Help, My Collaborator Goes Bayesian! Why And How To Apply Bayesian Data Analysis

Wouter de Nooy[1,2]

[1] Amsterdam School of Communication Research ASCor
[2] University of Amsterdam

Abstract

This paper introduces researchers to a Bayesian approach to statistical analyses. Starting with the problems of null hypothesis significance testing, it outlines some fundamental advantages of a Bayesian statistical analysis. Two types of Bayesian approaches are illustrated using two different software packages: JASP and the R package `rstanarm`.

Keywords: R; problems null hypothesis significance testing; Bayesian statistical analysis;
Word count: 6755

1 What Is The Problem With Null Hypothesis Significance Testing?

Imagine that you have executed a regression analysis and one of the regression coefficients is not statistically significant, so you do *not* reject the null hypothesis of no effect in the population. I am nearly sure that you commit a *Type II error* here: rejecting a false null hypothesis. Why? Because the regression coefficient may be very close to zero in the population but it will hardly ever be exactly zero. So the null hypothesis of zero effect is nearly always false (Bakan, 1966: 425-6).

Why test a two-sided null hypothesis if we know that it is wrong? What information do we gain if we reject a null hypothesis that is nearly sure to be wrong? It is silly. From this perspective, it makes more sense to test *one-sided null hypotheses.* If we hypothesize that an effect can only be positive, for example, the null hypothesis of no effect or a negative effect can actually be true. In this case, we can fail to reject the null hypothesis without making a Type II error. Unfortunately, one-sided tests are unpopular in the field.

In a one-sided or two-sided test, can we conclude that the effect in the population is more different from zero, that is, stronger, for a rejected false null hypothesis than for a false null hypothesis that is not rejected? No, we can't for several reasons. Our sample can be more or less representative of the population: We may have drawn a sample with a large effect from a population with a very small effect or the other way around. In addition, a larger sample yields statistically significant results more easily than a smaller sample. Very small effects

Wouter de Nooy, Department of Communication Science, University of Amsterdam.
Correspondence concerning this article should be addressed to Wouter de Nooy, Nieuwe Achtergracht 166, 1018 WV Amsterdam. E-mail: w.denooy@uva.nl

can be statistically significant in very large samples, whereas large effects can be statistically non-significant in a very small sample.

It is a common mistake to interpret statistical significance as an indication that an effect is substantial or substantively interesting. This mistake was noted early on (for example, by Tyler, 1931). The book *Understanding the New Statistics* (Cumming, 2012) offers a lucid discussion of this mistake. The binary decision associated with a null hypothesis significance test incorrectly suggests a qualitative difference between a statistically significant effect and an effect that is not statistically significant.

It is easy to misunderstand and misinterpret p values and confidence intervals because the probabilities involved are counterintuitive. We would like to receive probabilities for population values, such as the probability that the effect is positive in the population or the probability that the population effect is within a particular range. Instead, we get probabilities of sample outcomes. Instead of asking for the probability that it will rain tomorrow, this approach reasons: "If it rains tomorrow, what is the probability that we are having at least as many clouds as we have today?"

The frequentist approach to statistics, which underlies null hypothesis significance testing, assumes that there is one true population value. In this approach, a population characteristic (parameter) does *not* have a probability. There is no probability distribution to a parameter. As a consequence, the following statements are not correct:

- A p value is the probability that the null hypothesis is true.
- A 95%-confidence interval contains the population values that have 95% probability of being true.

The practical value of this approach is quite limited. Imagine that you want to use your sample results for a practical decision such as fielding a new campaign. Based on your sample, you would like to calculate the probability that the effect of the campaign is above a particular threshold, for example, the minimum effect size that makes the investments in the campaign worthwhile. Your significance test only tells you whether you should reject the null hypothesis of no effect. Your confidence interval gives you plausible population values but not the probabilities of those values. It doesn't tell you the probability of a positive effect. Why? Because there is no probability distribution for a population characteristic in this approach.

Assuming (in the null hypothesis) a true population value, very efficient methods have been invented for generalizing results from samples to populations. In the pre-computer and early-computer era, this was the only option for statistical generalization. Thanks to our present-day computing power, however, we have alternatives, such as Bayesian statistics.

There are more problems with null hypothesis significance testing, some of which will be touched upon in later sections of this paper. Let us conclude this section with some references. You may want to know more about these problems or your reviewer may ask you why you do not test a null hypothesis. Consult or refer to publications like the *Understanding the New Statistics* book referenced above, (Ziliak & McCloskey, 2008) or (Morrison & Henkel, 1970) for an older survey of the discussion. Carver (1993) offers some suggestions on how to get articles published without significance tests. Levine et al. (2013; 2008) survey the problems of null hypothesis significance testing in published communication research.

2 What Are The Benefits Of A Bayesian Approach?

This paper does not pretend to present an introduction to Bayesian statistics. It won't even present Bayes' theorem, which really is inadmissible in an introduction. There are several

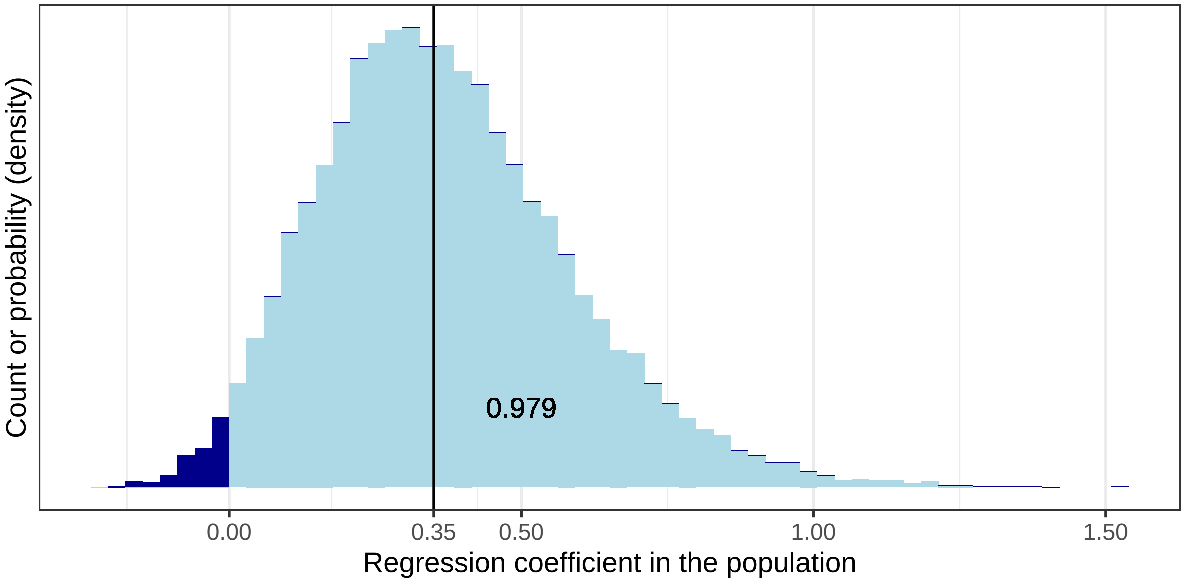


Figure 1. An example of a probability distribution for a parameter such as the regression coefficient in the population.

introductions to Bayesian statistics (Berry, 1996; Clark, 2014; Kurt, 2019; Lambert, 2018; Lee, 2012), among which I particularly recommend Erickson (2017) as a first read. Here, I will just point out some benefits of a Bayesian approach in comparison to the standard frequentist approach.

2.1 Intuitive results

In Bayesian statistics, population values have a probability distribution. In other words, Bayesian statistics allow us to speak of the probability that it will rain tomorrow given our knowledge of today's weather. It gives us probabilities like the ones that we use in ordinary life.

In the example of a regression coefficient estimating the effect of campaign exposure, a Bayesian analysis gives us a probability distribution of the effect in the population. Practically speaking, we obtain a long list of regression coefficient estimates, which is similar to a bootstrapped distribution. Usually, the median value (vertical line in Figure 1) of this distribution is used as the point estimate.

The point estimate is a very poor summary of the distribution. The probabilities implied by the distribution are more interesting. The proportion of the estimates that are above zero, for example, can be interpreted as the probability that the effect is positive in the population. If you like, you can use this probability to test the one-sided hypothesis that the effect is positive. We can actually conclude that the probability of a positive effect in the population is 98 per cent in this example. In contrast, a significance test of the null hypothesis that the effect is not negative only allows us to conclude that our sample is very likely *if the null hypothesis is true*. We cannot reject the null hypothesis but we don't know the probability of a positive effect in the population.

A probability distribution of the regression coefficient in the population helps us to understand that it does not make sense to focus on one value as in a two-sided null hypothesis significance test. In theory, a regression coefficient, like most statistics, can have an infinite number of decimal places. The probability of one exact value such as 0.00000000000000000000(. . . and so on) must be close to zero. In other words, this one value is highly improbable. We know that beforehand, so we are wasting our time if we test against this one value.

Instead, we have to look at a range of values for the statistic (Kruschke, 2011: 302), the regression coefficient in the current example, if we want to use a probability in our conclusion. In line with the concept of confidence intervals, Bayesians use the *credible interval* as the population values that are most likely to be true. The 95% credible interval, for example, contains the 95 per cent estimates in the center of the probability distribution. Here, we can say that the population value has 95 per cent probability (or chance) to be within this credible interval. This is how we would like to (and sometimes erroneously do) interpret the traditional 95 per cent confidence interval.

Let us return to the practical decision on launching a campaign based on a pretest in a sample. The Bayesian probability distribution of the regression coefficient for the effect of campaign exposure tells us the probability of a positive effect versus the probability of a negative effect in the population. If the campaign is intended to have a positive effect, we may advise to launch the campaign if the probability of a positive effect is much larger than the probability of a negative effect. What constitutes “much larger” is a substantive decision, not a statistical decision. If we can quantify the costs and benefits of the campaign, we can use the probability distribution to calculate the expected (averaged over the probability distribution) benefits and compare them to the costs.

2.2 Other distributional assumptions

The sampling distribution is central to null hypothesis significance testing and confidence intervals. We usually approximate the sampling distribution by a theoretical probability distribution, such as the normal, t, or F distribution. The approximation only works if the data meet requirements such as a normal distribution in the population, equal population variances, or minimum sample size. If your data does not meet the requirements, the results are not to be trusted.

The sampling distribution, however, does not play a role in a Bayesian approach. In the latter approach, we are estimating probabilities of population values, not probabilities of samples. As a consequence, we do not care about requirements for approximating the sampling distribution with a theoretical probability distribution. The probability distribution resulting from a Bayesian analysis can have any shape.

This freedom comes at a cost. In Bayesian analysis, the researcher has to specify a probability distribution for each population value that is to be estimated. This probability distribution is called the *prior distribution.* The sample data update the prior distribution into the *posterior distribution*, which is the probability distribution that the researcher interprets and reports. A Bayesian approach, then, comes with a different type of distributional assumptions, namely the choice of prior distributions.

How do we chose prior distributions? In the ideal case, the posterior distributions of previous research can be used as prior distributions in the analysis of new data. This is called an *informative prior distribution.* It represents more or less established knowledge about the statistic that we estimate, which should not change much if it is updated by the data in our new sample. This is a wonderful way of using existing information and accumulating knowledge.

If we do not have this type of information, we may resort to *weakly informative prior distributions.* These are theoretical probability distributions like the normal or t distribution centered at a possible and not too unlikely value. As prior distributions, they express our belief that the population value of the statistic is most likely to be found in a relatively restricted range. But we do not rule out that the true values are much higher or much lower than expected.

These prior distributions are called weakly informative because they are easily overridden by the sample data if the sample is not very small. As a consequence, the posterior distribution is usually hardly influenced by the choice of a prior distribution. A plot of a weakly informative prior distribution and the posterior distribution should show that the two are quite different.

It is relatively safe to chose a weakly informative prior distribution. Still, it is wise to check that the results do not change substantially if you change the prior settings (robustness check). For guidelines, see Depaoli and van de Schoot (2017)], who recommend to compare with a model estimated from *uninformative prior distributions*, which more or less give equal prior probabilities to all possible values of the statistic in the population.

In addition, a Bayesian approach requires another type of check. The posterior distribution is usually estimated with an iterative (MCMC) algorithm. This algorithm requires many steps (iterations) to arrive at a stable level of the parameter estimate. This phase is called the burn-in or warm-up phase. The algorithm is said to have converged when the estimates stabilize within a fixed range. When convergence has been achieved, an additional set of steps is required to find and count the values that the parameter estimate can take.

The estimates from the steps in this second phase constitute the posterior distribution of the parameter, which is the analysis result. Posterior distributions can only be trusted if the algorithm has converged. Convergence can be checked visually or numerically; examples are provided in Section 3. If these checks give rise to doubts about convergence, the posterior distributions should not be interpreted. The model may have problems, for example, it can be too complicated to be estimated from the available data, or we simply need more step (iterations) to achieve convergence.

Both phases usually require thousands of steps, so estimation may take quite some computer time for complicated latent variable or multilevel models. For complicated models, a computer with ample memory and a fast processor comes in handy. Cloud computing is another option. For common statistical models such as analysis of variance or linear regression, however, ordinary computers produce results quickly.

Much work has be done and is still being done to develop safe weakly informative priors for different types of models. But Bayesian estimation is more complicated than approximating a sampling distribution with a theoretical probability distribution. Especially for complicated models, things may go wrong. For an extensive checklist, see *Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist* (Depaoli & van de Schoot, 2017). The upside: Complicated models can be estimated.

2.3 Estimation of complicated models

As stated in the preceding section, the prior distribution brings our prior knowledge to bear in the estimation process. This knowledge can be quite superficial, for example, we just know that very large effects are impossible or highly unlikely, so we specify a weakly informative prior distribution. Why would we want to take quite superficial prior knowledge as our starting point?

A very practical reason for including this knowledge is that it helps the estimation algorithm. The prior information nudges the estimation algorithm in a plausible direction—even if it is not necessarily the right direction—so estimation will converge more easily and more quickly. For relatively simple models, this nudge hardly makes a difference, but complicated models, such as multilevel models with random intercepts and random slopes, can sometimes only be estimated with the help of these nudges.

2.4 Precautions against overfitting the sample data

A focus on statistical significance is at the expense of attention to model fit: How well does the model fit the current data? More importantly, how well would the model fit new data? Would the results replicate?

Traditional frequentist statistics gives us the point estimate that fits the current sample best. For example, we get the value of the regression coefficient that best predicts the dependent variable scores from the independent variable cores in our sample. Another sample will most probably be different, so our current regression coefficient estimate will not be the best fitting estimate for the new sample. It is overfitted to our current sample.

The problem of overfitting a model to the data increases if the model becomes more complex. Every additional predictor adds to overfitting the model to the sample. The more complicated the model, the higher the risk of overfitting the model to the sample at hand. More complicated models, however, are more realistic because empirical reality tends to be complex. We end up with a conundrum, which is known as the bias-variance trade-off (James, Witten, Hastie, & Tibshirani, 2017: 34-5): We value realistic, hence complex models (little bias) but we also value replication, which is compromised by overfitting the model to the current sample (variance).

How can we identify and correct for overfitting? It is important to realize that p values are not related to the probability that results will replicate (e.g., see R. Carver, 1978: 385-6). Adjusted R^2 for a regression model downsizes the proportion of variance predicted (“explained”) by the model in proportion to the number of predictors used, but the estimated regression coefficients are still overfitted to the sample data.

We need another way to tell us the probability that our results will replicate. The best option is to actually replicate our study. If sufficient data are at hand, a split-half approach can be used, which fits a model to the first half of the data (training set) and then tests it against the second half of the data by calculating how well the fitted model predicts the dependent variable in the test set (validation set). With insufficient data for a split-half approach or if we do not want to ignore half of the information when fitting and validating the model, resampling techniques can be used. These techniques repeatedly sample from the original sample and fit the model to each new sample (cross-validation), for instance, leave one randomly chosen observation out or leave out a randomly selected subset of observations (k-fold cross-validation) (Efron & Hastie, 2016: 213ff).

A Bayesian approach estimates a probability distribution (posterior) for a population value, not the single value produced in a frequentist approach. If we use the posterior distribution to predict outcomes in a new or re-sampled data set, the predictions will generally be better than using the point estimate obtained in a frequentist approach. There are additional techniques for reducing overfitting in a Bayesian approach, some of which are briefly discussed in the next section.

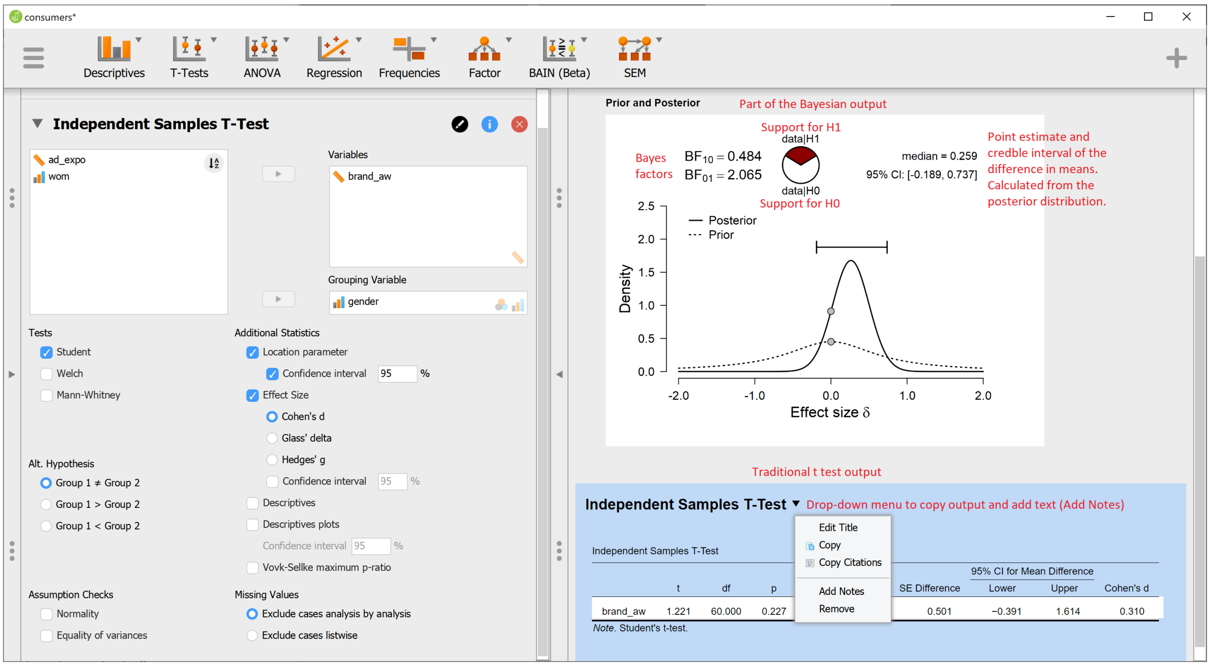
3 How Can I Use A Bayesian Approach?

Statistical software for Bayesian analysis is increasingly available. For example, Bayesian alternatives to many types of statistical models are available as R packages; consult the task view Bayesian or search for "Bayes" in other task views.

Here, I focus on two software packages for Bayesian analysis: a menu-driven software package (*JASP*) with a work-flow like SPSS for the most common statistical models and a command-driven R package (`rstanarm`), which is very good at estimating complicated models, such as non-linear multilevel models. The two software packages represent two different Bayesian approaches: a focus on null hypothesis testing versus a focus on parameter estimation (Kruschke, 2011).

3.1 JASP

JASP is free statistical software developed and maintained at the University of Amsterdam by Eric-Jan Wagenmakers. It offers an SPSS-like interface featuring both traditional frequentist analyses and Bayesian analyses. In contrast to SPSS, the output is very much like a real document, which can be edited, so the user can add comments, interpretations, and conclusions. If you open a data set, the last saved output document is also opened. Results tables are formatted in APA style and they can be copied and pasted into Word or LaTeX documents. Figures can be copied and exported and, finally, there are provisions for interacting with the Open Science Framework. It takes very little time to learn using this interface if you are used to SPSS.



Independent Samples T-Test

ad_expo
wom

Variables: brand_aw

Grouping Variable: gender

Tests
- ☑ Student
- ☐ Welch
- ☐ Mann-Whitney

Alt. Hypothesis
- ● Group 1 ≠ Group 2
- ○ Group 1 > Group 2
- ○ Group 1 < Group 2

Assumption Checks
- ☐ Normality
- ☐ Equality of variances

Additional Statistics
- ☑ Location parameter
  - ☑ Confidence interval 95 %
- ☑ Effect Size
  - ● Cohen's d
  - ○ Glass' delta
  - ○ Hedges' g
  - ☐ Confidence interval 95 %
- ☐ Descriptives
- ☐ Descriptives plots
  - Confidence interval 95 %
- ☐ Vovk-Sellke maximum p-ratio

Missing Values
- ● Exclude cases analysis by analysis
- ○ Exclude cases listwise

Part of the Bayesian output

Bayes factors $BF_{10} = 0.484$ $BF_{01} = 2.065$

Support for H1

Support for H0

median = 0.259
95% CI: [-0.189, 0.737]

Point estimate and credble interval of the difference in means. Calculated from the posterior distribution.

Traditional t test output

Independent Samples T-Test ▾ Drop-down menu to copy output and add text (Add Notes)

Independent Samples T-Test

	t	df	p	SE Difference	95% CI for Mean Difference		Cohen's d
					Lower	Upper	
brand_aw	1.221	60.000	0.227	0.501	−0.391	1.614	0.310

Note. Student's t-test.

Figure 2. Annotated (red) screen shot of the JASP interface after executing a Bayesian and a traditional independent-samples t test.

Traditional (frequentist) estimation is available for all standard tests (t tests, ANOVA, regression, chi-squared tests) as well as more advanced tests, namely, MANOVA, repeated measurements ANOVA, exploratory factor analysis, confirmatory factor analysis for one factor, SEM (provided that you can specify the model in lavaan syntax), meta analysis, and basic network analysis. Note that the SEM menu has a separate entry for parallel mediation models, which does not require writing syntax. It has an option to draw the mediation model, although this figure does not include the estimate of the indirect effect.

At the time of writing this text (November 2019), a Bayesian approach is mainly available for comparing group averages, for example, using data from experiments. JASP offers Bayesian analysis for t tests on one or two means, ANOVA, Repeated measures ANOVA, and ANCOVA. In addition, Bayesian estimation is available for correlations and for linear regression but not for regression models with dichotomous or categorical predictors (although the beta-version *BAIN* module can compare models with categorical predictors). Finally, binomial and multinomial tests, contingency tables, and log-linear models can be executed in a Bayesian way.

JASP's Bayesian approach focuses on *Bayes factors*. A Bayes factor quantifies the support in the sample for one model in comparison to another model. The more a model is supported rather than the other model, the more evidence we have for the former model. Model comparison with Bayse factors takes into account the model complexity as a guard against overfitting the model to the sample data (Kass & Raftery, 1995). Rules of thumb have been developed for interpreting Bayes factor values as strength of the evidence in favor of a model. Check out the Wikipedia page on the Bayes factor for rules of thumb.

One of the two models that are compared is usually the model of no effect representing the null hypothesis. Thus, Bayes factors offer an alternative to traditional null hypothesis significance tests. A Bayes factor gives us a value expressing how much better the model allowing for an effect (H_1) is supported than the model of the null hypothesis (indicated by BF_{10} in JASP output). At the same time, however, we can see how much better the null hypothesis is supported by the data (BF_{01} in JASP output). In Figure 2, the null hypothesis of equal population averages is more strongly supported by the data than the alternative hypothesis. In contrast, a null hypothesis significance test only tells us whether or not we should reject the null hypothesis. It doesn't provide us with evidence to accept it.

Bayes factors for a comparison between models representing a null and alternative hypothesis depend on the prior distribution that is used (Kruschke, 2011). The choice of prior can make a world of difference here. Therefore, it is strongly recommended to execute a sensitivity test: How does the choice of prior distribution affect the Bayes factor? This type of test is available in JASP as the *Bayes factor robustness test* option.

3.2 R `rstanarm` package

The other Bayesian approach focuses on parameter estimation, modelling, and prediction. Here, we are usually dealing with more complicated regression models: non-linear and multilevel models. Estimation of these models is not trivial, requiring attention and effort. Estimation results and problems help us to reflect on the nature of the phenomena that we are trying to model. Null hypothesis testing plays, if any, a minor role in this approach. The book *Statistical Rethinking* (McElreath, 2015) offers an accessible introduction to this Bayesian approach and *Bayesian Data Analysis* (Gelman et al., 2013) a more technical introduction.

The focus here is on the R `rstanarm` package for Bayesian applied regression modelling (Gabry & Goodrich, 2019) and some auxiliary packages. This package was developed by Andrew Gelman's team at Columbia University. It combines the power of the stan modelling language with the ease of applying regression models offered by R packages such as `lme4` (Bates, Maechler, Bolker, & Walker, 2019). In other words, you can run a Bayesian regression model with syntax that is not more complicated than running the same regression model in a frequentist approach. But yes, you have to work with R.

Let us illustrate how to apply Bayesian regression models using `rstanarm`. Consult (Muth, Oravecz, & Gabry, 2018) for a more detailed step-by-step tutorial. Our example data concern students in a US dormitory (Madan, Cebrian, Lazer, & Pentland, 2010), who submitted daily health reports including their sadness status (sad versus not sad, variable `sad.depressed`) during a few months. The data set can be downloaded here. We will try to predict their sadness status from the (logarithm of the) number of contacts with sad friends within the dormitory using spatial contacts (variable `prox_24_friend`) and mediated contacts (variable `media_24_friend`) in the 24 hours preceding health report submission. We control for the student's sadness status on the previous day (variable `prev_sad`). Does sadness spread via contacts and communication?

3.2.1 Estimation. Because of the dichotomous nature of the dependent variable, we use logistic regression. We have multiple health reports per student, so we use a multilevel model with health reports nested within students. As a first step, we estimate a variance components model, that is, a model with random (varying) intercepts but without random (varying) slopes and student-level predictors (code not reproduced here). A random (varying) intercept captures a student's propensity for sadness. Some students are just more often sad than other students.

Next, we estimate a model with varying slopes of spatial contact (code not reproduced here). This model allows spatial contact to have different effects on the odds of being sad for different students. Finally, we add two student-level predictors, namely popularity as the (square root of) the number of friendship nominations received within the dorm (variable `friends_received_sqrt`) and the student's study phase (variable `phase_adv`, 1 : advanced study phase, 0: early study phase). The latter student-level variable is used to predict variation in the effect (slopes) of spatial contact with sad friends. The R code for this model is depicted below.

```r
# R code for estimating a multilevel logistisc regression model with random
# intercepts and random slopes with student characteristic 'adv_phase'
# predicting slope variation.

# Load the rstanarm and rstan packages and set stan options.
library(rstanarm)
library(rstan)
#Set options for multicore CPU with excess RAM.
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())

# Load the data (first download it to your R working directory).
reports <- read.csv("sadness.csv")

# Estimate the model.
reports_fit3 <- stan_glmer(formula = sad.depressed ~  #outcome
  prev_sad + log(media_24_friend+1) + log(prox_24_friend+1) + #level-1 predictors
  phase_adv + friends_received_sqrt + #level-2 predictors (for intercept)
  phase_adv * log(prox_24_friend+1) + #level-2 predictor for slope
  (1 + log(prox_24_friend+1) | user_id), #intercept varying across students
  data = reports, #data frame stored in sadness.csv
  family = binomial(link = "logit"), #logistic link
  # prior_intercept = normal(0,10,autoscale=TRUE), #intercept prior
  # prior = normal(0, c(2.5, 2.5, 2.5),autoscale=TRUE), #regression coefficients
  # prior_covariance = decov(regularization = 1, concentration = 1, shape = 1,
  # scale = 1), #covariance prior
  # iter = 2000, #default number of iterations (including warmup iterations)
  # warmup = floor(iter/2), #default number of warmup runs: half of all iterations
  # adapt_delta = 0.95, #default target average acceptance probability
  # max_treedepth = 10, #max steps within an iteration
  seed = 9122) #randomiser seed for reproducibility
```

The `stan_glmer()` function in the `rstanarm` package mimics the `glmer()` function in the `lme4` package, which estimates **g**eneral **l**inear **m**odels with a multilevel structure (**m**ixed **e**ffects). The regression model is specified as a formula with the dependent variable to the left and the predictors to the right of the tilde. Random (varying) intercepts and varying slopes are specified between parentheses using the symbol `|`, which separates the varying parameters (`1` represents the intercept) from the level at which they are allowed to vary (`user_id` is the student ID number). The blog Using R and lme/lmer to fit different two- and three-level longitudinal models (Magnusson, 2015) (Magnusson, 2015) shows how to specify more complicated multilevel models.

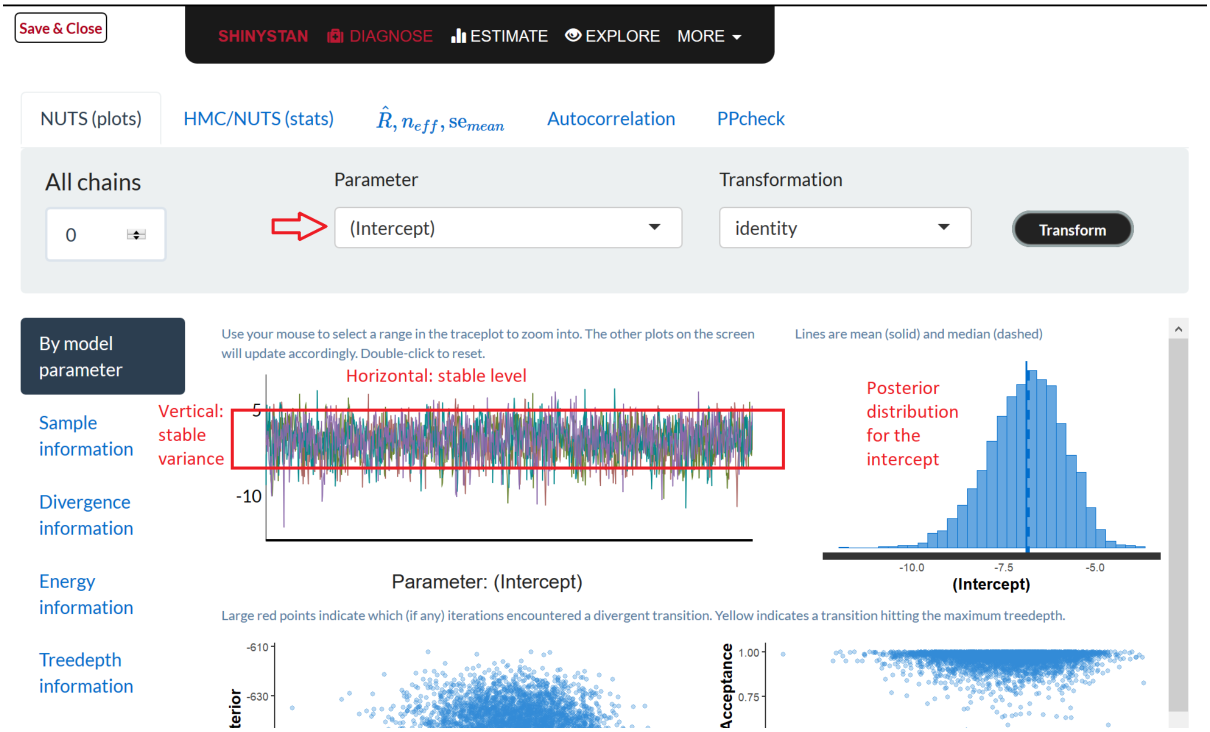
The `family =` argument is followed by arguments that are specific to Bayesian inference. They are commented out to indicate that they represent the default values at the time of writing this paper. The default values for the prior are good starting values, but you would have to change them if you want to compare the results to a model with another prior (robustness check). Note that the default values may change in new releases of the `rstanarm` package due to new insights. For research to be reproducible, it is recommended to specify the priors. You can get the default prior distributions if you apply the `prior_summary()` function to the fitted model.

The `iter =` argument is the most important argument for us. It specifies the number of samples (iterations) that are used to estimate the posterior distributions. On a computer with a multiple core processor and with the option `options(mc.cores = parallel::detectCores())` set, this number of samples will be drawn by each available core. With a processor containing four cores, four times 2,000 samples will be drawn by default. Half of these are used for burning in (option `warmup =`), that is, for converging to a stable range for the parameter estimates. These samples are not used for the posterior distribution. The second half of the samples are used to estimate the parameters (the posterior distributions). If the estimation process does not converge, the main solution is to increase the number of iterations.

3.2.2 Convergence checks. As argued in Section 2.2, it is of paramount importance that convergence has been achieved before we interpret and use the posterior distributions. The R package `shinystan` (Gabry, 2018) offers an interactive user interface for checking the estimation process and inspecting the results.

```
# Package for interactive inspection of stan estimation process and results.
library(shinystan)

# Display convergence checks for the third fitted (rstanarm) model.
# May take some time!
launch_shinystan(reports_fit3)
```



All chains
0

Parameter
(Intercept)

Transformation
identity

Use your mouse to select a range in the traceplot to zoom into. The other plots on the screen will update accordingly. Double-click to reset.

Lines are mean (solid) and median (dashed)

Parameter: (Intercept)

Large red points indicate which (if any) iterations encountered a divergent transition. Yellow indicates a transition hitting the maximum treedepth.

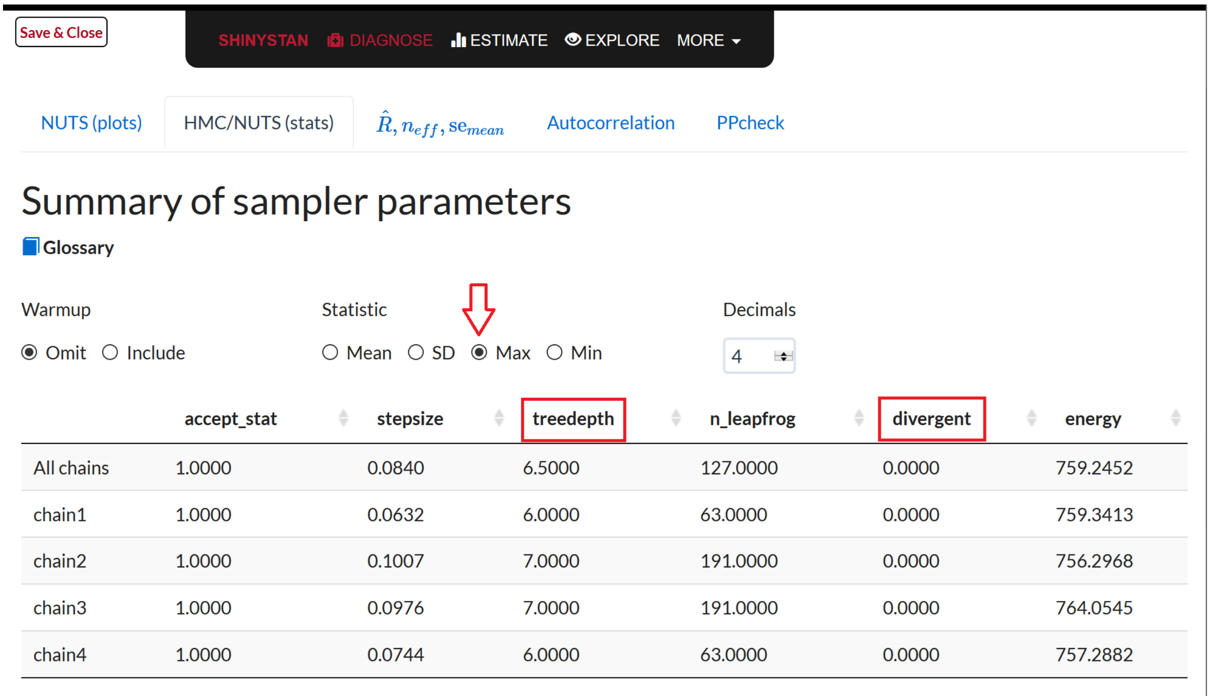
Figure 3. Trace plot and posterior distribution for each parameter. Annotations are in red.

Shinystan opens a window in your web browser with four pages: Diagnose, Estimate, Explore, and More. The Diagnose tab contains convergence checks and other checks on the estimation process. Figure 3 shows the first tab on the Diagnose page, displaying the estimation process for the intercept of the regression model.

The top-left graph is the *trace plot*, showing the intercept estimates (vertical axis) for all subsequent samples (iterations, horizontal axis) for all cores (chains, represented by line color). These are the estimates for the second half of the iterations, so after warm-up/burn-in has completed. The estimation process must have converged in the first half, so the average estimated value for the intercept should not change (increase or decrease) during the second half of the estimation process. This seems to be the case here. In addition, the variation in estimates (vertical differences) should have stabilized. Again, this seems to be the case here.

The top-right figure shows the posterior distribution of the intercept. This is actually the analysis result, which is summarized on the Estimate page. It is shown here because it should display a rather regular, smooth shape. If we have too few iterations, this histogram will look jagged.

These are just checks on the intercept. The drop-down list under *Parameter* allows us to check the other parameters. It is good practice to check all parameters. A deviant pattern may indicate that a parameter could not be estimated in a satisfactory way. Perhaps we just have insufficient data to estimate a particular parameter or this parameter is superfluous. In the present example, the variance of the random slopes of spatial contact (parameter `Sigma[user_id:log(prox_24_friend + 1),log(prox_24_friend + 1)]`) seems to be hard to estimate (if you run the code, you can check this).



Summary of sampler parameters

Warmup
● Omit ○ Include

Statistic
○ Mean ○ SD ● Max ○ Min

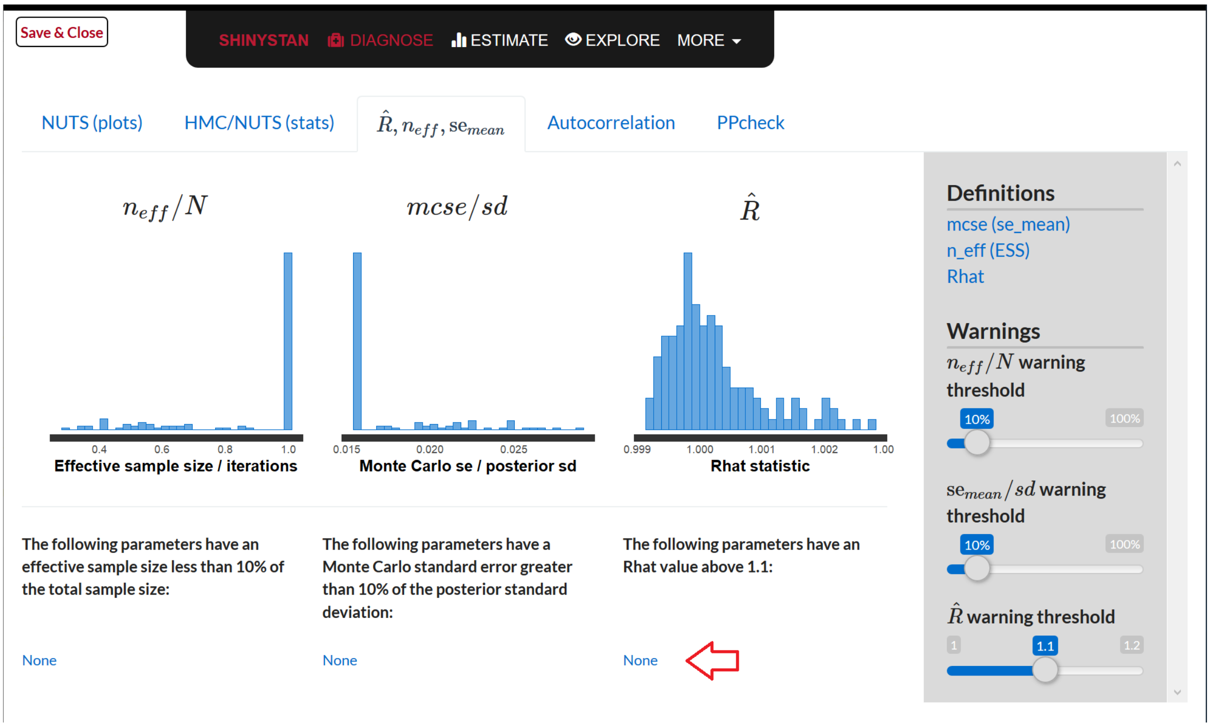
Decimals
4

	accept_stat	stepsize	treedepth	n_leapfrog	divergent	energy
All chains	1.0000	0.0840	6.5000	127.0000	0.0000	759.2452
chain1	1.0000	0.0632	6.0000	63.0000	0.0000	759.3413
chain2	1.0000	0.1007	7.0000	191.0000	0.0000	756.2968
chain3	1.0000	0.0976	7.0000	191.0000	0.0000	764.0545
chain4	1.0000	0.0744	6.0000	63.0000	0.0000	757.2882

Figure 4. Summary diagnostics for the sampler. Annotations are in red.

Figure 4 shows the *HMC/NUTS (stats)* tab on the Diagnose page. It summarizes some characteristics of the sampler (**N**o **U**-**T**urn **S**ampler) used. I recommend to select the *Max* option and pay attention to the *treedepth* and *divergent* columns. Tree depth represents the number of steps within an iteration (sample). Maximum tree depth is 10 by default. If this maximum appears in this summary of sampler parameters, it is wise to re-estimate the model with a higher value for the `max_treedepth =` argument.

The *divergent* column should only contain zeros. Divergence during the estimation process is a serious problem (see Brief Guide to Stan's Warnings). Sometimes this problem can be solved by setting the argument `adapt_delta =` closer to 1.00. This will increase estimation time.



The following parameters have an effective sample size less than 10% of the total sample size:

None

The following parameters have a Monte Carlo standard error greater than 10% of the posterior standard deviation:

None

The following parameters have an Rhat value above 1.1:

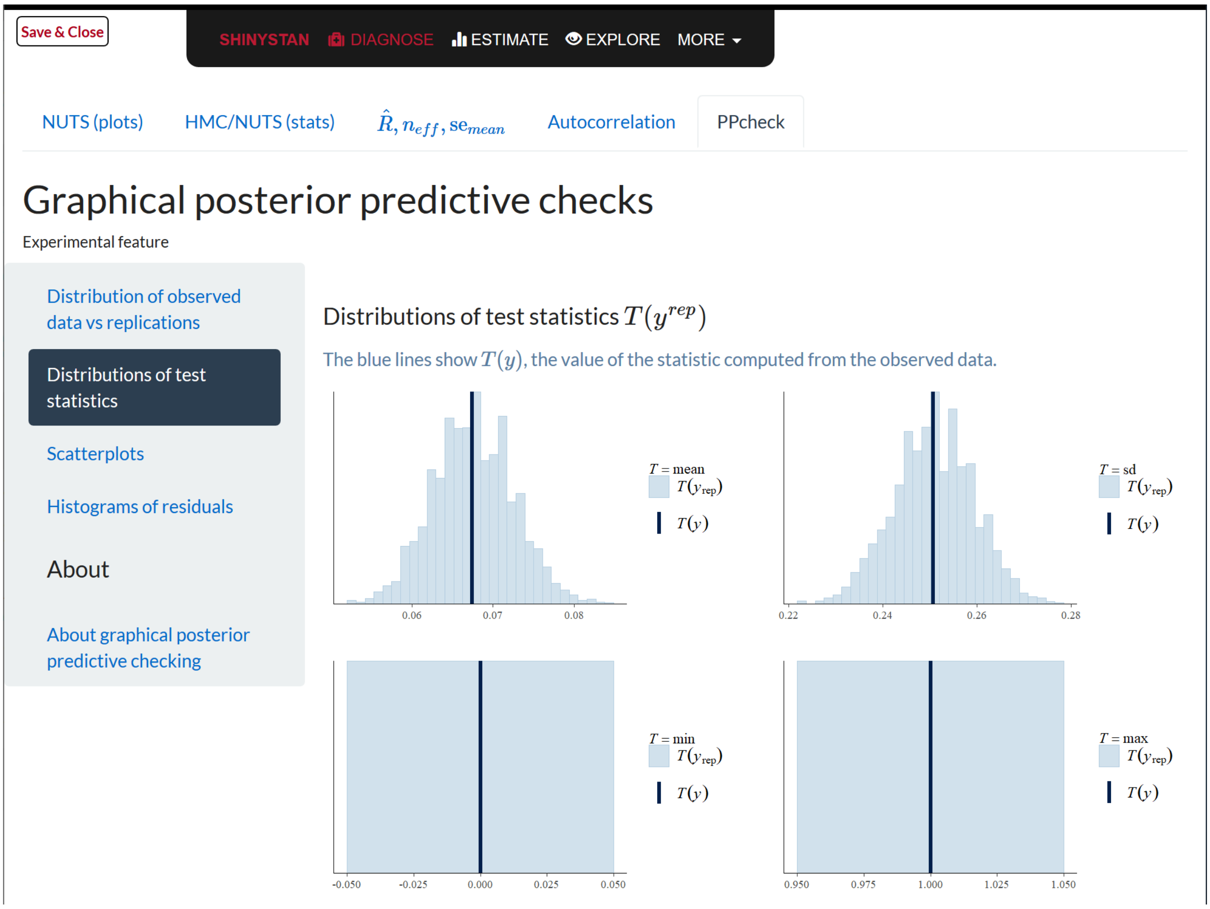
None

Figure 5. Success of the sampler for parameters in terms of the standard error of the mean of the posterior draws, the effective number of samples, and potential scale reduction. Annotations are in red.

It can be more or less difficult for the sampler to find valid new values for a parameter. It can be difficult if iterations are highly correlated. If the sampler has substantial problems in this department, relatively few values are found that can be regarded as truly different samples. As a consequence, the number of effective samples (n_{eff}) is low, which increases the standard error of the mean of the estimates ($mcse$). In this situation, the posterior distribution is based on relatively little information. More iterations are needed to create a reliable posterior distribution.

In addition, the Rhat ($\hat{R}$) statistic indicates if estimation had converged. If convergence has been achieved, the value of this statistic is near 1.0. A value close to 1.0 does not prove that convergence has been achieved, but a value that is not close to 1.0 indicates convergence problems. In the latter case, we had better re-estimate the model with a larger number of iterations or with a larger share of the iterations as warm-up phase.

Parameters for which estimation is problematic according to the rules of thumb (which can be adjusted in the right-hand panel) are listed below the graphs. In Figure 5, it only says "none" below the graphs, so there are no parameters with estimation problems according to these three checks.



Graphical posterior predictive checks

Experimental feature

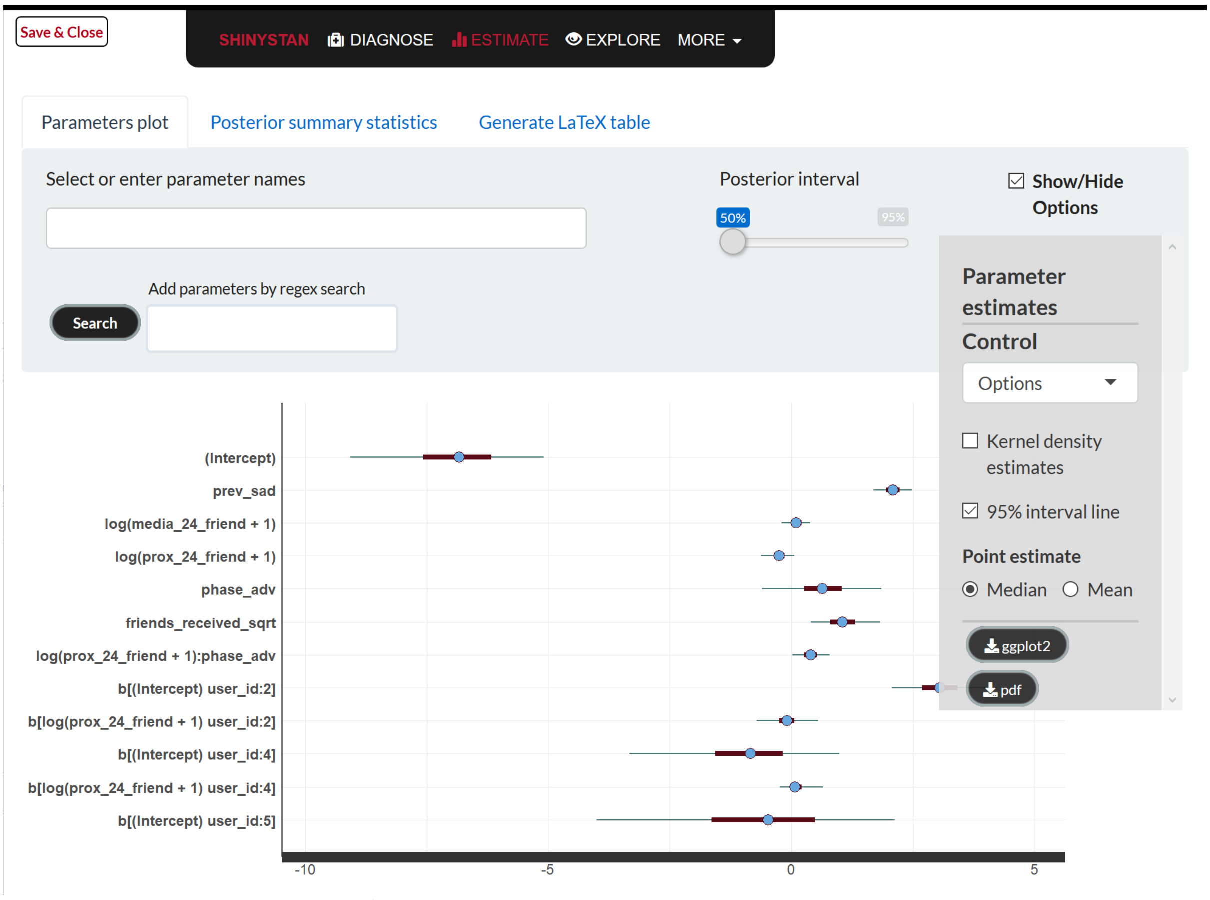
Distributions of test statistics $T(y^{rep})$

The blue lines show $T(y)$, the value of the statistic computed from the observed data.

Figure 6. Posterior predictive checks.

If the estimation process is successful, the estimated model should predict the observed outcome scores well. We can use the (joint) posterior distributions of the model parameters to predict outcome scores. The PPcheck tab of the Diagnose page applies this principle many times and calculates the mean and standard deviation of the outcome scores. The results are represented by the light-blue histograms in Figure 6. In the present example, mean outcome is the proportion or probability of reporting sadness. The dark vertical line represents the proportion of sad reports in the sample. This value should be somewhere near the center of the predicted values (the histogram). This is the case here.

In addition to the mean, the standard deviation, minimum and maximum value of the outcome variable are predicted and compared to the values in the sample. In a logistic regression model, these values are not very interesting. The standard deviation of a dichotomous variable (like sad versus non-sad) is a function of the mean. The minimum and maximum values can only be zero or one (if the outcome variable is coded this way). For other types of regression models, however, these statistics can be of interest.



Posterior interval: 50% (95%)

☑ Show/Hide Options

☐ Kernel density estimates

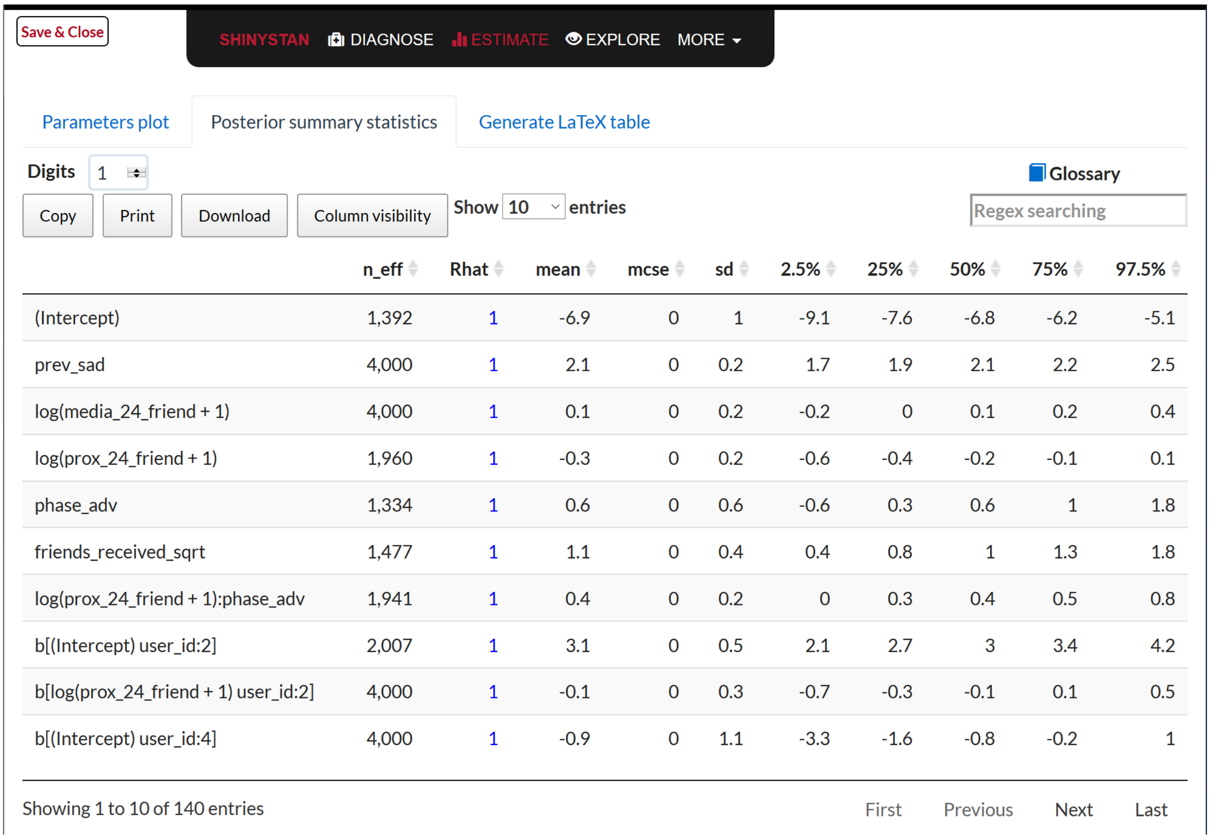
☑ 95% interval line

Point estimate: ● Median ○ Mean

Figure 7. Visual summary of the results.

The Estimate tab presents the estimation results both as a graph and a table. The plot in Figure 7 summarizes the posterior distributions of parameters. By default, the median of the posterior distribution is used as the point estimate (blue dot) and both the 50 and 95 per cent credible intervals are shown as a fat and thin line segment. These settings can be adjusted with *Posterior interval* and *Show/Hide Options*. The latter dialog also allows color adjustment and exporting the graph as a `ggplot2` plot, which can be included and further modified in R, or as a PDF picture.

The plot shows the first (twelve) model parameters but you can select the parameters to be displayed. This does not work well in all browsers. Also note that the web page may become unresponsive if you select many parameters.



Digits 1

Show 10 entries

	n_eff	Rhat	mean	mcse	sd	2.5%	25%	50%	75%	97.5%
(Intercept)	1,392	1	-6.9	0	1	-9.1	-7.6	-6.8	-6.2	-5.1
prev_sad	4,000	1	2.1	0	0.2	1.7	1.9	2.1	2.2	2.5
log(media_24_friend + 1)	4,000	1	0.1	0	0.2	-0.2	0	0.1	0.2	0.4
log(prox_24_friend + 1)	1,960	1	-0.3	0	0.2	-0.6	-0.4	-0.2	-0.1	0.1
phase_adv	1,334	1	0.6	0	0.6	-0.6	0.3	0.6	1	1.8
friends_received_sqrt	1,477	1	1.1	0	0.4	0.4	0.8	1	1.3	1.8
log(prox_24_friend + 1):phase_adv	1,941	1	0.4	0	0.2	0	0.3	0.4	0.5	0.8
b[(Intercept) user_id:2]	2,007	1	3.1	0	0.5	2.1	2.7	3	3.4	4.2
b[log(prox_24_friend + 1) user_id:2]	4,000	1	-0.1	0	0.3	-0.7	-0.3	-0.1	0.1	0.5
b[(Intercept) user_id:4]	4,000	1	-0.9	0	1.1	-3.3	-1.6	-0.8	-0.2	1

Showing 1 to 10 of 140 entries

Figure 8. Tabular summary of the results.

The table of parameter estimates (Fig. 8) can be browsed and adapted more easily than the graphical representation. Select the number of digits, columns, and rows (entries) to be shown. Note that at most 100 rows can be shown. The table as shown can be copied, printed, or downloaded as CSV or PDF. A separate tab on the Estimate page offers the option to construct a LaTeX results table, which can be copied and pasted into a (La)TeX word processor.

3.2.3 Using the results. The main result of this type of Bayesian analysis consists of a (long) list of estimated regression equations. Each iteration (sample) during the estimation phase yields an estimate for each parameter in the model. Table 1 shows a small part of this list.

```
# Display part of the joint posterior distribution.

# Requires the rstanarm, knitr and kableExtra libraries.
library(rstanarm)
library(knitr)
library(kableExtra)

# Extract the parameter estimates from a fitted rstanarm model.
posterior_model3 <- data.frame(reports_fit3, check.names = FALSE)

# Show the first 5 estimates for the (seven) fixed effects.
kable(posterior_model3[1:5, 1:7], digits = 2, booktabs = TRUE,
      caption = "The estimated fixed effects for the first five entries
      in the joint posterior distribution of the model with random slopes
      for spatial proximity and study phase as slope predictor.") %>%
  kable_styling(bootstrap_options = c("basic", "condensed"),
                full_width = T, font_size = 12,
                latex_options = c("scale_down"))
```

Table 1

The estimated fixed effects for the first five entries in the joint posterior distribution of the model with random slopes for spatial proximity and study phase as slope predictor.

(Intercept)	prev_sad	log(media_24_friend + 1)	log(prox_24_friend + 1)	phase_adv	friends_received_sqrt	log(prox_24_friend + 1):phase_adv
-5.63	2.29	0.29	-0.24	0.01	0.68	0.35
-5.94	1.88	0.21	-0.45	-0.09	0.77	0.57
-7.29	1.96	0.17	-0.41	-0.63	1.30	0.51
-6.23	1.88	0.07	-0.33	-0.01	0.90	0.55
-7.18	2.24	-0.19	-0.18	0.42	1.27	0.35

The posterior distribution of a parameter is just the column of this parameter in the list described above. The results presented by `shinystan` (Section 3.2.2) calculate the median (or mean) as point estimate and the 2.5 and 97.5 percentiles as limits of the credible interval. It is easy to calculate the probability that a regression coefficient is larger than a particular value, for example, that the effect of mediated exposure to sad friends is larger than zero:

```
mean(posterior_model3$`log(media_24_friend + 1)` > 0)
```

The list is a joint posterior distribution, meaning that you can also extract probabilities for combinations of parameters. For example, the probability that the effect of mediated exposure is larger (more positive or less negative) than the effect of exposure through spatial proximity is calculated as:

```
mean(posterior_model3$`log(media_24_friend + 1)` >
       posterior_model3$`log(prox_24_friend + 1)`)
```

The result is 0.95.

In a similar way, we can find out more about the effect of spatial proximity to sad friends for students who are in an advanced study phase. Due to the interaction effect between study phase and spatial proximity in the model, the estimated effect of spatial proximity on sadness is the effect for the reference group scoring zero on study phase. This is the effect for students in an early study phase, who are coded as zero on variable `adv_phase`.

The effect of spatial proximity for students in an advanced study phase is the sum of the effect for early-phase students (`log(prox_24_friend + 1)`) and the interaction effect (`log(prox_24_friend + 1)`). Is this effect more likely to be positive or negative? The probability that the sum of the two effects is positive, is 0.86. This is calculated with the statement below:

```
mean(posterior_model3$`log(prox_24_friend + 1)` +
       posterior_model3$`log(prox_24_friend + 1):phase_adv` > 0)
```

Finally, the posterior distribution can be used to visualize the uncertainty about a regression coefficient. Just sample a number of entries from the joint posterior distribution and use the regression coefficien and the intercept to draw a regression line for each entry. Figure 9 shows the median regression line (point estimate) and 100 sampled regression lines for the effect of exposure via spatial proximity for advanced-level students. Note that the plot shows regression lines for an average student, not taking into account variation across students modeled by random intercepts and slopes.

3.2.4 Model comparison and weighting. In the `rstanarm` approach to Bayesian analysis, model comparison serves to optimize predictions on new data sets rather than estab-

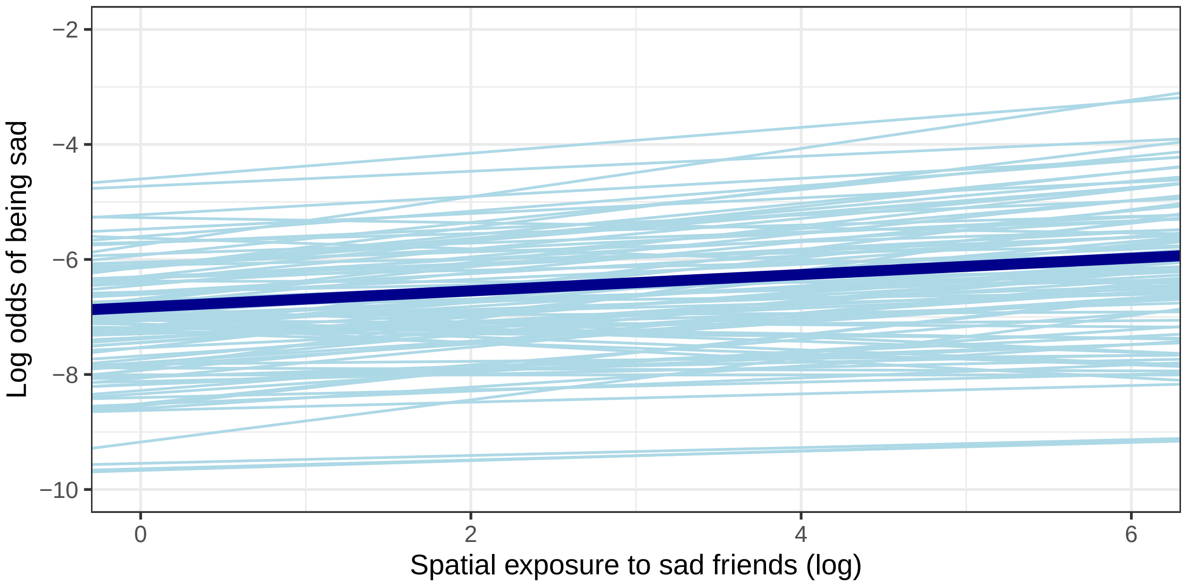


Figure 9. The median regression line (dark blue) and 100 sampled regression lines (light blue) for the effect of exposure via spatial proximity for advanced-level students. Other predictors are set to zero.

lishing the evidence for two competing hypotheses. If the simultaneous use of competing models improves predictive accuracy, why stick to one model and discard the other models (McElreath, 2015: 201)? If we acknowledge that we cannot be certain that one model is the correct model, we had better not stick to one model.

Uncertainty about models is expressed by Akaike weights. Akaike weights compare the information criterion scores of two or more models, assigning weights to models that sum to one over all compared models. If we want to predict new data using a set of models (a *model ensemble*), we weight the predictions by the model's Akaike weight. Akaike model weights can be interpreted as the probability that a model makes the best predictions within the set of compared models (Burnham & Anderson, 2001; Wagenmakers & Farrell, 2004).

This approach to model comparison, then, uses information criteria rather than Bayes factors. A well-known information criterion is Akaike's Information Criterion (AIC) (Hirotuge Akaike, 1973; Hirotugu Akaike, 1974). In Bayesian analysis, however, other information criteria are preferred because they make no distributional assumptions like AIC. Examples are the Widely Applicable Information Criterion or Watanabe–Akaike Information Criterion (WAIC) (Watanabe, 2010) and the Leave-One-Out cross-validation Information Criterion (LOOIC), which is recommended over WAIC (Vehtari, Gelman, & Gabry, 2017).

The R code below illustrates the use of Leave-One-Out cross-validation Information Criterion (LOOIC) and Akaike weights using this information criterion for the three models predicting sadness introduced above.

```
# Model comparison using LOOIC and Akaike weights.

# Assumes that 3 rstanarm models have been fitted and saved as reports_fit1,
# reports_fit2, and reports_fit3.

# Requires the loo library.
library(loo)
```

```
# Create LOO objects for all three models.
# If the models contain problematic observations, the k_threshold argument deals
# with those by re-estimating the model. This can be time consuming!
reports_loo1 <- loo(reports_fit1, k_threshold = 0.7)
reports_loo2 <- loo(reports_fit2, k_threshold = 0.7)
reports_loo3 <- loo(reports_fit3, k_threshold = 0.7)

# Model comparison.
# Tabulate LOOIC results for the three models.
reports_compare <- loo_compare(reports_loo1, reports_loo2, reports_loo3)

# Calculate Akaike model weights.
reports_weights <- loo_model_weights(list(reports_loo1, reports_loo2, reports_loo3))
```

Table 2

Three models arranged in descending order of predictive accuracy according to the approximate leave-one-out cross-validation information criterion (LOOIC).

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
Plus student-level predictors	0.0	0.0	-463.4	27.8	41.4	4.2	926.8	55.6
Plus varying slope spatial contact	-1.6	1.7	-465.1	27.7	43.4	4.3	930.1	55.5
Varying intercepts	-2.6	2.9	-466.0	27.7	37.6	3.8	932.1	55.4

Table 2 shows the results of comparing three models: a model with varying intercepts only, a model adding varying slopes for spatial contacts, and a third model also adding the two student-level predictors study phase and number of friendship nominations received. The models are sorted from low to high LOOIC values (corrected deviance) or, equivalently, from high to low values of the model's *expected log predictive density* (elpd) or expected log-likelihood.

The model with the highest predictive accuracy according to the LOOIC index (elpd_loo), that is, with the lowest corrected deviance (LOOIC), is shown in the top row of the table. In this application, the model with varying effects of spatial contact and student-level predictors performs best (LOOIC = 926.80). This model is compared to the other models and the difference in predictive accuracy (elpd_diff) is shown in the first column. The model with varying slopes but without the student-level predictors performs slightly worse (predictive accuracy difference is -1.60), and the model without varying slopes and student-level predictors performs worst (-2.60).

The sizes of the differences, however, are hard to interpret. If we want to make sense of them, we should evaluate them in relation to the standard errors of the predictive accuracy (se_elpd_loo). In this example, the predictive accuracy differences are small in comparison to the standard error of the predictive accuracy, so our uncertainty about the predictive accuracy is much larger than the differences between the models. Based on this information, no model clearly stands out as a better or worse model.

Table 3 shows the weights for the three models (Yao, Vehtari, Simpson, & Gelman, 2018). The model including student-level predictors clearly has the highest weight (0.82). It has a much higher probability of yielding the best predictions for new observations than the model

Table 3

Model weights for predictive accuracy according to the approximate leave-one-out cross-validation information criterion (LOOIC) with stacking.

Model	Weight
Varying intercepts	0.18
Plus varying slope spatial contact	0.00
Plus student-level predictors	0.82

with only varying intercepts (0.18).
The former probability is more than
four times the latter probability, so this is a substantial difference. Nevertheless, the latter model has a substantial probability of yielding the best predictions. The random slopes may slightly overfit the sample data for which a model without random slopes corrects. So let us not discard the variance components model. Instead, let us combine the two models, weighting the model with their Akaike weights. The model with varying effects but without student-level predictors can be ignored. It is very unlikely to have the best predictions and it does not contribute to the predictions in a model ensemble (model weight is 0.00).

4 Conclusion

Bayesian analyses are flexible and they provide results that are intuitive.

5 References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov & F. Csáki (Eds.), *2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, eptember 2-8, 1971* (pp. 267–281). Budapest: Akadémiai Kiadó.

Akaike, H. (1974). A new look at the statistical model identification. *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control*, *19*(6), 716–723. doi:https://doi.org/10.1109\%2FTAC.1974.1100705

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2019). *Lme4: Linear mixed-effects models using 'eigen' and s4*. Retrieved from https://CRAN.R-project.org/package=lme4

Berry, D. A. (1996). *Statistics: A Bayesian Perspective*. Belmont, CA: Duxbury Press.

Burnham, K. P., & Anderson, D. R. (2001). Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research*, *28*(2), 111–119. doi:10.1071/wr99107

Carver, R. (1978). The Case Against Statistical Significance Testing. *Harvard Educational Review*, *48*(3), 378–399. doi:10.17763/haer.48.3.t490261645281841

Carver, R. P. (1993). The Case against Statistical Significance Testing, Revisited. *The Journal of Experimental Education*, *61*(4), 287–292.

Clark, M. (2014). *Bayesian Basics. A Conceptual Introduction with Application in R and Stan.* Center for Statistical Consultation and Research, University of Michigan.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*, *22*(2), 240–261. doi:10.1037/met0000065

Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science* (1 edition.). New York, NY: Cambridge University Press.

Erickson, T. (2017). Beginning Bayes. *Teaching Statistics*, *39*(1), 30–35. doi:10.1111/test.12121

Gabry, J. (2018). *Shinystan: Interactive visual and numerical diagnostics and posterior analysis for bayesian models*. Retrieved from https://CRAN.R-project.org/package=shinystan

Gabry, J., & Goodrich, B. (2019). *Rstanarm: Bayesian applied regression modeling via stan*. Retrieved from https://CRAN.R-project.org/package=rstanarm

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. CRC Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning: With Applications in R* (1st ed. 2013, Corr. 7th printing 2017 edition.). New York: Springer.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical*

Association, *90*(430), 773–795. doi:10.2307/2291091

Kruschke, J. K. (2011). Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. *Perspectives on Psychological Science*, *6*(3), 299–312. doi:10.1177/1745691611406925

Kurt, W. (2019). *Bayesian Statistics the Fun Way*. San Francisco: No Starch Press.

Lambert, B. (2018). *A Students Guide to Bayesian Statistics* (1 edition.). Los Angeles: SAGE Publications Ltd.

Lee, P. M. (2012). *Bayesian Statistics: An Introduction, 4th Edition: An Introduction, 4th Edition* (4th edition.). Chichester, West Sussex ; Hoboken, N.J: Wiley.

Levine, T. R. (2013). Quantitative Communication Research: Review, Trends, and Critique. *Review of Communication Research*, *1*, 69–84. doi:10.12840

Levine, T. R., Weber, R., Hullett, C., Park, H. S., & Lindsey, L. L. M. (2008). A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research. *Human Communication Research*, *34*(2), 171–187. doi:10.1111/j.1468-2958.2008.00317.x

Madan, A., Cebrian, M., Lazer, D., & Pentland, A. (2010). Social sensing for epidemiological behavior change. In J. E. Bardram, M. Langheinrich, K. N. Truong, & P. Nixon (Eds.), *Proceedings of the 12th ACM international conference on Ubiquitous computing* (pp. 291–300). ACM.

Magnusson, K. (2015, April). Using R and lme/lmer to fit different two- and three-level longitudinal models - R Psychologist. *R Psychologist*.

McElreath, R. (2015). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton: CRC Press.

Morrison, D. E., & Henkel, R. E. (1970). *The Significance Test Controversy: A Reader* (1st ed.). New York: Routledge.

Muth, C., Oravecz, Z., & Gabry, J. (2018). User-friendly Bayesian regression modeling: A tutorial with rstanarm and shinystan. *The Quantitative Methods for Psychology*, *14*(2), 99–119. doi:10.20982/tqmp.14.2.p099

Tyler, R. W. (1931). What Is Statistical Significance? *Educational Research Bulletin*, *10*(5), 115–142.

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. doi:10.1007/s11222-016-9696-4

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. doi:10.3758/BF03206482

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

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using Stacking to Average Bayesian Predictive Distributions. *Bayesian Analysis*. doi:10.1214/17-BA1091

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance : How the standard error costs us jobs, justice, and lives.* Ann Arbor, MI: University of Michigan Press.