

Some Prior(s) Experience Necessary

Templates for Getting Started with Bayesian Analysis

Chanda Phelan

University of Michigan School of Information
Ann Arbor, Michigan
cdphelan@umich.edu

Matthew Kay

University of Michigan School of Information
Ann Arbor, Michigan
mjskay@umich.edu

Jessica Hullman

Northwestern University
Evanston, Illinois
jhullman@northwestern.edu

Paul Resnick

University of Michigan School of Information
Ann Arbor, Michigan
presnick@umich.edu

ABSTRACT

Bayesian statistical analysis has gained attention in recent years, including in HCI. The Bayesian approach has several advantages over traditional statistics, including producing results with more intuitive interpretations. Despite growing interest, few papers in CHI use Bayesian analysis. Existing tools to learn Bayesian statistics require significant time investment, making it difficult to casually explore Bayesian methods. Here, we present a tool that lowers the barrier to exploration: a set of R code templates that guide Bayesian novices through their first analysis. The templates are tailored to CHI, supporting analyses found to be most common in recent CHI papers. In a user study, we found that the templates were easy to understand and use. However, we found that participants without a statistical background were not confident in their use. Together our contributions provide a concise analysis tool and empirical results for understanding and addressing barriers to using Bayesian analysis in HCI.

CCS CONCEPTS

• **Human-centered computing** → *User studies; HCI theory, concepts and models;*

KEYWORDS

Bayesian statistics, statistics, code templates, tutorials, hypothesis testing, evaluation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-5970-2/19/05...\$15.00

<https://doi.org/10.1145/3290605.3300709>

ACM Reference Format:

Chanda Phelan, Jessica Hullman, Matthew Kay, and Paul Resnick. 2019. Some Prior(s) Experience Necessary: Templates for Getting Started with Bayesian Analysis. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk*. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3290605.3300709>

1 INTRODUCTION

The Bayesian approach to statistical analysis has been the focus of increasing interest in recent years as an alternative to the traditional statistical approach, called alternately *frequentist statistics* or *null hypothesis significance testing* (NHST). Partly, this interest is driven by its increased accessibility: computationally expensive fitting procedures used in Bayesian statistics can now be run on many personal computers [6, 27]. Because of its ability to integrate prior knowledge to reduce the chance of overfitting to extreme observations, Bayesian analysis has also been proposed as a way to address some of the issues raised by the replication crisis occurring in scientific research. A number of high-profile replication failures have called attention to weaknesses in the dominant approach to analyzing and reporting statistical results.

Bayesian analysis is particularly well-suited for the needs of researchers who publish at CHI. For one, through the use of priors, the Bayesian approach provides a formal way to build on evidence. Small-n studies are common in HCI research, a product of the field's focus on early evaluations of novel systems. The frequentist approach provides no way to build on this knowledge other than running more studies with larger samples sizes, a requirement that is often impractical in HCI research [27]. In an ideal Bayesian analysis, each successive study builds on the knowledge gained in previous work, affording more precise estimates of effect sizes.

In addition, Bayesian analysis produces results that are more intuitively interpretable by people with a range of statistical knowledge. (This claim is discussed in more detail in Section 2, Background.) NHST relies on some measures

that are frequently misinterpreted, even by experienced researchers. Studies have shown that two of the most important statistics in NHST—the confidence interval and the p-value—are commonly misunderstood. For CHI, which prioritizes making the results of research accessible to a wide audience, it is particularly important that statistical results are reported in a way that can be understood intuitively.

Interest in Bayesian methods is visible in CHI through sessions and papers that argue for increased adoption of Bayesian methods in CHI research (e.g. [7, 25, 27]). However, very few CHI papers have utilized Bayesian analysis in interpreting the results of their research. Barriers to adoption exist, not least that individuals must bear switching costs to make the change, investing significant time and effort into learning an entirely new approach to statistics in order to complete the analysis. Online tutorials and courses exist to ease the transition (e.g. [33, 35]), but are still require substantial investment of effort before one can independently complete a Bayesian analysis of their own data.

In this paper, we present a set of user-friendly R Markdown templates written specifically for a CHI audience that guide users step-by-step through a Bayesian analysis.¹ The analyses in the templates are Bayesian analogues for some of the statistical tests we identified as most common in CHI, such as one- and two-way ANOVA and t-tests. The templates walk the user through analysis in straightforward prose that explains each chunk of code, while also minimizing the amount of new code a user has to write in order to successfully complete an analysis. The format allows the user to explore Bayesian analysis with more ease and confidence than code written from scratch or adapted from tutorials. The templates are intended for users who have some experience with NHST statistics, but requires minimal experience with R and no experience with Bayesian statistics.

We present a user study that demonstrates the templates can be used even by people who have no familiarity with Bayesian analysis. The user study also provides evidence to better understand the barriers to adopting Bayesian methods. Some users, especially those without a background in statistics, tended to lack confidence in the analyses they produced. The user study presents preliminary evidence for strategies to increase user confidence and analysis accuracy.

2 BACKGROUND

What is Bayesian statistics?

The core idea of Bayesian analysis is that Bayesian inference is a "reallocation of credibility across possibilities," where the "possibilities" are parameters in a statistical model [33]. In other words, Bayesian inference is a formalized way of making a best guess of what one expects to see in the data,

and then using new data to adjust the expectations of that best guess. A Bayesian model has three components [33, 35]:

- (1) *A likelihood function*: Determined by the statistical model. It is a function of the parameters in the model, and computes the probability of some observation occurring, given some set of parameter values.
- (2) *Parameters*: As in frequentist models, these are the variables in the model. Parameter values are unknown and are estimated from the data.
- (3) *Priors*: The probability distributions for each of the parameters in the model. These are the "best guesses" of what one expects to see in the data, before examining the data. Priors are used to constrain parameter values to reasonable ranges. Bayesian priors come in a variety of forms. A *strongly informed* prior might be one drawn from the results of similar past observations, allowing one to combine previous results directly with new data. Strongly informed priors are not always available, and may be harder to defend; more commonly, weakly informed priors are used. These priors may be informed by prior literature or prior experience of a domain, but are chosen to only minimally restrict the posterior to a range of plausible values. Priors that do not restrict the parameter values are called *flat* priors.

Using these three components—the likelihood function, the parameters, and the priors—together with a set of observed data points, the Bayesian estimation process *updates* the prior distribution to produce a *posterior distribution*. The posterior is a probability distribution describing the likely values of the parameters, given the prior, the model, and the observed data. Put another way, this process combines observed data from the current study with prior knowledge, reallocating credibility across the parameters, and produces a new best guess for the relative credibility of each parameter value. This new best guess is the posterior distribution.

Priors are the most distinctive component of Bayesian analysis, and for many the most difficult to understand. Prior-setting is an inherently subjective process; Bayesian textbooks often describe "correct" priors with descriptions such as: "The prior must pass muster with the audience of the analysis, such as skeptical scientists" [33]. Bayesian proponents have argued that priors are not any more inherently subjective than other model assumptions in NHST [2, 35]. The Bayesian prior forces one to make assumptions about prior knowledge explicit, which in turn makes it easier to examine and question those assumptions. This often involves experimenting with different priors (for example, when there is no strong prior available), and evaluating how sensitive an analysis is to different prior beliefs.

¹Code templates at www.github.com/cdphelan/bayesian-template.

Why use Bayesian statistics?

The debate between proponents of Bayesian and frequentist statistics is not settled. In this section, we provide a brief overview of some of the advantages of choosing Bayesian statistics; the sources in this section address the debate in more detail (particularly [2, 6, 32]).

One of the great strengths of Bayesian analysis is the use of priors, which serves as a formal method for building on knowledge across studies. This knowledge accrual affords a more precise estimation of true effect sizes, a statistic that single studies often estimate poorly [22]. This is a particularly important feature in HCI, as other formal methods of knowledge accrual (i.e. meta-analyses) are rarely used [27].

Bayesian analysis can also produce results that are more intuitive to understand than frequentist results, which are frequently misunderstood even by experts. For example, a frequentist 95% confidence interval, is often interpreted as meaning that the true parameter value lies somewhere within the interval [18]. In fact, it means that if the experiment was repeated many times, 95% of those intervals (but not necessarily for the experiment at hand) would contain the true parameter value. However, for the Bayesian analogue, the credible interval, the intuitive interpretation is the correct one: there is a 95% probability that the true parameter value lies within the credible interval, conditional on the prior.

Bayesian analysis avoids using p-values, instead relying on credible intervals and effect sizes. P-values, like confidence intervals, are prone to misinterpretation: a survey of 70 experienced researchers found that only two understood the meaning of statistical significance captured by p-values [40]. For example, p-values do not indicate if the size of an effect is important, only that an effect probably exists [25, 43].

Because of these weaknesses, a number of authors have argued for discarding p-values entirely in favor of effect sizes [9, 25, 32, 43]. Frequentist statistics can produce effect sizes, but often under-emphasize them [27]. Bayesian analysis brings effect sizes to the foreground and provides extra tools to increase the accuracy of effect size estimates.

This emphasis on effect sizes is one of the reasons Bayesian analysis has been raised as a way to address the replication crisis occurring in many scientific fields [1], where a number of highly cited studies have failed to replicate (e.g. behavioral priming in psychology [8]). Evidence suggests that many published studies are in fact false positives [22].

Small-n studies, which are common in HCI, are particularly susceptible to false positives [22]. Bayesian analysis can be used to correct this by setting skeptical priors (i.e. strong priors) that shrink overestimated effect sizes that small-n studies tend to produce [27]. Additionally, Bayesian methods are especially useful for small-n studies because they do not rely on asymptotic properties of the sampling distribution, a

foundational assumption of frequentist methods that small sample studies often fail [36]. For these reasons and more, authors have argued that Bayesian methods are well-suited for research in HCI (e.g. [25, 27]) and other fields (e.g. [6, 32]).

Existing tools for learning Bayesian methods

A number of tutorials and textbooks exist that teach Bayesian statistics. One popular textbook is *Statistical Rethinking* by Richard McElreath, which includes an R package for Bayesian analysis, *rethinking*, that was custom-built for the text [35].

Another popular textbook is *Doing Bayesian Data Analysis* by John K. Kruschke [33]. A particular advantage of this textbook is that it includes a table of Bayesian analogues of many NHST analyses, such as t-tests and logistic regression. There are also many code examples, available as a downloadable zip file, that work through many different types of analyses.

Both textbooks prioritize laying down a solid conceptual grounding of Bayesian statistics, but the tradeoff is that they require a significant investment of time and effort before one can actually complete a Bayesian analysis with one's own data. In McElreath, for example, the text provides few shortcuts for anyone who wishes to easily "translate" a given frequentist test to Bayesian terms. Kruschke provides more guidance for "translations" through extensive code examples; however, the example code is complex and difficult to parse without referencing the textbook heavily.

Though both of these textbooks are strong examples of how to teach Bayesian concepts, both have significant "start up costs" before a user can perform an independent Bayesian analysis. For casual explorers, there are few options.

Active learning and statistics pedagogy

Instead of asking users to invest large amounts of time and effort up front to establish a knowledge base in Bayesian concepts, the templates presented here drop users directly into a Bayesian analysis and break apart the process, so users can work backward in their understanding.

This strategy is drawn from active learning, a philosophy of teaching that prioritizes students becoming actively involved in the learning process [3, 10]. Active learning encompasses a wide variety of approaches, including interactive demonstrations [42] and replacing lectures with workbooks or tutorials [13]. In traditional classroom settings, active learning strategies have been repeatedly demonstrated to improve student performance [4, 13, 16, 31]. This has been found to be true across academic fields, including statistics. For example, Carlson and Winquist found that students outperformed a control group when a "workbook" class structure was employed: instead of lectures, students read very short introductory text and then worked through example problems on an online homework portal [4]. Similar findings exist for the use of active learning in MOOCs. Koedinger et al.

found that students learn more with "learn by doing" activity modules, over simply reading or watching videos [31].

Experts in statistics pedagogy have argued that focusing on theory and rule memorization fails to engage students meaningfully. Instead, they argue for alternate strategies such as having students work through analyses that use real data, and automating computations and visualization generation as much as possible in order to avoid complications that interfere unnecessarily with student learning [5, 37]. We employ these strategies in our templates, and present a usability study to assess whether people appear to benefit from being "dropped in" to Bayesian analysis.

3 TEMPLATE DESIGN

Motivation

Despite the advantages of Bayesian statistics, very few papers published in CHI use Bayesian statistics in the analysis of their results. Searching papers in the last five years of CHI, we found only seven [11, 23, 26, 28, 34, 38, 44].

This is at least partly because of switching costs: learning a new approach to statistics can be time-consuming and difficult. To address this barrier, we propose a set of R code templates where each template guides the user through a different Bayesian analysis, allowing them to explore simple Bayesian analyses with no advance preparation outside of reading through the template. Researchers can use their own datasets, so that they can explore the analysis using data that is meaningful to them. The templates also support more advanced users who wish to produce analyses for publication.

We had three main design goals for the templates:

- (DG1) Allow users to easily complete a Bayesian analysis, even with no prior knowledge of Bayesian statistics.
- (DG2) Communicate statistical results in a way that matches people's intuitive interpretations of uncertainty. This is particularly important for a CHI audience, which comprises people with a range of expertise in statistics.
- (DG3) Prioritize the particular needs of CHI researchers. The templates should support the analyses that are most relevant to CHI researchers.

In the following section, we describe the design of each component of the template, and how each component supports these design goals.

Design

Template format & structure. The code templates are R Markdown files, a type of notebook interface that combines narrative text and chunks of code into a single document. This choice was informed primarily by the first design goal (DG1): the template walks through every step in plain words, explaining the code chunks that are interspersed throughout the text. This structure makes it easy to follow along with the analysis without having to interpret large blocks of code. In addition, the narrative text is written in a style that is intended to be clear and encouraging, with minimal jargon; it assumes little knowledge of R or technical statistical terms.

To further simplify understanding (DG1), we created the templates with the goal of minimizing the need to write new code, or even to parse the existing code. Where changes to the code are necessary, the code chunk is preceded by a numbered list of items and descriptions of what need to be changed. The workflow of the template is diagrammed in Fig. 1. As shown in the figure, only four steps require user input, under the headings "Set up" and "Set model."

Finally, all the templates are pre-filled with an example analysis, using a concrete example to make it easier to follow along (DG1). In keeping with DG3, prioritizing the needs of CHI researchers, the example analyses were all done with data from a 2016 CHI paper with Bayesian analysis by Moser et al. about choice overload in an e-commerce context [38].

Supported analyses. The first step in Bayesian analysis is to create a model. To simplify this process, we created a set of templates where the model is already created; users then simply select from the available options and customize the model to their data. These are the steps under the "Set model" heading in Fig. 1. To support casual exploration of Bayesian analysis, the models we selected for the templates all have the same assumptions as their frequentist analogues, so that a user can make a post-hoc shift from a frequentist to Bayesian model without threatening the model's validity.

When choosing which models to support in the templates, the most important implication of DG3 was to ensure that the templates were compatible with the most common simple statistical tests present in CHI research papers. Using the Digital ACM database, we searched through papers from recent years of CHI to get an estimate of which analyses

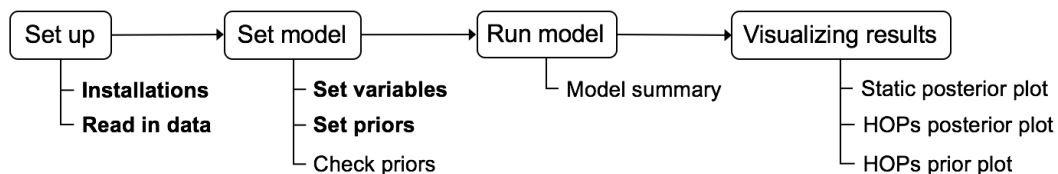








Figure 1: Workflow of the code templates. Sections that require user input are bolded.

Table 1: Analyses supported by the code templates.

One independent variable				Interaction of two independent variables			
#	Sample plot	Variable	Compatible with:	#	Sample plot	Variables	Compatible with:
1		Categorical	t-tests one-way ANOVA	4		Categorical Categorical	two-way ANOVA
2		Ordinal	t-tests one-way ANOVA	5		Categorical Ordinal	two-way ANOVA
3		Continuous	linear regressions	6		Continuous Categorical	linear regressions with interaction

were most popular in recent CHI research. We first skimmed a sample of CHI 2018 papers to obtain a broad sample of statistical tests present at CHI. Using the identified search terms, we searched 2014-2018 CHI proceedings to get a rough estimate of prevalence. The test in the most papers by far was ANOVA (781 total; 133 one-way; 58 two-way; 298 repeated measures; 292 unspecified). About half of those were followed up by post-hoc tests, most of which were t-tests: t-tests appeared in 540 papers, 240 of which also had an ANOVA. Less common but also notable were regressions (338 total, 52% linear). Other tests we searched for included correlations (256 papers) and Wilcoxon tests (296).

It would be difficult to give a single, generic model that works across many repeated measures designs without customization. For example, random effects models—commonly used for repeated measures—may include varying slopes or intercepts, and the correct model depends on the experimental design and/or assumptions about the data generating process. A template for a random effects model would require heavy customization and thus introduces a higher risk of errors. Therefore, in order to support the goal of creating an easy-to-use template for a Bayesian novice (DG1), we excluded analyses that involved repeated measures.

We also excluded all regressions except linear regression, in order to define a small set of simple models that nevertheless cover a large number of use cases in CHI (approximately 55% of analyses published since 2014). Once that was done, we were left with six types of analyses to support with the code templates. The templates are summarized in Table 1.

Though the code for each analysis is fairly similar, in order to support DG1, we split each analysis into separate template documents; we were then able to explain each analysis precisely and avoid complicated conditional statements in the code that would make the code harder to follow.

To confirm that these analyses were compatible with DG3, we used six datasets drawn from recent CHI papers to run

11 analyses through the templates. This ensured that the templates were flexible to different datasets. Where the full datasets were not publicly available, we simulated data using descriptive statistics from the paper.

Bayesian analyses. Bayesian analyses are performed using the R package `rstanarm`. This package uses syntax that is similar to standard regression functions in R such as `lm()` and `glm()`, making the code easier to follow for those users who are already familiar with regressions in R (DG1). A model summary is produced at the "Run model" step (Fig. 1).

Setting priors. Setting priors (under the "Set model" heading in Fig. 1) is the most conceptually difficult part of completing a Bayesian analysis with our template. Thus, in order to stay consistent with DG1, we simplified certain aspects of setting priors. For one, we limited all priors to a normal distribution. As a result, the templates may not be compatible with analyses that have parameters with non-normal distributions, but in exchange the conceptual and technical difficulty of setting priors is much reduced. In addition, we limited users to setting two priors: the "control condition" prior, and the "effect size" prior. This follows from our evidence for common CHI analyses (DG3), which found that the majority of studies test for differences between experimental conditions. For example, in an experiment testing a novel interaction method (treatment) against current best practice (control), the effect size would be the expected difference in performance between control and treatment. This means that users cannot set different effect sizes for different treatment conditions, but again significantly reduces the conceptual and technical difficulty of setting priors.

To further support users in exploring their choices of priors, a section called "Check priors" was included in the template (under the "Set model" heading in Fig. 1). This section generates a set of graphs that visualize a number of draws from the priors. See Fig. 2 for an example output of this

section. This encourages users to test priors, checking that their choices generate reasonable values for parameters and examining them critically to see if the assumptions are valid.

Visualizing results. DG2, supporting intuitive interpretation of results, came into play most in the last section of the template, "Visualizing results" (Fig. 1). To facilitate reasoning about distributions in the Bayesian analysis process, our templates use hypothetical outcome plots (HOPs) [21, 24]. HOPs are visualizations of uncertainty created by randomly sampling draws from a distribution and presenting each sample as a separate frame in an animation (Fig. 3). Rather than requiring users to reason about statistical constructs like confidence intervals to infer probabilities, HOPs encode probability via frequency [21].

Simply framing probabilities as frequencies (3 out of 10 vs. 30%) has been shown to improve performance on classical Bayesian reasoning problems [14, 19], and recent research in uncertainty visualization shows benefits of frequency framings for static plots [12, 20, 26]. More importantly, when probability is encoded as frequency, a viewer needs only watch the visualization to extract the frequency information automatically [17]. HOPs have been demonstrated to be more effective in communicating uncertainty when compared to standard static representations such as error bars or violin plots [21]. Studies have shown that compared to static plots, HOPs improve multivariate probability estimates [21], increase sensitivity to underlying trends in data [24], and improve Bayesian reasoning [30].

One of the strengths of HOPs is that they afford more accurate interpretations even among people who have no background in statistics and have received little training on

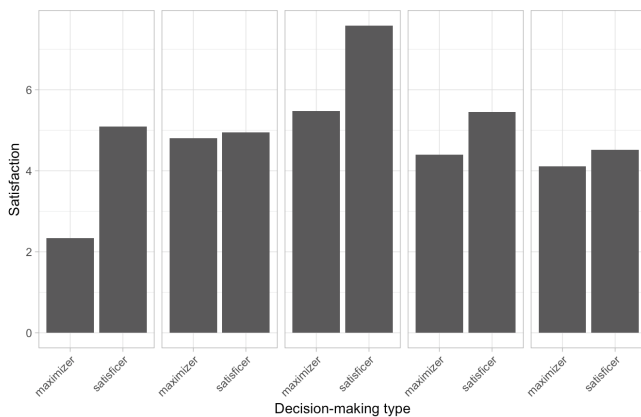


Figure 2: An example of the output of the "Checking priors" section of the template, which plots five sample draws from the priors set by the user. These plots are used to check the priors' plausibility. This example uses priors from an example analysis modeled after Moser et al. [38].

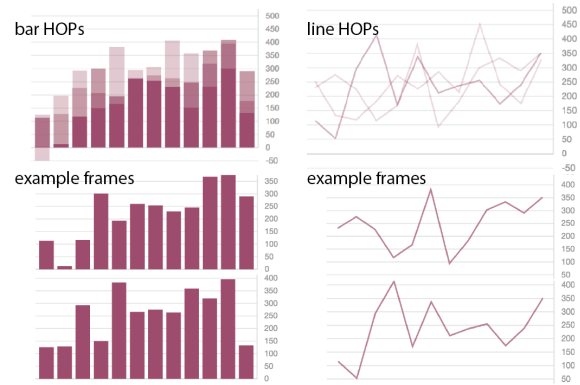


Figure 3: Example of bar and line HOPs using dummy data. In this static example, individual frames are made transparent and combined into a single figure. In the animated version, each individual frame appears sequentially. This affords more accurate interpretations of uncertainty.

how to use the plots [21, 24]. This makes HOPs an especially powerful tool for communicating results of CHI research to an audience with a range of statistical backgrounds.

The templates also produce a static graph with credible interval bars (see Fig. 4 for an example). This was done to stay consistent with standard practice in CHI, but with a Bayesian twist: the bars represent the 95% credible interval, which are more intuitive and easier to understand (DG2) than frequentist plots with error bars [18].

The plots in the template are generated using the ggplot2 package, which most R users are familiar with; consistent with DG1, this makes the code more readable and supports easily customization of the plot aesthetics. HOPs animations are generated using the gganimate package.

4 USER STUDY

We conducted a user study to evaluate if the templates were able to guide participants through their first Bayesian analysis. We were most interested in if the template was considered easy to use, and if the minimal conceptual background provided in the template was sufficient for users to complete the analysis confidently. We also wanted to learn more about what barriers remained to adopting Bayesian analysis.

Method

We recruited 13 participants to complete two Bayesian analyses with our templates, using datasets we provided. We expect that most of the users of the template will be CHI researchers who want to explore a Bayesian analysis using a dataset they have already analyzed using frequentist methods. We therefore designed the user study to parallel this expected use case as closely as possible, and recruited primarily information researchers as participants.

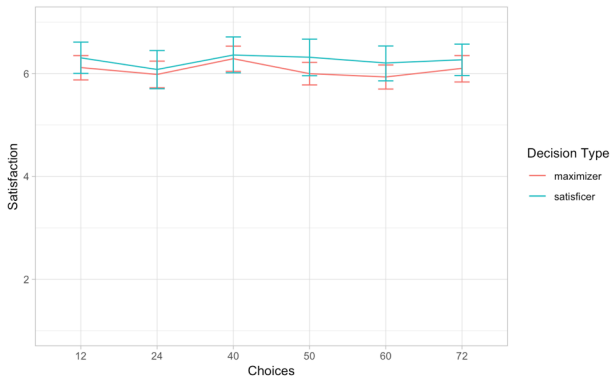


Figure 4: An example static output of the template with credible interval bars. This example plots a result from Moser et al. [38] showing that one group from the study (maximizers) tend to be slightly less satisfied than the second group (satisficers), but with considerable overlap in the credible bars.

Participants. User study participants had two eligibility requirements: they had never performed a Bayesian analysis before, and they could use R at least well enough to do simple statistical analyses and plotting. Participants were recruited from a convenience sample of graduate students and faculty. We paid participants \$50 to attempt the task. Participants were not required to successfully complete the template, but were asked to try for at least two hours.

Protocol. The user study was conducted across three rounds of testing (R1, R2, R3). Between the rounds, feedback from the previous round of user testing was used to iterate on the template design. Some minor changes (e.g. text or formatting changes) were also pushed out during the rounds.

The protocol of the user study changed slightly after R1: R2 and R3 used different datasets with slightly different analyses. Once results from R1 indicated that participants could complete the analyses much faster than expected, we switched one of the datasets to more precisely fit the templates. Otherwise, the protocol in the two rounds was very similar.

In both rounds, qualifying participants were given a written protocol and told to complete the tasks at their convenience. They were told to email study members with any questions or issues. They were also asked to rate their knowledge of R on a 5-point Likert scale.

In the protocol, participants were given a link to a github repository with all the template files and instructed to download the repository as a zip file. A README for the repo acted as a table of contents for the templates, describing which template did which analysis.

In addition to the link to the code repository, participants were given the datasets they would use to complete the analysis. All datasets were derived from papers published in

CHI 2018; participants were also given the corresponding published paper. One dataset, used in both rounds of user testing, was a subset of data from Nordhoff et al., published publicly on the researchers’ website [39]. The other datasets used the user study were simulated from results published in two other CHI 2018 papers with honorable mention awards. These datasets were selected because the original analyses were high-quality and used variables that did not require extensive background knowledge to understand.

In R1, two datasets were used; participants were given one of the two. Then, they were told they would be reproducing two versions of a particular figure from the published paper. This task was modeled to match the behavior of what we assume to be the most common use case: researchers who have already performed a frequentist analysis on their data and are interested in comparing the plotted results. One set of participants was asked to complete a one-variable analysis first, then a two-variable analysis. To control for ordering effects, the analysis order was reversed for the other set of participants. Participants were randomly assigned to one of the two orders. This afforded testing of two templates (Templates 1 and 4) with two levels of difficulty.

When R1 showed that participants completed the analyses faster than expected, the study protocol was changed for R2 and R3 so that each participant did two analyses with two different datasets, instead of two analyses with the same dataset. As in R1, all participants completed both a one- and two-variable analysis, this time testing Templates 1 and 5.

To keep the task close to natural use, the protocol did not specify whether participants should use weakly informative priors or strong priors drawn from previous studies. Because setting weak priors is a simpler process, it was expected that most participants would use weakly informative priors.

After completing (or attempting to complete) the two tasks, participants were asked to knit the template, an R Markdown document, and return the HTML output file. In addition, they answered two free-response questions. One question asked to compare the results of the Bayesian and frequentist analysis; the purpose of this question was to assess their understanding of Bayesian methods. In the second question, they were asked to report on what they perceived to be the pros and cons of the template.

Results

User testing was conducted in three rounds. In most cases, we report results aggregated over all three rounds of testing. The exception is the prior-setting section of the template. As that section required multiple iterations, we report the results of different iterations separately.

Rounds of testing. Testing rounds were characterized by major changes in the template or study protocol. After R1, the

datasets used in the testing protocol were changed, and edits were made in the template to the installation and prior-setting instructions. After R2, additional changes were made to the prior-setting section. Changes to the prior-setting section are discussed in more detail in the "Setting priors" section of these results.

Participants. Over three rounds of user testing, 13 people participated (6 women). The first round had 5 participants; the second and third each had 4. All participants had at least some knowledge of R code and frequentist statistics. Twelve were graduate students or faculty; of those, ten were information researchers. Seven rated themselves "slightly knowledgeable" in R (a rating of 2 out of 5), and the remaining participants rated themselves as "moderately" or "very" knowledgeable (3 and 4 out of 5, respectively). None had performed a Bayesian analysis before. Participants varied in their experience of frequentist statistics, ranging from relative beginners who had only done analyses in statistics classes, to more advanced users who had extensive publication records of statistical analyses.

Ease of use. All participants were able to complete at least one of the two assigned analyses; all but one participant was able to complete both analyses.² All participants selected the correct templates for their analyses from the six available.

After completing the installation process, most participants were able to complete the analyses faster than we expected. Most participants were able to complete two analyses in about one hour, with several people completing both tasks in about 30 minutes.

Overall, participants said they found the template easy to use. Many participants commented on how easy it was to use, like B-1³: *"The template was super easy to use when you figured it out and knew what parts to change and what types of information it was looking for."* D-1 had a similar comment: *"The template is pretty self-explanatory especially after going through it once."* L-3 was even more positive: *"This is fun!"*

Specifically, some of the most common positive feedback related to the structure of the template. F-2 said one of their favorite parts of the template was *"the encouraging narrative style,"* while C-1 said, *"I liked how it read from beginning to end like a narrative."*

Several said the template structure contributed to their ease of understanding. F-2 said the *"not-too-long [code] chunks"* helped him understand what part of the analysis was being conducted in each chunk. Similarly, E-1 said, *"the notes about what to change and what each chunk is meant to do were excellent & helpful!"*, sentiment echoed by several others.

²In R2, several users completed the two-variable analysis but were unable to produce a knitted document because of a compatibility error with a package used in the template. The issue has since been addressed.

³The numeral in participant IDs corresponds to their testing round.

The most difficult and time-consuming part of using the templates for all users was installing the required R packages. A list of troubleshooting options was added to the template during the first round of testing. Though the installation issues persisted, there were fewer reports of installation problems that the users could not resolve themselves.

Confidence of use. Though participants reported the template easy to use, some of the participants also said they were unsure of the accuracy of their results. As expected, the majority of participants said that their choice of priors was the main source of their uncertainty. G-2 said, *"Although that following your code was pretty simple and easy to complete the task, I think I still don't get what roles each parameter plays out there [in the analysis]."* C-1 had a similar comment: *"The only section I found a bit hazy was when I had to set the prior parameters. I'm not completely confident that I interpreted the prior effect size parameter correctly at first."*

High levels of confidence were not expected, as the template design prioritizes simplicity over deep understanding of priors. However, that tradeoff had meaningful consequences for usability, as L-3 pointed out: *"It will take some degree of trust to use this template."*

Like several participants, F-2 appeared to have a generally positive opinion of the template but wanted more information before relying on their own analysis: *"If I learn the proper usage and how to come up with priors myself, I would definitely consider using [Bayesian analysis] in future experiments."* These participants mentioned specific concepts they would need to learn to bolster their understanding.

It appeared that more advanced frequentist statistical knowledge helped people be confident in their priors. Participants with a more extensive background in statistics were more likely to set priors correctly and less likely to express a lack of confidence in their analysis.

Of those who reported not being confident about their choice of priors, most pointed out some aspect of the template that aided their understanding. In addition to the clear instructions, participants liked the "Checking priors" section of the template, which visualized several results drawn from the prior distribution and helped participants think about their choices of priors. A prior-setting quizlet in R3 (see following section) also received positive comments. However, these features did not completely resolve the lack of confidence for all participants.

Even so, for many participants the templates seemed to lower the barrier for future Bayesian analyses even if they were not confident in their specific results. A number of participants said they were more likely to use Bayesian analysis in the future. Several participants said this was because the templates broke apart the steps of the analysis, making it easy to follow along. For example, E-1 said that though they

were uncertain about whether they had performed the analysis correctly, *"I did understand the *process* much better than I had previously, and thus would be more likely to consider Bayesian analysis in the future."*

Setting priors. As expected, a number of participants found it difficult to set priors for the Bayesian analysis. All set priors that were "reasonable," which we defined as priors that produced results that were not meaningfully different from the expected results. However, 6 of 13 showed some evidence of misunderstanding how to set priors, all in R1 and R2. Consequently, this section of the template was iterated on more than any other section.

One common issue in R1 of user testing was misunderstanding how to determine the standard deviations of each parameter's prior. To address this, we changed the template after R1 so that the user inputs the mean and the maximum plausible value of the parameter, instead of the mean and SD. Then, the template calculates the correct SD using those two numbers. After this change, every participant who was able to set reasonable prior means also set prior SDs correctly.

In R1 and R2, a number of participants used data from their current analysis to set priors. This is incorrect: priors should be set using only information that would have been available before gathering the current data. After R1 and during R2, we made changes in the text of the template to emphasize this principle. We also added an additional example in the template of how to set strong priors. These changes reduced the likelihood that participants would use their dataset to set priors, but did not eliminate it. Participants also continued to express a lack of confidence with their choice of priors.

As a consequence, to further aid understanding, for R3 we created an optional six-question quizlet for users to check their understanding of priors. The quiz included four examples of how to set both strong and weak priors, with the intention of providing additional practice and building users' confidence. It also emphasized the principle of not using current data during prior-setting.

Four participants tested the template after the quizlet was added. All four set priors correctly, and two mentioned the quiz specifically: *"I had no idea what priors were, and [the quizlet] was really helpful"* (L-3), and *"the quizlet was helpful in improving my understanding of priors"* (K-3). The other two said the information in the main template was sufficient and they did not need to use the quiz. By chance three of the four users in R3 had more advanced statistical knowledge than the average usability tester, which may have helped them set priors correctly. It is therefore not certain if the quiz would be broadly helpful for building accuracy and confidence.

It also appeared that there was some risk of over-explaining priors. Though some participants said they wanted more information on priors, several also said that the information

in the template caused them to expect prior-setting to be harder than it was. One example was K-3, who said that the quizlet was helpful but caused them to be *"predisposed to think [setting priors] was more complicated."*

5 DISCUSSION

The purpose of the code templates is to support CHI researchers in exploring Bayesian analysis, even as Bayesian novices. The templates serve as a shortcut to more CHI researchers producing high-quality Bayesian analyses, in service of the overall goal of lowering the barrier to adopting Bayesian analyses in CHI research.

To facilitate exploration, we made simplicity and ease of use our top priority. For the templates, this meant 1) minimizing the amount of code-writing necessary and 2) providing the minimum amount of background information on Bayesian concepts necessary for users to complete an analysis. Prioritizing simplicity allows users to complete a Bayesian analysis quickly and easily, even with no prior Bayesian knowledge.

This strategy is consistent with the principles of active learning. Instead of asking for a large up-front investment of effort into learning Bayesian concepts, the templates allow users to drop straight into the activity without any knowledge and then work backward. Even in one's first use, the templates simplify analysis enough that users are protected from many possible pitfalls, and will in most cases produce reasonable results even with no prior knowledge. For example, every user in our study produced results very similar to the original analysis, even when they made errors in prior setting. For users who want to investigate the conceptual background more deeply, the templates' step-by-step breakdown of the process better equips them to direct and focus their learning.

Evidence from the user study indicates the templates succeeded in being simple and easy to use. Most participants found the analyses easier to complete than they expected, and commented positively on how the template instructions were clear and easy to follow. Even people with no experience in R were able to complete the analyses successfully.

User study results also suggest the templates were largely successful in helping users understand the *process* of Bayesian analysis. Even those participants who were not confident in their specific analyses often reported they were more confident in their understanding of the process. They also appeared to have specific and directed goals for learning more, suggesting that the templates were able to help participants establish concrete goals for learning.

For users of the template who are interested in taking the next steps in learning more about Bayesian analysis, we recommend Richard McElreath's *Statistical Rethinking* [35]

and John K. Kruschke's *Doing Bayesian Analysis* [33], both excellent introductory textbooks for Bayesian analysis.

The user study also provided insight into barriers to adopting Bayesian methods. Across multiple iterations of the template, there was persistent confusion over how to set priors. Though we were able to address these issues in later iterations of the template, some participants in late rounds of testing still expressed some uncertainty about their choices of priors. One particularly promising aid was the check-your-understanding quizlet added in R3 of user testing. Though participant response was positive, not enough users were exposed to the quizlet in the user study to make confident conclusions about its usefulness in this template; however, similar quizzes deployed in the MOOC context have been demonstrated to improve learning outcomes [31].

Importantly, even participants who set reasonable priors often expressed some level of confusion or uncertainty over them. This suggests that many users may need additional guidance before they can set priors confidently. The amount of guidance needed may depend on the user: evidence from the user study suggests that people with a more extensive background in frequentist statistics were better prepared to understand priors using only the guidance from the template.

Overall, however, feedback from the user study indicates that the templates were largely successful in guiding Bayesian novices easily through their first analysis. It is important to build tools that lower the barrier for adoption of Bayesian methods at CHI. Bayesian analysis provides a formal method of knowledge accrual that is particularly powerful for small-n studies. For a field such as HCI, which produces many small-n studies of novel systems but publishes few meta-analyses [27], having such a tool is critical.

6 LIMITATIONS AND FUTURE WORK

The code templates we designed are inherently limited and cover only a portion of the many types of analyses that can be done using Bayesian methods. To keep analyses as simple as possible, we excluded analyses that would be useful to the CHI research community. As explained in "Supported analyses" in template design, this was the reason for excluding repeated-measures ANOVA, a common test in CHI papers. Repeated measures tests often require users to make different assumptions about how the data is generated, so choosing a Bayesian analysis post-hoc may introduce errors.

Also for the sake of simplicity, we require that priors are chosen from a normal distribution, which restricts the variety of possible shapes of priors participants might wish to be able to provide. In future work, we plan to expand the number of templates to include more, and more complex, types of analysis, and a wider variety of possible priors.

The choice to provide minimal conceptual background information also had tradeoffs. Though it makes the template

simple and easy to follow, for some users, the information provided in the template may not be sufficient for them to confidently complete an analysis. Some users may need to augment the template with external information in order to complete an analysis confidently, particularly in relation to setting priors. More research is necessary to determine how to best guide users confidently through prior-setting. Future research might benefit from results suggesting what types of elicitation techniques for subjective uncertainty (see, e.g., [41] for a review) best allow people of varying levels of background articulate a subjective distribution, such as graphical frequency-based and sample-oriented (i.e., HOPs-like) representations [15, 20, 29, 30].

The user study also had limitations. The study only tested three of the six available templates, as the three unused templates are extremely similar to the tested templates. In addition, the feedback measures in the survey were vulnerable to demand characteristics, possibly causing overly positive feedback. Though the questions were designed to minimize demand characteristics, and participants were reminded to be honest, some effect likely remained. Lastly, most participants chose weakly informative priors in the usability task. Thus, though the templates do provide guidance on how to set strong priors, the user study did not evaluate whether this guidance was adequate. In future work, we will investigate in more detail how to guide Bayesian novices through the process of setting both weakly and strongly informed priors.

More generally, there is a need for more empirical research to understand strategies for lowering the barrier to adopting Bayesian statistics. Though there is much work in statistics pedagogy (e.g. [5, 37]), most focuses on frequentist statistics; there has been little work done to identify the best strategies for teaching Bayesian methods.

7 CONCLUSION

Bayesian methods for statistical analysis have been gaining attention in CHI in recent years, but adoption of Bayesian methods has lagged behind. To address this gap between interest and adoption, we have presented a set of code templates that walk users through simple Bayesian analyses. The templates allow users to explore Bayesian statistics, and break down the process of a Bayesian analysis to help interested users better understand and direct their learning. Results from a user study demonstrate the templates' usability; further, they provide preliminary evidence to advance understanding of barriers to adoption of Bayesian methods.

8 SUPPLEMENTARY MATERIALS

Frozen versions of the templates are available in a zip file in the supplementary materials. The most updated version of the code templates, as well as documentation, are available at www.github.com/cdphelan/bayesian-template.

REFERENCES

- [1] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature News* 533, 7604 (2016), 452.
- [2] James O Berger and Donald A Berry. 1988. Statistical analysis and the illusion of objectivity. *American Scientist* 76, 2 (1988), 159–165.
- [3] Charles C Bonwell and James A Eison. 1991. *Active Learning: Creating Excitement in the Classroom*. 1991 ASHE-ERIC Higher Education Reports. ERIC.
- [4] Kieth A Carlson and Jennifer R Winquist. 2011. Evaluating an active learning approach to teaching introductory statistics: A classroom workbook approach. *Journal of Statistics Education* 19, 1 (2011).
- [5] Ben-Zvi Dani and Garfield Joan. 2004. Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In *The challenge of developing statistical literacy, reasoning and thinking*. Springer, 3–15.
- [6] Zoltan Dienes. 2011. Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science* 6, 3 (2011), 274–290.
- [7] Alan Dix. 2017. Making Sense of Statistics in HCI: From P to Bayes and Beyond. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, New York, NY, USA, 1236–1239. <https://doi.org/10.1145/3027063.3027109>
- [8] Stéphane Doyen, Olivier Klein, Cora-Lise Pichon, and Axel Cleeremans. 2012. Behavioral priming: it's all in the mind, but whose mind? *PloS one* 7, 1 (2012), e29081.
- [9] Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI*. Springer, 291 – 330. https://doi.org/10.1007/978-3-319-26633-6_13
- [10] Jim Eison. 2010. Using active learning instructional strategies to create excitement and enhance learning. (2010).
- [11] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 144, 12 pages. <https://doi.org/10.1145/3173574.3173718>
- [12] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. In *Conference on Human Factors in Computing Systems - CHI '18*. <https://doi.org/10.1145/3173574.3173718>
- [13] Scott Freeman, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences* 111, 23 (2014), 8410–8415. <https://doi.org/10.1073/pnas.1319030111> arXiv:<http://www.pnas.org/content/111/23/8410.full.pdf>
- [14] Gerd Gigerenzer and Ulrich Hoffrage. 1995. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review* 102 (1995), 684–704.
- [15] Daniel G Goldstein and David Rothschild. 2014. Lay understanding of probability distributions. *Judgment & Decision Making* 9, 1 (2014).
- [16] Richard R Hake. 1998. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American journal of Physics* 66, 1 (1998), 64–74.
- [17] L Hasher and R T Zacks. 1984. Automatic processing of fundamental information: the case of frequency of occurrence. *The American psychologist* 39, 12 (1984), 1372–1388. <https://doi.org/10.1037/0003-066X.39.12.1372>
- [18] Rink Hoekstra, Richard D. Morey, Jeffrey N. Rouder, and Eric-Jan Wagenmakers. 2014. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review* 21, 5 (01 Oct 2014), 1157–1164. <https://doi.org/10.3758/s13423-013-0572-3>
- [19] U. Hoffrage and G. Gigerenzer. 1998. Using natural frequencies to improve diagnostic inferences. *Academic Medicine: Journal of the Association of American Medical Colleges* 73, 5 (May 1998), 538–540.
- [20] Jessica Hullman, Matthew Kay, Yea-Seul Kim, and Samana Shrestha. 2018. Imagining Replications: Graphical Prediction & Discrete Visualizations Improve Recall & Estimation of Effect Uncertainty. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* (2018). <http://idl.cs.washington.edu/papers/imagining-replications>
- [21] Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one* 10, 11 (2015), e0142444.
- [22] John PA Ioannidis. 2005. Why most published research findings are false. *PLoS medicine* 2, 8 (2005), e124.
- [23] Yvonne Jansen and Kasper Hornbæk. 2018. How Relevant Are Incidental Power Poses for HCI?. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 14, 14 pages. <https://doi.org/10.1145/3173574.3173588>
- [24] Alex Kale, Francis Nguyen, Matthew Kay, and Jessica Hullman. 2018. Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data. *IEEE transactions on visualization and computer graphics* (2018).
- [25] Maurits Kaptein and Judy Robertson. 2012. Rethinking Statistical Analysis Methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1105–1114. <https://doi.org/10.1145/2207676.2208557>
- [26] Matthew Kay, Tara Kola, Jessica Hullman, and Sean Munson. 2016. When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI '16)*.
- [27] Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. 2016. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4521–4532. <https://doi.org/10.1145/2858036.2858465>
- [28] Matthew Kay, Shwetak N. Patel, and Julie A. Kientz. 2015. How Good is 85%?: A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 347–356. <https://doi.org/10.1145/2702123.2702603>
- [29] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2017. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 1375–1386.
- [30] Yea-Seul Kim, Logan Walls, Peter Krafft, and Jessica Hullman. 2019. A Bayesian Cognition Approach to Improve Data Visualization. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019).
- [31] Kenneth R Koedinger, Jihee Kim, Julianna Zhuxin Jia, Elizabeth A McLaughlin, and Norman L Bier. 2015. Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In *Proceedings of the second (2015) ACM conference on learning@ scale*. ACM, 111–120.
- [32] John K Kruschke. 2010. What to believe: Bayesian methods for data analysis. *Trends in cognitive sciences* 14, 7 (2010), 293–300.
- [33] John K. Kruschke. 2015. *Doing Bayesian Data Analysis: A Tutorial with R and BUGS* (2nd ed.). Academic Press, Inc., Orlando, FL, USA.

- [34] Jisoo Lee, Erin Walker, Winslow Burleson, Matthew Kay, Matthew Buman, and Eric B. Hekler. 2017. Self-Experimentation for Behavior Change: Design and Formative Evaluation of Two Approaches. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17). ACM, New York, NY, USA, 6837–6849. <https://doi.org/10.1145/3025453.3026038>
- [35] Richard McElreath. 2016. Statistical Rethinking: A Bayesian Course with Examples in R and Stan (1st ed.). CRC Press, Boca Raton, FL, USA.
- [36] Daniel McNeish. 2016. On using Bayesian methods to address small sample problems. Structural Equation Modeling: A Multidisciplinary Journal 23, 5 (2016), 750–773.
- [37] David S Moore. 1997. New pedagogy and new content: The case of statistics. International statistical review 65, 2 (1997), 123–137.
- [38] Carol Moser, Chanda Phelan, Paul Resnick, Sarita Y Schoenebeck, and Katharina Reinecke. 2017. No such thing as too much chocolate: evidence against choice overload in e-commerce. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM, 4358–4369.
- [39] Manuel Nordhoff, Tal August, Nigini A Oliveira, and Katharina Reinecke. 2018. A Case for Design Localization: Diversity of Website Aesthetics in 44 Countries. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, 337.
- [40] M.W. Oakes. 1986. Statistical inference: a commentary for the social and behavioural sciences. Wiley. <https://books.google.com/books?id=OhFHAAAAMAAJ>
- [41] Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. 2006. Uncertain judgements: eliciting experts' probabilities. John Wiley & Sons.
- [42] David R Sokoloff and Ronald K Thornton. 1997. Using interactive lecture demonstrations to create an active learning environment. The Physics Teacher 35, 6 (1997), 340–347.
- [43] Gail M Sullivan and Richard Feinn. 2012. Using effect size - or why the P value is not enough. Journal of graduate medical education 4, 3 (2012), 279–282.
- [44] Fengpeng Yuan, Xianyi Gao, and Janne Lindqvist. 2017. How Busy Are You?: Predicting the Interruptibility Intensity of Mobile Users. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17). ACM, New York, NY, USA, 5346–5360. <https://doi.org/10.1145/3025453.3025946>