Meaningful results for meaningful hypotheses: A tutorial on hypothesis testing with Bayes factors using ROPEs

Timo B. Roettger[1] and Michael Franke[2]

[1]Department of Linguistics & Scandinavian Studies, University of Oslo

[2]Department of Linguistics, University of Tübingen

Abstract

Recent times have seen an increase of interest in Bayesian inference across the behavioral sciences. However, the process of testing hypotheses is often conceptually challenging or computationally costly. This tutorial provides an accessible, non-technical introduction to a technique that is both conceptually easy to understand and computationally cheap, and that also covers many common scenarios in the experimental sciences: Quantifying the relative evidence for a pair of interval-based hypotheses using Bayes factors through the Savage Dickey approximation.

Keywords: statistics, Bayes, Bayes factor, Savage Dickey, hypothesis testing, ROPE

1 Introduction

One of the most common scenarios in experimental research is to measure one or more outcomes (dependent variables) in an experiment with one or more predictors (independent variables). Usually, if we are testing hypotheses, we want to statistically test whether the predictors affect the measured variables. Traditionally, these statistical tests have been done within the *null hypothesis significance testing* (NHST) framework.[1] While extensions of the NHST framework exist, in its basic form, NHST only allows us to *reject* the null hypothesis, but not to provide evidence in favor of it. Over the last decade or so, however, there has been rising interest in statistical approaches within an alternative inferential framework using *Bayesian inference*. One of the main reasons for this rising interest is that Bayesian inference allows to not only quantify evidence *against* an assumed null hypothesis, but also to yield quantitative evidence *in favor of* the null hypothesis.

Unfortunately, there are several approaches to hypothesis testing within the Bayesian framework, and many of them are either conceptually challenging, computationally (too) costly, or both. For example, there are good conceptual arguments that support Bayesian hypothesis testing through *model comparison* using Bayes factors (Kass & Raftery, 1995; Morey et al., 2016; Vandekerckhove et al., 2015), but the computation

[1] Not strictly necessary for this tutorial but, in case you need a reminder, the logic of NHST goes something like this: we assume —for the sake of argument— that a null hypothesis is correct, i.e., that there is no effect of a relevant predictor. We then ask ourselves how likely different observations would be based on that assumption, and use this so-called *sampling distribution* to quantify how surprising the observed data is under the assumed null hypothesis. If the observed data are very unlikely, we *reject* the null hypothesis and conclude that the predictor affects the dependent variable.

Timo B. Roettger https://orcid.org/0000-0003-1400-2739

The authors have no conflict of interest to declare.

Correspondence concerning this article should be addressed to Timo B. Roettger, Email: timo.roettger@iln.uio.no

of Bayes factors can be quite costly, especially for complex models. Yet, for some of the most common use cases, there are some simple and computationally cheap approaches to Bayesian hypothesis testing with Bayes factors that are easy to understand and implement. One such method is the *Savage-Dickey density ratio* (Dickey & Lientz, 1970; Wagenmakers et al., 2010). While prior work has prominently documented how to use this method for the case of point-valued null-hypotheses (Wagenmakers et al., 2010), this method can be hard to estimate reliably with posterior sampling, which is the most prevalent method for approximating Bayesian computation at the moment. This tutorial therefore focuses on the use of the Savage-Dickey density ratio for testing hypotheses that are grounded in *regions of practical equivalence* (ROPEs) (Kruschke, 2018) using the so-called *encompassing priors* approach (Klugkist et al., 2005; Klugkist & Hoijtink, 2007; Oh, 2014; Wetzels et al., 2010), which is both conceptually more meaningful and computationally more robust than point-valued hypothesis testing. While this method does not seem to be widely known, it is conceptually simple and easy to apply, e.g., through implementation in the package `bayesfactorR`. This tutorial therefore provides an accessible, non-technical introduction to this method of Bayesian hypothesis testing, which is easy to understand, computationally cheap and widely applicable.

2 Motivation and intended audience

This tutorial provides a very basic introduction to hypothesis testing with Savage-Dickey density ratios using R (R Core Team, 2025). We wrote this tutorial with a particular reader in mind. If you have used R before and if you have a basic understanding of linear regression and Bayesian inference, this tutorial is for you. We will remain mostly conceptual to provide you with an accessible tool to approach hypothesis testing within Bayesian inference. The form of hypothesis testing that we would like to introduce to you is, however, different from the traditional null hypothesis significance testing in that it requires more thinking about the quantitative nature of your data. This is not a bug but, at least for us, a feature that will allow you to understand both your data and what you can learn from them better.

If you don't have any experience with regression modeling, you will probably still be able to follow, but you might also want to consider doing a crash course. To bring you up to speed, we recommend the excellent tutorial by Bodo Winter (2013) on mixed effects regression in a non-Bayesian paradigm. To then make the transition to Bayesian versions of these regression models, we shamelessly suggest our own tutorial on "Bayesian Regression for Factorial Designs" (Franke & Roettger, 2019) which uses the same example data set as Winter's tutorial. In a sense, the present tutorial on hypothesis testing could be considered the long-awaited sequel (!?) of the series started by Winter.

To actively follow this tutorial, you find all code and data in this repository: https://github.com/michael-franke/bayes_factors_intervals_tutorial. You should have R (R Core Team, 2025) installed on your computer (https://www.r-project.org). Unless you already have a favorite editor for tinkering with R scripts, we recommend to try out RStudio (https://www.rstudio.com). You will also need some packages, which you can import with the following code:

```
# package for Bayesian regression modeling
library(brms)

# package for BF calculation and plotting
library(bayestestR)
```

3 Data, research questions & hypotheses

In this section, we introduce the data set that we will use throughout this tutorial, the research question that we want to address, and how to formulate meaningful hypotheses in a way that allows us to test them with Bayes factors using so-called Regions of Practical Equivalence (ROPEs), to be introduced below.

3.1 The data set: Voice pitch in Korean across social contexts

This tutorial looks at a data set relevant for investigating whether voice pitch differs across social contexts in Korean. Korean is a language in which the social distance between speakers plays a central role to the way utterances are pronounced. The way Korean speakers talk depends for example on whether they are in a formal context (e.g. during a job interview) or an informal context (e.g. chatting with a friend about the holidays). To investigate this phenomenon, our data set contains pitch measurements of utterances in different social contexts (Winter & Grawunder, 2012). To load and inspect the data into your R environment, run the following code:

```r
polite <-

  # load data set & cast strings to factors
  read_csv("https://tinyurl.com/yu5zskbp", col_types = "ffffd")

head(polite)
```

```
# A tibble: 6 x 5
  subject gender sentence context  pitch
  <fct>   <fct>  <fct>    <fct>    <dbl>
1 F1      female S1       formal    215.
2 F1      female S1       informal  211.
3 F1      female S2       formal    285.
4 F1      female S2       informal  266.
5 F1      female S3       formal    211.
6 F1      female S3       informal  286.
```

This data set contains anonymous identifiers for 16 individual speakers stored in the variable `subject`. Voice pitch is dependent on speakers' `gender`, which we need to take into account as well. Speakers produced 7 different `sentences`, and the experiment manipulated whether the sentences were produced in a `formal` or an `informal` social `context`. Crucially, each row contains a measurement of pitch in Hz stored in the variable `pitch`.

For most analyses of behavioral experiments, researchers are interested in whether an outcome variable is meaningfully affected by at least one manipulated variable and if so how the outcome variable is affected by it. In this case, Winter and Grawunder (2012) wanted to test whether voice pitch is meaningfully affected by the social context of the utterance.

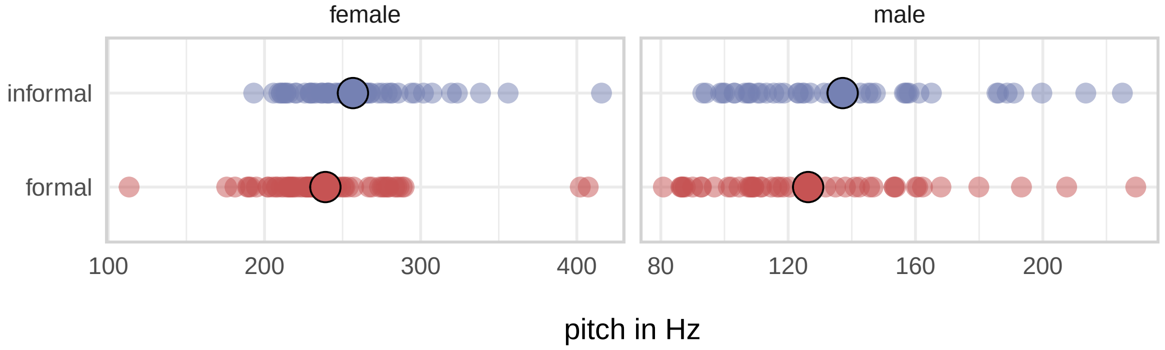
As a first step, we can explore this question visually. Figure 1 displays the pitch values for all utterances in the dataset across contexts (semi-transparent points). The solid points indicate the average pitch values across all sentences and speakers. Looking at the plot, we can see that voice pitch from utterances in formal contexts are on average slightly lower than those in informal contexts: The red distribution is slightly shifted to the left of the blue distribution by 10-ish Hz for male speakers and 30-ish Hz for female speakers. In other words, speakers tend to slightly lower their voice pitch when speaking in a formal context. But there is also a lot of overlap between the two contexts. Now as Bayesians, we would like to translate the data into an expression of evidence: Does the data provide evidence for our research hypotheses?

3.2 A Bayesian regression model to address our research question

Let us build a Bayesian linear model to approach an answer to this question. Using the package `brms` (Bürkner, 2018), our first step is to specify the model formula and check which priors need to be specified:

Figure 1

Empirical distribution of speakers' pitch values across contexts and sex



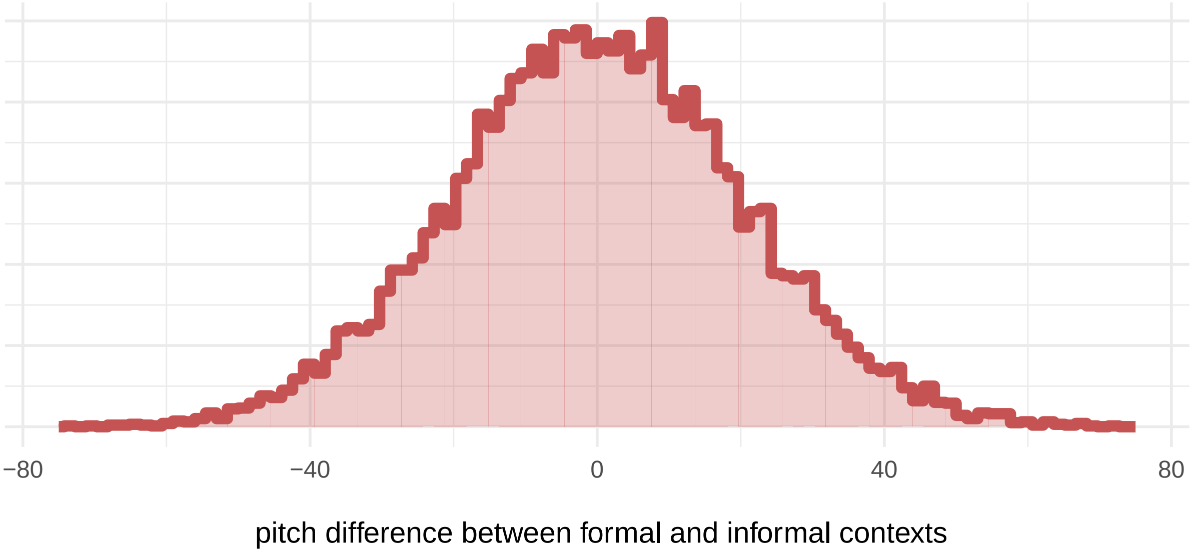
```
# contrast code predictors
contrasts(polite$context) <- c(-0.5,0.5)
contrasts(polite$gender) <- c(-0.5,0.5)

# define linear model formula
# predict pitch by context and gender
# and allow for context to vary between subjects and sentences
formula <- bf(pitch ~ context +
                      gender +
                      (1 + context | subject) +
                      (1 + context | sentence))

# get information about priors that are set per default for this model
# NB: no prior for `context1` is set per default (!)
as_tibble(get_prior(formula, polite)) |>
  select(class, prior) |> filter(prior != "")
```

```
# A tibble: 4 x 2
  class     prior
  <chr>     <chr>
1 cor       lkj(1)
2 Intercept student_t(3, 211.6, 80.2)
3 sd        student_t(3, 0, 80.2)
4 sigma     student_t(3, 0, 80.2)
```

The default priors that brms picks for the Intercept and the variance parameters are mostly reasonable as they are derived from the data. They are weakly informative and symmetrical. However the default choice for our critical parameter context1 is to not specify a prior at all, i.e., to assume a flat (improper) prior, so that it is not even listed in the table above. Yet, there are good arguments why it should also receive a weakly informative prior (Gelman et al., 2017), i.e., the prior assumption about the difference between informal and formal contexts should be that we don't know, but our best guess is that it is zero in expectation and equally likely to be more or less than zero. So we specify a normal distribution centered on zero for this parameter (and we do the same for gender). Since we used contrast coding, the prior for context1 reflects our prior

Figure 2

Prior probability of the difference in pitch between contexts, i.e., before seeing the data



belief about the difference between formal and informal contexts. Note that we use default priors for the other parameters for convenience here, but you should always critically reflect on all of your priors.

```
# define a weakly informative prior for context and gender
priors <- c(prior(normal(0, 20), coef = "context1"),
            prior(normal(0, 100), coef = "gender1"))
```

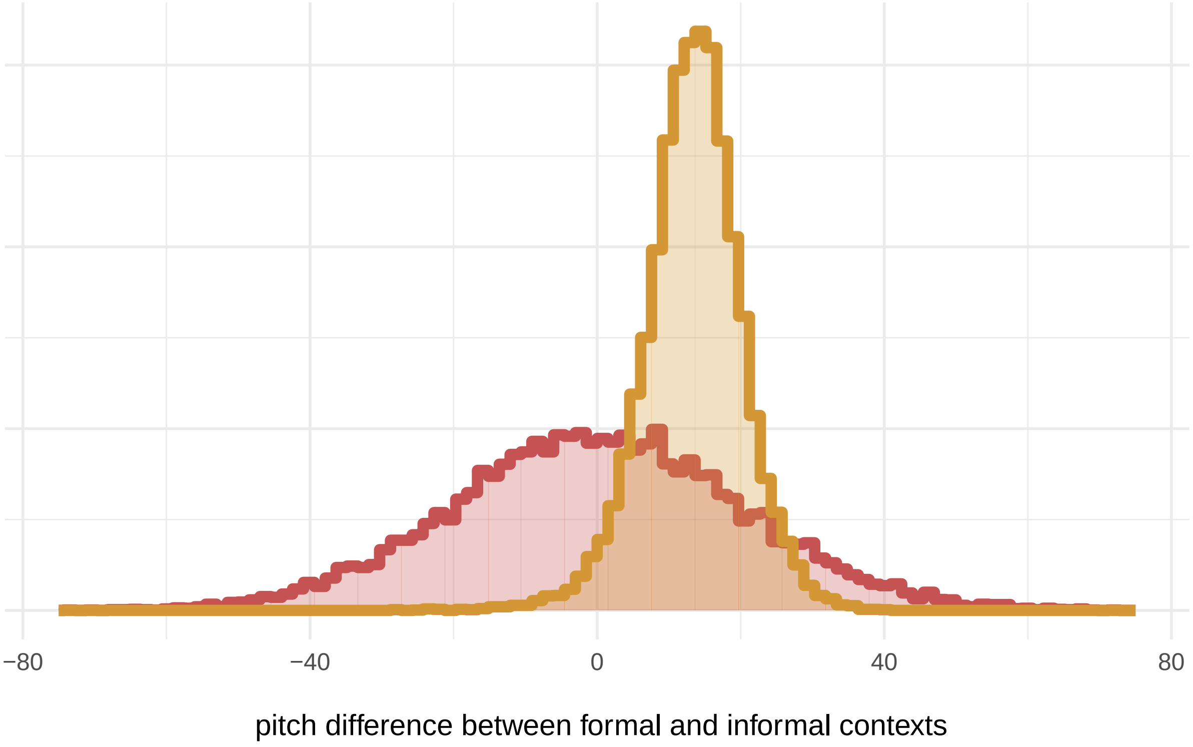
Now we inspect how the parameter distribution looks like *before* having seen the data, based on the priors only. This is a useful exercise to make sure that the priors result in reasonable quantitative assumptions. We usually do it for all parameters, but here we will focus only on the critical parameter `context1`, i.e., the difference between formal and informal contexts. Let us also have a look at the predictions for the prior-only model.

```
# run the prior-only model
fit_prior <- brm(formula, prior = priors, data = polite,
           # sample prior only
           sample_prior = "only",
           # common sampling specifications
           seed = 1234, iter = 8000)
```

Looking at the distribution in Figure 2, the prior for the effect of context on pitch seems sensible. The most plausible value is zero. Values that are smaller or larger than zero become less plausible the further they are away from zero and values being smaller or larger than zero are equally likely. Good. Before we have seen the data, our model is somewhat pessimistic about the effect of context on pitch. Now we can run the full model that integrates the likelihood (our data) with the priors and visualize the posteriors for the critical parameter.

Figure 3

Posterior probability of the effect of context on pitch, i.e., after seeing the data



```
# run the model
fit <- brm(formula, prior = priors, data = polite,
           seed = 1234, iter = 8000)
```

Figure 3 shows the prior (red distribution) and posterior (gold distribution) probability of the effect of context on pitch. The distribution of posterior samples suggests that the majority of plausible values after seeing the data are positive, or in other words, informal contexts elicit larger pitch values. Negative values are not very plausible under the posterior distribution, but also not completely implausible. Compared to our prior probability (red distribution) for which roughly 50% of posteriors are negative, this decrease in plausibility of negative values is quite noteworthy already.

What we have done here should be quite familiar. We care about whether values are positive or negative because we compare our model predictions to a reference point: the single point value zero. But do we really care that much for such point hypotheses? Is zero really that special? We might think so because years of using null hypothesis significance testing has conditioned us to think that way. But we can also go beyond point-values if we see reason to do so.

3.3 Grounding hypotheses in Regions of Practical Equivalence

Above we claimed that we wanted to test “whether pitch is **meaningfully affected** by the social context of the utterance”. We snuck the word “meaningfully” in there for a reason. But what does “meaningful” mean? This is an interesting yet deep questions and (un)fortunately requires some thinking. What a meaningful difference really constitutes depends on the context of the data. So let’s have a closer look at our data.

This tutorial deals with speech data. Speech is, in spoken languages at least, *the* vehicle to transmit linguistic information in order to communicate with each other. Speech is also very complex and very noisy: Not everything that can be measured in the acoustic signal matters for a listener. For example, if something cannot be perceived reliably, it is at least conceivable that it might play little to no role in communication. While the speech sciences have a rich research tradition to estimate what can and what cannot be reliably heard, exact estimation depends on a lot of moving parts. Approximate thresholds of what can be reliably heard are often referred to as *Just Noticeable Differences* (JNDs). This terminology has been argued to be dangerously misleading (Sanford & Halberda, 2023) as it implies a hard and absolute threshold of perceptibility. However, many researchers work with a more lenient and more practical conceptualization of a threshold: The JND is traditionally defined as the point at which the probability that our perceptual system registers a stimulus or a stimulus difference reaches as certain value. By convention, the JND is often defined as the point where listener accuracy exceeds the arbitrary threshold of 75%. When we say "just noticeable difference" in the following we mean that latter probabilistic notion, as we would like to use the idea of regions at which the perceptual system is unlikely to reliably register a stimulus difference as a means of justifying a Region of Practical Equivalence (ROPE). For example, while classic work by Klatt (1973) suggest JNDs ranging from 0.3 to 4 Hz, more modern treatments such as Turner et al. (2019) report on JNDs between 17 and 25 Hz for non-speech stimuli and between 35 and 40 Hz for speech stimuli. While these studies are hard to compare, their effect magnitudes are grounded in a comparable probability threshold (75% accuracy). Thus, they give us at least an idea about the rough order of magnitude for JND values to work with when it comes to speech data like the data set at hand.

Based on these considerations, we could interpret the original hypothesis the following way: If a pitch difference is below the JND, it is not (practically) meaningful. So, instead of testing against a point-valued hypothesis, we can test against a range of parameter values that are equivalent to the null value for practical purposes. In our case, let us begin with a JND value that is somewhat conservative in relation to Klatt (1973) and somewhat liberal in relation to Turner et al. (2019): We assume that pitch values between `-10` and `10` are negligible. Bear with us, we will later revisit this assumption. Said differently, in order to be convinced that a difference in pitch is meaningful, it should be reliably greater than 10 Hz. Such ranges are sometimes called *Regions of Practical Equivalence* (ROPEs), range of equivalence, equivalence margin, smallest effect size of interest, or good-enough belt (see Kruschke, 2018).

```r
# define our ROPE
rope <- c(-10,10)
```

With a ROPE being defined, we can now test our hypothesis "whether pitch is **meaningfully affected** by the social context of the utterance" using Bayes factors.

4 Testing hypotheses using Bayes factors

4.1 What are Bayes factors?

We often consider two hypotheses H_0 and H_1, and want to know which of these is correct. We do so by looking at some observed data D. As Bayesians, the first most obvious thing to look at is how likely each hypothesis is after seeing the data, i.e., something like $P(H_0 \mid D)$ and $P(H_1 \mid D)$. Now, it turns out that these *posterior probabilities of hypotheses* are problematic, because they depend on the prior probabilities of the hypotheses $P(H_0)$ and $P(H_1)$, which are often hard to justify. To see this, imagine that the hypotheses to compare are polarizing issues like contrasting Darwinian evolutionary theory and a dull form of creationism, referred to here as *arbitrary design*. Proponents of either view would have a hard time agreeing on priors for these hypotheses, but may find it much easier to agree on whether a given observation D is more likely under the assumption that one of the two hypotheses is correct, rather than the other. Therefore, Bayes factors are defined as the *likelihood ratio* of the data given each hypothesis:

$$\text{Bayes factor in favor of hypothesis 1 over hypothesis 0} \; : = \; \frac{P(D \mid H_1)}{P(D \mid H_0)}$$

To see how this is a more objective and actually quite intuitive measure of observational evidence in scientific reasoning, consider the case of Darwinian evolution (H_1) versus arbitrary design (H_0) again. Let's assume that the observed data D consist of (i) measurements of the beak shapes of finches on two different islands, and (ii) information about the food sources available on these islands. Let's assume for simplicity that on island A the predominant food source are insects found deep in wood or earth, and that on island B it is insects with hard shells living on the surface. We also observe that finches on island A have long and narrow beaks, while finches on island B have short and thick beaks. (This here is a crude simplification of an actual case that led Darwin to formulate his theory, but bear with us for the sake of the example.) What is a better explanation of data D, evolutionary selection or arbitrary design? To begin with, let's notice that D is *not* ruled out by either hypothesis. But the probability of observing D is higher under Darwinian evolution (H_1) than under arbitrary design (H_0). This is because, before seeing the data, a proponent of evoluationary theory H_1, if asked to make a prediction about beak shapes under the given food sources, would have considered it *more* likely that beaks are adapted to their function of exploiting the dominant food source, so making the actually observed data D more likely *ex ante* than at least some other possible observations. In contrast, a proponent of arbitrary design H_0 would not have had any reason to expect that beak shapes are adapted to food sources, and would like to be able to rationalize *ex post* also any apparent violation of this expectation as the inscrutable way of the arbitrary designer. To the extent that there are many alternative observations which arbitrary design would consider reasonably likely *ex ante* but evoluationary selection would rule out (or deem very unlikely), the probability of the observed data is much higher under Darwinian evolution than under arbitrary design, so that $P(D \mid H_1) > P(D \mid H_0)$, irrespective of what we initially believed is the more plausible hypothesis. This is what corroborates the intuition that the observation D is an argument in favor of H_1 over H_0. This intuition is exactly what the Bayes factor quantifies.

Concretely, a Bayes factor of 1 corresponds to the case of $P(D \mid H_1) = P(D \mid H_0)$, i.e., the data is equally likely under both hypotheses, so the data does not provide any evidence for or against either hypothesis. Any Bayes factor larger than 1 indicates that the data is more likely under H_1 than under H_0, and the larger the Bayes factor, the stronger the evidence in favor of H_1. Conversely, any Bayes factor smaller than 1 indicates that the data is more likely under H_0 than under H_1. Notice that the Bayes factor is symmetric in the sense that a Bayes factor of 3 in favor of H_1 over H_0 corresponds to a Bayes factor of 1/3 in favor of H_0 over H_1. There are various conventions for interpreting the strength of evidence of Bayes factors, such as to consider Bayes factors smaller than 3 as "*anecdotal evidence*"; Bayes factors bigger than 3 as "*moderate evidence*"; and Bayes factors bigger than 10 as "*strong evidence*".

One way to interpret Bayes factors in absolute terms is this: A Bayes factor of n in favor of H_1 over H_0 means that after seeing the data, a rational researcher who initially thought both hypotheses were equally likely would consider H_1 to be n times more likely than H_0 after observing D.

4.2 Bayes factors for statistical models

After motivating Bayes factors in general, let's have a look at the definition of Bayes factors in the context of statistical models in this section. What follows in this section is a bit more technical, so you can skip ahead without missing out too much information for applying these methods.

In the context of statistical models, we can use Bayes factors to compare two statistical models M_0 and M_1 that instantiate two competing hypotheses (or assumptions) H_0 and H_1. A Bayesian statistical model M consists of:

1. a *likelihood function* $P(D \mid \theta, M)$ that specifies how likely the observed data D is given the model M and the model's parameters θ, and
2. a *prior distribution* $P(\theta \mid M)$ that specifies how likely different parameter values are before seeing the data.

The probability of some observed data $P(D \mid M)$ under a model M is then obtained by integrating over all possible parameter values θ:

$$P(D \mid M) = \int P(D \mid \theta, M) \;\; P(\theta \mid M) \, \mathrm{d}\theta$$

This is called the *marginal likelihood* of the data under the model M. We can think of this quantity as obtained from sampling parameter values from the prior and then sampling, for each of the sampled parameter values, a potential data observation. (Notice that this is the *prior predictive data distribution* of the model.)

Putting things together, the resulting definition for Bayes factors in statistical models is:

$$\text{Bayes factor in favor of model 1 over model 0} \;:=\; \frac{P(D \mid M_1)}{P(D \mid M_0)} = \frac{\int P(D \mid \theta, M_1) \;\; P(\theta \mid M_1) d\theta}{\int P(D \mid \theta, M_0) \;\; P(\theta \mid M_0) d\theta}$$

4.3 Bayes factor for point-valued hypotheses (the Savage-Dickey method)

While Bayes factors are a very intuitive and useful measure of evidence, they are often hard to compute. There are various approximation methods, such as bridge sampling (Gronau et al., 2017), which can be used for any arbitrary pair of models, but these can still be computationally costly and sometimes hard to implement. However, for the special case of *nested models*, there is a simple and computationally cheap approximation method called the *Savage-Dickey density ratio* (Dickey & Lientz, 1970; Wagenmakers et al., 2010).

What are nested models? Intuitively speaking, model M_0 is nested in model M_1 if M_0 can be obtained from M_1 by setting one or more parameters to a specific value. (More precisely, by conditioning on a specific value of one or more parameters.) For example, take the regression model M_1 for the Korean speech data we introduced at the beginning of this tutorial. We suggested a normal distribution on the `context` coefficient as a prior. A model M_0 nested under it would be one that is exactly like M_1 except that M_0's prior for the `context` coefficient allows only one value, e.g., that the slope coefficient is equal to zero. That model M_0 would then correspond to the (standard, point-valued) null hypothesis that there is no effect of `context` on `pitch`. This process might sound familiar to people who have generated p-values for linear mixed effects models before. One way to check if a predictor significantly affects a dependent variable is by comparing a full model to a null model. The null model is the full model minus the critical predictor, which is clamped to a specific value, so to speak.

So, suppose that M_0 is nested in M_1 by fixing a critical parameter θ^* to a specific value x. Then, the Savage-Dickey density ratio states that the Bayes factor in favor of M_1 over M_0 can be computed as the ratio of the prior and posterior density of $\theta^* = x$ from M_1's point of view:

$$\text{Bayes factor in favor of model 0 over model 1} \quad = \quad \frac{P(\theta^* = x \mid D, M_1)}{P(\theta^* = x \mid M_1)}$$

Let's unpack this. First of all, this seemingly magical result is actually not that magical, but follows directly from the definition of Bayes factors and Bayes' theorem. Don't worry. We won't bother you with the derivation here. But that means that in practice we do not have to calculate or approximate any integrals at all, but we can simply look at the more complex model M_1 and its prior and posterior parameter distributions, like we routinely do with `brms`, for example. Look at the formula above: we would only need to run one model, M_1, and then look at the prior and posterior density of the critical parameter θ^* at the point value x. The prior us usually easily determined because it is in our hands to specify it. The posterior can by estimated from the samples that are returned by software like `brms` …

… well, at least in principle. One problem here is that estimating $P(\theta^* = x \mid D, M_1)$ from posterior samples is fickle. We can do it with some mathematical methods, but we may need a lot of samples and do some post-processing (e.g., using splines). But one technical wrinkle is that posterior samples are less reliable for estimating densities at specific points, but are usually more reliable for estimating probabilities over wide-enough intervals of values.[2] Moreover, point-valued hypotheses may often not be that interesting or meaningful in practice anyway, as argued above. Fortunately, there is a generalization of the Savage-Dickey density ratio that works for ranges of values, too!

4.4 Bayes factors for interval-based hypotheses (like ROPEs)

The Savage-Dickey density ratio can be generalized to the case where the null hypothesis H_0 is not a point-valued hypothesis, but a hypothesis that the critical parameter θ^* lies in some interval I_0, such as our ROPE from above. There are several different ways to define the alternative hypothesis H_1 in this case, but the most common one is to define it as the complement of H_0, i.e., that θ^* lies in the interval that contains all values that are not in I_0, so that:

$$H_0 = \theta^* \in I_0 \qquad H_1 = \theta^* \notin I_0$$

An efficient way of computing Bayes factors for such a setting is to use the so-called *encompassing priors approach* (Klugkist et al., 2005; Klugkist & Hoijtink, 2007; Oh, 2014; Wetzels et al., 2010). According to this approach, we consider an *encompassing model* M_e that contains both the null and the alternative hypothesis as special cases. Concretely, the encompassing model M_e could be just a regression model like the model we used above for the Korean speech data, with a prior distribution on the critical parameter θ^*, such as the normal distribution on the slope coefficient for `context`. The null model M_0 would then be the nested model that is obtained from M_e by conditioning on $\theta^* \in I_0$, and the alternative model M_1 would be the nested model that is obtained from M_e by conditioning on $\theta^* \notin I_0$. An alternative intuition can be gained by visualizing this principle. In Figure 4, blue parts of the sampled distributions fall within the ROPE representing the null model M_0, grey parts fall outside the ROPE representing the alternative model M_1. We get a null and an alternative model for both the model before observing the data, and after observing the data.

Based on this setup, the Bayes factor in favor of M_0 over M_1 can be computed as the ratio of the posterior and prior odds of θ^* being in I_0 (blue parts of the distributions) versus being in I_1 (grey parts), where I_1 is the complement of I_0:

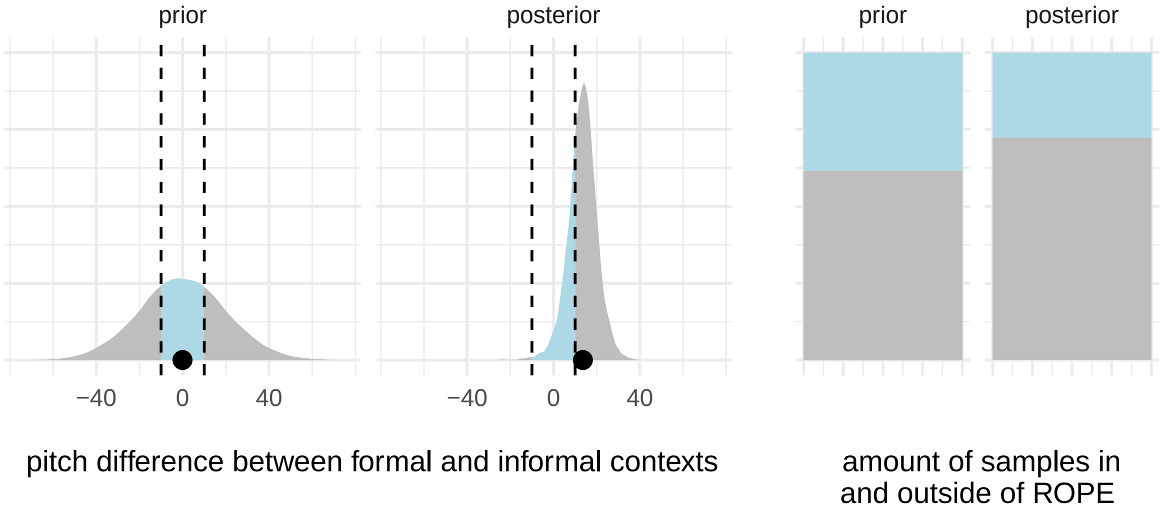
$$BF_{01} = \frac{P(\theta \in I_0 \mid D, M_e)}{P(\theta \in I_1 \mid D, M_e)} \, \frac{P(\theta \in I_1 \mid M_e)}{P(\theta \in I_0 \mid M_e)}$$

4.5 Manually calculating the Bayes factor for our ROPE hypothesis

To calculate the Bayes factor for our ROPE hypothesis, we can use the formula above using samples from the prior and the posterior based on our encompassing model. For the case of the Korean speech data, we already obtained prior samples above in the `fit_prior` model, and posterior samples in the `fit` model. So,

[2] The problem is one of a class of so-called **vanishing measures problems**. The probability of a point value is a probability *density*. Markov Chain Monte Carlo (MCMC) methods, which are used by `brms` and many other Bayesian software packages, generate samples from the posterior distribution. It is extremely unlikely that you will ever get a sample that is exactly equal to the point value x. So you would need to estimate something like the probabilities of a very small interval around x, and put that in relation to similarly small intervals for all other possible values of θ^* to get a reliable estimate of the density at x.

Figure 4



we can use these to extract the proportion of samples that fall inside and outside of our ROPE and do the calculations by hand. Let's do this first.

```
prior_ROPE <- fit_prior |>
  spread_draws(b_context1) |>
  summarise(prior_ROPE = mean(b_context1 >= rope[1] & b_context1 <= rope[2])) |>
  pull()

post_ROPE <- fit |>
  spread_draws(b_context1) |>
  summarise(post_ROPE = mean(b_context1 >= rope[1] & b_context1 <= rope[2])) |>
  pull()
```

Using these numbers, we can now calculate the Bayes factor in favor of the null hypothesis that the effect of `context` on `pitch` is in the ROPE versus the alternative hypothesis that it is outside of the ROPE:

```
BF_favoring_Null <- (post_ROPE / (1 - post_ROPE)) /
                    (prior_ROPE / (1 - prior_ROPE))
BF_favoring_Null
```

```
[1] 0.6159867
```

The Bayes factor in favor of the alternative hypothesis is simply the inverse of this number:

```
BF_favoring_Alt <- 1 / BF_favoring_Null
BF_favoring_Alt
```

```
[1] 1.623412
```

With a Bayes factor of around 1.62, the data does not provide noteworthy evidence in favor of the alternative hypothesis that the effect of `context` on `pitch` is outside of the ROPE.

In Figure 4, the Bayes factor is the amount of change between (i) the ratio of samples *within* and *outside* the ROPE in the prior, and (ii) the same quantity for the posterior. So, to see evidence in favor of the null hypothesis (the ROPE), we would want to see the ratio of points shift in favor of the points *inside* of the ROPE as we go from prior to posterior. In the plot above, this does not seem to be the case. Rather, we see a small shift that *more* probability mass is located *outside* the ROPE for the posterior distribution as opposed to inside of it, as compared to the prior. This is why, at least in direction, the BF tells us to favor the alternative hypothesis. However, the shift is not so very pronounced, so that we would not speak of noteworthy evidence in favor of the alternative hypothesis, let alone strong or decisive evidence.

4.6 Calculating ROPE-ed Bayes factor with the `bayesfactorR` package

Instead of doing these calculations by hand, we can more conveniently calculate the Savage Dickey ratio with the `bayesfactor_rope()` function from the `bayestestR` package (Makowski et al., 2019). The function takes as input the posterior and prior fit objects (you can also only provide the posterior fit, in which case the function will sample from the prior for you). The function then computes the ratio for the specified rope for the specified parameter. Notice that the method implemented in this package is slightly different from the naive one we used above, in that it uses a method that provides more stable and precise estimates for smaller sets of samples. (The package uses logsplines.)

```
BF_1 <- bayesfactor_rope(posterior = fit,
                         prior = fit_prior,
                         null = rope,
                         parameter = "b_context1")
BF_1
```

```
Bayes Factor (Null-Interval)

Parameter |   BF
----------------
context1  | 1.60

* Evidence Against The Null: [-10.000, 10.000]
```

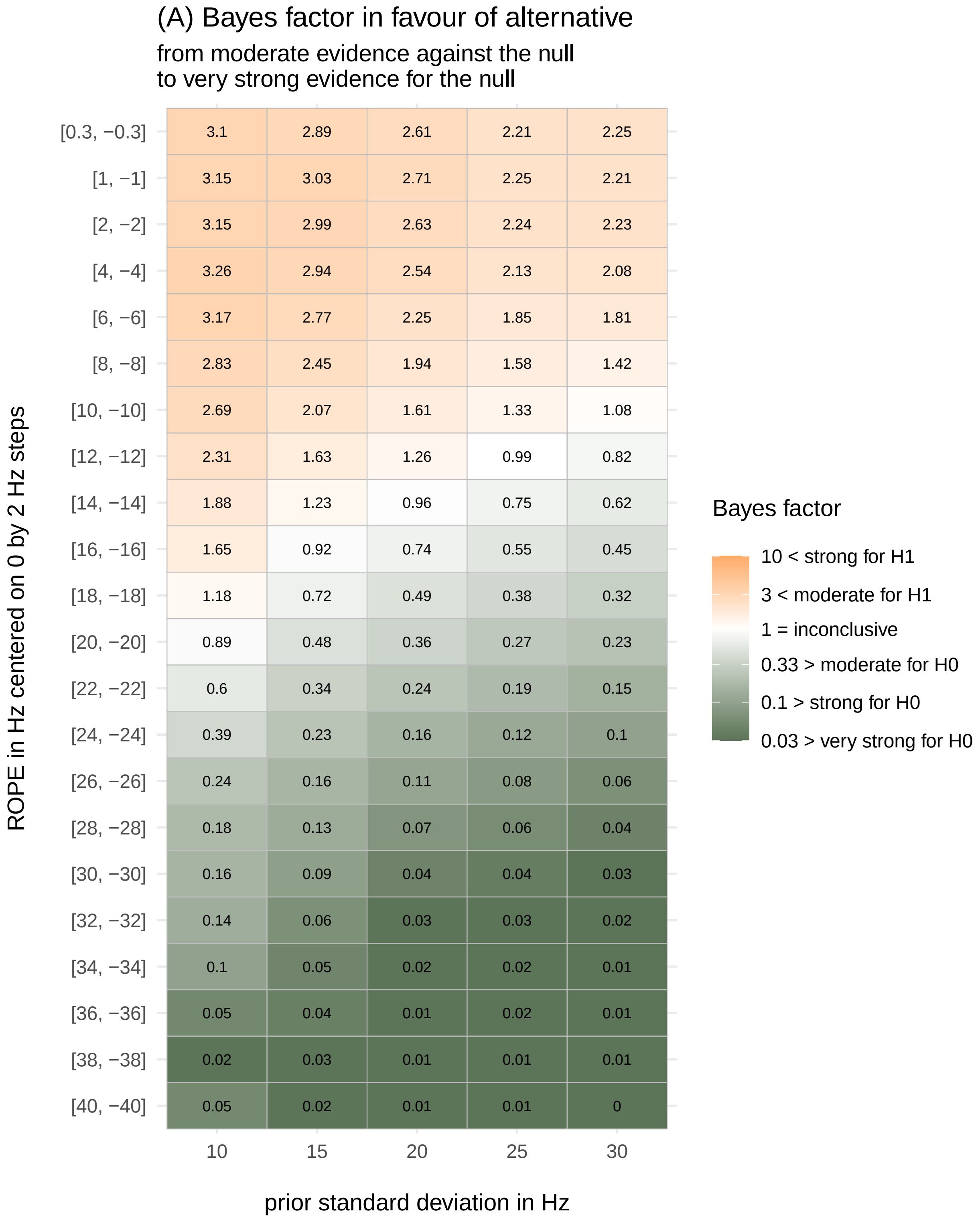
We obtain (almost) the same result in this way: the Bayes factor in favor of the alternative hypothesis for the given ROPE is around 1.6.

4.7 Sensitivity analysis for different priors and ROPEs

Now as you probably have guessed already, all these probabilities are very much dependent on the priors of the model, so it is important to evaluate the robustness of our Bayes factor-based interpretation across a range of sensible priors. And as long as we are not a 100% sure about what a meaningful difference is, we might as well explore the robustness of the Bayes factor across different ROPEs, i.e., in our case different JNDs. We won't bore you with the code for that process, but you can follow it along in our scripts. Let us explore the following ROPE intervals as informed by the two studies cited above on pitch perception: we test a range of ROPE intervals from 0.3 Hz to 40 Hz. We also assume the following five prior values for the standard deviation of the critical parameter (centered on zero): 10, 15, 20, 25, 30. These are all sensible prior widths assuming that medium to strong pitch effects in either direction are plausible.

Figure 5

Bayes factors for a range of priors and a range of ROPEs



(A) Bayes factor in favour of alternative

from moderate evidence against the null
to very strong evidence for the null

The combination of Bayes factors is visualized in Figure 5. Orange cells indicate evidence for the alternative. Green cells indicate evidence for the null. It becomes clear that the conclusions we can draw from our data are rather dependent on the choices we made along the way.

By comparing the Bayes factors along the y-axis, we can see that they are heavily dependent on the chosen ROPE. We here chose (theoretically speaking) a quite large range of ROPEs, all of which are informed by psychoacoustic studies of what pitch differences can be reliably heard and thus likely are meaningful for communication. In light of this range of possible definitions of what constitutes meaningful differences, our data do not seem very robust, as illustrated by the shift from orange to green. Even the smallest ROPE intervals provide only anecdotal to moderate evidence for the alternative. And the most conservative ROPEs, following Turner et al. (2019), leads to moderate to very strong evidence against the alternative hypothesis.

Additionally, when comparing the Bayes factors along the x-axis, we can see that they are comparatively consistent for different standard deviations of the critical prior. However, we can also see that the Bayes factors decrease with the width of the priors (from left to right).[3]

Combined, we can see that the larger the ROPE and the wider the priors, the more likely becomes the null hypothesis. In an ideal world, the evidence provided by the data should be robust across these choices.

This exploration of our inference is a fantastic opportunity to assess the boundaries of our conclusions. In this case, the conclusions of the original study by Winter and Grawunder (2012) was based on the null hypothesis significance testing framework, traditionally testing the compatibility of the data with a point-null hypothesis. Given his framework, it is reasonable to conclude that in formal speech, Korean speakers lower their average fundamental frequency. However, thinking more deeply about the theoretical consequences of differences in pitch, it might be less clear that these differences are truly meaningful. Given our proposed assumptions about what constitutes a meaningful effect, our analyses suggest that we neither find robust evidence for nor against the hypothesis that Korean speakers meaningfully lower their average fundamental frequency in formal speech.

5 How to write this inferential procedure up?

Here is a possible way to write up our analysis, following Kruschke's catalog of best practices (Kruschke, 2021). We first have to describe our model structure, including the priors of all parameters, and then the inferential procedure combining ROPEs with Bayes factor.

5.1 Model structure

The data were modeled using a hierarchical linear model predicting the continuous variable pitch (in Hz) by both the categorical predictor gender (male vs. female, contrast-coded) and social context (informal vs. formal, contrast-coded) and the maximal random-effects structure justified by the study's design (Barr et al., 2013), including by-subject random slopes (n = 16), and by-sentence random slopes (n = 7) for social context. Parameter estimation and inference is performed within the Bayesian framework. The model was fit using `brms` (Bürkner, 2018) in R (R Core Team, 2025). We used regularizing, weakly informative priors for the models (Gelman et al., 2017). Concretely, we used a Student's *t* distribution for the prior for the intercept (df = 3, mean = 211.6, df = 80.2), corresponding to the grand mean of the empirical data, a Student's *t* distribution for the prior for all random effect variance components as well as residual variance (df = 3, mean = 0, df = 80.2), and a Lewandowski-Kurowicka-Joe distribution (LKJ, shape = 1) for all correlation parameters. These priors were default priors, estimated from the data by brms. We specified a reasonable weakly informative prior for the predictor gender (normal, mean = 0, sd = 100) and specified a range of

[3] This is not surprising and a known phenomenon, often discussed under the Jeffreys-Lindley paradox (Lindley, 1957): The more diffuse the priors are (i.e., wider priors), the more probability they put on extreme values that (usually) tend to make the data extremely unlikely.

reasonable weakly informative priors for the predictor of interest: social context (normal, mean = 0, sd = [10,15,20,25,30]).

We fit this model with four chains of Hamiltonian Monte Carlo sampling for the estimation of the joint posterior distribution using the No U-Turn Sampler as implemented in Stan (Carpenter et al., 2017), and 8000 iterations (of which 4000 for warm-up) per chain.

5.2 Inferential assessment via Bayes factor and ROPEs

Using the Bayesian framework, we aim to quantitatively evaluate the evidence for a perceptually meaningful effect of social context on pitch values based on our data against the background of our model and chosen priors. We combine two statistical concepts to make this evaluation: First, we define a region of practical equivalence (ROPE) that represents a reasonable range of pitch values around zero that we consider to be not meaningful (Kruschke, 2018). In our case, the ROPE is perceptually defined by studies on just noticeable differences in pitch perception (Klatt, 1973; Turner et al., 2019).

Subsequently, we calculate the Savage-Dickey density ratio (Dickey & Lientz, 1970; Wagenmakers et al., 2010), i.e., we relate the amount of evidence (the proportion of posterior samples) within the ROPE for the model based on the priors only (i.e., before seeing the data) to the amount of evidence within the ROPE for the model based on both priors and likelihood (i.e., after seeing the data) (e.g., Wetzels et al., 2010). The Savage-Dickey method lets us assess evidence for and against a null hypothesis using Bayes factor (BF), since we are dealing with nested models. Since BFs can depend on both the defined ROPE and the priors of the model, we assessed the sensitivity of the results through calculating BFs for a range of (sensible) ROPE values and a range of (sensible) priors. We assumed the ROPE intervals centered on zero from 0.3 Hz to 40 Hz. We assumed the following five prior values for the width of the context parameter (centered on zero): 10 Hz, 15 Hz, 20 Hz, 25 Hz, and 30 Hz. These are all sensible prior choices assuming reasonable pitch differences in either direction.

Based on this setup, we consider each pair of ROPE and prior. For each pair, we speak of moderate evidence for either hypothesis if the Bayes factor in its favor is as least 3. The overall conclusion from this sensitivity analysis is a nuanced assessment of the interplay of ROPE and prior, drawing on empirical knowledge of the magnitudes involved (here: pitch).

6 Some words of encouragement

Bayesian inference in general and this form of hypothesis testing in particular require much more thinking than we might be used to. We believe this is a good thing. Many voices have criticized the lack of engagement that we behavioral scientists invest into thinking how our theoretical ideas connect to concrete predictions in the quantitative systems under investigation (Coretta et al., 2023; Scheel, 2022; Woensdregt et al., 2024). The presented form of hypothesis testing is easy to understand, but does require to think deeply about prior quantitative assumptions as well as what it means for observations to be meaningfully different. That is neither trivial nor easy. But we would like to encourage everybody to engage in exactly this thinking to better understand the relevant data in front of us and how it might link to our understanding of cognition and behavior.

7 Other Resources

There are many fantastic resources out there to help you learn about the wonderful world of statistics in general and Bayesian inference in particular. Here are a few recommendations. A very accessible introduction to linear models in R, using a non-Bayesian frequentist approach, is (Winter, 2019). A good and gentle general first introduction to Bayesian statistics is (Kruschke, 2015). Another accessible, but slightly more technical introduction is (Lambert, 2018). A fairly technical but very comprehensive introduction to Bayesian statistics is (Gelman et al., 2014). If you already have some background in statistics, a great resource is (McElreath, 2016/2020).

8 References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

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

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*, 1–32. https://doi.org/10.18637/jss.v076.i01

Coretta, S., Casillas, J. V., Roessig, S., Franke, M., Ahn, B., Al-Hoorie, A. H., Al-Tamimi, J., Alotaibi, N. E., AlShakhori, M. K., Altmiller, R. M., et al. (2023). Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human-speech analyses. *Advances in Methods and Practices in Psychological Science*, *6*(3). https://doi.org/10.1177/25152459231162567

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*(1), 214–226.

Franke, M., & Roettger, T. (2019). *Bayesian regression modeling (for factorial designs): A tutorial*. OSF. https://doi.org/10.31234/osf.io/cdxv3

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd edition). Chapman; Hall.

Gelman, A., Simpson, D., & Betancourt, M. (2017). The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy*, *19*(10), 555. https://doi.org/10.3390/e19100555

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97. https://doi.org/doi.org/10.1016/j.jmp.2017.09.005

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Klatt, D. H. (1973). Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception. *The Journal of the Acoustical Society of America*, *53*(1), 8–16. https://doi.org/10.1121/1.1913333

Klugkist, I., & Hoijtink, H. (2007). The bayes factor for inequality and about equality constrained models. *Computational Statistics and Data Analysis*, *51*(12), 6367–6379. https://doi.org/10.1016/j.csda.2007.01.024

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neelandica*, *59*(1), 57–69. https://doi.org/10.1111/j.1467-9574.2005.00279.x

Kruschke, J. K. (2015). *Doing Bayesian data analysis* (2nd edition). 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. https://doi.org/10.1177/2515245918771304

Kruschke, J. K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behaviour*, *5*(10), 1282–1291. https://doi.org/10.1038/s41562-021-01177-7

Lambert, B. (2018). *A student's guide to bayesian statistics*. Sage Publications.

Lindley, D. V. (1957). A Statistical Paradox. *Biometrika*, *44*(1/2), 187–192. https://doi.org/10.2307/2333251

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. https://doi.org/10.21105/joss.01541

McElreath, R. (2016/2020). *Statistical rethinking*. Chapman; Hall.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18. https://doi.org/10.1016/j.jmp.2015.11.001

Oh, M.-S. (2014). Bayesian comparison of models with inequality and equality constraints. *Statistics and Probability Letters*, *84*, 176–182. https://doi.org/10.1016/j.spl.2013.10.005

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Sanford, E. M., & Halberda, J. (2023). A shared intuitive (mis)understanding of psychophysical law leads both novices and educated students to believe in a just noticeable difference (JND). *Open Mind*, *7*, 785–801. https://doi.org/10.1162/opmi_a_00108

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development*, *31*(1), e2295. https://doi.org/10.1002/icd.2295

Turner, D. R., Bradlow, A. R., & Cole, J. S. (2019). Perception of pitch contours in speech and nonspeech. *Interspeech*, 2275–2279. https://doi.org/10.21437/Interspeech.2019-2619

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. Busemeyer, J. Townsend, Z. J. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 300–319). Oxford University Press.

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*, 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the savage–dickey density ratio. *Computational Statistics and Data Analysis*, *54*, 2094–2102. https://doi.org/10.1016/j.csda.2010.03.016

Winter, B. (2013). *Linear models and linear mixed effects models in R with linguistic applications* (arXiv:1308.5499). arXiv. https://doi.org/10.48550/arXiv.1308.5499

Winter, B. (2019). *Statistics for linguists: An introduction using r*. Routledge.

Winter, B., & Grawunder, S. (2012). The phonetic profile of Korean formal and informal speech registers. *Journal of Phonetics*, *40*(6), 808–815. https://doi.org/10.1016/j.wocn.2012.08.006

Woensdregt, M., Fusaroli, R., Rich, P., Modrák, M., Kolokolova, A., Wright, C., & Warlaumont, A. S. (2024). Lessons for theory from scientific domains where evidence is sparse or indirect. *Computational Brain & Behavior*, *7*(4), 588–607. https://doi.org/10.1007/s42113-024-00214-8