut, G., and Gross, J. (2011). Cracking the Code of Oscillatory Activity. PLoS Biol *9*, e1001064.

Smith, P.L., and Little, D.R. (2018). Small is beautiful: In defense of the small-N design. Psychon Bull Rev *25*, 2083–2101.

Smith, M.L., Gosselin, F., and Schyns, P.G. (2006). Perceptual Moments of Conscious Visual Experience Inferred from Oscillatory Brain Activity. PNAS *103*, 5626–5631.

Stansbury, D.E., Naselaris, T., and Gallant, J.L. (2013). Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex. Neuron *79*, 1025–1034.

Thiebaut de Schotten, M., and Shallice, T. (2017). Identical, similar or different? Is a single brain model sufficient? Cortex *86*, 172–175.

Yarkoni, T. (2019). The Generalizability Crisis (PsyArXiv).

Zhan, J., Ince, R.A.A., van Rijsbergen, N., and Schyns, P.G. (2019). Dynamic Construction of Reduced Representations in the Brain for Perceptual Decision Behavior. Current Biology *29*, 319-326.e4.