An Introduction to Bayesian Data Analysis for Sport Scientists

Jorge R Fernandez-Santos[1,2], Jesus G Ponce-Gonzalez[2,3], Jose Castro-Piñero[1,2], & Jose Luis Gonzalez-Montesinos[4]

[1] GALENO reasearch group, Faculty of Education Sciences, University of Cádiz, Cádiz, Spain.

[2] Biomedical Research and Innovation Institute of Cádiz (INiBICA) Research Unit, Cádiz, Spain.

[3] MOVE-IT reasearch group, Faculty of Education Sciences, University of Cádiz, Cádiz, Spain

[4] Deparment of Physical Education, Faculty of Education Sciences, University of Cádiz, Cádiz, Spain.

Author Note

Correspondence concerning this article should be addressed to Jorge R Fernandez-Santos, Deparment of Physical Education, Faculty of Education Sciences, University of Cádiz, 11519, Puerto Real, Spain.. E-mail: jorgedelrosario.fernandez@uca.es

Abstract

There is a concern in scientific research about the misuse and misinterpretation of traditional methods of statistical inference based on confidence intervals and p-values. As an alternative, Bayesian data analysis (BDA) is a method that uses probability to quantify uncertainty in inferences based on statistical data analysis. However, current sports scientists are not trained in BDA despite the fact that easy-to-use software like the R package brms makes BDA an accessible tool. Therefore, this manuscript introduces different key concepts like the Bayes‘ rule, hierarchical modeling, Markov Chain Monte Carlo techniques, Bayesian workflow, and sensitivity analysis. In addition, an example of BDA using brms is also performed to help sports scientists understand how to apply the previous concepts from a practical point of view and how to interpret and report the obtained results.

Keywords: Bayesian data analysis, statistical modeling, Sport Science.

An Introduction to Bayesian Data Analysis for Sport Scientists

1. Introduction

BDA is already a well-established method of statistical inference in many different disciplines like psychology, ecology, economy or health science (Greenberg, 2012; Kéry, 2010; Lee & Wagenmakers, 2014; Lesaffre & Lawson, 2012). Briefly, BDA make use of the probability for quantify uncertainty in inferences based on statistical data analysis (Gelman et al., 2013). This approach has some advantages over the traditional methods (also known as frequentist statistics) like: 1) the incorporation of prior knowledge to the statistical model via prior distribution; 2) the result obtained is only based on the specific data under consideration; 3) regardless of model complexity the final estimation is always a posterior probability distribution not depend on the stopping or testing intentions of the analyst and 4) the straightforward interpretation of results (Dienes & Mclatchie, 2018; J. K. Kruschke & Liddell, 2018; Wagenmakers et al., 2018).

Although the use of Bayesian statistics in sport analytics has increased substantially in the last years (Santos-Fernandez et al., 2019), it has been argued recently that the current statistical practices in sport science are based on the null hypothesis significant testing under the frequentist approach and that this approach is flawless so sport scientists should shift towards alternative statistical methods (Bernards et al., 2017). In fact, p-values and 95% confidence intervals commonly reported in the scientific literature are misinterpreted in Bayesian terms (Baldwin & Larson, 2017; McElreath, 2020). Another popular method of statistical analysis in sport science is the magnitude-based inference and it encourage the use of confidence intervals and effect sizes to make a decision about the true or population value of that effect statistic (Batterham & Hopkins, 2015; Hopkins & Batterham, 2018). However, this approach also has several problems like an incorrectly interpretation of frequentist statistics and an increasing risk of finding spurious effects, especially when using small samples (Sainani, 2018; Sainani et al., 2019). Therefore, several authors have proposed the use of Bayesian statistics to overcome the aforementioned issues (Bernards et al., 2017; Borg et al., 2018). Nevertheless, the major drawback is that most current sport scientists are not trained in BDA despite a wide range of popular statistical software already implements Bayesian computation.

R is a programming language for statistical computing used by many scientists for which different packages have been developed in recent years for Bayesian modeling (Mai & Zhang, 2018). Of all of them, the package brms gather some characteristics that make it an ideal starting point to learn BDA: 1) it is user-friendly; models are specified using lme4-like formula syntax; 2) It can be used to fit from single-level linear regression to multivariate or non-linear multilevel models; 3) It uses the probabilistic programming language Stan to fit the models; and 4) it has a large and growing user community (Bürkner, 2017, 2018).

Therefore, the primary aim of this paper is to provide both theorical and practical introduction to BDA for sport scientists. There is no intention to be exhaustive rather to give an overview of the key concepts (highlighted in bold) and practical recommendations for data analysis. This paper is structured in two main sections: 1) a brief introduction to BDA fundamental ideas and 2) an application of the BDA workflow using an example. Excellent introductory texts like J. Kruschke (2014) or McElreath (2020) are recommended to those readers who want to continue learning BDA after reading this manuscript. Throughout this paper it is only assumed that the reader is familiar with the regression analysis and the R programming language for data analysis (R Core Team, 2020). Manuscript's data and reproducible code can be found at https://github.com/JorgeDelro/Intro_Bayesian.

2. Fundamentals of Bayesian data analysis

2.1. Bayes‘ theorem: the engine of Bayesian statistics

The core idea of Bayesian inference is to draw a probabilistic estimate of the parameters of the statistical model (posterior) by combining all the background knowledge (priors) with the new data obtained (likelihood). This process is obtained via Bayes‘ theorem and it generates a reallocation of credibility across possibilities (i.e., An updating of the knowledge toward the information provided by the new data) (J. K. Kruschke & Liddell, 2018). Suppose θ represents the parameters of the model and D represents the observed data, then the basic form of the Bayes‘ theorem can be written as:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{1}$$

where $p(\theta|D)$ is the **posterior probability distribution** and it contains all the information about the model parameters with the data D taken into account. It is important to note that the posterior distribution is a compromise between the data we have at hand and the prior information.

$p(D|\theta)$ is the **likelihood function** as it described the generative process of D given the parameters θ. Usually, researchers choose one of the member of the exponential family to describe the likelihood of the outcome. Sport scientists may be familiar with some of the members of this family like the normal distribution to describe a continuous outcome in linear regression, the binomial distribution for a binary outcome or the Poisson distribution for a count outcome in generalized linear regression.

$p(\theta)$ represents the **prior distribution** and it contains all the information we have about the parameters θ from previous studies and/or opinion from an expert on the subject matter. Generally, three different classes of prior distributions can be distinguished related the amount of (un)certainty they incorporate to the model. *Non-informative* priors (also known as vague prior) have been used commonly on parameters where the researcher has no knowledge about its possible values. However, they should be replaced by a more informative prior to improves inferences due to theorical and computational reasons. *Weakly informative priors* encoded information to restrict the plausible range of values of a specific parameter but still leave a wide range of values to be cover (Gelman et al., 2013). This class of prior distribution has been recently proposed as default prior when there is no information about a parameter of the model. A prior is considered to be *informative* when a researcher includes all the available information in a prior distribution restricting considerably the parameter space. Prior distributions play a key role in BDA, especially when dealing with small sample size due to we can increase the precision of the estimated model parameters by excluding values that are not plausible through the use of informative priors (Zondervan-Zwijnenburg et al., 2017). Guidelines about the construction of informative priors and practical applications have been recently published in the field of phycological research (Koenig et al., 2021).

Finally, $p(D)$ is the **marginal likelihood** or "evidence" which is computed for by summing up (for discrete-valued variables) or integrate (for continuous-valued variables) the product between the likelihood of each value in θ (θ^*) and its prior probability of occurrence. Therefore, the expression to calculate p(D) for continuous-valued variables is:

$$p(D) = \int_\theta p(D|\theta^*)p(\theta^*)d\theta^* \tag{2}$$

To illustrate how the Bayes‘ theorem works consider the Puranen-Orava test which is a clinical test commonly used for the diagnosis of hamstring tendinopathy and strain in athletes (Ahmad et al., 2013; Cacchio et al., 2012). This test has a sensitivity (i.e., probability of a positive diagnostic test when the athlete is indeed positive) of 76% and a specificity (i.e., probability of a positive diagnostic test when the athlete is indeed negative) of 82% (Reiman et al., 2013). The data obtained from the test *(D)* can have two possible values: a positive result *(+)* or negative *(-)*. In this case, the parameter θ represents the real presence in the athlete of hamstring strain *(HS)* or not having hamstring strain *(HSC)*. As an example, the prevalence of hamstring strain in elite football players is 40%, *p(HS)*, and the probability of not having hamstring strain is 60%, *p(HSC)*. Therefore, the probability of having a hamstring strain for an elite football player who is tested positive in this test is:

$$\begin{aligned} p(HS|+) &= p(+|HS)p(HS|)/p(+) \\ &= 0.76 * 0.40/p(+) \end{aligned}$$

According to equation 2 for discrete-valued variables, the marginal likelihood can be computed as:

$$\begin{aligned} p(+) &= [p(+|HS)p(HS)] + [p(+|HSC)p(HSC)] \\ &= [0.76 * 0.40] + [1 - p(-|HSC) * 0.60] \\ &= [0.76 * 0.40] + [(1\breve{}0.85) * 0.60] \\ &= 0.454 \end{aligned}$$

Finally, the value of the posterior is computed by substituting the result of equation 4 into equation 3: *p(HS/+)* $= 0.67$. As conclusion, an elite football player who test positive in the Puranen-Orava test has a probability of 67% of having a hamstring strain.

Although the previous example is the classic demonstration of Bayes‘ theorem, real world application of BDA are much complex for several reasons (Tso et al., 2021): 1) the probability of the parameter is unknow in almost all the data analyses; 2) databases contain multiple variables and subjects; 3) More than one parameter has to be estimated simultaneously; 4) The marginal likelihood is usually too complex to be calculated analytically for continuous parameters due to the high number of combinations in the joint parameter space. Modern Bayesian software implement a sampling technique called Markov chain Monte Carlo (MCMC) (section 2.2) to solve the previous issues and compute a representation of the posterior distribution for the parameters of the statistical model. Therefore, real world estimation of parameters using BDA are calculated with the following equation:

$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{3}$$

where the posterior probability distribution is *proportional* to the likelihood function times the prior distribution. This means that, from a practical point of view, scientists have to specify the likelihood function of the data and the prior distributions on the parameters to compute the posterior distribution.

2.2. Markov chain Monte Carlo methods

This method is the combination of two different techniques, Markov Chains and Monte Carlo simulation (Gill, 2014). The former is a stochastic process (i.e., set of random quantities) where the probability of change to a new state at time t + 1 is dependent only of the current state of the process at time t and conditionally independent of the previous values. The latter is a powerful computational method used to generate independent random samples from a sampling distribution. This empirical samples could be used to summarize the distribution without using analytical calculations. Therefore, a MCMC is a process where random samples are drawn sequentially from the approximate posterior distribution of each model parameter simultaneously. At each step of the sequence, the algorithm corrects the draws using the Markov property of the chain to better approximate the posterior distribution. The key point is that if we run the chain long enough it will converge to a stationary posterior distribution (Gelman et al., 2013). Metropolis-Hastings and Gibbs sampling are probably the most widely known algorithms implemented both in BUGS and JAGS (Lunn et al., 2012). Recently, a probabilistic programming language called Stan have been developed (Carpenter et al., 2017). This software makes use of the No-U-Turn sampler, a variant of Hamiltonian Monte Carlo to compute the posterior distribution (Hoffman & Gelman, 2014). Hamiltonian Monte Carlo sampling have been showed to outperforms Metropolis and Gibbs sampling for complex multiparameter models (Monnahan et al., 2017).

2.2.1 Understanding MCMC methods: the Metropolis algorithm. Due to its simplicity and elegance to obtain samples from the posterior distribution in Bayesian statistics, we believe that the Metropolis algorithm (a special case of the Metropolis-Hastings algorithm) is the ideal starting point to understand how MCMC works. Broadly speaking, this algorithm performs a random walk along all the possible values of the parameter remaining longer on those values with higher posterior distribution. As an example, let assume that our target distribution is $\theta \sim Normal(10, 1)$. This means that, if we run the algorithm long enough, then the samples obtained from the Markov chain will converge in θ. The main steps of the Metropolis algorithm are::

1. Initialize the algorithm within the range of values of θ. In this examples, θ is a continuous parameter so θ_1 could be any real number. Lets say $\theta_1 = 0$.

2. A jump is proposed within the range of values of θ randomly (hence the name random walk). To do this, we will add or subtract a number to the current position of θ. The simplest way to generate a new position is to obtain a random number form a stardardized normal distribution so the increase/decease in the position will be given by $\Delta\theta \sim Normal(0, 1)$. Therefore, the proposed position will be $\theta_{proposed} = \theta_{current} + \Delta\theta$.

3. The probability of accepting the proposed move is calculated. Recall that the algorithm always wants to move to parameter values of higher posterior probability

distribution, so if the proposed position is higher than the current one, then the movement must be accepted. Conversely, if the proposed values is lower then the movement is accepted probabilistically. A random value is drawn from a uniform distribution [0, 1] and if the value obtained is less than the proposed value, then the moment is accepted. Otherwise, the proposed movement is rejected and the algorithm will remain one more iteration in the current position. The probability of accepting the proposed move is given by:

$$p_{accept} = min(1, \frac{P(\theta_{proposed})}{P(\theta_{current})})$$

Where $P(\theta)$ is the probability density of the target distribution at position θ. When $P(\theta_{proposed})$ is higher than $P(\theta_{current})$ then the value of $p_{accept} = 1$ and therefore the proposed move is always accepted. We are going to code the algorithm in R, run it for 10,000 iterations, store the results and display it graphically:

```r
# Initialize the algorithm
theta_init <- 0
# Number of iterations
n_iterations <- 10^4
# Vector to store the results
theta <- rep(0,n_iterations)
# First value of the vector is the init
theta[1] <- theta_init


# Run the algorithm
for(i in 2:n_iterations){
  # Store the result at every iteration
  theta_current <- theta[i-1]
  # Propose a move
  theta_proposed <- theta_current + rnorm(1, mean = 0, sd = 1)
  # Probability of accept the move
  p_accept <- min(1, dnorm(x = theta_proposed,
                               mean = 10,
                               sd = 1) / dnorm(x = theta_current,
                                                mean = 10,
                                                sd = 1))

  # Generates a random value form Uniform(0,1)
  accept_value <- runif(1)
  # Accept or reject the proposed move
if(accept_value < p_accept){
    theta[i] <- theta_proposed
  } else {
    theta[i] <- theta_current
  }
}
```

INSERT FIGURE 1 HERE

The Metropolis algorithm has worked for this simple example (Figure 1). However, in higher dimensional and complex modeling situations, it has a computational limitation and it often takes too long to get an image of the posterior distribution of the parameters.

2.2.2. Advanced MCMC: the Hamiltonian Monte Carlo algorithm.

Hamiltonian Monte Carlo is a more complex algorithm that uses the gradient (i.e. the direction in which the distribution increases) of the log posterior to direct the Markov chain towards regions of higher posterior density (Thomas & Tu, 2021). Suppose $P(\theta)$ is the target distribution and $-logP(\theta)$ has the shape of a reverse bell. To generate samples in regions of high posterior density, the algorithm needs to obtain samples corresponding to low values of $-logP(\theta)$. The moves of the algorithm mimics a marble moving from one side of a valley to other, remaining longer at the bottom of the valley (lower values) and occasionally at the ends (higher values). In physics, such movements are described as a Hamiltonian system where the horizontal and vertical moves are dictated by θ and p, where p is known as the *momentum* variable.

In this algorithm, both θ and p are sampled together and the proposed jump for θ is determined largely by p. This simulation is carried out over time through the *Hamiltonian equations*. Another algorithm called the leapfrog method is used to solve these equations efficiently. This algorithm has 2 important tuning parameters L, the number of iterations of the leapfrog method or the number of steps, and ϵ the step size of every iteration for θ and p. These parameters control how θ and p are both updated so if they are not setting correctly the algorithm could lead to erroneous proposal distributions. As an example, suppose we generate 100 samples from a bivariate Normal distribution $z \sim Normal(\mu, \Sigma)$:

$$\mu = [0, 0]$$

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Then, we are going to sample 5 points by setting L = 10 leapfrog steps and the step size $\epsilon = 0.03$. The algorithm begins at the black diamond (step 0) and continues with random direction and *momentum.* The red dots represent the leapfrog steps while the width of the blue line displays the total kinetic energy at each step (Figure 2). If we run the algorithm long enough we will get an excellent representation of the posterior distribution for variable z.

INSERT FIGURE 2 HERE

At this point is when the No-U-Turn sampler implemented by Stan prevent inefficiencies in this algorithm. The main steps of the Hamiltonian Monte Carlo algorithm are too technical for an introduction to BDA so interested readers are referred to Thomas and Tu (2021) for a detail description and Gelman et al. (2013) and McElreath (2020) to get an implementation in R code.

MCMC methods are implemented by default in the Bayesian software so researchers do not have to worry about manually code it. However, it is essential to assess the representativeness of the posterior distribution and that the estimates of central tendency and limits are accurate and stable using numerical and graphical convergence diagnostics (J. Kruschke, 2014).

2.3. Bayesian data analysis workflow

There is a concern about the lack of reproducibility and efficiency in scientific research. The improvement of statistical and methodological practices is one of the proposed measures to optimize the scientific process (Munafò et al., 2017). In this sense, BDA could lead to erroneous conclusion if it is not use properly. Therefore, several recognized authors have proposed a workflow specific for BDA which includes the steps of model building, checking, inference and reporting. Two of these checklists are the *when to Worry and how to Avoid the Misuse of Bayesian Statistics* (WAMBS-v2) (van de Schoot et al., 2021) and the *Bayesian analysis reporting guidelines* (BARG) (J. K. Kruschke, 2021). Based on the aforementioned workflow guidelines, we are going to summarize the key steps of BDA along with some practical considerations.

2.3.1. Gather prior information. BDA starts even before analyzing the database. Researchers can use the results reported in previous studies to get an idea of the possible values that parameters of interest may have. These values can be incorporated to our analysis via prior distributions and thus exclude values that are not possible to reach. Therefore, to include informative prior in our model is going to provide us the possibility of increase the precision of the result even if the sample size is small. If previous information is not available maybe researchers are able to specify the limit of the parameter space. Some practical guidelines to construct informative priors are (Zondervan-Zwijnenburg et al., 2017): 1) research in high quality scientific literature and ask experts on the subject matter; 2) to use a good method to gather information systematically; 3) to specify where you got the information and 4) always visualize the prior distribution.

2.3.2. Definition of the statistical model. BDA involves the formulation of a full probability model starting from the likelihood function of the data to the prior distribution of the parameters. This mathematical formulation of the model where the values of some parameters depend on the values of other parameters is known as **hierarchical modeling** and represent the **parametization** of the model. Consider the following example with one outcome y and two predictor variables (x_1 and x_2):

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_1 x_{1[i]} + \beta_2 x_{2[i]} \\
\alpha &\sim \text{Normal}(\mu_\alpha, \sigma_\alpha) \\
\beta_1 &\sim \text{Normal}(\mu_{\beta 1}, \sigma_{\beta 1}) \\
\beta_2 &\sim \text{Normal}(\mu_{\beta 2}, \sigma_{\beta 2}) \\
\sigma &\sim \text{HalfCauchy}(\mu_\sigma, \sigma_\sigma)
\end{aligned}
$$

This formulation is the classical linear model where every observation of the outcome variable y is assumed to be distributed according to a Gaussian probability distribution with mean μ and standard deviation σ. Additionally, the mean μ is assumed to be equals a linear combination of the parameters α (i.e., the intercept), the coefficients of x_1 (β_1) and x_2 (β_2). The novel part is that prior probability distribution has been set on the model parameters α, β_1, β_2 and σ. In fact, these priori distributions also have parameters (known as **hyperparameters**) that are also estimated from the data.

2.3.3. Model checking. Two key steps must be considered: Markov chain behavior checking and predictive checking. The most common method to check the behavior Markov chain is by visualizing the one-dimensional **trace plots**. These plots display the value of a parameter at each iteration of the Markov chain on the y axis against the iteration number on the x axis (van de Schoot et al., 2021). We should look for 3 characteristics in a trace plot: 1) stationary, the mean value of the chain is stable during all the iterations; 2) good mixing, the chain fully explores the posterior distribution very quickly, and 3) Convergence, multiple independent chains stick around the same region of high probability (McElreath, 2020). Additionally, The **potential scale reduction factor** ($\hat{R}$) and the **effective sample size** (ESS) are probably the numerical converge diagnostics most used in the Bayesian software. $\hat{R}$ is a measure of how much variance there is between the chains relative to how much variance there is within chains and its value is 1.0 the chains are fully converged or greater if they are not converged to a common distribution. ESS is a measure of how much independent information there is in autocorrelated chains. Recently, an improved version of these numerical diagnostics has been developed and implemented in the probabilistic programming language Stan (Vehtari et al., 2021). Stan output reports for every parameter estimated the maximum of rank normalized $split - \hat{R}$ and rank normalized $folded - split - \hat{R}$ which work for thick tailed distributions and is sensitive also to differences in scale. Moreover, the bulk effective sample size (bulk-ESS) and tail effective sample size (tail-ESS) are reported. The former informs about the sampling efficiency in the bulk of the distribution (related to efficiency of mean and median estimates) whereas the latter is a measure for sampling efficiency in the tails of the distribution (related to efficiency of variance and tail quantile estimate). It is recommended from a practical point of view to run at least four chains by default to estimate the posterior distribution of model parameters using MCMC and use 1.01 (or lower) and 400 (or greater) as thresholds for $\hat{R}$ and ESS respectively to trust in the posterior distribution estimated.

Regarding predictive checking, it is a method to assess how similar is the observed data with the data generated under the fitted model. This is possible by simulating values from the joint predictive distribution and comparing these samples with the observed data (Gelman et al., 2013). In this sense, **prior predictive checking** assess that the prior distributions defined in the model really generates simulated data (from the prior predictive distribution) according to the prior knowledge *before observing the data* while **posterior predictive checking** is used to check whether the simulated data (from the posterior predictive distribution) resemble the observed data *after observing the data* (J. K. Kruschke, 2021; van de Schoot et al., 2021) Therefore, any deviation from true prior knowledge and/or data generating process could be considered a model misfit and a reformulation of the model should be performed. Posterior predictive checks can be also done numerically and with model comparison purpose as we explain in section 2.3.4.

2.3.4. Model comparison and predictive accuracy. Once the model is fitted sport researchers assess how well the model fit to the sample. Probably, the most common measure of goodness-of-fit is R^2 or “variance explained”. This measure has the problem that it increases when more predictors are added to the model even when the variables you add are random numbers. Moreover, while models with many parameters fit the data better, they tend to *overfit* more than simple models. **Overfitting** occurs when the model learns too much from the sample which leads to poor out-of-sample predictions. In contrasts, when a model has too few parameters, they are inaccurate both within and out-of-sample producing a statistical error called **underfitting**. To deal with the overfitting/underfitting dichotomy we can use two different approaches: cross-validation and information criteria.

The first approach consists basically on leave out a small part of our sample to test the model´s predictive accuracy. Therefore, the sample is divided into chunks (i.e., folds) which the statistical model predicts one by one using the remaining chunks of the sample. Then, an average score of the out-of-sample accuracy if obtained. There is a special cross-validation method for Bayesian models called the **Pareto smoothed importance sampling cross-validation** (PSIS-LOO) to estimate the model´s out of sample accuracy (Vehtari et al., 2017). This method computes the **expected log pointwise predictive density** which it is a useful measure to compare models and the **Pareto *k* diagnostics** which informs us about the reliability of the estimate by pointing to influential observations. Specifically, those data points associated with a k value higher than 0.7 are supposed to have a negative on PSIS-LOO score. A difference less than 4 in the expected log pointwise predictive density is considered small from a practical point of view when comparing models with a number of observations larger than 100. On the other hand, an information criterion is an estimate of the relative out-of-sample divergence. Thus, the model with a smaller deviance has better fit when comparing several models. The **widely applicable information criterion** (WAIC) is an information criterion that is invariant to parametrization, it uses entire posterior distribution and approximates the deviance for new samples (McElreath, 2020; Watanabe, 2010).

2.3.5. Posterior probability distribution analysis and hypothesis testing. Sport scientists maybe know the model of the section 2.2 as ANCOVA where the interest resides in estimate mean difference among groups by using some kind of planned contrasts or post-hoc analysis while adjusting the model with a continuous variable. In this way, the decision of whether or not there is a statistical significant difference among training groups is based on the computation of a p-value and if it is less than an established threshold (traditionally if $p < 0.05$). However, several publications have alarmed about the misuse and misinterpretation of *p*-values as an index of significance (Amrhein et al., 2019; Greenland et al., 2016). As an alternative, Bayesian inference offers three different overlapping approaches to analyze the presence or absence of an effect: the **Region of Practical Equivalence** (ROPE)-based indices, posterior indices and **Bayes factors** (BF) (Makowski, Ben-Shachar, Chen, et al., 2019).

The first approach uses an interval called a **credible interval** (CrI) which define a percentage of values that are found in the central portion of the posterior distribution. Although its aim is similar, CrI should not be confused with confidence intervals since its computation and meaning are different. A special CrI is the **highest-density interval** (HDI) which summarizes the uncertainty of the parameter estimated in such way that any parameter value inside a 95%HDI are the 95% most credible values. Then, it is calculated what percentage of the HDI falls inside a ROPE that represent a range of parameter values that equivalent to the null value for practical purposes (J. K. Kruschke, 2018). Thus, if for example the 95%HDI falls completely inside the ROPE means that the most credible values of the parameter a practically equivalent to the null value. An obvious drawback of this method is that the ROPE has had to be established by the researcher. Posterior indices indicate objective characteristics of the posterior distribution like the probability that a estimated parameter is strictly positive or negative. In fact, this index called **probability of direction** (pd) has been recome3nded recently as an objective index of effect *existence* for its simple interpretation and numeric proximity with the *p*-values (Makowski, Ben-Shachar, Chen, et al., 2019). The third approach is based on the comparison of two probability distributions: A prior distribution where all the probability is allocated over the null value (or ROPE), and a posterior distribution where the probability mass has shifted away from the null value once the observed data have been taken into account. Therefore, a BF indicates the degree to which the posterior distribution has move further away or closer to the null value, or in other words, it tells us how much the data is consistent with one hypothesis compared to other. A BF can only be a positive number and its interpretation ranges from “no evidence” (BF = 1) to “extreme” (BF > 100, extreme evidence for H1; BF < 1/100, extreme evidence for H0) (Lee & Wagenmakers, 2014).

It is important to note some advantages of using these approaches to perform post-hoc analysis: 1) there is no need to correct for multiple tests due to type I error rate inflation due to BDA does not rely on sampling distributions; 2) Conversely to frequentists statistics results, the interpretation of a HDI or pd is intuitive; and 3) BF assess both evidence in favor or against an effect (in contrasts to *p*-values).

2.3.6. Sensitivity analysis. In a BDA context, a sensitivity analysis means to assess how sensible is the estimated posterior distribution of the parameters to the choice of priors. This analysis is especially important when analyzing small datasets with informative priors however it should be performed regardless of the amount of information provided by the prior distributions and the sample size (J. K. Kruschke, 2021). Although there is no consensus about the way to assess differences among posterior distributions from different priors, we are going to follow the recommendations of J. K. Kruschke (2021) who suggest to plot density curves along with numerical tables showing the central tendency and credible interval of the estimated parameters.

3. Applied Bayesian data analysis example

We are going to consider as an example the study of Humberstone-Gough et al. (2013) to illustrate the aforementioned workflow. Briefly, they compared the effects of three different training regimens “Live High Training Low” altitude training (LHTL, n = 7), “Intermittent Hypoxic Exposure” (IHE, n = 7), and “Placebo” (n = 7) on different variables using a randomized control trial design. For the sake of simplicity, the difference in the concentration of hemoglobin mass in grams (Hbmass) is going to be the outcome of our example while the percentage change in weekly training load (ChangeWtr, %) and training group membership (Group, three levels: LHTL, IHE and Placebo) are the predictor variables. Our interest as researchers lies in analyzing differences among the training groups.

A box plot of the Hbmass by group show us that the outcome follows a Gaussian distribution in each group, the presence of an outlier in the IHE group and a possible effect of group membership (figure 3).

INSERT FIGURE 3 HERE

There are only 7 participants in each group so prior information about the parameters can help us to get a reliable estimate. In this case, an informative prior about the effect of LHTL was placed based on a meta-analysis about training regimens on Hbmass (Gore et al., 2013).

Note that this data has been already analyzed using Bayesian methods by Mengersen et al. (2016). However, throughout this example we are going to perform BDA showing the computer code at every step of the analysis using a more accessible software for sports scientists in addition to following a modern approach of analysis.

3.1. Model definition

We assumed that the Hbmass is distributed according a Gaussian distribution, where its mean (μ) is model as a linear combination of the effect of ChangeWtr and Group, while its standard deviation (σ) follow a half-Student´s T distribution (only positive values of the distribution). Therefore, the statistical model can be described as follow:

$$
\begin{aligned}
Hbmass_i &\sim Normal(\mu_i, \sigma) \ [likelihood] \\
\mu_i &= \alpha + \beta_1 ChangeWtr + \beta_2 Group \ [linear\ model] \\
\alpha &\sim StudentT(12, 7, 3) \ [\alpha\ prior] \\
\beta_1 &\sim Normal(0, 2) \ [\beta_1\ prior] \\
\beta_{2-IHE} &\sim Normal(0, 2) \ [\beta_{2-IHE}\ prior] \\
\beta_{2-LHTL} &\sim Normal(2.6, 0.5) \ [\beta_{2-LHTL}\ prior] \\
\sigma &\sim Half - StudentT(0, 15, 3) \ [\sigma\ prior]
\end{aligned}
$$

Under this model definition, the intercept (α) of the regression model represents the average *Hbmass* in the placebo group whereas β_{2-IHE} and β_{2-LHTL} represents the difference between placebo and IHE and between placebo and LHTL, respectively.

Before fitting the model, It is a good practice to plot prior distribution to check the range of plausible values for each parameter (Figure 4).

INSERT FIGURE 4 HERE

3.2. Prior predictive checking

Once the model is defined, the prior predictive distribution can be computed to check whether the model and prior distributions are consistent with domain expertise removing extreme but not impossible parameters values (section 2.3.3.). Hence, adding information via prior distribution allows the Bayesian computation and interpretation of the parameters estimated. The function `prior()` allow to define prior distribution on model parameters. Prior predictive distribution can be computed via brms by using the function `brm()` and setting the argument `sample_prior = "only"`. The function `brm()` can be considered the main function of the package since is the one used to fit the models. Consider special attention to arguments of the function related to the MCMC, `warmup` to set the number of iterations used by the MCMC algorithm to figure out how to explore the posterior distribution efficiently; `chains` to specify the number of Markov chains and `iter` to set the number of iterations per chain. In our example, we create an object called `bmod1_prior` which will store all the information about the model. Additional arguments like `data` to select a data frame that contains all the variables in the model; `family` to set the likelihood function of the outcome and `prior` to use the prior distribution on parameters previously defined, are mandatory Note that our model assume that the outcome follows a Gaussian distribution with an identity link function (`family = gaussian(link = "identity")`).

```
bmod1Priors <- c(prior(normal(0, 2), class = "b", coef = "ChangeWtr"),
                 prior(normal(0, 2), class = "b", coef = "GroupIHE"),
                 prior(normal(2.6, 0.5), class = "b", coef = "GroupLHTL"),
                 prior(normal(0, 2), class = "b", coef = "Intercept"),
                 prior(student_t(3, 0, 15), class = "sigma"))

bmod1_prior <- brm(formula =  HMabs ~ 0 + Intercept + ChangeWtr + Group,
                data = dbHb,
                family = gaussian(link = "identity"),
                warmup = 1000,
                iter = 2000,
                chains = 4,
                seed = 1234,
                prior = bmod1Priors,
                sample_prior = c("only"))
```

After fitting the model, multiple draws can be computed from the prior predictive distribution by using the function `posterior_predict()`. In this case, we are going to simulate 50 draws:

```
bmod1_prior %>%
posterior_predict(draws = 50) %>%
ppc_dens_overlay(y = dbHb$HMabs) +
    xlim(-750, 750)
```

Prior predictive distribution is showed in figure 5. In this figure y represent the distribution of Hbmass and y_{rep} the distribution of simulated sets using only information from prior distributions. Note that most of the distribution area is over the value 0 and values ± 100 g for Hbmass are very unlikely.

INSERT FIGURE 5 HERE

3.3. Model updating

Next, we are going to add the observed data to the model by changing the argument `sample_prior = "yes"`. Once brms fits the model we should check that the parameters have been estimated correctly (see section 2.2). the traceplot of each Markov chain used in the MCMC estimation and a histogram of the values estimated for every parameter (figure 6). These traceplot are concentrated around the estimated value for each parameter. Moreover, all the parameters have an $\hat{R}$ of 1 and both ESS > 400 so we can trust that these results have been obtained with accuracy (table 2).

```
bmod1 <- brm(formula =  HMabs ~ 0 + Intercept + ChangeWtr + Group,
             data = dbHb,
             family = gaussian(link = "identity"),
             warmup = 1000,
             iter = 2000,
             chains = 4,
             seed = 1234,
             prior = bmod1Priors,
             sample_prior = c("yes"))
```

INSERT FIGURE 6 HERE

3.4. Posterior predictive checking

We are going to simulate data sets (y_{rep}) to compare with the distribution of the observed data (y), like in section 3.2, but in this case the data is simulated from the posterior predictive distribution. Recall that this method is used to asses model adequacy. Figure 7 shows the posterior predictive distribution of our model. Look like the fit is reasonable but there is a high variation that it is not capture by model´s prediction.

```
ppc_dens_overlay(y = dbHb$HMabs,
         yrep = posterior_predict(bmod1, draws = 50)) +
  xlim(-150, 150)
```

INSERT FIGURE 7 HERE

3.5. Model selection

Figure 1 showed the presence of an outlier in IHE group so perhaps we could improve the model if we use a likelihood function that allows the presence of extreme values. This kind of method is commonly known as robust regression and makes use of the Student-t distribution. Like the Gaussian distribution, the Student-t distribution is defined by the mean μ and the scale σ parameters, but it has also the shape parameter ν that controls the thick of the tails of the distribution. Robust regression can be easily performed using brms by changing the argument `family = student(link = identity)`.

```
bmod2 <- brm(formula = HMabs ~ 0 + Intercept + ChangeWtr + Group,
        data = dbHb,
        family = student(link = "identity"),
        warmup = 1000,
          iter = 2000,
        chains = 4,
          seed = 1234,
         prior = bmod1Priors,
       sample_prior = "yes")
```

Once the model is fitted, we can compare the predictive accuracy of both model (section 2.4). First, the PSIS-LOO is estimated for each model via `loo()` function setting the argument `save_psis = T` and then function `loo_compare()` is used to compare the estimates. This function computes pairwise comparisons between the model with the largest expected predictive density (first row, better accuracy). In our case, the difference can be considered insignificant due to the small numbers computed (table 2). Interestingly, the Gaussian model has better accuracy so we are going to use that model to perform contrasts.

```
loo1 <- loo(bmod1, save_psis = TRUE)
loo2 <- loo(bmod2, save_psis = TRUE)
loo_compare(loo1, loo2)
```

INSERT TABLE 1 HERE

3.6. Posterior distribution analysis and hypothesis testing

To illustrate the approaches commented in section 2.3.5., we are going to analyze the effect of group membership on Hbmass by using both ROPE and Bayes factors approaches but independently. Readers must be aware that these approaches are not mutually exclusive and they could be used together to analyze the presence and significance of an effect (Makowski, Ben-Shachar, Chen, et al., 2019).

3.6.1. ROPE approach

As an example, we define a ROPE from -0.5 to 0.5 grams. This ROPE is the null value for practical purposes in our study. The percentage of the *full* posterior distribution that lies inside this ROPE is going to be the decision rule (J. K. Kruschke, 2018). If less than 2.5% of the full posterior distribution of the parameter lies outside the ROPE then the null hypothesis is “rejected”. On the other hand, if more than 97.5% is inside the ROPE then the null is “accepted”. The functions `describe_posterior()` and `equivalence_test` from the package *bayestestR* calculate the percentage of the full posterior distribution inside the ROPE and perform a test for practical equivalence with these results (Makowski, Ben-Shachar, & Lüdecke, 2019). The arguments of these functions allow to define the ROPE´s lower and upper bound (`rope_range` and `range`, respectively), the type of credible interval (`ci_method`) and the percentage of the credible interval to be evaluated inside the ROPE (`ci`). The probability of direction is also calculated as an index of *existence* of the effect.

```r
describe_posterior(bmod1, rope_range = c(-0.5, 0.5), ci_method = "HDI", ci = 1)
equivalence_test(bmod1, range = c(-0.5, 0.5), ci = 1)
```

INSERT TABLE 2 HERE

INSERT FIGURE 8 HERE

We can conclude from these results that *the effect of LHTL training has a probability of 100% [pd] of being positive (mean = 2.66, 95%CrI[1.71, 3.59]) and significant (0.00% inside ROPE)* (Table 2 and figure 8).

3.6.2. Bayes factors approach

As an example of Bayes factors computation and interpretation, we are going to test a planned contrast (also known as a priori contrast) about mean differences. Formally, we could test three different null hypothesis regarding mean differences among the levels of the group variable:

$$H_0^1 : \mu_{placebo} - \mu_{IHE} = 0$$
$$H_0^2 : \mu_{placebo} - \mu_{LHTL} = 0$$
$$H_0^3 : \mu_{IHE} - \mu_{LHTL} = 0$$

Where H_0^1 represent the null hypothesis for placebo VS IHE; H_0^2, placebo VS LHTL and H_0^3, IHE VS LHTL.

These specific contrasts (i.e. pairwise differences) should be encoded as a character string by using the name of the model parameters. The function `hypothesis()` allows to perform multiple non-linear hypothesis test for model parameters. In our example, to test H_0^1 (`"Intercept = Intercept + GroupIHE"`), H_0^2 (`"Intercept = Intercept + GroupLHTL"`) and H_0^3 (`"Intercept + GroupIHE = Intercept + GroupLHTL"`) respectively we should use the following code:

```
hypothesis(bmod1, c("Intercept + GroupIHE = Intercept ",
                    "Intercept + GroupLHTL = Intercept ",
                    "Intercept + GroupIHE = Intercept + GroupLHTL"))
```

This function computes a Bayes factor between the hypothesis and its alternative and is expressed as BF_{10}, and for two-sided hypothesis the BF is computed via the Savage-Dickey density ratio method which is merely the ratio between the height of the posterior distribution and the height of the prior distribution at the point of interest (Wagenmakers et al., 2010). This result refers to the evidence of H_1 (i.e., alternative hypothesis = significant difference) over H_0 (i.e., null hypothesis = no significant difference). For hypothesis 1, 2 and 3 the BF_{10} is 1,05 , >100 and 0.91 respectively. This evidence can be classified as anecdotical for hypothesis 1 and 3 and extreme for hypothesis 2 (table 3). Additionally, we also computed a standardized effect size for mean differences called Cohen´d which can be calculated using the following formula:

$$Cohen\ d = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}}$$

where μ and σ^2 represents the mean and the variance of the groups to be compared.

INSERT TABLE 3 HERE

From the Bayes factor analysis we can report that *there is extreme evidence (BF > 100) supporting an effect of LHTL training group over placebo group (mean difference = 2.65, 95%CrI[0.60, 4.35], Cohen´d = 1.30)* (table 2).

3.7. Sensitivity analysis results

We refit the model setting non-informative priors on regression parameters to compare to each of our original priors and understand the impact of different priors on the posterior distribution. These non-informative priors were set as follow:

$$\begin{aligned}
\alpha &\sim Normal(0, 10^5) \\
\beta_1 &\sim Normal(0, 10^5) \\
\beta_{2-IHE} &\sim Normal(0, 10^5) \\
\beta_{2-LHTL} &\sim Normal(0, 10^5) \\
\sigma &\sim Half - Normal(0, 10^5)
\end{aligned}$$

After the model was refitted (named “bmod3”) with non-informative prior, we used the function `sensitivity_analysis()` from the package *introbayes* which compares the posterior densities of the selected parameters graphically and computing the percentage of deviation ((mean – original mean)/ original mean * 100) between the fitted models (Kery, 2010).

```
sens_analysis <- sensitivity_analysis(bmodels = list(original_model = bmod1,
                                      alternative_prior = bmod3),
                                      params = c("b_Intercept", "b_ChangeWtr",
                                                 "b_GroupIHE", "b_GroupLHTL",
                                                 "sigma"))
```

INSERT TABLE 4 HERE

INSERT FIGURE 9 HERE

Sensitivity analysis showed *a high impact of prior distribution on the results obtained. Specifically, there is a significant effect on the posterior distributions of α (-839%), on β_{2-IHE} (736%) and on β_{2-LHTL} (1058%). Additionally, the variance of the posterior distributions of α, β_{2-IHE} and β_{2-LHTL} reduce drastically after incorporate the data into the model. Regarding the 90%HDI, zero is always inside the HDI for β_{2-IHE} and for β_{2-LHTL}* (table 4 and figure 9).

4. Conclusions

BDA offers a very interesting alternative for sport scientists who want to overcome the limitation of traditional statistics, especially those who need to analyze databases with low sample size. Obviously, there are lot of concepts and methods that have not been treated in this introduction. However, through this manuscript the basic concepts, benefits, workflow and a practical example are presented as a starting point for those who are interested in learn how to perform Bayesian inference.

Acknowledgments

Different R packages have been used together with *brms* to perform the analysis of this manuscript: *tidyverse* and *ggpubr* for data manipulation and plotting and *bayesplot*, *bayestestR* and *loo* for further analysis (Gabry & Mahr, 2017; Kassambara & Kassambara, 2020; Makowski, Ben-Shachar, & Lüdecke, 2019; Vehtari et al., 2020; Wickham et al., 2019).

References

Ahmad, C. S., Redler, L. H., Ciccotti, M. G., Maffulli, N., Longo, U. G., & Bradley, J. (2013). Evaluation and management of hamstring injuries. *The American Journal of Sports Medicine*, *41*(12), 2933–2947.

Amrhein, V., Greenland, S., & McShane, B. (2019). *Scientists rise up against statistical significance*. Nature Publishing Group.

Baldwin, S. A., & Larson, M. J. (2017). An introduction to using Bayesian linear regression with clinical data. *Behaviour Research and Therapy*, *98*, 58–75.

Batterham, A. M., & Hopkins, W. G. (2015). The case for magnitude-based inference. *Medicine and Science in Sports and Exercise*, *47*(4), 885.

Bernards, J. R., Sato, K., Haff, G. G., & Bazyler, C. D. (2017). Current Research and

Statistical Practices in Sport Science and a Need for Change. *Sports (Basel)*, *5*(4). https://doi.org/10.3390/sports5040087

Borg, D., Minett, G., Stewart, I., & Drovandi, C. (2018). Bayesian methods might solve the problems with magnitude-based inference. *Medicine and Science in Sports and Exercise*, *50*(12), 2609–2610.

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Cacchio, A., Borra, F., Severini, G., Foglia, A., Musarra, F., Taddio, N., & De Paulis, F. (2012). Reliability and validity of three pain provocation tests used for the diagnosis of chronic proximal hamstring tendinopathy. *British Journal of Sports Medicine*, *46*(12), 883–887.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, *25*(1), 207–218.

Gabry, J., & Mahr, T. (2017). Bayesplot: Plotting for Bayesian models. *R Package Version*, *1*(0).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gill, J. (2014). *Bayesian methods: A social and behavioral sciences approach* (Vol. 20). CRC press.

Gore, C. J., Sharpe, K., Garvican-Lewis, L. A., Saunders, P. U., Humberstone, C. E., Robertson, E. Y., Wachsmuth, N. B., Clark, S. A., McLean, B. D., Friedmann-Bette, B., et al. (2013). Altitude training and haemoglobin mass from the optimised carbon monoxide rebreathing method determined by a meta-analysis. *British Journal of Sports Medicine*, *47*(Suppl 1), i31–i39.

Greenberg, E. (2012). *Introduction to Bayesian Econometrics* (Second). Cambridge University Press. https://doi.org/10.1017/CBO9781139058414

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(47), 1593–1623.

Hopkins, W. G., & Batterham, A. M. (2018). The vindication of magnitude-based inference. *Sportscience*, *22*, 18–30.

Humberstone-Gough, C. E., Saunders, P. U., Bonetti, D. L., Stephens, S., Bullock, N., Anson, J. M., & Gore, C. J. (2013). Comparison of live high: Train low altitude and intermittent hypoxic exposure. *Journal of Sports Science & Medicine*, *12*(3), 394.

Kassambara, A., & Kassambara, M. A. (2020). Package "ggpubr." *R Package Version 0.1*, *6*.

Kery, M. (2010). *Introduction to WinBUGS for Ecologists: Bayesian approach to regression, ANOVA, mixed models and related analyses.* Academic Press, Inc.

Kéry, M. (2010). *Introduction to WinBUGS for ecologists: Bayesian approach to regression, ANOVA, mixed models and related analyses.* Academic Press.

Koenig, C., Depaoli, S., Liu, H., & Van De Schoot, R. (2021). Moving beyond non-informative prior distributions: Achieving the full potential of bayesian methods for psychological research. *Frontiers in Psychology*, *12*.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.* Academic Press.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation.

Advances in Methods and Practices in Psychological Science, *1*(2), 270–280.

Kruschke, J. K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behaviour*, *5*(10), 1282–1291.

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155–177.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Lesaffre, E., & Lawson, A. B. (2012). *Bayesian biostatistics*. John Wiley & Sons.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. CRC press.

Mai, Y., & Zhang, Z. (2018). Software Packages for Bayesian Multilevel Modeling. *Null*, *25*(4), 650–658. https://doi.org/10.1080/10705511.2018.1431545

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

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, *4*(40), 1541.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

Mengersen, K. L., Drovandi, C. C., Robert, C. P., Pyne, D. B., & Gore, C. J. (2016). Bayesian Estimation of Small Effects in Exercise and Sports Science. *PLoS One*, *11*(4), e0147311. https://doi.org/10.1371/journal.pone.0147311

Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, *8*(3), 339–348. https://doi.org/10.1111/2041-210X.12681

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du

Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9.

Reiman, M. P., Loudon, J. K., & Goode, A. P. (2013). Diagnostic accuracy of clinical tests for assessment of hamstring injury: A systematic review. *Journal of Orthopaedic & Sports Physical Therapy*, *43*(4), 222–231.

Sainani, K. L. (2018). The problem with" Magnitude-based Inference". *Medicine and Science in Sports and Exercise*, *50*(10), 2166–2176.

Sainani, K. L., Lohse, K. R., Jones, P. R., & Vickers, A. (2019). Magnitude-based inference is not Bayesian and is not a valid method of inference. *Scandinavian Journal of Medicine & Science in Sports*, *29*(9), 1428.

Santos-Fernandez, E., Wu, P., & Mengersen, K. L. (2019). Bayesian statistics meets sports: A comprehensive review. *Journal of Quantitative Analysis in Sports*, *15*(4), 289–312. https://doi.org/10.1515/jqas-2018-0106

Thomas, S., & Tu, W. (2021). Learning hamiltonian monte carlo in R. *The American Statistician*, *75*(4), 403–413.

Tso, I. F., Taylor, S. F., & Johnson, T. D. (2021). Applying hierarchical bayesian modeling to experimental psychopathology data: An introduction and tutorial. *Journal of Abnormal Psychology*, *130*(8), 923.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., et al. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1), 1–26.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Vehtari, A., Gelman, A., Gabry, J., Yao, Y., Bürkner, P., Goodrich, B., Piironen, J., & Magnusson, M. (2020). Loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models.(2016). *R Package Version*, *2*(0).

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Burkner, P.-C. (2021).

Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC. *Bayesian Anal.* https://doi.org/10.1214/20-BA1221

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychon Bull Rev*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *J. Mach. Learn. Res.*, *11*, 3571–3594.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Zondervan-Zwijnenburg, M., Peeters, M., Depaoli, S., & Van de Schoot, R. (2017). Where Do Priors Come From? Applying Guidelines to Construct Informative Priors in Small Sample Research. *Null*, *14*(4), 305–320. https://doi.org/10.1080/15427609.2017.1370966