Describing posterior distributions of variance components: Problems and the use of null distributions to aid interpretation

Joel L. Pick[1,2,3,12,*], Claudia Kasper[4], Hassen Allegue[5,12], Niels J. Dingemanse[6,12], Ned A. Dochtermann[7,12], Kate L. Laskowski[8,12], Marcos R. Lima[9,12], Holger Schielzeth[10,12], David F. Westneat[11,12], Jonathan Wright[1,12], Yimen G. Araya-Ajoy[1,3,12]

[1] Centre for Biodiversity Dynamics (CBD), Department of Biology, Norwegian University of Science and Technology (NTNU), N-7491 Trondheim, Norway.

[2] Institute of Ecology and Evolution, University of Edinburgh, Charlotte Auerbach Road, Edinburgh, EH9 3FL, UK

[3] Both authors contributed equally

[4] Animal GenoPhenomics group, Agroscope, Tioleyre 4, CH-1725 Posieux, Switzerland.

[5] Département des Sciences Biologiques, Université du Québec à Montréal, Montréal, QC, Canada

[6] Behavioural Ecology, Faculty of Biology, Ludwig-Maximilians University of Munich, Planegg-Martinsried, Germany

[7] Department of Biological Sciences, North Dakota State University, Fargo, ND, USA

[8] Department of Evolution and Ecology, University of California Davis, Davis, CA, USA

[9] Departamento de Biologia Animal e Vegetal, Centro de Ciências Biológicas, Universidade Estadual de Londrina, Londrina, Brazil

[10] Institute of Ecology and Evolution, Friedrich Schiller University Jena, Jena, Germany

[11] Department of Biology, University of Kentucky, Lexington, KY, USA

[12] Members of the SQuID working group

[*] Corresponding author, email address: joel.l.pick@gmail.com

Running title: Null distributions for variance components

Keywords: hierarchical models, variance, null distribution, permutation, simulations, squidSim

Abstract

1. Assessing the biological relevance of variance components estimated using MCMC-based mixed-effects models is not straightforward. Variance estimates are constrained to be greater than zero and their posterior distributions are often asymmetric. Different measures of central tendency for these distributions can therefore vary widely, and credible intervals cannot overlap zero, making it difficult to assess the size and statistical support for among-group variance. Statistical support is often assessed through visual inspection of the whole posterior distribution and so relies on subjective decisions for interpretation.

2. We use simulations to demonstrate the difficulties of summarising the posterior distributions of variance estimates from MCMC-based models. We then describe different methods for generating the expected null distribution (i.e. a distribution of effect sizes that would be obtained if there was no among-group variance) that can be used to aid in the interpretation of variance estimates.

3. Through comparing commonly used summary statistics of posterior distributions of variance components, we show that the posterior median is predominantly the least biased. We further show how null distributions can be used to derive a p-value that provides complementary information to the commonly presented measures of central tendency and uncertainty. Finally, we show how these p-values facilitate the implementation of power analyses within an MCMC framework.

4. The use of null distributions for variance components can aid study design and the interpretation of results from MCMC-based models. We hope that this manuscript will make empiricists using mixed models think more carefully about their results, what descriptive statistics they present and what inference they can make.

Introduction

Estimating variance components using mixed-effects models is common in ecology and evolution (Bolker *et al.*, 2009; Dingemanse & Dochtermann, 2013; Harrison *et al.*, 2018). Mixed-effect models are a flexible statistical tool used to study hierarchically structured data, with extensions facilitating quantitative genetic (animal models; Henderson, 1988; Kruuk, 2004) and comparative (meta-analysis and phylogenetic mixed models; Hadfield & Nakagawa, 2010) analyses. Markov chain Monte Carlo (MCMC) algorithms are increasingly used to fit mixed-effects models due to their flexibility and the availability of open-source software (e.g. winBUGS (Gilks *et al.*, 1994), JAGS (Plummer, 2003), MCMCglmm (Hadfield, 2010) and Stan (Stan Development Team, 2022b)). MCMC algorithms are a collection of probabilistic simulation methods for generating observations from designated statistical distributions and are typically implemented within a Bayesian framework (Gelman *et al.*, 2021).

MCMC methods have many advantages. Derived metrics (such as standardised measures of variance like repeatability, heritability and evolvability; Nakagawa & Schielzeth, 2010; Houle, 1992) can be easily estimated using the posterior distributions of their components, propagating uncertainty within and among analyses. In contrast, in a maximum likelihood framework, the methods to estimate the uncertainty of derived metrics (for example, the delta method) can be laborious and biased with small sample sizes (O'Hara *et al.*, 2008). Data in ecological and evolutionary studies are also commonly non-Gaussian, for example counts (e.g. number of offspring), binary and ratio data (e.g. survival, presence/absence, sex ratio) and categorical data (e.g. colour morphs). The performance of MCMC algorithms in generalized linear mixed-effects models (GLMMs) has been found to be superior in terms of accuracy and precision compared with Restricted Maximum Likelihood (REML) approaches (O'Hara & Merilä, 2005; de Villemereuil *et al.*, 2013). Bayesian methods also allow existing information to be incorporated as a prior distribution, although this has rarely been used in ecological or evolutionary studies (Lemoine, 2019).

Despite these advantages, empiricists face several issues when using MCMC mixed-effect models. Here we focus on the difficulties of describing and interpreting variance estimates and their uncertainty. We highlight two problems, both of which centre around the difficulty of describing the posterior distribution of variance components using summary statistics: (i) finding an appropriate measure of central tendency; and (ii) assessing the statistical support for non-zero among-group variance. These problems arise as variance estimates are constrained to be greater than zero, and so their posterior distributions are often asymmetric.

When describing posterior distributions, we typically present some measure of central tendency alongside some measure of uncertainty (quantile-based intervals or Highest Posterior Density (HPD) intervals). The posterior mean, median and mode have all been used as measures of central tendency, and recent works have advocated the general use of the posterior median (Gelman *et al.*, 2020; McElreath, 2020). There is, however, no clear guidance on which measure provides an appropriate summary statistic for variance components; in our experience the mode and mean are most commonly reported. When the posterior distribution of a variance component is far away from zero and is symmetric, then the mean, median and mode are approximately equal (Figure 1a) and inferences are robust to the choice of central tendency metric. However, when variances are small (relative to the total variance) and/or sample sizes are small (both common in ecology and evolution), the posterior distributions can be close to zero. As variances are constrained to be greater than zero, these posterior distributions are typically asymmetric and can even be bimodal, with one mode close to zero (e.g. Figure 1b). Consequently, there can be a considerable difference between the mean, median and mode (Figure 1b), making it difficult to draw inferences about the magnitude of the posterior variance estimate.

The use of the posterior mode is often justified as being the closest to the maximum likelihood estimate (MLE) when uninformative priors are used. However, this comparison refers to the joint posterior mode, rather than the marginal mode that is typically estimated and reported. In more complex models, the joint and marginal modes may differ (Held & Sabanés Bové, 2020, Section 6.5.4), meaning that the marginal mode and MLE are no longer the same. As shown in Figure S1, the convergence of the posterior mode and MLE also requires the use of uninformative improper priors on the variance, which are generally not advised (Gelman *et al.*, 2021), in part because 'uninformative' priors can be uninformative on one scale but not another (e.g. priors on standard deviation versus variance). The posterior mode is also hard to estimate; it is typically done using kernel density estimation and different methods may provide quite different estimates (Figure 2), thereby providing another source of hidden ambiguity. Furthermore, the mode requires a larger number of samples in the posterior distribution to be reliably estimated, and will show greater variation between models/chains run on the same dataset (Kruschke, 2015). In contrast, the mean is strongly affected by extreme values, and so by the long tail of an asymmetric distribution.

It is also often important to assess statistical support for among-group variance at a particular level. Typically 95% credible intervals (CRIs) are presented as a measure of uncertainty in parameter estimates derived from MCMC models. As variance components cannot overlap zero, CRIs give no information about the compatibility of the estimates with a null hypothesis (e.g. no among-group variance). Posterior distributions are commonly inspected visually as density plots; a right skewed distribution with a mass near 0 is often assumed to signify that the estimated variance is not different from zero. What is seldom appreciated, however, is that the degree of smoothing that is applied in such plots (via the binning interval or bandwidth) can alter these conclusions. The same distribution can be seen as uni- or bimodal, or peaking at zero or away from zero depending on the degree of smoothing (Figure 2). Such assessments are therefore subjective and lack a proper quantitative basis.

To address this, several metrics for assessing the confidence in a result (such as p-values) have been suggested in a Bayesian framework (reviewed in Makowski *et al.*, 2019a). Two of these, Region of Practical Equivalence (ROPE) and Bayes Factors, can be used for variance components. The ROPE approach identifies a range of values considered too small to be of any practical relevance (i.e. the Region of Practical Equivalence), and quantifies the proportion of overlap between the posterior distribution and the ROPE. This is similar to equivalence testing in a frequentist framework, specifically to the two one-sided tests approach (Lakens *et al.*, 2018). Bayes Factors are analogous to frequentist likelihood ratios, comparing different models (for example with and without the random effects of interest). Unlike likelihood ratios, they incorporate information from the prior distributions of the parameters into the comparison of the models, and are evaluated using the marginal likelihood rather than at the maximum likelihood. Additionally, Bayesian models can be compared using information criteria which aim to provide out-of-sample prediction accuracy, of which LOO-CV (Leave-One-Out Cross-Validation; Browne, 2000; Gelman *et al.*, 2014) has been suggested as the most suitable for complex hierarchical models (Gelman *et al.*, 2021). These metrics (ROPE, Bayes Factors, LOO-CV) can be used to provide a measure of statistical support for estimates of variance components, but their implementation is complicated. ROPE requires the definition of a threshold, incorporating further subjectivity into the analysis, whilst the computation of Bayes Factors and LOO-CV can be challenging, and even not implementable in some commonly used programs in ecology and evolution (e.g. MCMCglmm). The use of Bayes Factors and LOO-CV is also the topic of active debate (Gronau & Wagenmakers, 2019a,b; Chandramouli & Shiffrin, 2019; Vehtari *et al.*, 2019; Navarro, 2019; Gelman *et al.*, 2021). We address these metrics further in the discussion.

Here, we suggest a complementary method to assess statistical support in mixed-effect models, which compares variance estimates to a null distribution in order to aid statistical inference. This involves creating a distribution of effect sizes that would be expected under the null hypothesis (no among-group variance), and comparing this null distribution with the observed among-group variance. This method has several advantages. Null distributions can be used to generate a p-value describing the probability that the observed estimate is as or more extreme than expected under the null hypothesis. Although often criticised through their association with Null Hypothesis Significance Testing (NHST; Wasserstein & Lazar, 2016; Amrhein *et al.*, 2017; McShane *et al.*, 2019; Amrhein *et al.*, 2019), p-values have well understood and useful properties. When correctly interpreted, these statistics provide a continuous measure of statistical support, indicating how inconsistent an observed effect size is with a scenario in which there is no among-group variance. In contrast to ROPE, creating null distributions requires no subjective decisions about thresholds and, in contrast to Bayes Factors and LOO-CV, they can be implemented using the output from any mixed model.

We present two methods, permutation and simulation, for generating null distributions for variance components. To generate a null distribution using permutation, some feature of the data is randomised to produce a new dataset with the structure of the original dataset, but with no relationship between the response variable and the variable of interest. This randomization is repeated a large number of times to create many different permuted datasets. The same analysis is then carried out on the permuted datasets as on the original dataset, and a test statistic of interest (e.g. the estimate of among-group variance) is used to create a null distribution of test statistics (Figure 1c,d). A (one-tailed) p-value can then be derived as the proportion of permuted datasets with a test statistic greater than or equal to the test statistic observed with the real data set. Permutation tests have already been suggested as an alternative to likelihood ratio tests for frequentist analyses (Fitzmaurice *et al.*, 2007; Samuh *et al.*, 2012), although they are not commonly utilized in ecology and evolution (but see Araya-Ajoy & Dingemanse, 2017; Stoffel *et al.*, 2017). Permutation tests are a subclass of nonparametric tests (Pesarin & Salmaso, 2010; Lehmann & Romano, 2005) and do not rely on specific probability distributions, and so make few assumptions. However, as we show later in the manuscript, datasets can be permuted in several different ways when the data structure is complex, and the consequences of the choices involved in such cases are often not immediately obvious. An alternative method of creating a null distribution is using simulations. This process is similar to permutation, but instead of generating permuted datasets we can simulate datasets from the observed model parameters (similar to parametric bootstrapping), whilst setting the variance in question to zero. Again, the same analysis is carried out on the simulated datasets, and the test statistics of interest used to create a null distribution. This simulation method makes more assumptions about the data and model, but allows for more control of the manipulated features of the simulated datasets compared with permutations.

Finally, a crucial part of designing experiments and statistical analyses is assessing the power to detect an effect size of interest. Power is defined as the probability of rejecting the null hypothesis for a given effect size at a specified alpha level. Although power typically relates to NHST and the often criticized alpha level (Wasserstein & Lazar, 2016; Amrhein *et al.*, 2017; McShane *et al.*, 2019; Amrhein *et al.*, 2019), it remains an important tool for study design regardless of statistical philosophy, by providing a quantitative approach to calculating optimal sample sizes and designing sampling regimes. Power may also provide a more useful metric than precision when considering variance components. As their distributions are bounded at zero, standard errors will always decrease when distributions are close to zero (see Figure S2). The concept of power for variance components in MCMC models is not well developed, however. As null distributions can be used to generate p-values, they provide a convenient way of conducting power analysis.

Here, we first compare the metrics of central tendency that are commonly used as summary statistics of posterior distributions of variance components. We then demonstrate the utility of null distributions to generate a complementary p-value statistic and aid the interpretation of the variance components, and compare two methods of generating them. Null distributions can provide a continuous, quantitative measure of confidence that the observed variance component is larger than what might be expected under the null hypothesis (no among-group variance), given the data structure and priors used.

Importantly, we are not advocating that this approach should replace the presentation and use of effect sizes and credible intervals, but rather that it should be used as an additional and complementary statistic. Finally, we show how null distributions can be used to perform power analysis within an MCMC framework.

Methods

All simulations were carried out in R (version 4.1.0; R Core Team, 2022) using the squidSim R package (version 0.1.0; Pick, 2022).

Generation of 'observed' datasets

We generated a series of datasets with known parameters, which we will refer to as our 'observed' datasets (to distinguish them from the 'null datasets' in following sections). We first simulated Gaussian data with one hierarchical level and varied the number of observations per group (2 and 4) and the number of groups (20, 40 and 80). We simulated a total variance of 1 and varied the among-group variance (0, 0.1, 0.2 and 0.4; also representing the intra-class correlations (ICCs)/repeatabilities). We simulated every combination of these parameters (24 parameters sets) and for each set we simulated 500 'observed' datasets. Power to detect among-group variance is known to be determined by effect size and sample size both within and among groups. We chose these parameter values and sample sizes to explore scenarios where power is low (Dingemanse & Dochtermann, 2013) to understand the impact on posterior distributions. These sample sizes also correspond to typical experimental designs in behavioral ecology or life history data collected on wild populations (Bell *et al.*, 2009).

We analysed each 'observed' dataset with a linear mixed-effect model specifying group level random effects in a Bayesian framework, using Stan with the rstan package (version 2.21.3; Stan Development Team, 2022a). We specified weakly informative priors on the among-group and residual standard deviations (half-Cauchy distribution with scale 2; a commonly used and recommended prior for variance components (Gelman, 2006)), and ran one chain for each model with 5000 iterations and a warm-up period of 2000 iterations. This ensured an effective sample size in the posterior distribution of the among group variance of >500 across the majority of models (95%). For comparison, we also ran REML models using the lmer function of the lme4 package (version 1.1-29; Bates *et al.*, 2015), the results of which are shown in Figure S3. To ensure that our results were not affected by the choice of the prior, we ran additional models on a subset of the data with a range of different priors (see Supplementary Materials). Changing the prior on the among-group standard deviation did not affect our results, whilst using uninformative priors on the among-group *variance* led to a concordance between REML estimates and posterior mode, as might be expected (Figure S1).

As a demonstration that our findings hold with more complex data, we additionally simulated Bernoulli (binomial with one observation) and Poisson data. Bernoulli data were simulated with 80 groups and 4 observations per group. Among-group effects were simulated from a Gaussian distribution on the latent scale, with a mean of 0 and among-group variances of 0 and 0.2, 0.4 and 0.8. The latent scale response variable was then transformed using the inverse logit function to provide the probabilities, and sampled with a Bernoulli process. Poisson data were simulated with 80 groups and 2 observations per group, with a mean of 2 and a total variance of 0.2 on the latent scale, with among-group variances of 0, 0.02, 0.04 and 0.08 (ICCs of 0 and 0.1, 0.2 and 0.4 on the latent scale). The mean and total variance were chosen based on a literature survey of provisioning data in Pick *et al.* (2023). We took the exponent of the latent scale response variable to provide expected values, and sampled them with a Poisson process. We simulated 500 'observed' datasets for each variance, and analysed the data using GLMMs as outlined above.

Comparison of posterior distribution summary statistics

From the posterior distributions of the among-group variances, we calculated the posterior mean, median and mode, and compared these estimates with the true values.

While calculating the mean and median of the posterior distribution is straightforward, estimating the posterior mode involves some (hidden) assumptions. Typically the mode is estimated using kernel density estimation, which involves fitting a model to the distribution of posterior samples to estimate a density function. The maximum of this function is then calculated over a series of predicted values, to give the estimated mode. One key parameter in kernel density estimation is the bandwidth, which describes the amount of smoothing and is analogous to the number of breakpoints in a histogram. As shown in Figure 2, with the degree of smoothing can affect where the posterior mode is estimated. To explore this further, we estimated the posterior mode using two bandwidth scalings (0.1 and 1; low and high smoothing, respectively), which are the defaults in two commonly used R functions for estimating the mode (MCMCglmm (Hadfield, 2010) and the ggdist and bayestestR packages (Kay, 2022; Makowski *et al.*, 2019b), respectively). Further details about the differences between these functions are presented in the Supplementary Materials. In both cases, the kernel density was estimated using the SJ algorithm (Sheather & Jones, 1991), and the mode was estimated using 512 predicted values with a cut-off point at zero.

To compare these different measures of central tendency, we calculated measures of bias, precision and accuracy. Because variance components are limited by 0, deviations from the mean or simulated values will be smaller at smaller effect sizes. To account for this, we also calculated relative measures. We calculated the bias as $\frac{1}{n}\sum\hat{\theta}_i - \theta$ (where θ is the true value, $\hat{\theta}_i$ is the model estimate from ith simulation in a parameter set, and n is the number of simulations). For the non-zero effect sizes, we also calculated relative bias as $\frac{1}{n}\sum\frac{\hat{\theta}_i-\theta}{\theta}$, and mean absolute error as $\frac{1}{n}\sum\frac{|\hat{\theta}_i-\theta|}{\theta}$. Note this is a also relative measure. Mean absolute error is similar to root mean squared error, and combines bias and precision. We also calculated the precision as $1/\sqrt{\frac{1}{n}\sum(\hat{\theta}_i - \bar{\hat{\theta}})^2}$, and relative precision as $\bar{\hat{\theta}}/\sqrt{\frac{1}{n}\sum(\hat{\theta}_i - \bar{\hat{\theta}})^2}$, where $\bar{\hat{\theta}}$ is the mean of the model estimates across all simulations in a parameter set. Precision is presented in Figure S2.

Creation of null distributions and p-values

We created null distributions for each 'observed' dataset using two methods to generate 'null datasets'. First, we permuted the 'observed' datasets by shuffling the group indices (IDs) to create 100 new permuted null datasets per 'observed' dataset, in which among-group variance is expected to be zero. Second, we used simulations to create 100 null datasets with the same data structure but no among-group variance for each 'observed' dataset. To determine the value of the residual variance for these simulations, we added together the posterior distributions of the among-group variance and residual variance from the models of each original 'observed' dataset, and used the median of the resulting distributions. This ensured that the total variance in the simulated null datasets was the same as in the 'observed' datasets. The choice of the median for this step should have little consequence, as this derived distribution will be estimated with much less uncertainty and so will be symmetric, meaning that the three measures of central tendency will be equivalent. It is important that any fixed effects, including the intercept, are included in the simulations, especially for GLMMs as the expectations will affect the stochastic variance on the data scale. Each of these null datasets was analysed with the same model as the original 'observed' dataset, and the same parameters (the central tendency estimates of the posterior distribution of the among-group variance) were extracted to create the corresponding null distributions. Although we recommend using null distributions with more samples for empirical studies (e.g. 1000), here we used 100 permutations/simulations for each 'observed' datasets in order to reduce the computational burden (500 simulations for 24 parameter sets is 12000 Gaussian datasets, each with 100 permutations and 100 simulations). We calculated a p-value for each 'observed' dataset, as the proportion of estimates in the null distribution that were higher than the estimate from that 'observed' data. We calculated p-values using each central tendency measure, which are compared in Figure S4.

Power analysis and comparison with bias and precision

Power is defined as the probability of rejecting the null hypothesis (no among-group variance in this case) for a given effect size and data structure at a specified alpha level (typically 0.05). Although power is typically interpreted in the context of NHST, power can also be seen as a description of the distribution of p-values expected for a given effect size and data structure (it is the cumulative density at 0.05 for a given p-value distribution). Other descriptions of the p-value distribution (e.g. the mean) would be simple functions of the power (Figure S5). We chose to present power as a description of the distribution of p-values as it is conceptually well understood and frequently used rather than because of any philosophical alignment with NHST.

Using the 'observed' datasets described above, we compared two ways by which power can be calculated. In both methods, power was calculated for the parameter sets where the true value was greater than zero, as the proportion of 'observed' datasets in which the p-value was below a nominal threshold of 0.05. In the first ('full') method, we used the p-values generated above, through comparison of the 'observed' datasets with their null distributions from both permutation and simulation approaches. In the second ('reduced') method, we generated p-values by using model estimates from the 'observed' datasets with zero among-group variance for each data structure (combination of among- and within-group sample sizes) as a null distribution, against which the estimates from 'observed' datasets simulated with among-group variance could be compared. This method of generating p-values is similar to the simulation method of generating null distributions, but uses one null distribution for *all* 'observed' datasets with the same data structure, instead of null distributions for *each* 'observed' dataset. It is therefore massively less computationally intensive for power analyses; exploring power within the parameter space presented here required 12,000 models, rather than 1,212,000.

We were also able to calculate the false positive rate for the 'full' method in the same way as power, using the parameter sets where the simulated value was zero. It was pointless to calculate a false positive rate for the 'reduced' method; by comparing the null distribution with itself, the false positive rate is, by definition, 5%.

As stated above, posterior distributions are expected to be asymmetric when power is low, which is also when we expect biases in the different measures of central tendency. We therefore examined how well power predicts the relative bias of the different measures of central tendency. During the review process, it was suggested that we could use relative precision to account for the appearance of higher precision in effect sizes near zero. We therefore also compared this metric with power, as it may provide an alternative measure to power for study design.

Worked example - Random slopes

Empiricists commonly encounter more complex questions and data structures in their studies than we have presented above. Here we outline a more realistically complex example where the permutation of datasets require some careful decisions.

Random slope models (where group-specific intercepts and slopes are modelled, also known as random regression) provide a good example of this complexity. We will focus here on generating a null distribution for the estimate of among-group variance in slopes. This estimate is based upon the relationship between the predictor variable and response, the distribution of the response variable across groups and the distribution of the predictor variable within and across groups. This structure provides four ways to generate null distributions via permutation that retain different relationships in the observed dataset (illustrated in Figure S6). The first two are general to variance components, and the second two are specific to random regression models.

1. Permuting the response variable. This is the most unspecific permutation. It retains data structure and breaks all relationships with the response, removing the effects of all random factors and predictors, and allows for testing multiple components at the same time. It is a full null model of all biological processes described by the model.

2. Permuting the group identities. This is a more specific permutation. It breaks the relationship between a specific group and both the response and other predictors, and retains associations between predictors and response (including any other random effects linked to different grouping variables). It will remove the effects of both random intercepts and random slopes associated with the grouping factor in question.

3. Permuting the predictor. This is even more specific, targeting random slopes specifically. It retains the group data structure, but breaks link between predictor and response, and the distribution of the predictor across groups. By breaking the link between predictor and response, there is no relationship that can vary between groups (i.e. random slopes).

4. Permuting the predictor within groups. This is the most specific permutation. It is similar to 3) but retains the distribution of the predictor across groups, whilst breaking the link between predictor and response within group.

We can also generate null distributions through simulation. Here we have multiple variance components (intercepts and slopes), and so the simulations can either test one component at a time or both together. In this example, we can either simulate no among-group variance in slopes (adding the variance generated by the random slopes to the residual to ensure the same total phenotypic variance), or simulate no variance in either intercepts or slopes (adding the variance generated by both random intercepts slopes to the residual). We explore these six null distributions using a simulated and a real dataset. They provide a useful contrast between a dataset where we know the true values, and one where the true values are unknown with the potential for greater complexity.

To generate our simulated dataset, we imagined a hypothetical researcher who wants to test whether there is variation among individuals in how temperature affects their body mass. The dataset was simulated with 300 individuals measured 4 times each. Body mass and temperature were both normally distributed. Temperature was scaled to have a mean of 0 and variance of 1, and has an effect on body mass of 0.2 for the average individual. The simulated among individual variance in the intercepts was 0.2 and the phenotypic variance generated by variation in slopes was 0.1 (with no correlation among random slopes and intercepts), while the residual variance was set to 0.7 to ensure a total phenotypic variance not explained by the average effect of the environment was 1. Formulas to estimate the total phenotypic variance in random slope models can be found in Allegue *et al.* (2017). There were no systematic differences in the average temperature experienced by the different individuals.

For our real data example, we employed a study on the aggressive response of great tits (*Parus major*) to intruders in a nestbox population in southern Germany (Araya-Ajoy & Dingemanse, 2017). Data were collected over a 6-year period (2010–2015) for all males during their first breeding attempt each year. A male great tit model was presented with a playback song 1m away from the subject's nest box. Aggression was measured as the minimum distance of the focal male to the model (Araya-Ajoy & Dingemanse, 2014). Territorial intrusions were performed twice during the egg-laying stage and twice during the egg-incubation stage of each focal nest, with males responding, on average, to 2.8 out of the 4 intrusions. Males were also sampled across years, with an average of 1.4 reaction norms per male. In total there was 2854 aggression tests performed to 1042 breeding attempts of 679 individuals. Full details are provided in Araya-Ajoy &

Dingemanse (2014); Araya-Ajoy & Dingemanse (2017).

Both datasets were analysed using random slope mixed-effects models, specifying the environmental predictor (temperature and breeding stage, respectively) as a fixed covariate, and random intercepts and environment slopes across individuals. Breeding stage was coded as zero (for egg-laying) versus one (for incubation), and then mean centred and standardized to standard deviation units (Schielzeth, 2010). We generated six null distributions of posterior medians for each dataset (four permutations and two simulations), as outlined above, with which we compared the estimate of among individual variance in slopes from the observed data. Null distributions were generated based upon the analyses of 1000 null datasets. Models were fitted in a Bayesian framework, using Stan with the rstan package (version 2.21.3, Stan Development Team, 2022a). We specified weakly informative priors on the among-group and residual standard deviation. We ran three chains for the models of the simulated and real observed datasets, and a single chain the models for the null datasets, all with 5,500 iterations and a warm-up period of 500 iterations.

Results

Comparing summary statistics of the posterior distribution

When the simulated among-group variance was zero, all summary statistics were upwardly biased to some extent as the posterior distribution cannot include 0 (Figure 3a; full sampling distributions are shown in Figure S7). Predictably, the posterior mean and median from datasets with zero variance were considerably more upwardly biased for small sample sizes then the mode, with the mean being the most biased. Note that this upward bias was also present in frequentist analyses (see Figure S3), and was not just a feature of Bayesian analyses.

When the simulated among-group variance was non-zero, then the mean, median and mode all appeared to be consistent estimators, in that any bias occurred only at small sample and/or effect sizes. The posterior median generally converged on the simulated value at lower effect and sample sizes (Figure 3b) with a slight tendency to be biased downwards, as compared with the posterior mean, which was upwardly biased, and the posterior mode that was biased towards zero (Figure 3b). Consistent with Figure 2, the bias in the mode depended upon the chosen bandwidth, with the higher smoothing bandwidths showing less bias. We found similar overall patterns in the Poisson and Bernoulli simulations (Figure S8).

In terms of relative precision (Figure 3c), the mean was the most precise estimator, with both estimates of the mode showing considerably lower precision than either median or mean. Similar to the bias, the precision of the different estimators converged as sample size and effect size increased.

In terms of mean absolute error (Figure 3d), a (relative) measure of accuracy that combines bias and precision, the mean and median were very similar, with exception of the lowest sample and effect size combination where the mean was less accurate. The mode was consistently less accurate than the other measures (Figure 3d), although this lower accuracy disappeared at higher sample and effect sizes.

Performance of the null distributions

A p-value is defined as the probability that an estimate equal to or more extreme than the observed estimate would occur under the null hypothesis (i.e. if the true among-group variance is zero). When the null hypothesis is true, we expect a uniform distribution of p-values (we expect 5% of values to be ≤ 0.05, 50% to be ≤ 0.5 etc). When the null hypothesis is false, we expect smaller p-values to become more likely, in line with the power we have to detect an effect. We find exactly these patterns when considering the p-values generated by null distributions. Both permutation and simulation methods produced a uniform distribution of p-values when the simulated among-group variance was zero (Figures 4), and the distributions of p-values from both permutation and simulation methods shift towards zero as the sample size and the magnitude of the variance increase (Figure 4). We found similar results in the Bernoulli and Poisson simulations (Figure S9).

Importantly, although the mean, median and mode were often quite different in magnitude (reflecting skew in the posterior distribution), the inference based upon the p-values did not differ between the different metrics. There were strong correlations between p-values derived with the different central tendency metrics, except when the mode was estimated with less smoothing which produced less consistent p-values (see Figures S4 and S10). P-values were also strongly correlated between simulation and permutation methods (see Figures S11).

Power analyses and comparison with bias and precision

When we used the full method of estimating power, both ways of generating null distributions (permutation and simulation) gave very similar results (Figure 5), with marginally higher power for the permutation method. These power estimates were very similar to previous published estimates for frequentist models (Dingemanse & Dochtermann, 2013). When the among-group variances was simulated as zero, these methods displayed the expected false positive rates of 5% (black points in Figure 5). The reduced method for estimating power, using the same null distribution for all datasets with an effect size > 0 within a particular data structure, generally showed similar power to the other methods (Figure 5). As with the p-values, power was not particularly sensitive to the measure of central tendency used, the highest power being seen in the mode with higher smoothing and the lowest power for the mode with the least smoothing (Figure S12).

As shown in Figure 6, relative bias in all measures of central tendency decreased as power increased. This pattern was similar across Gaussian, Poisson and Bernoulli traits.

Power was also closely related to relative precision (Figure S13) and consequently also to relative bias (Figure S14).

Random slope worked example

In both the simulated and real datasets, the different types of null distributions (generated using two simulations and four permutations; Figure S6) provided the same qualitative results, supporting the conclusion that there was among-individual variation in slopes (Figure 7). For both of these datasets, permuting individual identity created null distributions with a larger mean value of random slope variance that the other permutations. Note that these results should be considered in the context of random regression, and may not generalize to other types of model (see discussion). We therefore generally recommend exploring the particular consequences of different types of permutations for specific datasets where possible, as this may reveal patterns in the data that warrant further exploration.

Discussion

We demonstrate the difficulties of summarising the posterior distributions of variance estimates from MCMC-based models. We describe different methods for generating null distributions that provide useful complementary information alongside the presentation of central tendency and uncertainty that are generally reported. We also show a way in which null distributions could be used to derive a p-value, which is a simple addition to the statistics presented when summarizing a posterior distribution and also facilitates power analysis. Importantly we show that biases in central tendency measures are functions of power.

Summary statistics

Our experience in ecology and evolution is that both posterior mean and mode are commonly, but inconsistently, presented without justification. For fixed effect parameter estimates, this is typically inconsequential, as the posteriors are usually symmetrically distributed. For estimates of variance components, however, our simulations show that depending upon the underlying parameter value, both mean and mode can show large biases in opposite directions. When posterior distributions were close to zero and there *was* among-group variance, the posterior mode was very biased towards zero, whereas the posterior median and mean performed better. On the other hand, if there was no among-group variance, the mode was the least biased. The mode, however, is more subjective as its estimation depends on the choice of underlying algorithm (results shown here), it requires larger posterior distributions for reliable estimation, and will show greater variation between models/chains (Kruschke, 2015). Unfortunately, the method of mode estimation is rarely justified or even stated in empirical papers. Therefore, we cautiously recommend the presentation of the posterior median, or both median and mean, as a measure of central tendency for variance components. This recommendation is based upon the median being generally less biased than the mean when power is low (Figure 6). Presenting both allows any discrepancy to be seen, which would indicate that the distribution is near to zero and not symmetric, and further emphasize the uncertainty in these measures.

Upward biases in variance components have been seen before when power is low, but the dependence on the choice of the central tendency metric has not been highlighted. For example, Fay *et al.* (2022) note overestimation of variance components in Bernoulli models, with this overestimation decreasing in size as sample size and effect size increase. Fay *et al.* (2022) use the posterior mean as a summary statistic, and (as we show in Figure S15) this bias will decrease (although not disappear completely) through the use of a posterior median. This is not just a bias in Bernoulli models, or in fact MCMC models (Figure S3), but a general property of variance components estimated with low power (Figure 6, or low relative precision - Figure S14).

We urge caution in interpreting our results in terms of absolute sample sizes or effect sizes alone. Different types of data and data structures will contain different amounts of information and so vary in power, meaning that the same bias might not result from the same sample size or variance in a different context. GLMMs make this more complex, as similar variances on the latent scale can equate to different variances and so different effect sizes on the expected and observed scales, depending on the link function and the form of stochastic variance (de Villemereuil *et al.*, 2018). For example, we found a comparable range of powers for our Poisson and Bernoulli examples, despite very different simulated variances on the latent scale (0.02, 0.04 and 0.08 versus 0.2, 0.4 and 0.8, respectively). Similarly, Bonnet & Postma (2015) found very different power to detect the same latent scale variances in Bernoulli and Poisson traits. Given the strong relationship between these biases and power (or relative precision), considering the potential bias in variance estimates in relation to power (or relative precision) may be a productive way forward, as this is comparable across models, distributions, effect and sample sizes.

It is commonly argued that rather than presenting summary statistics, we should present and interpret the whole posterior distribution, typically portrayed using density plots. However, the underlying parameters of the kernel density estimation are not given alongside density plots, meaning the amount of smoothing is unknown. A large degree of smoothing can hide asymmetry and/or bi-modality, and so change inferences. We therefore suggest the use of histograms over density plots in the presentation of posterior distributions, because although histograms are subject to the same smoothing problems, the degree of smoothing is at least explicit. Alternatively, other plots that explicitly show the raw posterior samples can used (e.g. beeswarm plots; Figures 4 and 7).

Null distributions

The null distribution approaches outlined here are relatively easy to use, although computationally intensive (see section 'Computational burden'). They allow the quantification of confidence that the estimated group level variance is not simply a consequence of the choice of priors and data structure. Importantly, the p-values based upon null distributions are not dependent upon which measure of central tendency is used. Such inferential statistics comparing the observed estimates with the null distributions can provide quantitative measures that can be reported alongside the observed estimates and uncertainty, and provides a useful tool for assessing the probability that variance components are non-zero and thereby supplement visual inspections of posterior distributions, or comparison of posterior mode, median and mean. Furthermore, inferential statistics can serve as an objective and easy-to-communicate assessment of the biological relevance of an estimated variance component to the general public and policy makers, or for the statistical support of non-zero values for derived statistics like heritability, repeatability or evolvability. A common criticism of p-values is that they are often misinterpreted. We therefore recommend those using the null distribution approach to acquaint themselves with the relevant literature (useful examples include: Wasserstein & Lazar, 2016; Amrhein *et al.*, 2017; McShane *et al.*, 2019; Amrhein *et al.*, 2019). Importantly, p-values cannot demonstrate absence of effect, just confidence in difference from the null hypothesis. We believe generating null distributions will help empiricists understand these concepts, as they give a visual representation of what p-values signify.

Increasing the complexity of the data structure and model will create more ways to permute datasets. In our random slope examples, we showed how these permutations can become increasing specific to target particular components of the model, from permuting the response to permuting the environmental predictor within individuals. Here, these different permutations led to qualitatively similar results, although whether they always or usually do so would require a much broader set of simulations. Interestingly, permuting individual identity created null distributions with noticeably larger values of random slope variance. We believe this is due to the existence of random slopes generating heterogeneous residuals (i.e. variance in response changed with the environmental predictor) that were confounded with random slope variation in the analyses of the null data sets (similar effects were also shown in Ramakers *et al.*, 2020), whereas the other permutation methods broke up the relationship between the predictor and response. Comparing the results of the different methods of null distributions generation may provide insights that help inform statistical inference or highlight the need for further exploration.

The bulk of the simulations presented here do not directly consider how to assess multiple variance components. In our random slope examples, it made little difference whether we simulated no variance in random slopes and intercepts or just random slopes. However, this may differ between model structures. Generating null distributions for all components at once (for example by permuting the response variable, or setting all random effect variances to 0 in simulations) makes the assumption that the different variance components do not affect each other. If this assumption is reasonable (it is typically being made when a given model structure is chosen to be appropriate), then generating null distributions for all components at once would be reasonable. If there is a reason to think that they may affect each other, then null distributions are better generated for each random term at a time.

In some instances, generating a null distribution using permutations may not be possible. For example, in event-history models of survival (where individuals have a a sequence of 0/1 (survived/died) for each time point where they are observed), permuting the individual identifiers would fundamentally alter the data structure, meaning that some individuals had multiple deaths. However, this could work in the context of an animal model, where 0's and 1's could be interchanged between individuals, retaining the same structure across individuals, but breaking the link with the pedigree. This demonstrates that the suitability of permutations and how they impact the data structure needs to be carefully assessed on a case-by-case basis. Overall, we are not advocating a specific recipe for permutations - it is likely context and question dependent. We instead advocate a simulation approach at the planning stage to check in advance that the permutation design gives desired properties with your likely data structure.

Generating null distributions through simulation avoids many of the issues with the permutation approach, although it may not account so well for the particularities of each dataset, (for example, the heteroskedasticity in the random regression examples above). Simulations allow the structure of the data to be fully retained, allow a more fine-scale alternation of the variances in question, and make no additional assumptions than those already made by the statistical model. A simulation approach also simplifies the simultaneous generation of null distributions for multiple variance components whilst retaining the data structure. Reassuringly, in our random regression examples, the null distributions generated using two simulation methods were similar, and the results were similar to those obtained using the permutation methods. We therefore cautiously recommend the use of this simulation method, as it is the most flexible for complex models.

These null distribution approaches are computationally intensive and the suitability of their application will depend upon the model complexity, the amount of data and the available computational resources (see section 'Computational burden'). MCMC methods are often used for highly complex problems (e.g. double hierarchical GLMs; Cleasby *et al.*, 2015), where generating a large number of samples for a null distribution may not be an option. The number of samples affects the minimum p-value that can be calculated and its precision - a null distribution with 100 samples can have a minimum p-value of 0.01 and vary by intervals of 0.01. In addition, stochastic fluctuations in the p-value can have a large impact on inference. For this reason, we would recommend a higher number of samples in the null distributions than we used here. We remain neutral to the application of NHST, preferring the use of p-values as a continuous measure of statistical support. However, if NHST is employed, researchers must ensure that a large number of samples is used, to prevent inference being based on a handful of rare events. Note that, although not advisable for NHST, we were able to produce meaningful results with 100 simulations, which provided information (although much less reliably) of how incompatible the observed variance was with the range expected under the null hypothesis.

Alternative approaches

Use of a p-value relies upon the distribution of p-values being uniform when the null hypothesis is true, a property that is expected to be invariant to sample size (as we show in Figure 4). P-values therefore only provide support against the null hypothesis; they do not provide support for the null hypothesis. In contrast to p-values, the ROPE value and Bayes Factors aim to additionally assess support for the null hypothesis, and therefore depend upon sample size under both the null and alternative hypotheses. These alternatives are not always simple to implement, and below we outline some potential issues that empiricists may encounter.

The ROPE (Region of Practical Equivalence) introduces another source of subjectivity into the analysis through defining an arbitrary threshold. This is not trivial for variance components, as small variances can have large knock-on effects. For example, McFarlane *et al.* (2015) found that maternal genetic effects accounted for 2% of variation in fitness, which predicted a 56% increase in mean lifetime reproductive success under 10 generations. Bonnet *et al.* (2022) employed a ROPE approach, using simulations to demonstrate the biological relevance of the thresholds they use. ROPE is often discussed in a context where a cost-benefit analysis can be used to work out the minimum effect size that warrants the use of a particular (e.g. medical) intervention (Kruschke, 2018). Whilst there are clear applications for using ROPE in fields like conservation, where interaction with stakeholders requires thresholds over which decisions are made, for many empiricists, ROPE requires more subjective decisions to be made and justified.

Bayes Factors can be used to test the 'significance' of parameters in Bayesian mixed-effect models. However, the calculation of Bayes Factors is not straightforward. They require large posterior distributions for stable estimation (Schad *et al.*, 2022). They also depend on the marginal likelihoods of the two models which are sensitive to prior specification (Gelman *et al.*, 2021; Navarro, 2019; Schad *et al.*, 2022), even when there is little or no visible effect on the posteriors. Using Bayes Factors as a measure of posterior odds also assumes equal probability of the two models, and it is not clear whether this is a reasonable assumption as some would argue that some among-group variance always exists.

Bayesian models can also be compared using information criteria, in particular DIC (Deviance Information Criteria; Spiegelhalter *et al.*, 2002), WAIC (Widely Applicable Information Criteria; Watanabe, 2010) and LOO-CV (Leave-One-Out Cross-Validation; Browne, 2000; Gelman *et al.*, 2014), which aim to provide out-of-sample prediction accuracy. DIC has several problems which in part come from being based on a point estimate (Plummer, 2008), and provides poor estimates when posterior distributions are not well described by their means (Gelman *et al.*, 2021). WAIC addresses these issues by using the whole posterior. However, some assumptions of WAIC do not hold for hierarchical models with weak priors (Gelman *et al.*, 2014; Millar, 2018). LOO-CV may, therefore, be the most suitable information criteria for this purpose. It is also important whether these information criteria are generated using marginal or conditional likelihoods (Millar, 2018; Merkle *et al.*, 2019; Ariyo *et al.*, 2020) - although the marginal likelihood may be more appropriate for comparing hierarchical models, many software packages only (MCMCglmm) or by default (BUGS, JAGS, Stan) provide the conditional likelihood.

The use of both LOO-CV and Bayes Factors for complex models is currently the subject of intense debate. Regardless of the various intricacies of this debate, perhaps a more constraining factor is that Bayes Factors and LOO-CV are not implementable in all programs, including those commonly used for variance component estimation in ecology and evolution (i.e. MCMCglmm). Our approach provides an alternative to these methods, which is easily implemented and allows straightforward interpretation.

Power analysis and possible alternatives

Power analysis is controversial because of its link to NHST, and the misuse of NHST has been linked to scientific misconduct and the replication crisis (Wasserstein & Lazar, 2016; Amrhein *et al.*, 2017; McShane *et al.*, 2019; Amrhein *et al.*, 2019). Whilst these criticisms relate to the use of p-values *after* data collection and analysis, power analysis is typically conducted *pre*-analysis, and serves a clear purpose in aiding experimental design. Power can also be seen as a description of the distribution of p-values expected for a given effect size and data structure. Other descriptions of this distribution (e.g. the mean) would be simple functions of the power (Figure S5), but the common use of this metric makes it more widely understood. One suggested alternative, Type M error (absolute relative bias of significant estimates), also relies upon calculation of p-values and an arbitrary alpha value, and is a simple function of power (Gelman & Carlin, 2014). Unlike power, Type M error is affected by the measure of central tendency that is chosen (Figure S16). Another alternative to power is to design studies around a desired level of precision in estimates. Although this works for unbounded parameters, precision is difficult to interpret for variance components, because it increases as the true value gets closer to zero due to the constraint at zero (see Figure S2). Using relative precision (the inverse of the coefficient of variation of the sampling distribution) avoids this problem. It is strongly related to power (Figure S13), and so optimizing this value may provide an alternative target for planning optimal experimental designs. The relative precision is, however, also highly dependent on the measure of central tendency used. We would therefore suggest that power still provides a suitable metric for designing studies to estimate variance components.

We show two methods of power analysis based upon null distributions. The first (full) method involved generating p-values by creating a null distribution for each 'observed' dataset. This method is highly computationally intensive as it involves running a certain number of simulations multiplied by the number of permutations/simulations models, which could realistically be one million models per parameter. Our alternative (reduced) method involved generating p-values by comparing the parameter estimates from the 'observed' datasets to a single null distribution for each data structure. Whilst the two methods estimated similar power, the reduced method was massively less computationally intensive (requiring running 2000 models rather than a million for each set of parameters). The disadvantage is that a false positive rate cannot be calculated.

Even if power is not the intended use, these simulations can still serve an extremely useful purpose before studies are conducted. First, these simulations allow an empiricist to consider the distribution of p-values expected under a given effect size and design (power is essentially a description of the shape of this distribution). Second, the null distribution of point estimates can be visualised. Even if an empiricist does not want to calculate a p-value, creating a null distribution is a powerful way of considering the distribution of estimates that would be generated with no among-group variance, and would serve to encourage caution in how results that lie within that distribution are interpreted.

Computational burden

Null distribution approaches can be computationally intensive. When model complexity and/or sample sizes are high, applying them can take a long time, and may prohibit their use. There are several points in this regard that are worth noting. First, these computational constraints will become increasingly less problematic with advances in computing and software. For example, the introduction of Stan has led to a considerable decrease in computation time for many MCMC models, and the increased availability of computer clusters means that null distribution can be generated in parallel. Second, the mean and median require far lower effective sample size than credible intervals to be well estimated (Vehtari *et al.*, 2021). ‘Null’ models can therefore be run for much shorter times than the original model, as only the mean/median is needed. Third, other metrics are also computationally expensive. For example, the generation of Bayes Factors and LOO-CV requires running two models with much larger posterior distributions (1-2 orders of magnitude larger; Vehtari *et al.*, 2017; Gronau *et al.*, 2020), followed by additional computationally expensive steps. Finally, our suggested method will be the most efficient for power analysis. Whereas each Bayes Factors and LOO-CV require two models with large posteriors, in our method the same null distribution can be used for all simulated datasets with the same data structure, requiring models with much smaller posteriors. Relative precision is even less computationally intensive to generate, but perhaps slightly harder to interpret. Overall, the computational burden of generating a null distribution is perhaps not so high when compared to other alternatives.

There will be cases in which none of these methods (null distributions, Bayes Factors or LOO-CV) will be feasible for computational reasons. Are there any less computationally expensive alternatives? The ROPE method provides a clear advantage here as it requires no additional computationally expensive steps to generate, although it may be difficult to apply with variance components. We realised when considering the relative precision as a metric for the sampling distributions that for an individual posterior distribution this metric (mean/SD) is analogous to a z-ratio. Interpretation in this context is a little strange, and z-ratios are typically used to represent the potential overlap of the uncertainty of a parameter estimate with 0, which cannot occur here. However, this kind of method is used with variance components in frequentist models that report the SEs of the variance components (e.g. when estimating genetic variance/heritability in ASReml (Butler *et al.*, 2017)). Ultimately, we are looking for a usable statistic to describe the support for a difference between the variance component estimate and 0. These metrics would be considerably less computationally intensive to generate than a p-value from a null distribution, but may give similar information about the model estimates. Comparing them for individual models shows this appears to be true; the z-ratio correlates strongly with p-value (Figure S17a). This statistic (posterior mean/posterior SD) may therefore provide some inferences about the posterior distribution of variance components, although it is much more conservative than a p-value generated from null distributions (Figure S17b). Whilst this may provide an interesting solution to the problems of computational power, use of the z-ratio requires further exploration before being implemented.

Recommendations

1. We advocate using the posterior median as a measure of central tendency for posterior distributions of variance components from MCMC-based models. Our results show that the median is the least biased estimate, but will overestimate variances when power is low. Reporting multiple measures of central tendency allows any asymmetry in the posterior to be made obvious.

2. We advocate reporting of smoothing values in kernel estimation. Kernel density estimation is commonly used for estimating the posterior mode and creating density plots. The parameters used in this estimation are seldom reported, but can have a large impact on interpretation. We advise the reporting of parameters in the kernel density estimation, or the use of more explicit methods of plotting posterior distributions, such as histograms.

3. We recommend using null distributions for inference. Null distributions provide a way of putting the observed parameter estimates into a context expected under an explicitly defined null hypothesis (i.e. no among-group variance). Null distributions can be created in multiple ways, but they are most easily controlled when generated using simulations. As with many aspects of statistical analysis, there are many decisions relating to generating null distributions that may have an affect on the results. Therefore, these methods should be defined pre-analysis, in order to reduce researcher degrees of freedom (Simmons *et al.*, 2011).

4. We also advocate for using a null distribution to estimate power. As well as aiding *post-hoc* inference, null distributions can be used for power analysis. We provide details of a method for doing so that does not present a large computational burden.

Acknowledgments

This is fourth contribution of the Statistical Quantification of Individual Differences (SQuID) working group, and we'd like to thank the other members of SQuID and the Wild Evolution and Statistics in Ecology and Evolution Discussion groups at the University of Edinburgh for valuable feedback on the ideas presented here. We also thank Pierre de Villemereuil and Rémi Fay, whose reviews greatly improved the quality of the manuscript. Work on this project at SQuID workshops in 2022 and JLP were funded by a Research Council of Norway INTPART project number 309356 grant to JW. YGA was supported by the Research Council of Norway project number 325826. JW and YGA were also partially supported by the Research Council of Norway (SFF-III 223257/ F50). HS was supported by the German Research Foundation (DFG, 215/543-1, 316099922). DFW was supported by the U.S. National Science Foundation (NSF). KLL was supported by the NSF (IOS-2100625). HA was supported by the Natural Sciences and Engineering Research Council of Canada (CGSD3-504399-2017) and the Fond de Recherche du Québec - Nature et Technologies (FRQNT; 283511).

Conflict of Interest statement

The authors declare no conflict of interest.

Author Contributions

JLP, CK, NJD, DFW and YGAA conceived the ideas; JLP, YGAA, HS and NAD designed methodology; JLP and YGAA ran the simulations; All authors contributed to the interpretation of results; JLP and YGAA led the writing of the manuscript, and all authors contributed critically to the drafts and gave final approval for publication.

Data and code availability

All code and generated data for the simulated examples are deposited in `https://github.com/squidgroup/null_distributions`

References

Allegue, H., Araya-Ajoy, Y.G., Dingemanse, N.J., Dochtermann, N.A., Garamszegi, L.Z., Nakagawa, S., Réale, D., Schielzeth, H. & Westneat, D.F. (2017) Statistical Quantification of Individual Differences (SQuID): an educational and statistical tool for understanding multilevel phenotypic data in linear mixed models. *Methods in Ecology and Evolution*, **8**, 257–267. https://dx.doi.org/10.1111/2041-210X.12659.

Amrhein, V., Greenland, S. & McShane, B. (2019) Scientists rise up against statistical significance. *Nature*, **567**, 305–307. https://dx.doi.org/10.1038/d41586-019-00857-9.

Amrhein, V., Korner-Nievergelt, F. & Roth, T. (2017) The earth is flat ($p > 0.05$): Significance thresholds and the crisis of unreplicable research. *PeerJ*, **5**, e3544. https://dx.doi.org/10.7717/peerj.3544.

Araya-Ajoy, Y.G. & Dingemanse, N.J. (2014) Characterizing behavioural 'characters': an evolutionary framework. *Proceedings of the Royal Society of London Series B*, **281**, 20132645. https://dx.doi.org/10.1098/rspb.2013.2645.

Araya-Ajoy, Y.G. & Dingemanse, N.J. (2017) Repeatability, heritability, and age-dependence of seasonal plasticity in aggressiveness in a wild passerine bird. *Journal of Animal Ecology*, **86**, 227–238. https://dx.doi.org/10.1111/1365-2656.12621.

Ariyo, O., Quintero, A., Muñoz, J., Verbeke, G. & Lesaffre, E. (2020) Bayesian model selection in linear mixed models for longitudinal data. *Journal of Applied Statistics*, **47**, 890–913. https://dx.doi.org/10.1080/02664763.2019.1657814.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48. https://dx.doi.org/10.18637/jss.v067.i01.

Bell, A.M., Hankison, S.J. & Laskowski, K.L. (2009) The repeatability of behaviour: a meta-analysis. *Animal Behaviour*, **77**, 771–783. https://dx.doi.org/10.1016/j.anbehav.2008.12.022.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.S.S. (2009) Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology and Evolution*, **24**, 127–135. https://dx.doi.org/10.1016/j.tree.2008.10.008.

Bonnet, T., Morrissey, M.B., de Villemereuil, P., Alberts, S.C., Arcese, P., Bailey, L.D., Boutin, S., Brekke, P., Brent, L.J.N., Camenisch, G., Charmantier, A., Clutton-Brock, T.H., Cockburn, A., Coltman, D.W., Courtiol, A., Davidian, E., Evans, S.R., Ewen, J.G., Festa-Bianchet, M., de Franceschi, C., Gustafsson, L., Höner, O.P., Houslay, T.M., Keller, L.F., Manser, M., McAdam, A.G., McLean, E., Nietlisbach, P., Osmond, H.L., Pemberton, J.M., Postma, E., Reid, J.M., Rutschmann, A., Santure, A.W., Sheldon, B.C., Slate, J., Teplitsky, C., Visser, M.E., Wachter, B. & Kruuk, L.E.B. (2022) Genetic variance in fitness indicates rapid contemporary adaptive evolution in wild animals. *Science*, **376**, 1012–1016. https://dx.doi.org/10.1126/science.abk0853.

Bonnet, T. & Postma, E. (2015) Successful by chance? The power of mixed models and neutral simulations for the detection of individual fixed heterogeneity in fitness components. *American Naturalist*, **187**, 60–74. https://dx.doi.org/10.1086/684158.

Browne, M.W. (2000) Cross-Validation Methods. *Journal of Mathematical Psychology*, **44**, 108–132. https://dx.doi.org/10.1006/jmps.1999.1279.

Butler, D., Cullis, B., Gilmour, A., Gogel, B. & Thompson, R. (2017) ASReml-R Reference Manual.

Chandramouli, S.H. & Shiffrin, R.M. (2019) Commentary on Gronau and Wagenmakers. *Computational Brain & Behavior*, **2**, 12–21. https://dx.doi.org/10.1007/s42113-018-0017-1.

Cleasby, I.R., Nakagawa, S. & Schielzeth, H. (2015) Quantifying the predictability of behaviour: Statistical approaches for the study of between-individual variation in the within-individual variance. *Methods in Ecology and Evolution*, **6**, 27–37. https://dx.doi.org/10.1111/2041-210X.12281.

de Villemereuil, P., Morrissey, M.B., Nakagawa, S. & Schielzeth, H. (2018) Fixed-effect variance and the estimation of repeatabilities and heritabilities: issues and solutions. *Journal of Evolutionary Biology*, **31**, 621–632. https://dx.doi.org/10.1111/jeb.13232.

de Villemereuil, P., Gimenez, O. & Doligez, B. (2013) Comparing parent–offspring regression with frequentist and Bayesian animal models to estimate heritability in wild populations: A simulation study for Gaussian and binary traits. *Methods in Ecology and Evolution*, **4**, 260–275. https://dx.doi.org/10.1111/2041-210X.12011.

Dingemanse, N.J. & Dochtermann, N.A. (2013) Quantifying individual variation in behaviour: Mixed-effect modelling approaches. *Journal of Animal Ecology*, **82**, 39–54. https://dx.doi.org/10.1111/1365-2656.12013.

Fay, R., Authier, M., Hamel, S., Jenouvrier, S., van de Pol, M., Cam, E., Gaillard, J.M., Yoccoz, N.G., Acker, P., Allen, A., Aubry, L.M., Bonenfant, C., Caswell, H., Coste, C.F.D., Larue, B., Le Coeur, C., Gamelon, M., Macdonald, K.R., Moiron, M., Nicol-Harper, A., Pelletier, F., Rotella, J.J., Teplitsky, C., Touzot, L., Wells, C.P. & Sæther, B.E. (2022) Quantifying fixed individual heterogeneity in demographic parameters: Performance of correlated random effects for Bernoulli variables. *Methods in Ecology and Evolution*, **13**, 91–104. https://dx.doi.org/10.1111/2041-210X.13728.

Fitzmaurice, G.M., Lipsitz, S.R. & Ibrahim, J.G. (2007) A Note on Permutation Tests for Variance Components in Multilevel Generalized Linear Mixed Models. *Biometrics*, **63**, 942–946. https://dx.doi.org/10.1111/j.1541-0420.2007.00775.x.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. & Rubin, D.B. (2021) *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition.

Gelman, A., Hill, J. & Vehtari, A. (2020) *Regression and Other Stories*. Cambridge University Press, Cambridge.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515–533.

Gelman, A. & Carlin, J. (2014) Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, **9**, 641–651. https://dx.doi.org/10.1177/1745691614551642.

Gelman, A., Hwang, J. & Vehtari, A. (2014) Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, **24**, 997–1016. https://dx.doi.org/10.1007/s11222-013-9416-2.

Gilks, W.R., Thomas, A. & Spiegelhalter, D.J. (1994) A Language and Program for Complex Bayesian Modelling. *Journal of the Royal Statistical Society Series D (The Statistician)*, **43**, 169–177. https://dx.doi.org/10.2307/2348941.

Gronau, Q.F., Singmann, H. & Wagenmakers, E.J. (2020) bridgesampling: An R Package for Estimating Normalizing Constants. *Journal of Statistical Software*, **92**, 1–29. https://dx.doi.org/10.18637/jss.v092.i10.

Gronau, Q.F. & Wagenmakers, E.J. (2019a) Limitations of Bayesian Leave-One-Out Cross-Validation for Model Selection. *Computational Brain & Behavior*, **2**, 1–11. https://dx.doi.org/10.1007/s42113-018-0011-7.

Gronau, Q.F. & Wagenmakers, E.J. (2019b) Rejoinder: More Limitations of Bayesian Leave-One-Out Cross-Validation. *Computational Brain & Behavior*, **2**, 35–47. https://dx.doi.org/10.1007/s42113-018-0022-4.

Hadfield, J.D. (2010) MCMC methods for multi-response generalized linear mixed models: The {MCMCglmm} {R} package. *Journal of Statistical Software*, **33**, 1–22. https://dx.doi.org/10.1002/ana.23792.

Hadfield, J.D. & Nakagawa, S. (2010) General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, **23**, 494–508. https://dx.doi.org/10.1111/j.1420-9101.2009.01915.x.

Harrison, X.A., Donaldson, L., Correa-Cano, M.E., Evans, J., Fisher, D.N., Goodwin, C.E.D., Robinson, B.S., Hodgson, D.J. & Inger, R. (2018) A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, **6**, e4794. https://dx.doi.org/10.7717/peerj.4794.

Held, L. & Sabanés Bové, D. (2020) *Likelihood and Bayesian Inference*, volume 10. Springer.

Henderson, C.R. (1988) Theoretical basis and computational methods for a number ofdifferent animal models. *Journal of Dairy Science*, **71**, 1–16.

Houle, D. (1992) Comparing evolvability and variability of quantitative traits. *Genetics*, **130**, 195–204. https://dx.doi.org/citeulike-article-id:10041224.

Kay, M. (2022) *ggdist: Visualizations of Distributions and Uncertainty.* R package version 3.2.0.

Kruschke, J. (2015) *Doing Bayesian Data Analysis.* Acadmiec Press/Elsevier, second edition.

Kruschke, J. (2018) Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, **1**, 270–280. https://dx.doi.org/10.1177/2515245918771304.

Kruuk, L.E.B. (2004) Estimating genetic parameters in natural populations using the "animal model". *Philosophical Transactions of the Royal Society of London B*, **359**, 873–890. https://dx.doi.org/10.1098/rstb.2003.1437.

Lakens, D., Scheel, A.M. & Isager, P.M. (2018) Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, **1**, 259–269. https://dx.doi.org/10.1177/2515245918770963.

Lehmann, E.L. & Romano, J.P. (2005) *Testing Statistical Hypotheses.* Springer Texts in Statistics. Springer, New York, 3rd ed edition.

Lemoine, N.P. (2019) Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, **128**, 912–928. https://dx.doi.org/10.1111/oik.05985.

Makowski, D., Ben-Shachar, M.S., Chen, S.H.A. & Lüdecke, D. (2019a) Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, **10**, 2767. https://dx.doi.org/10.3389/fpsyg.2019.02767.

Makowski, D., Ben-Shachar, M.S. & Lüdecke, D. (2019b) bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, **4**, 1541. https://dx.doi.org/10.21105/joss.01541.

McElreath, R. (2020) *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, 2nd edition.

McFarlane, S.E., Gorrell, J.C., Coltman, D.W., Humphries, M.M., Boutin, S. & Mcadam, A.G. (2015) The nature of nurture in a wild mammal's fitness. *Proceedings of the Royal Society of London B*, **282**, 1–7.

McShane, B.B., Gal, D., Gelman, A., Robert, C. & Tackett, J.L. (2019) Abandon Statistical Significance. *The American Statistician*, **73**, 235–245. https://dx.doi.org/10.1080/00031305.2018.1527253.

Merkle, E.C., Furr, D. & Rabe-Hesketh, S. (2019) Bayesian Comparison of Latent Variable Models: Conditional Versus Marginal Likelihoods. *Psychometrika*, **84**, 802–829. https://dx.doi.org/10.1007/s11336-019-09679-0.

Millar, R.B. (2018) Conditional vs marginal estimation of the predictive loss of hierarchical models using WAIC and cross-validation. *Statistics and Computing*, **28**, 375–385. https://dx.doi.org/10.1007/s11222-017-9736-8.

Nakagawa, S. & Schielzeth, H. (2010) Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. *Biological Reviews of the Cambridge Philosophical Society*, **85**, 935–56. https://dx.doi.org/10.1111/j.1469-185X.2010.00141.x.

Navarro, D.J. (2019) Between the Devil and the Deep Blue Sea: Tensions Between Scientific Judgement and Statistical Model Selection. *Computational Brain & Behavior*, **2**, 28–34. https://dx.doi.org/10.1007/s42113-018-0019-z.

O'Hara, R.B., Cano, J.M., Ovaskainen, O., Teplitsky, C. & Alho, J.S. (2008) Bayesian approaches in evolutionary quantitative genetics. *Journal of Evolutionary Biology*, **21**, 949–957. https://dx.doi.org/10.1111/j.1420-9101.2008.01529.x.

O'Hara, R.B. & Merilä, J. (2005) Bias and precision in QST estimates: Problems and some solutions. *Genetics*, **171**, 1331–1339. https://dx.doi.org/10.1534/genetics.105.044545.

Pesarin, F. & Salmaso, L. (2010) *Permutation Tests for Complex Data.* John Wiley & Sons, Ltd, first edition.

Pick, J.L. (2022) *squidSim: a flexible simulation tool for linear mixed models.* R package version 0.1.0.

Pick, J.L., Khwaja, N., Spence, M.A., Ihle, M. & Nakagawa, S. (2023) Counter culture: causes, extent and solutions of systematic bias in the analysis of behavioural counts. *PeerJ*, **11**, e15059. https://dx.doi.org/10.7717/peerj.15059.

Plummer, M. (2003) Jags: A program for analysis of bayesian graphical models using gibbs sampling. *3rd International Workshop on Distributed Statistical Computing (DSC 2003); Vienna, Austria*, **124**.

Plummer, M. (2008) Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9**, 523–539. https://dx.doi.org/10.1093/biostatistics/kxm049.

R Core Team (2022) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ramakers, J.J.C., Visser, M.E. & Gienapp, P. (2020) Quantifying individual variation in reaction norms: Mind the residual. *Journal of Evolutionary Biology*, **33**, 352–366. https://dx.doi.org/10.1111/jeb.13571.

Samuh, M.H., Grilli, L., Rampichini, C., Salmaso, L. & Lunardon, N. (2012) The Use of Permutation Tests for Variance Components in Linear Mixed Models. *Communications in Statistics - Theory and Methods*, **41**, 3020–3029. https://dx.doi.org/10.1080/03610926.2011.587933.

Schad, D.J., Nicenboim, B., Bürkner, P.C., Betancourt, M. & Vasishth, S. (2022) Workflow techniques for the robust use of bayes factors. *Psychological Methods*, pp. No Pagination Specified–No Pagination Specified. https://dx.doi.org/10.1037/met0000472.

Schielzeth, H. (2010) Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, **1**, 103–113. https://dx.doi.org/10.1111/j.2041-210X.2010.00012.x.

Sheather, S.J. & Jones, M.C. (1991) A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society Series B (Methodological)*, **53**, 683–690.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Simmons, J.P., Nelson, L.D. & Simonsohn, U. (2011) False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, **22**, 1359–1366. https://dx.doi.org/10.1177/0956797611417632.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639. https://dx.doi.org/10.1111/1467-9868.00353.

Stan Development Team (2022a) RStan: the R interface to Stan. R package version 2.21.3.

Stan Development Team (2022b) Stan modeling language users guide and reference manual. Version 2.3.

Stoffel, M.A., Nakagawa, S. & Schielzeth, H. (2017) rptR: Repeatability estimation and variance decomposition by generalized linear mixed-effects models. *Methods in Ecology and Evolution*, **8**, 1639–1644. https://dx.doi.org/10.1111/2041-210X.12797.

Vehtari, A., Gelman, A. & Gabry, J. (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, **27**, 1413–1432. https://dx.doi.org/10.1007/s11222-016-9696-4.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. & Bürkner, P.C. (2021) Rank-Normalization, Folding, and Localization: An Improved Rˆ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, **16**, 667–718. https://dx.doi.org/10.1214/20-BA1221.

Vehtari, A., Simpson, D.P., Yao, Y. & Gelman, A. (2019) Limitations of "Limitations of Bayesian Leave-one-out Cross-Validation for Model Selection". *Computational Brain & Behavior*, **2**, 22–27. https://dx.doi.org/10.1007/s42113-018-0020-6.

Wasserstein, R.L. & Lazar, N.A. (2016) The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, **70**, 129–133. https://dx.doi.org/10.1080/00031305.2016.1154108.

Watanabe, S. (2010) Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, **11**, 3571–3594.

Figures

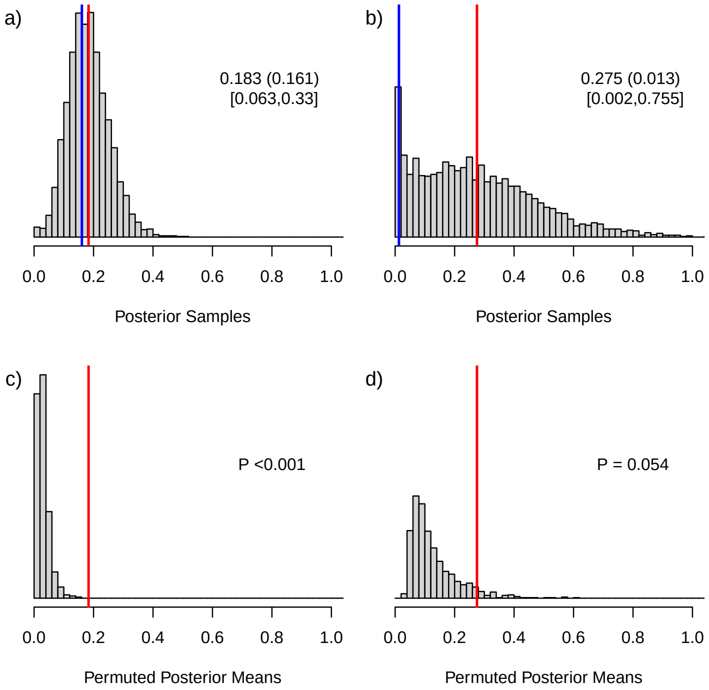


Figure 1: Posterior distributions of variance estimates for two different scenarios (a and b) and their respective null distributions (c and d) generated using permutations. Example a) shows a symmetric posterior distribution far away from zero with close agreement between the posterior mean (red lines) and mode (blue line), whilst b) shows an asymmetric posterior distribution close to zero, with clear divergence between the posterior mean and mode. Examples c) and d) show null distributions of posterior means generated through permuting the datasets, and corresponding p-values, of a) and b), respectively. The values given in a) and b) correspond to mean (mode) [CRIs]. Both datasets were simulated from Gaussian distributions with among-group variances of 0.2, but with differing sample sizes; a) with 80 groups and 4 observations per group; b) with 40 groups and 2 observations per group.

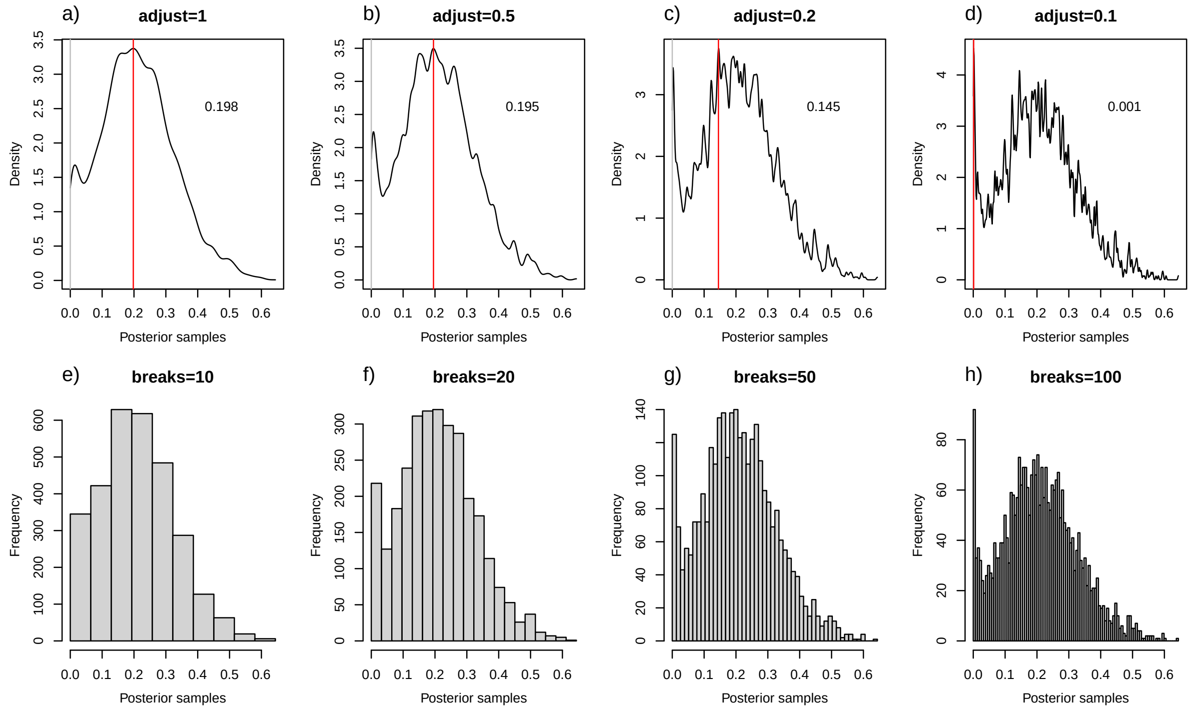


Figure 2: The effect of bandwidth choice on the estimation of the posterior mode. Top row shows kernel densities of the same posterior distribution, estimated with different bandwidth scalings, from 1 in a) to 0.1 in d). Red lines shows the posterior modes estimated from that scaling. Bottom row shows the equivalent histograms for comparison.

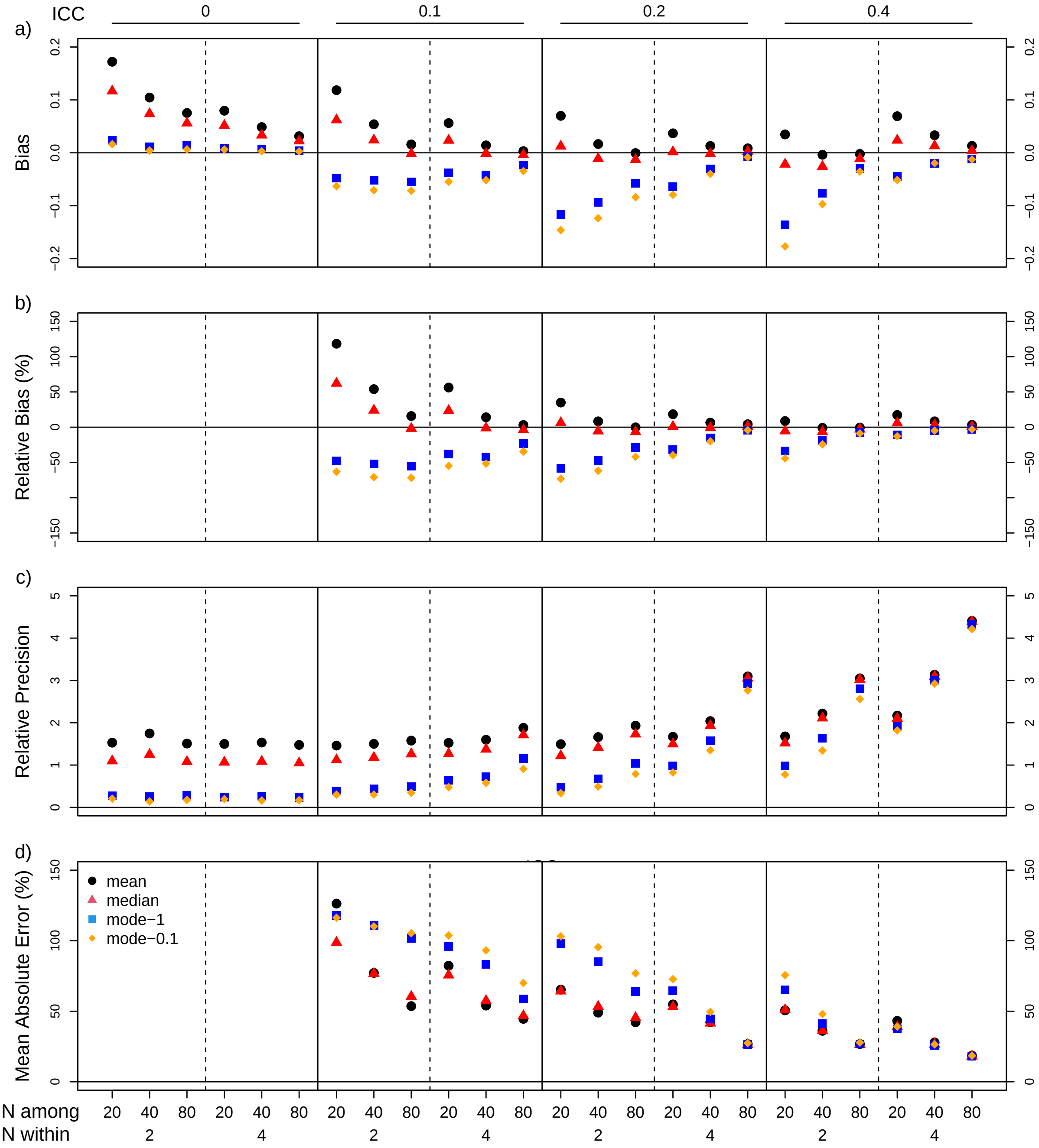


Figure 3: Bias (a), relative bias (b), relative precision (c) and mean absolute error (d) of posterior mean, median and mode of variance components from linear mixed effects models run on data simulated with a Gaussian distribution varying in among group variance (ICC - 0, 0.1, 0.2, and 0.4) and sample size within (2 or 4) and among (20, 40, 80) groups. Each point is based on the estimates from 500 datasets. Two posterior modes were estimated: mode-1 and mode-0.1 with more and less smoothing, respectively (see text for more details). Mean absolute error is also a relative measure, being standardised by the simulated value (see text for more details).

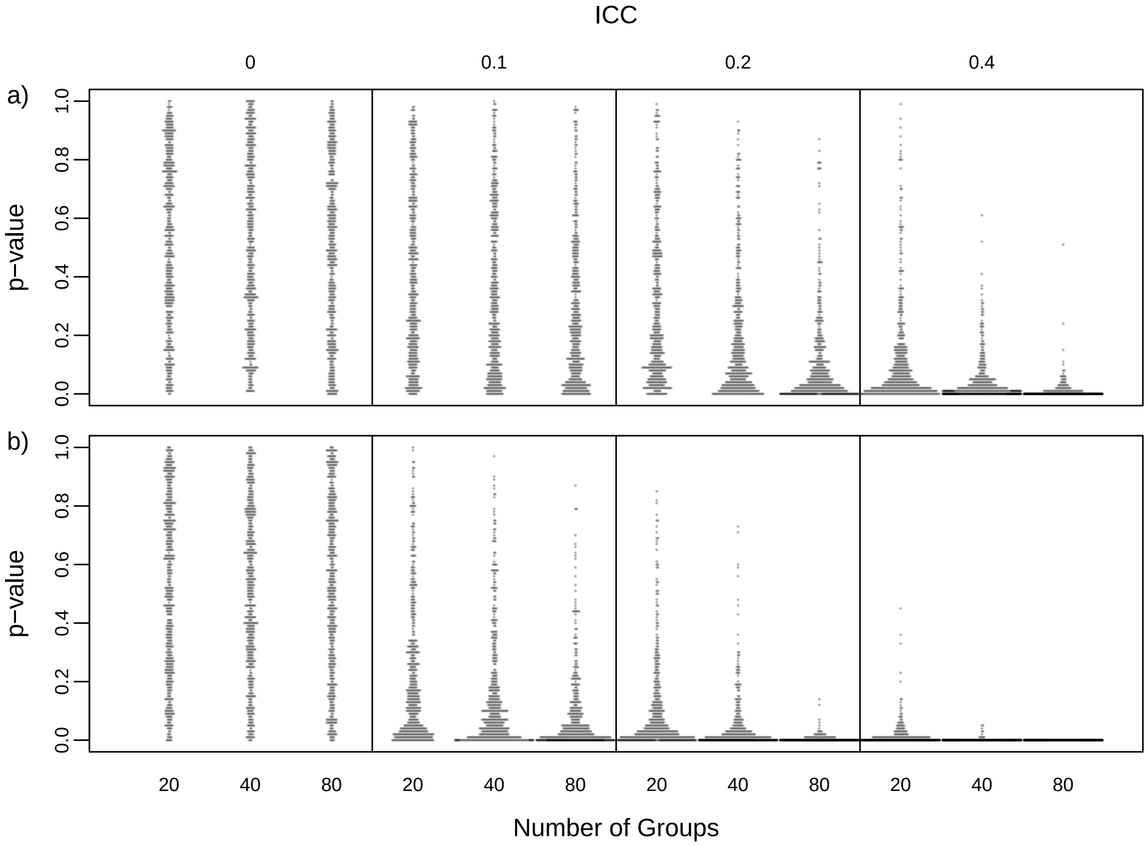


Figure 4: Distributions of p-values for the among-group variance, estimated used linear mixed effects models run on data simulated with a Gaussian distribution, varying in among-group variance (ICC - 0, 0.1, 0.2, and 0.4) and sample size among groups (20, 40, 80), with 500 datasets per combination. P-values were estimated using the posterior median and null distributions generated through simulations. a) shows a within group sample size of 2, and b) a within group sample size of 4.

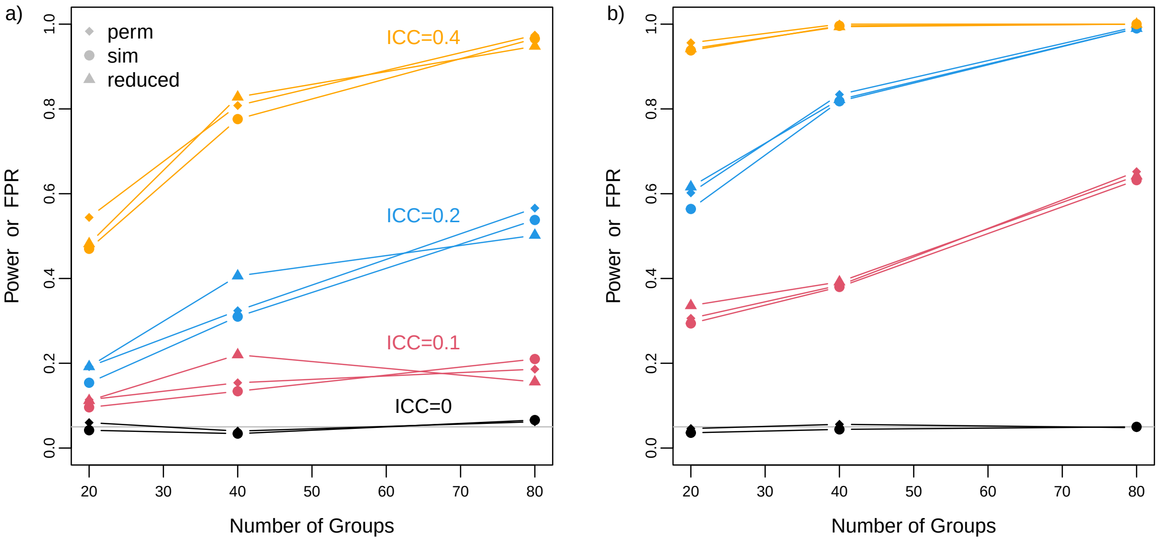


Figure 5: Comparisons of power (in colour) and false positive rate (FPR, in black) calculated using permutation (perm), simulation (sim) or a global null distribution (the 'reduced' method in the main text). For each within-group sample size of a) 2 and b) 4, we show results for four among-group variances (0 (representing FPR), 0.1 ,0.2 and 0.4) and three among-group sample sizes (20, 40 and 80), with 500 datasets per combination. All datasets were simulated with a Gaussian distribution. Power/FPR was calculated using posterior medians.

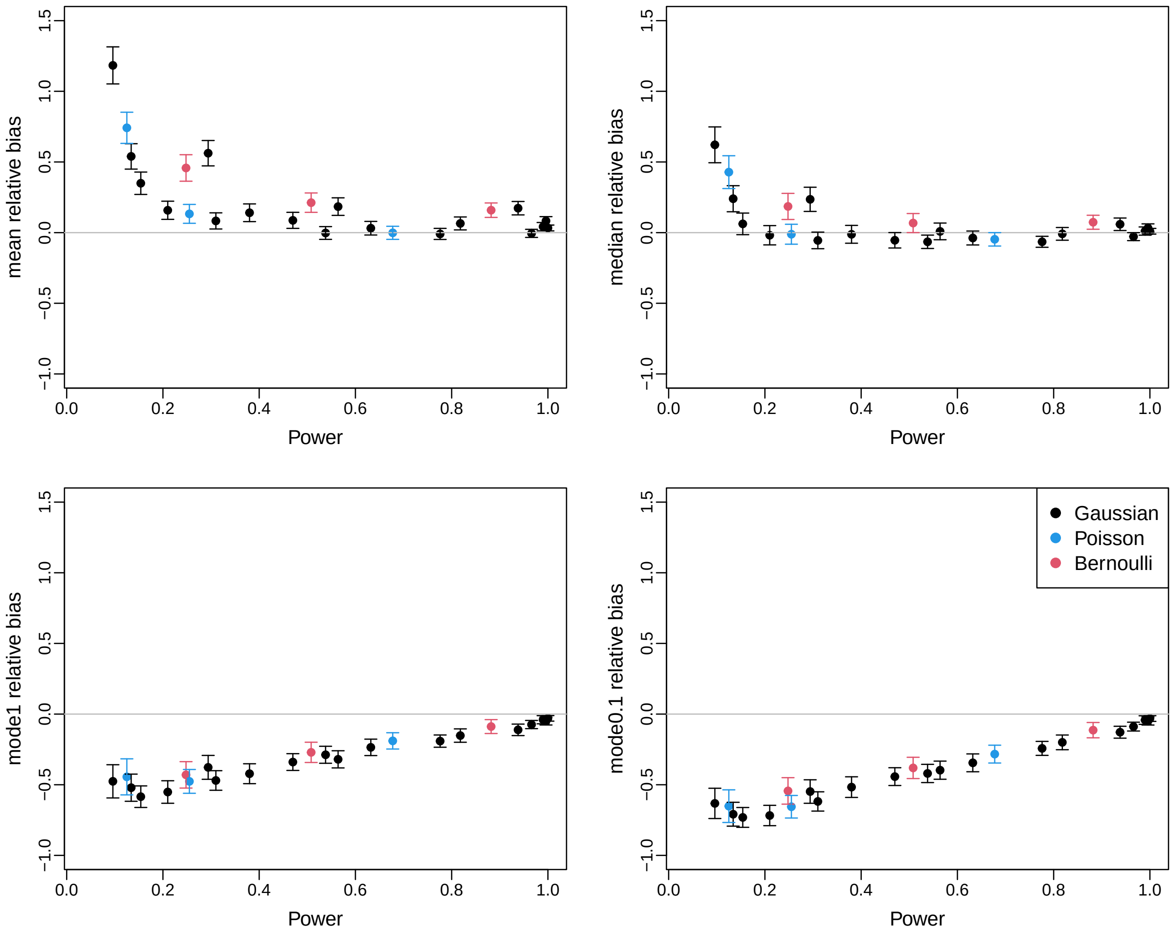


Figure 6: Relationships between power and relative bias, the latter being estimated across different measures of central tendency. Power was calculated using null distributions generated using the simulation method and the posterior median. Each point is based on 500 datasets, simulated with either a Gaussian, Bernoulli or Poisson distribution, with varying effect and sample sizes. Mean and 95% confidence intervals of the the relative bias are shown.

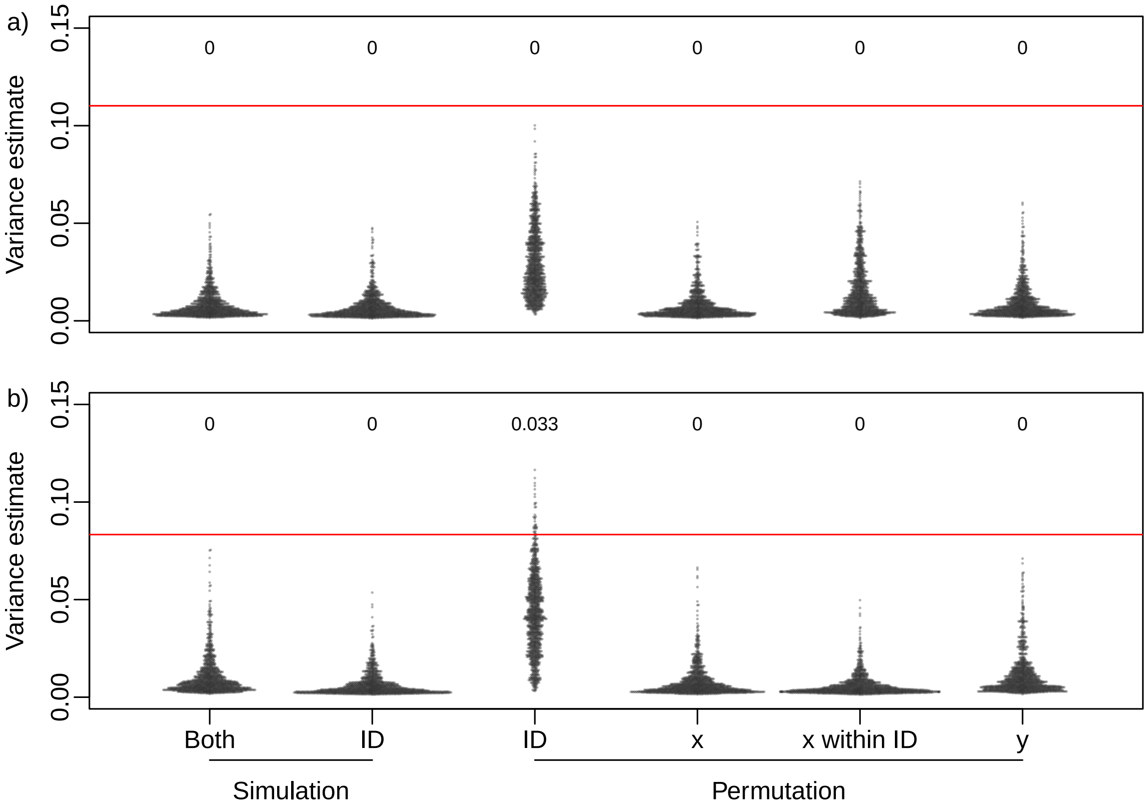


Figure 7: Null distributions of posterior medians generated with five different methods (see main text), from a) a simulated dataset, and b) a real dataset on aggressiveness in great tits. Red line represents posterior median estimated from original dataset. Values above the points represent the respective p-values.

Supplementary Materials

Supplementary Methods

Impact of prior choice on measures of central tendency

To ensure that our results, especially on the mode, were not driven by the choice of the prior, we ran additional models on a subset of the data (ICC=0.2, N groups=80, N within=2) with a range of weaker priors; half-Cauchy priors with scale 5 and 25, and uniform priors from 0 to 5 and 0 to 25 on the among-group standard deviation. The half Cauchy prior has been recommended for variance components (Gelman, 2006) and is commonly used (note it is equivalent to the commonly used parameterisation of the parameter expanded priors in MCMCglmm (V=1, nu=1, alpha.mu=0)). The different parametrizations of the half Cauchy and uniform priors resulted in no difference in the results (Figure S1). More recently the use of stronger priors has been suggested, for example a half normal prior with scale 1. The use of this prior also did not affect our results. For demonstration purposes, we also ran models in MCMCglmm specifying uninformative improper priors on the variance. Given the simplicity of these models, the posterior mode is expected to correspond to the REML estimate. For comparison, we also ran a wide uniform prior (U(0,25)) on the variance in Stan. As expected, using these uninformative priors on the variance led to a concordance between REML and posterior mode, although the strength of this similarity differed between the methods used to estimate the mode (Figure S1).

Methods of posterior mode estimation

Commonly used functions for estimating the mode of the marginal posterior distribution in R include the `posterior.mode` function in the MCMCglmm package (Hadfield, 2010), the `Mode` function in the ggdist package (Kay, 2022), and the `map_estimate` function of the bayestestR package (Makowski *et al.*, 2019b). Typically these functions estimate the mode by estimating the parameter value at which the kernel density is maximised. Kernel density estimation involves fitting a model to the distribution of posterior samples to estimate a density function. The maximum of this function (the estimated mode) is then calculated over a series of predicted values. One key parameter in kernel density estimation is the bandwidth, which describes the amount of smoothing and is analogous to the number of breakpoints in a histogram (Figure 2). Common methods generally generate the bandwidth using specific algorithms, which are then scaled. MCMCglmm scales the bandwidth generated by Silverman's 'rule of thumb' algorithm (nrd0; eqn 3.31 in Silverman, 1986) by 0.1 (i.e. it is much less smoothed; Figure 2d). In contrast, ggdist and bayestestR use the default values of the nrd0 and SJ algorithms (Sheather & Jones, 1991), respectively (the default bandwidth of the nrd0 algorithm is also used by `density` function in R; Figure 2a). The impact on the potential inferences caused by the choice of scaling is demonstrated in Figure 2, with the degree of smoothing affecting where the posterior mode is estimated. To explore this impact of bandwidth, we estimated the posterior mode using these two bandwidth scalings (0.1 and 1). The kernel density was estimated using the SJ algorithm (Sheather & Jones, 1991), and the mode was estimated using 512 predicted values with a cut-off point at zero. These additional parameters differ between commonly used functions, but have much less impact than the bandwidth, and so we held them constant.

Simulations based on Fay *et al.* (2022)

We simulated datasets based on Fay *et al.* (2022), but ran simplified models (univariate instead of bivariate), as the purpose was simply to demonstrate the effect of different measures of central tendency on the bias in these models. We simulated data with the same parameters of one set of simulation in Fay *et al.* (2022) - fast life history and low heterogeneity. We simulated the probability of survival as 0.5 and probability of reproduction as 0.7, standard deviations on the latent scale of 0.2 for both survival and reproduction and a correlation of 0.6 between the two. We simulated 100 datasets from sample sizes of 250, 500, 1000, 2000, 4000 individuals. For each simulated dataset we ran a binomial GLMM, with random effects of individual identity using Stan with the rstan package (version 2.21.3 Stan Development Team, 2022a). We specified weakly informative priors on the among-group standard deviations (half-Cauchy distribution with scale 2), and ran one chain for each model with 7500 iterations and a warm-up period of 2000 iterations. We then estimated the posterior mean, median and 2 modes as in the main text.

Supplementary Figures

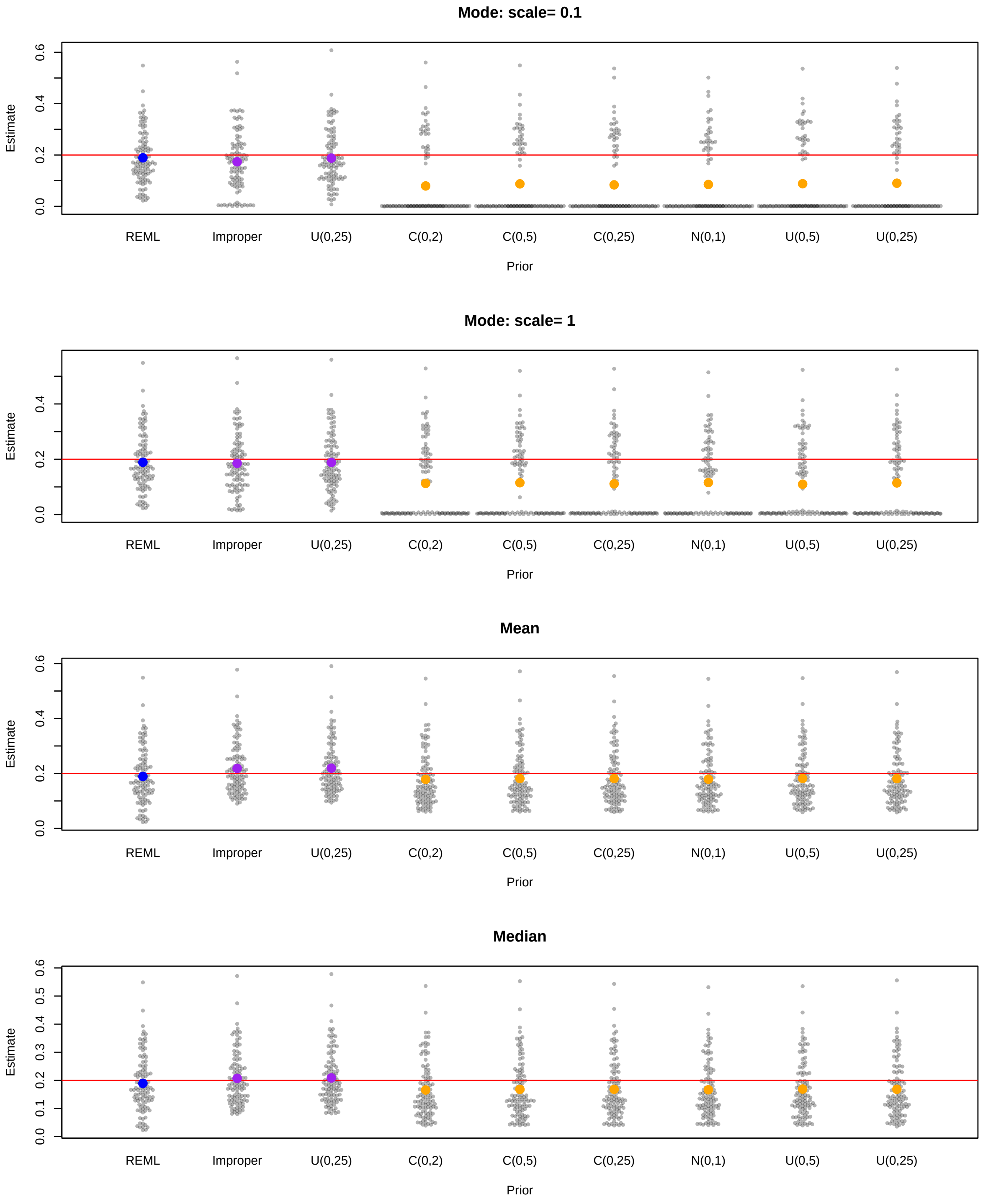


Figure S1: Impact of prior choice on measures of central tendency. 'C' represents half Cauchy priors, 'N' normal priors, 'U' uniform priors, and 'Improper' uninformative improper prior. Red lines shows simulated values. Blue points show the mean of the REML estimates across simulations, purple points show means of different point estimates from across the 100 simulations with priors on the variance, and orange points show means of different point estimates from across the 100 simulations with priors on the SD. Data was simulated from a Gaussian distribution, with a among-group variance of 1, with 80 groups and 2 observation within a group.

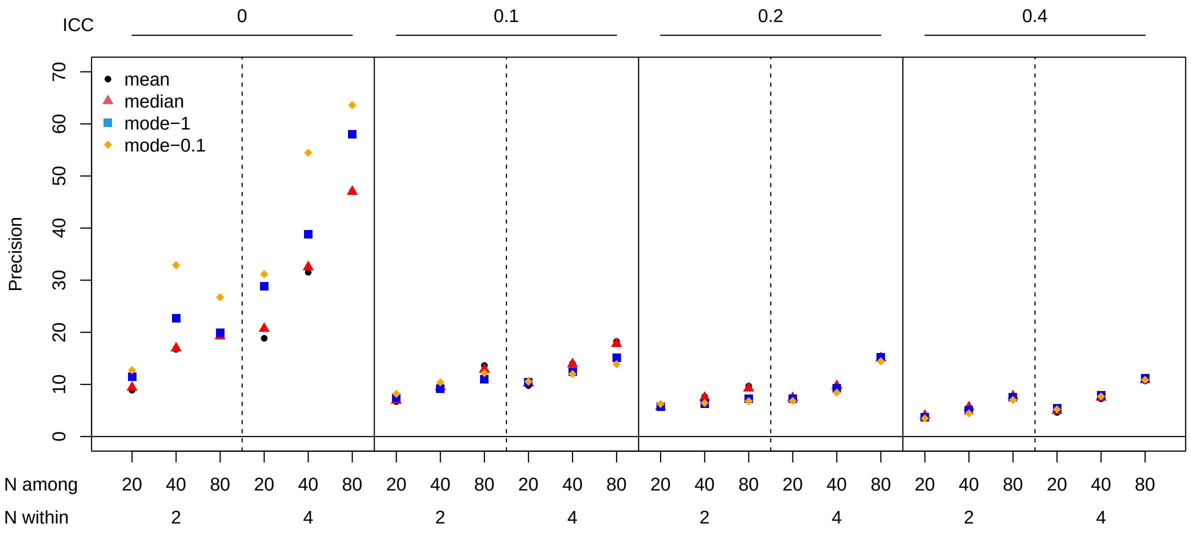


Figure S2: Precision increases with sample size, but decreases with effect size. The different panels show the precision of posterior mean, median and mode of variance components estimated using linear mixed models, from data simulated with a Gaussian distribution, varying in among-group variance (ICC - 0, 0.1, 0.2, and 0.4) and sample size within (2 or 4) and among (20, 40, 80) groups, with 500 datasets per ICC and sample size combination. Two posterior modes were estimated; mode-1 and mode-0.1 with more and less smoothing, respectively (see text for more details).

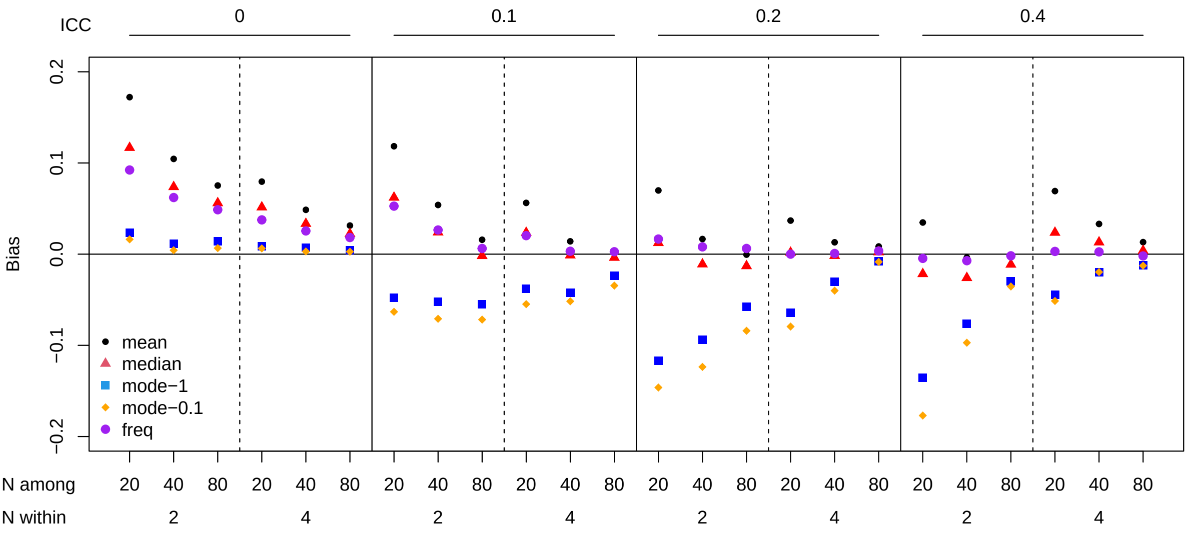


Figure S3: Bias of frequentist estimates of the among group variance alongside bias in the posterior mean, median and mode, estimated used linear mixed effects models run on data simulated with a Gaussian distribution, varying in among-group variance (ICC - 0, 0.1, 0.2, and 0.4) and sample size within (2 or 4) and among (20, 40, 80) groups. Two posterior modes were estimated; mode-1 and mode-0.1 with more and less smoothing, respectively (see text for more details).

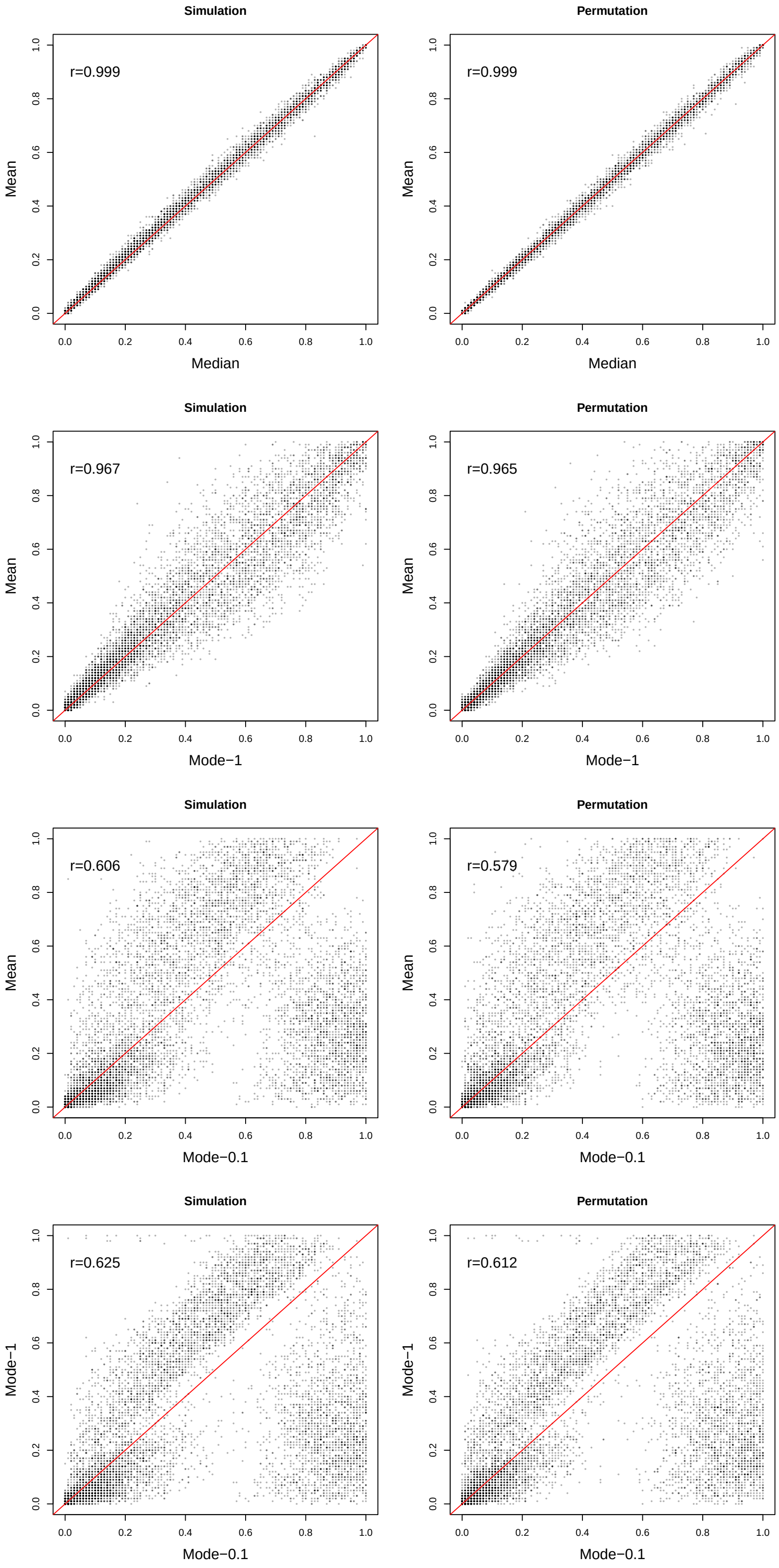


Figure S4: Comparison of p-values generated with different measures of central tendency estimated using linear mixed models, using null distributions generated from both simulation and permutation methods. Data were simulated with a Gaussian distribution.

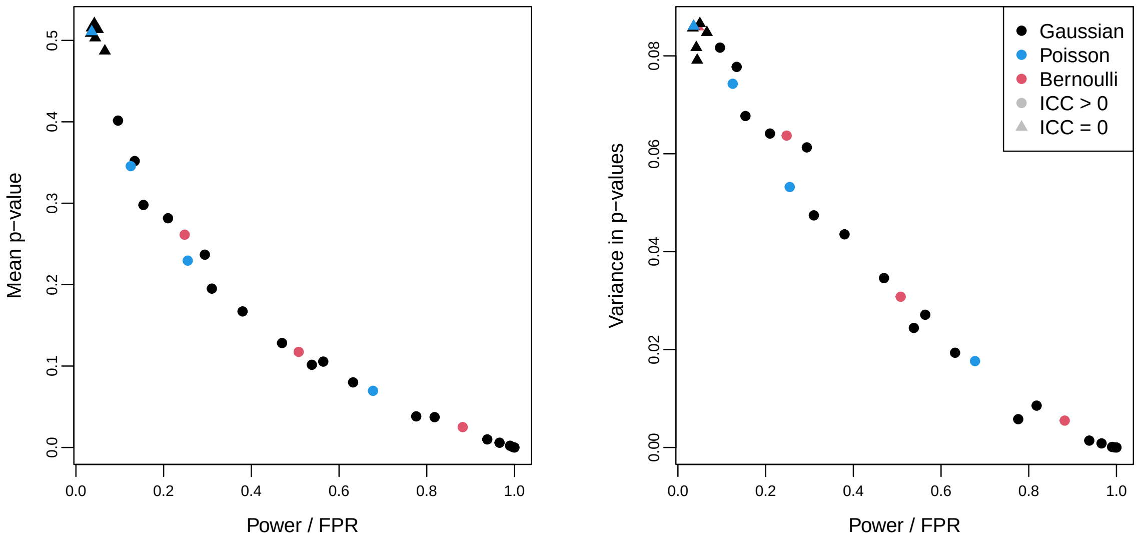
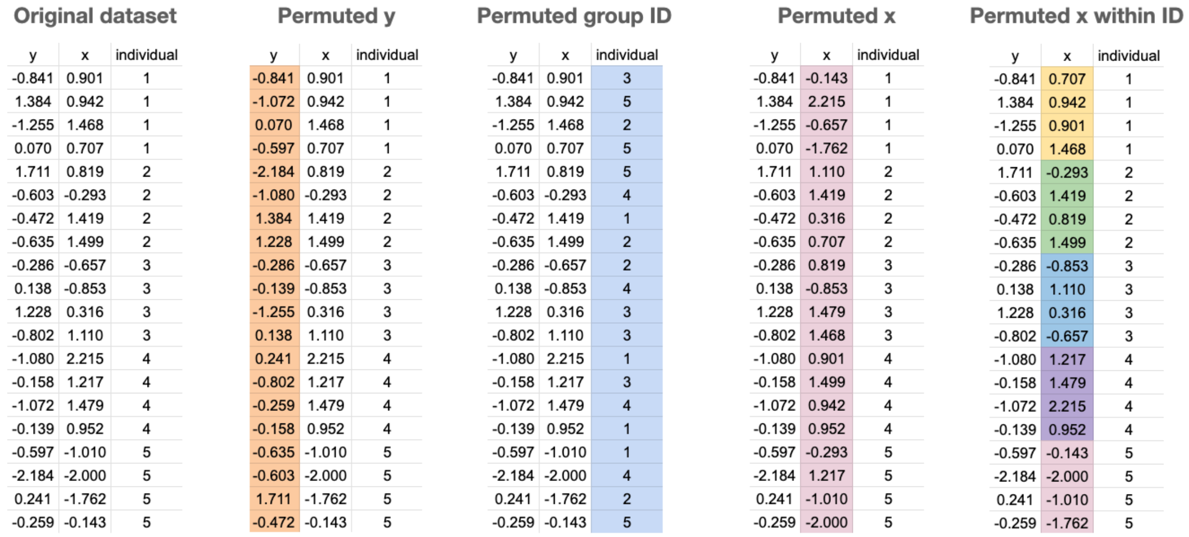


Figure S5: Relationships between power and false positive rate (FPR) and a) mean and b) variance in p-values. Power/FPR was calculated using null distributions generated using the simulation method and the posterior median. Each point is based on 500 datasets, simulated with either a Gaussian, Bernoulli or Poisson distribution, with varying effect and sample sizes.



Original dataset

y	x	individual
-0.841	0.901	1
1.384	0.942	1
-1.255	1.468	1
0.070	0.707	1
1.711	0.819	2
-0.603	-0.293	2
-0.472	1.419	2
-0.635	1.499	2
-0.286	-0.657	3
0.138	-0.853	3
1.228	0.316	3
-0.802	1.110	3
-1.080	2.215	4
-0.158	1.217	4
-1.072	1.479	4
-0.139	0.952	4
-0.597	-1.010	5
-2.184	-2.000	5
0.241	-1.762	5
-0.259	-0.143	5

Permuted y

y	x	individual
-0.841	0.901	1
-1.072	0.942	1
0.070	1.468	1
-0.597	0.707	1
-2.184	0.819	2
-1.080	-0.293	2
1.384	1.419	2
1.228	1.499	2
-0.286	-0.657	3
-0.139	-0.853	3
-1.255	0.316	3
0.138	1.110	3
0.241	2.215	4
-0.802	1.217	4
-0.259	1.479	4
-0.158	0.952	4
-0.635	-1.010	5
-0.603	-2.000	5
1.711	-1.762	5
-0.472	-0.143	5

Permuted group ID

y	x	individual
-0.841	0.901	3
1.384	0.942	5
-1.255	1.468	2
0.070	0.707	5
1.711	0.819	5
-0.603	-0.293	4
-0.472	1.419	1
-0.635	1.499	2
-0.286	-0.657	2
0.138	-0.853	4
1.228	0.316	3
-0.802	1.110	3
-1.080	2.215	1
-0.158	1.217	3
-1.072	1.479	4
-0.139	0.952	1
-0.597	-1.010	1
-2.184	-2.000	4
0.241	-1.762	2
-0.259	-0.143	5

Permuted x

y	x	individual
-0.841	-0.143	1
1.384	2.215	1
-1.255	-0.657	1
0.070	-1.762	1
1.711	1.110	2
-0.603	1.419	2
-0.472	0.316	2
-0.635	0.707	2
-0.286	0.819	3
0.138	-0.853	3
1.228	1.479	3
-0.802	1.468	3
-1.080	0.901	4
-0.158	1.499	4
-1.072	0.942	4
-0.139	0.952	4
-0.597	-0.293	5
-2.184	1.217	5
0.241	-1.010	5
-0.259	-2.000	5

Permuted x within ID

y	x	individual
-0.841	0.707	1
1.384	0.942	1
-1.255	0.901	1
0.070	1.468	1
1.711	-0.293	2
-0.603	1.419	2
-0.472	0.819	2
-0.635	1.499	2
-0.286	-0.853	3
0.138	1.110	3
1.228	0.316	3
-0.802	-0.657	3
-1.080	1.217	4
-0.158	1.479	4
-1.072	2.215	4
-0.139	0.952	4
-0.597	-0.143	5
-2.184	-2.000	5
0.241	-1.010	5
-0.259	-1.762	5

Figure S6: Illustration of the different permutation designs that can be used for a random regression analysis. The colours highlight what variables are permuted within each permutation.

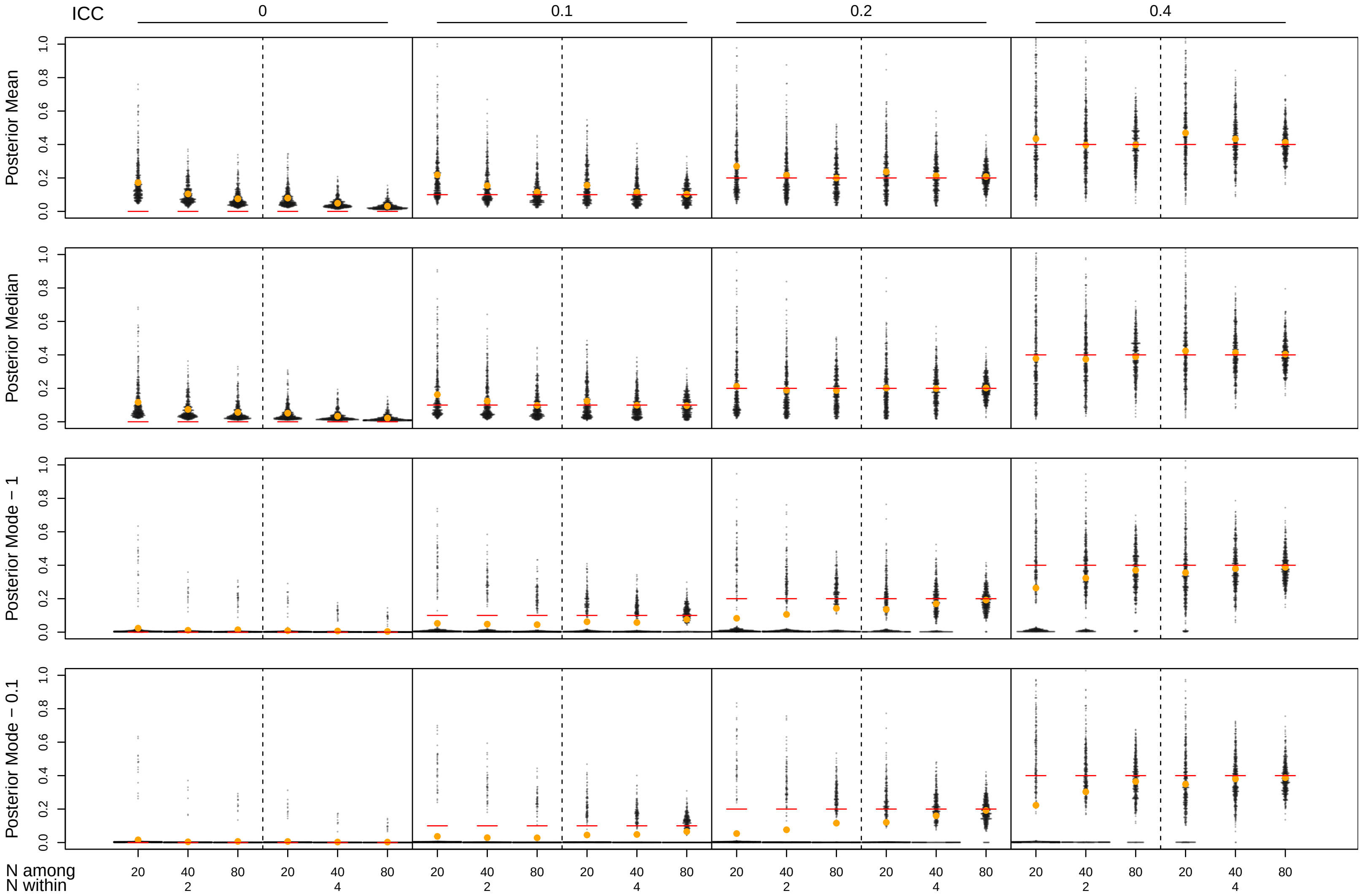


Figure S7: Sampling distributions of posterior mean, median and mode estimated using linear mixed models, from data simulated with a Gaussian distribution, varying in among-group variance (ICC - 0, 0.1, 0.2, and 0.4) and sample size within (2 or 4) and among (20, 40, 80) groups, with 500 datasets per ICC and sample size combination. Red lines show the simulated value and orange points the mean of the sampling distributions.

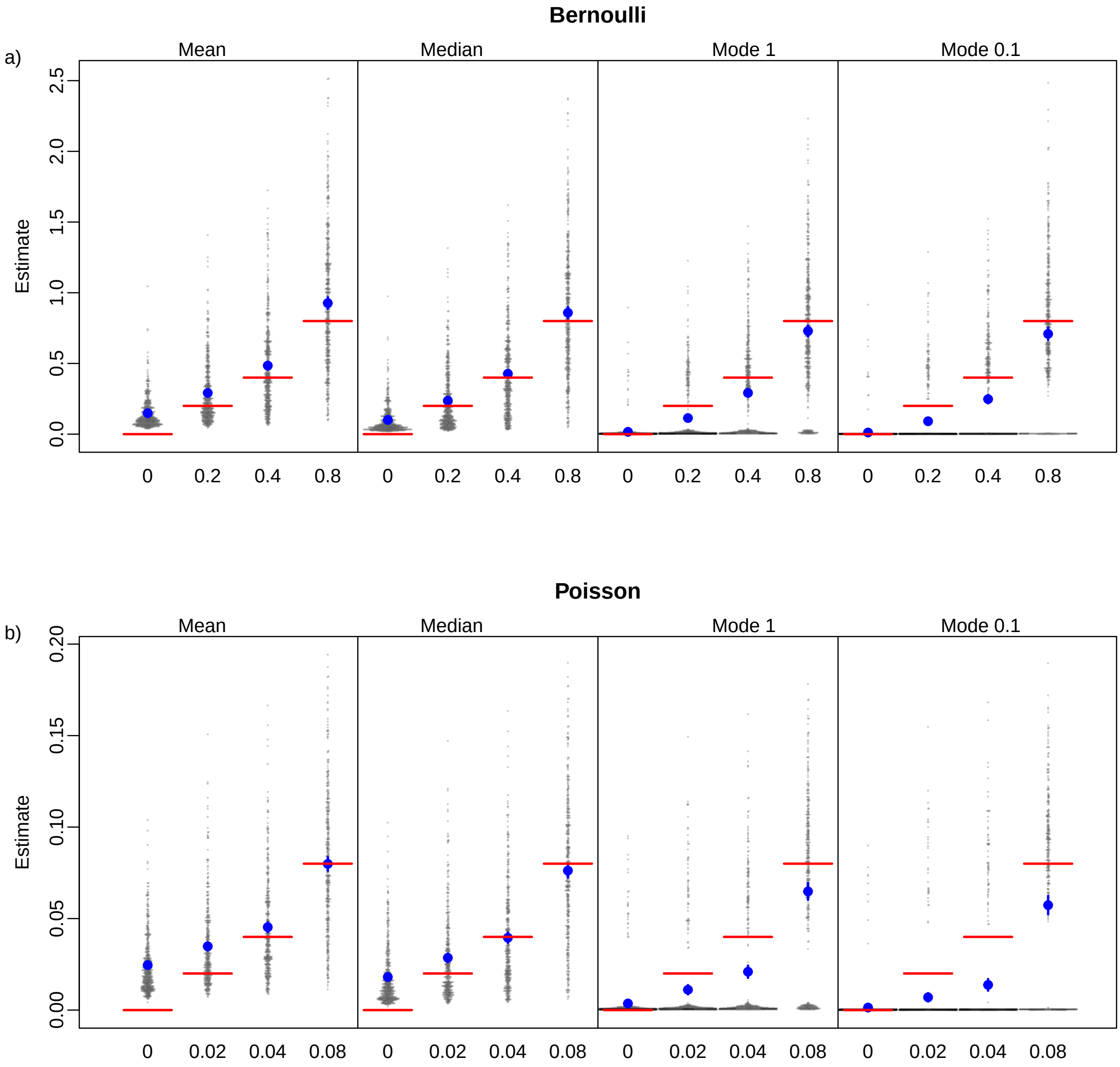


Figure S8: Sampling distributions of posterior mean, median and mode estimated using GLMMs, from data simulated with a) Bernoulli and b) Poisson distributions, varying in among-group variance, with 500 datasets per variance. Red lines show the simulated value and blue points and error bars show mean and 95% confidence intervals of the sampling distributions.

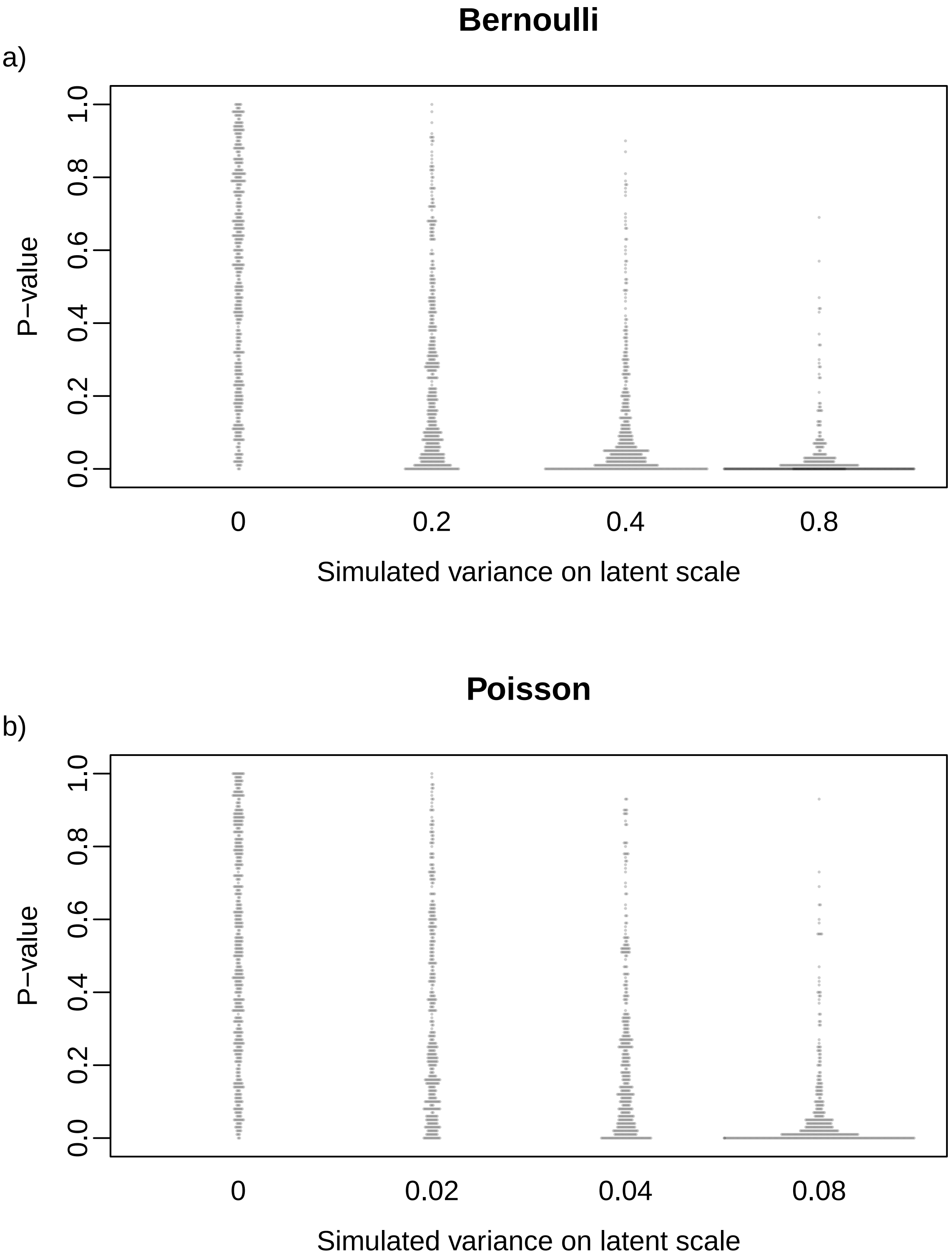


Figure S9: Distributions of p-values for the among group variance estimated used GLMMs run on data simulated with a) Bernoulli and b) Poisson distributions, varying in among-group variance, with 500 datasets per combination. P-values were estimated using the posterior median and null distributions generated through simulations.

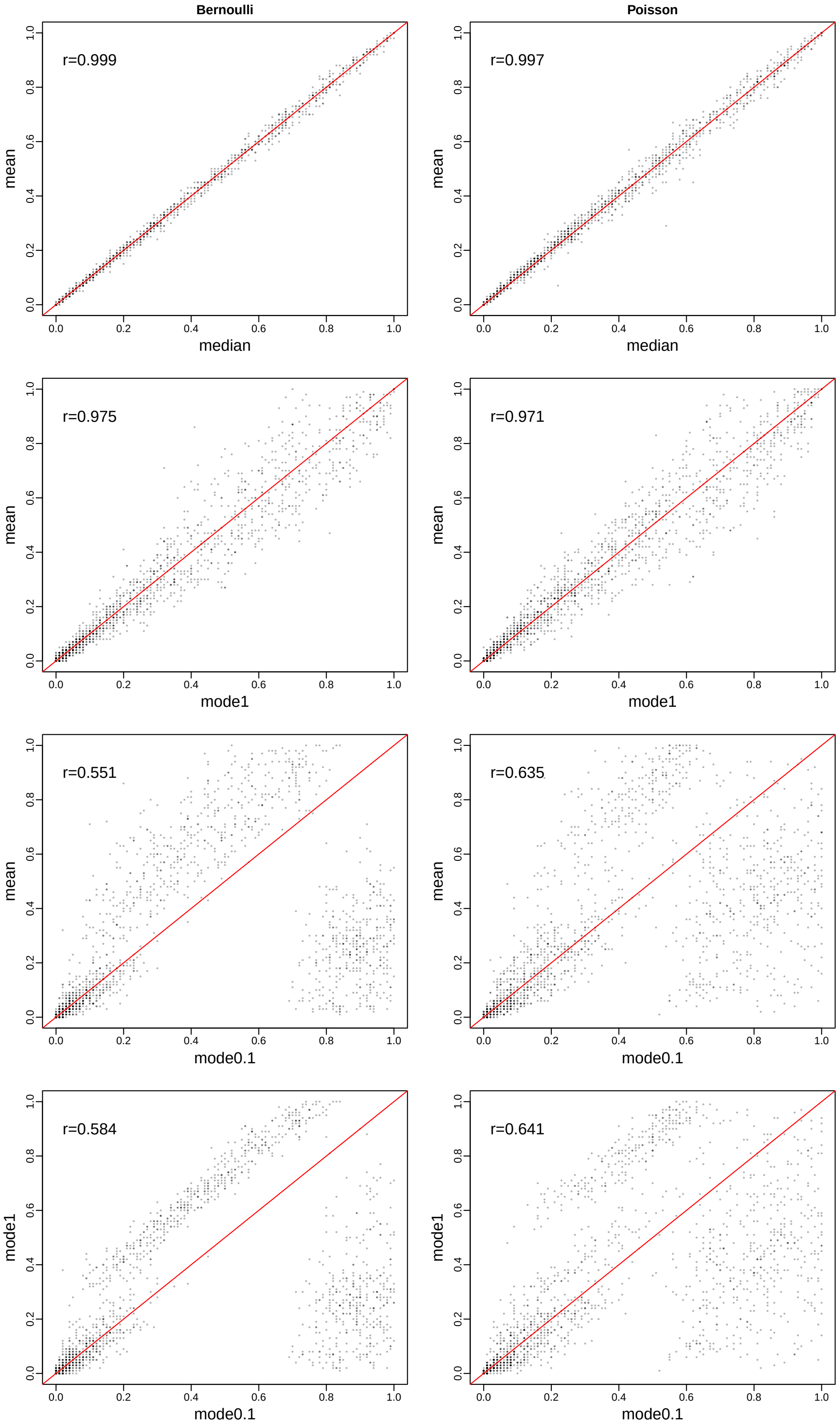


Figure S10: Comparisons of p-values generated with different measures of central tendency estimated using GLMMs, using null distributions generated by simulation. The left column shows comparison from data generated and analysed with a Bernoulli distribution and the right column with a Poisson distribution.

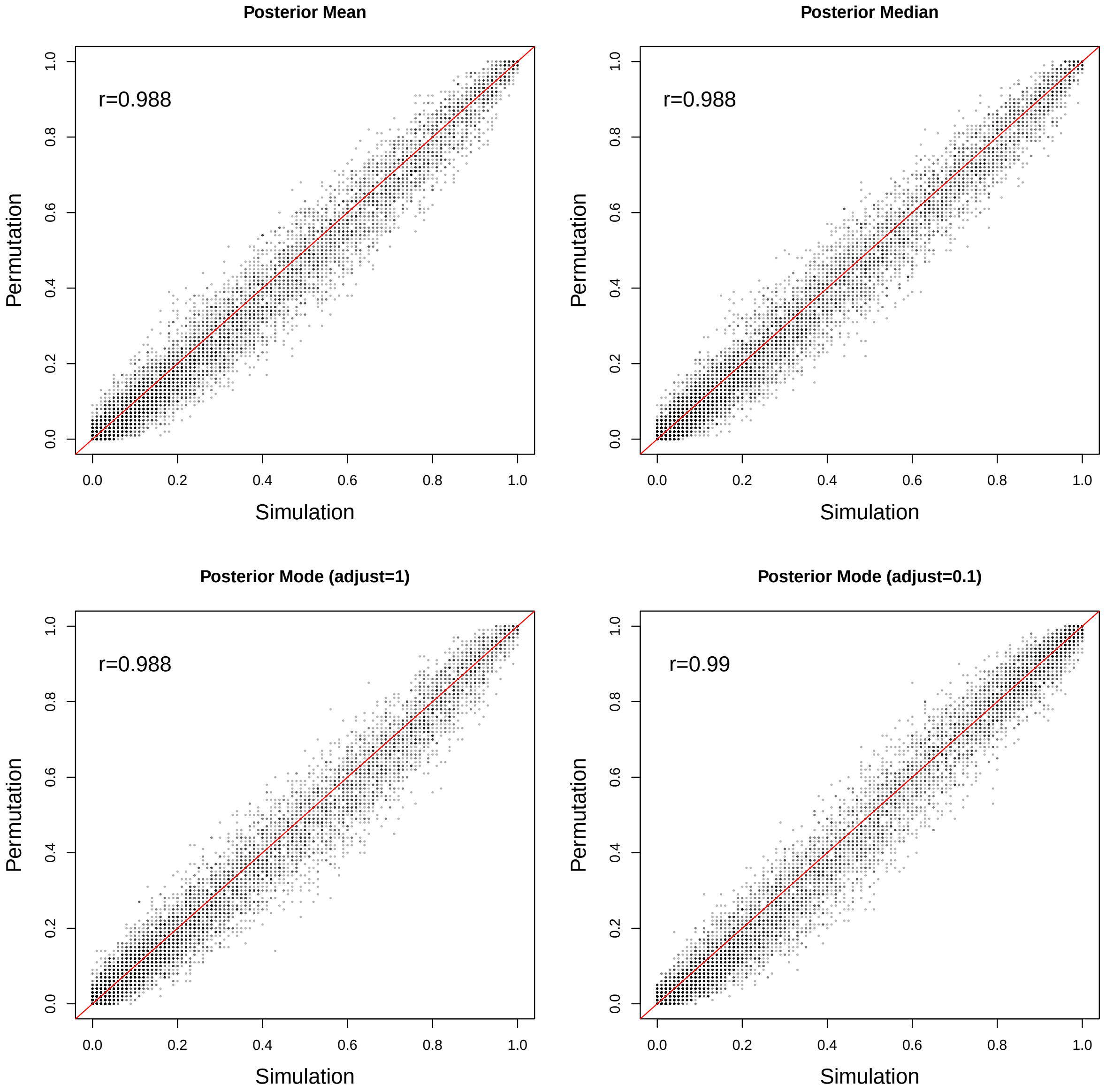


Figure S11: Comparisons of p-values generated from null distribution using permutation and simulation methods across all measures of central tendency estimated using linear mixed models. Data were simulated with a Gaussian distribution.

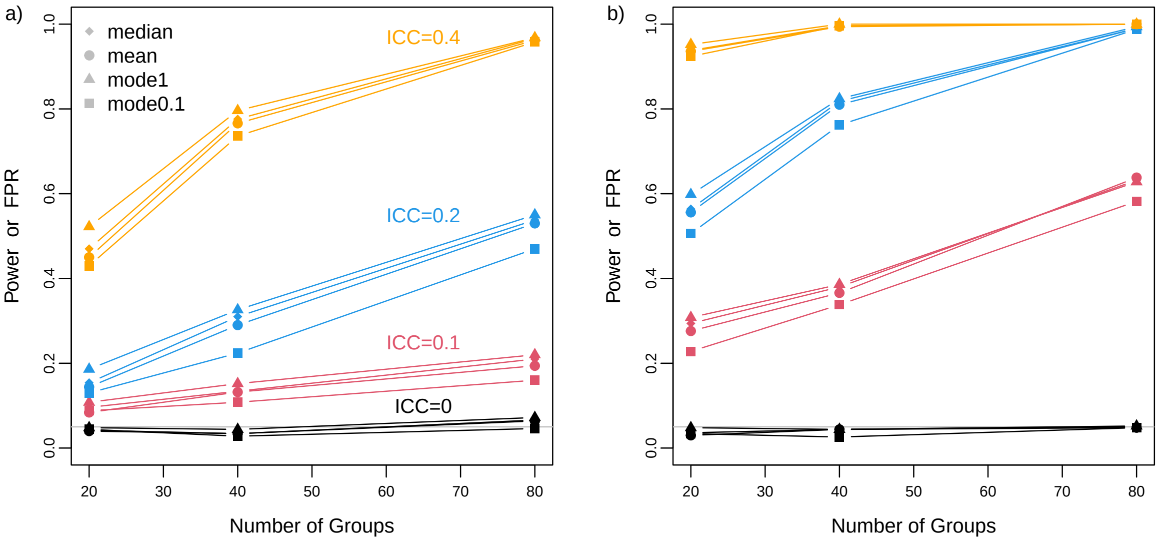


Figure S12: Comparisons of power (in colour) and false positive rate (FPR, in black) generated using different measures of central tendency. For each within-group sample size of a) 2 and b) 4, we show results for four among-group variances (0 (representing FPR), 0.1 ,0.2 and 0.4) and three among-group sample sizes (20, 40 and 80), with 500 datasets per combination. All datasets were simulated with a Gaussian distribution. Power/FPR was calculated using null distributions generated using the simulation method.

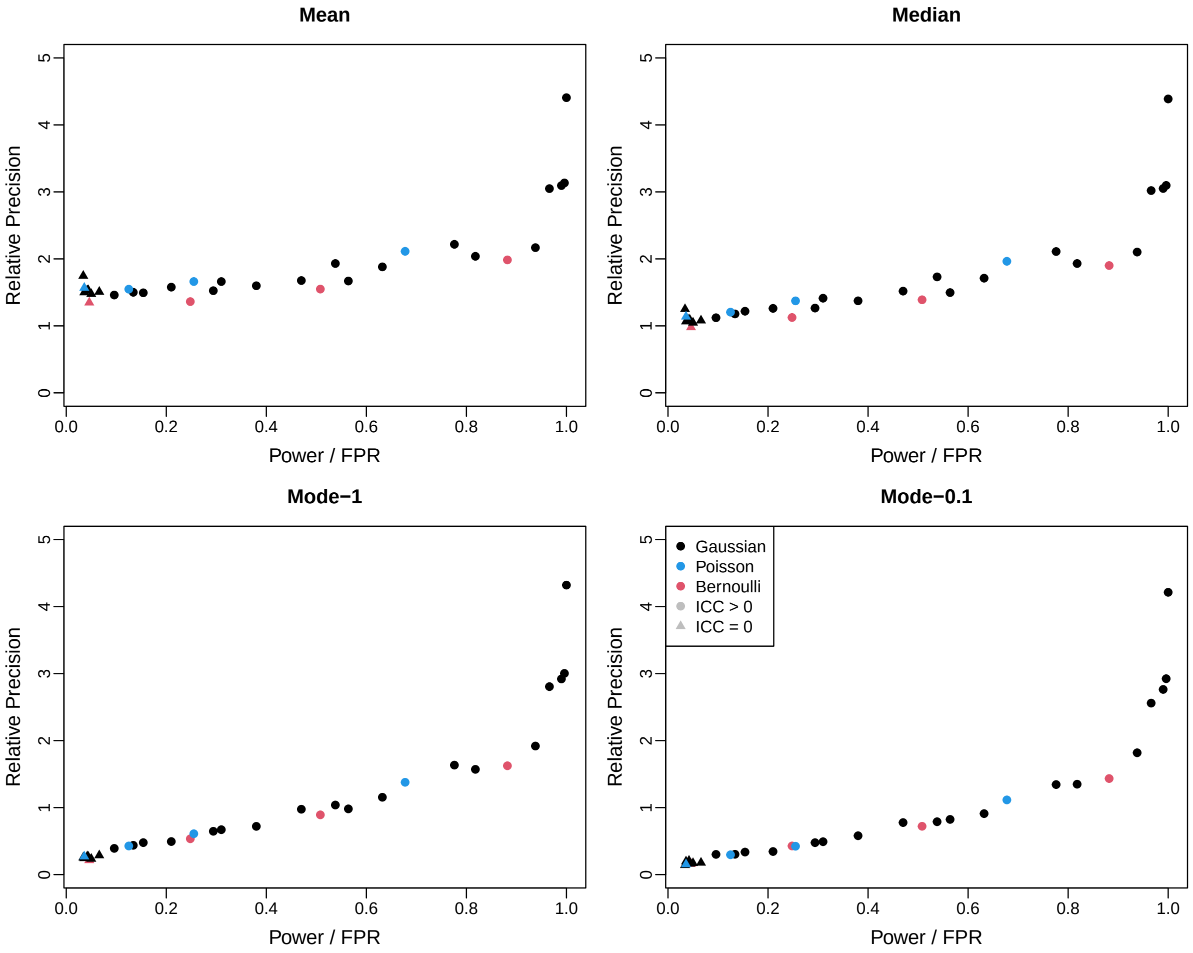


Figure S13: Relationships between Power/false positive rate (FPR) and relative precision, the latter being estimated across different measures of central tendency. Power/FPR was calculated using null distributions generated using the simulation method and the posterior median. Each point is based on 500 datasets, simulated with either a Gaussian, Bernoulli or Poisson distribution, with varying effect and sample sizes.

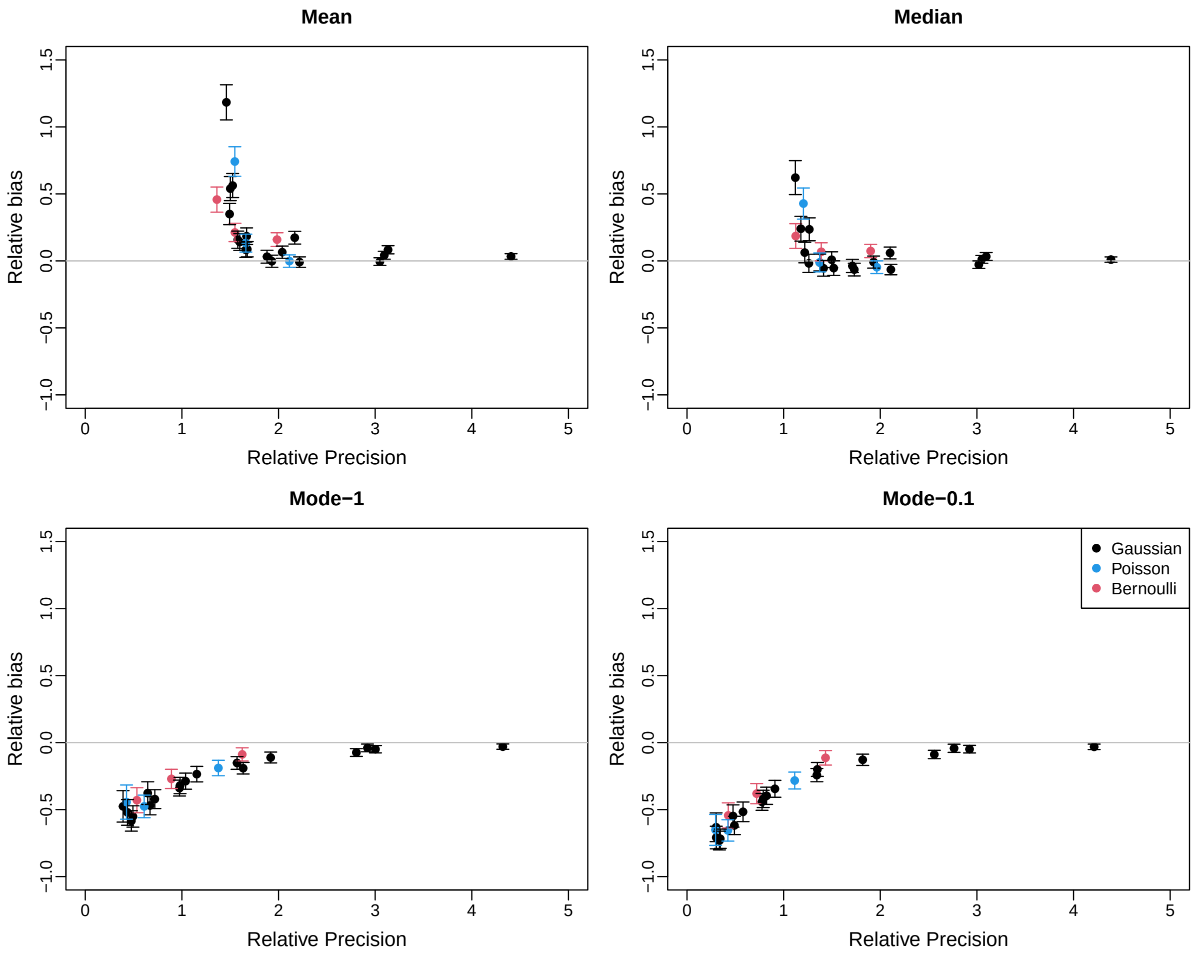


Figure S14: Relationships between relative bias and relative precision, estimated across different measures of central tendency. Each point is based on 500 datasets, simulated with either a Gaussian, Bernoulli or Poisson distribution, with varying effect and sample sizes. Mean and 95% confidence intervals of the the relative bias are shown.

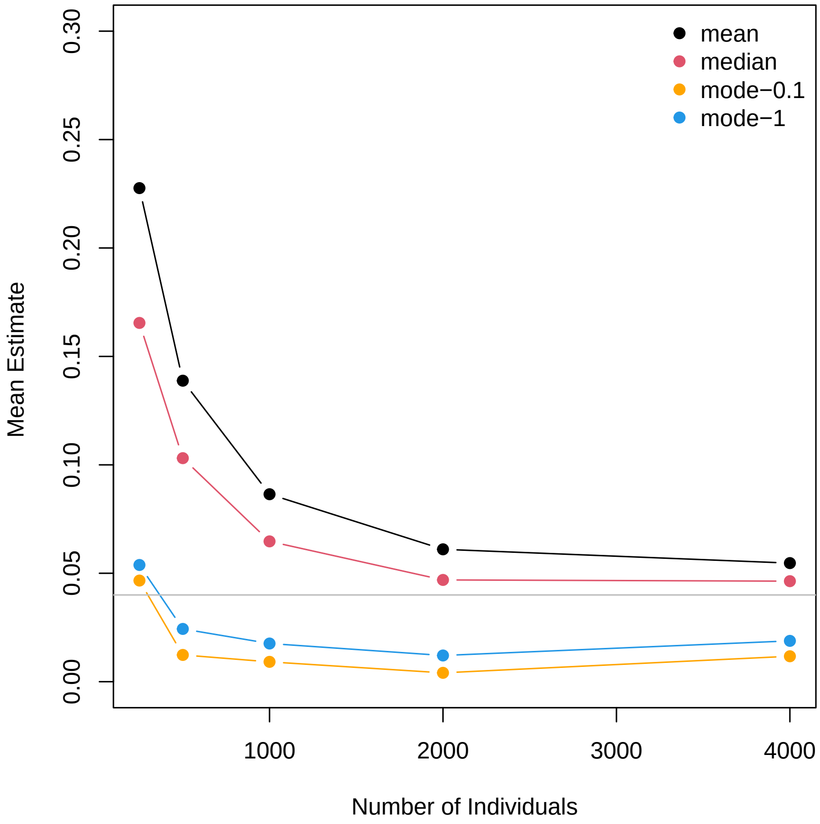


Figure S15: Mean posterior mean, median and mode of variance components from GLMMs, analysing simulated survival data with increasing number of individuals. Simulations were based upon Fay et al. *(2022) - see Supplementary Methods for more details of parameterisation.*

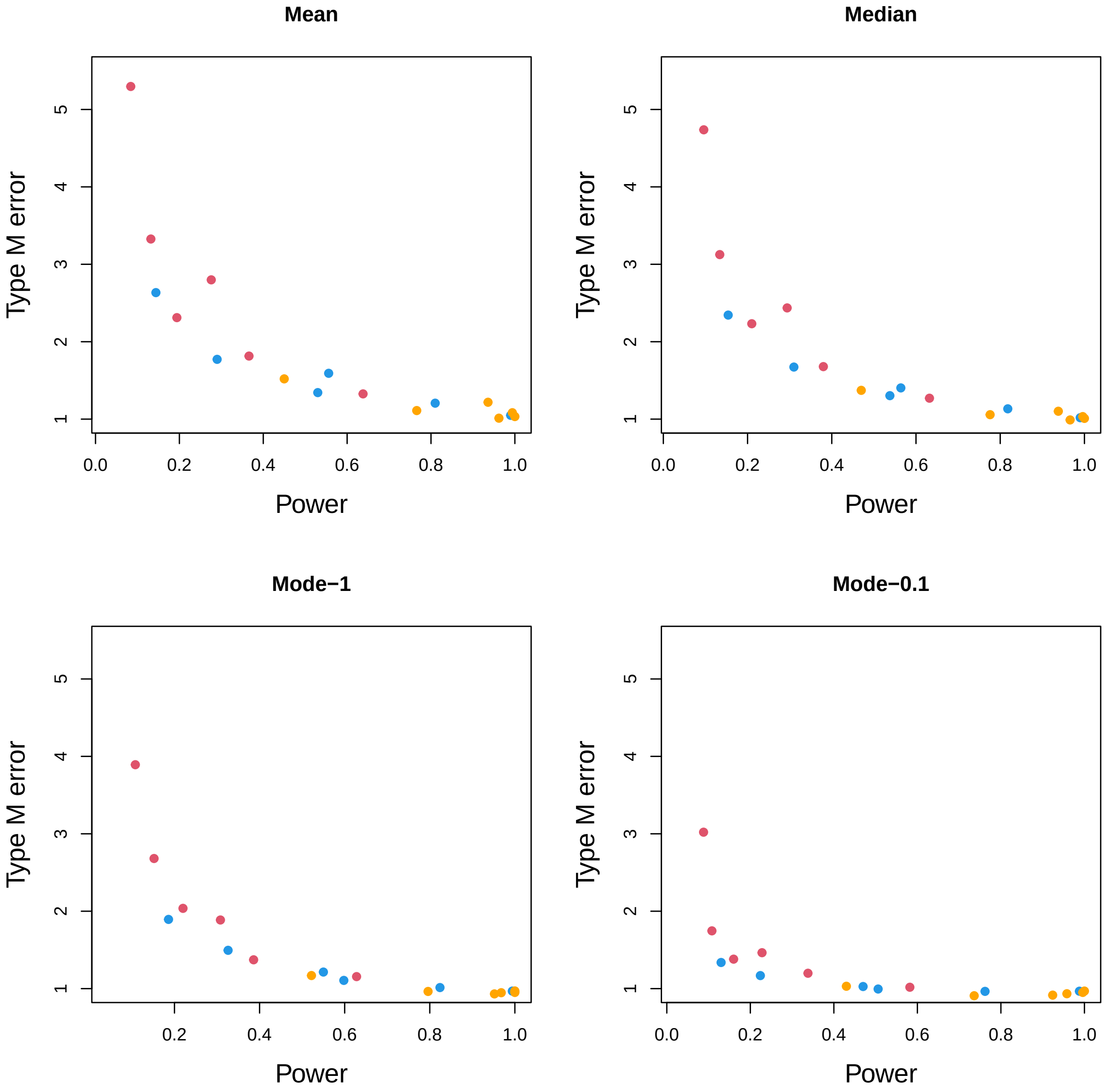


Figure S16: Type M error and power from posterior mean, median and mode calculated using null distribution generated through simulation. Colours represent simulated ICCs, red - 0.1, blue - 0.2, and orange - 0.4.

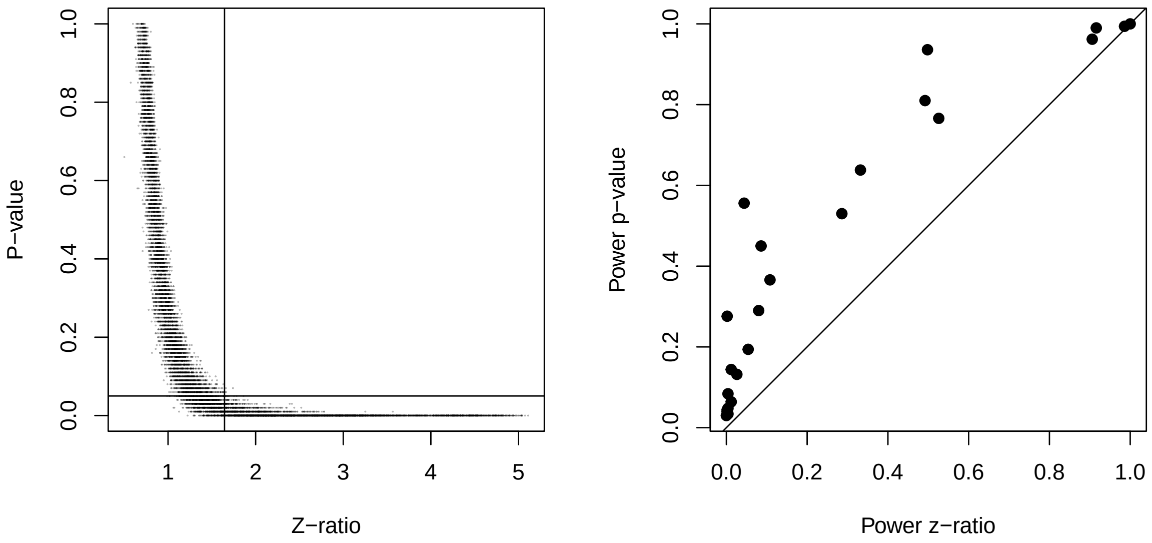


Figure S17: a) Relationship between z-ratio (posterior mean/posterior SD) and p-value. Grey lines represent $p = 0.05$ and $z = 1.64$, the later being equivalent to the z-ratio that would give $p = 0.05$ on a one-sided test. b) Relationship between power derived from z-ratio and and power derived from p-values. Power was calculated for the z-ratios as the proportion of datasets where $z > 1.64$. Each point is based on 500 datasets. All datasets were simulated with a Gaussian distribution, with varying effect and sample sizes.