Journal of User Experience

Vol. 21, Issue 1, November 2025 pp. 1–6

Beyond *P*-Values: Bayesian Approaches for User Experience Research

Mohsen Rafiei
Assistant Professor of Psychology
University of Arkansas at Little Rock
2801 S University Ave
Little Rock, AR, USA
mrafiei@ualr.edu

Iman Tahamtan
Lecturer of User Experience,
University of Tennessee, Knoxville
1345 Circle Park Dr.
Knoxville, TN, USA
tahamtan@vols.utk.edu

Abstract

Null hypothesis significance testing (NHST), using *p*-values and confidence intervals, has long been the standard in user research, particularly in large-sample settings like A/B testing. However, user experience studies often rely on smaller samples, rapid iterations, and design-driven outcomes, in which *p*-values can be difficult to interpret, and confidence intervals may offer limited practical guidance. This paper introduces Bayesian statistics as a complementary framework better suited to these conditions. Unlike the frequentist view, which treats parameters (such as satisfaction score) as fixed but unknown quantities—meaning there is one true value in the population that doesn't change—Bayesian methods treat parameters as uncertain and represent them through probability distributions, indicating which values are plausible given the data and any prior knowledge. Bayesian methods enable direct probability statements about parameters, integration of prior knowledge, and more interpretable results that align with iterative UX practices. In this paper, we introduce key Bayesian tools, such as Bayes factors and credible intervals, as more informative alternatives to *p*-values and confidence intervals that make it easier to compare different models and express uncertainty in a way that is more useful for iterative design decisions. Advantages include robustness with small samples (when using appropriately informative priors), flexibility in handling hierarchical models (for example, data in which tasks are nested within users or users are nested within groups), handling missing data (by estimating values from the posterior under assumed missingness), and decision-readiness in design contexts. By reframing statistical inference around probability, evidence, and prior knowledge, Bayesian methods provide UX researchers a more transparent and practical toolkit for guiding design decisions.

Keywords

Bayesian statistics, UX, statistical significance, Bayes factors, credible intervals, *p*-values

Background

In user research, null hypothesis significance testing (NHST) with *p*-values has long been the standard for assessing whether a statistically significant difference exists between two conditions. However, researchers, including UX researchers, often misuse *p*-value to decide which condition is better. It is important to note that *p*-values do not indicate anything about the presence or absence of a true effect or difference (Goodman, 2008); they merely reflect "the probability … that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value" (Wasserstein & Lazar, 2016, p. 131). Therefore, using *p*-values alone to judge the superiority of one condition, like a design, over another is inaccurate and conceptually flawed. Some researchers may rely on other statistics too, such as confidence intervals and effect sizes, which provide more nuanced and informative insights for decision-making than *p*-values. *P*-values, confidence intervals, and effect sizes fall under the umbrella of frequentist statistics.

Frequentist statistics are highly effective in many applied fields and are especially well-suited for large-sample experiments, such as online A/B tests, in which abundant data and opportunities for replication make them powerful and reliable. However, user research often works under different constraints. Studies may rely on relatively small sample sizes, rapidly evolving prototypes, and a need for the findings to directly inform design decisions. In these settings, *p*-values become more difficult to interpret: a non-significant result might stem from low statistical power rather than the absence of an effect, yet a statistically significant result might be misinterpreted as evidence of practical importance, even when the observed difference is trivial and negligible.

Bayesian statistics provide a complementary approach to frequentist statistics and are well-suited to the field of UX. The key difference lies in how each framework describes parameters, such as task completion time or satisfaction scores. In the frequentist view, parameters are considered fixed but unknown quantities. For example, it considers one true average satisfaction score for the population, and the data serve as samples used to estimate that fixed but unknown value (Greenland et al., 2016). In contrast, the Bayesian view treats parameters as uncertain and represents that uncertainty with a probability distribution (Bolstad & Curran, 2016). A probability distribution is a mathematical description of how plausible different parameter values are, given the observed data and any prior knowledge. In practice, rather than indicating that "the true average task time is exactly 60 s" Bayesian analysis might indicate "the task time could be between 55 and 70 s, with some values more likely than others." This shift in perspective has important implications for how we interpret statistical results and make inferences and decisions about the data (Gelman et al., 2013).

This shift leads to results that align more naturally with design decisions. For example, Bayesian analysis can provide a statement such as "there is a high probability that the true average satisfaction falls within this range," which directly communicates the likelihood of outcomes, rather than whether a difference exceeds an arbitrary threshold of statistical significance (typically ranging from 0.10 to 0.01 in the social sciences). This approach is grounded on Bayes' theorem, which provides a method for updating beliefs as new evidence becomes available by integrating prior information with new observed data (Bolstad & Curran, 2016; McElreath, 2018). For instance, if past usability studies show that users typically complete a checkout flow in about 90 s, this information can be used as a prior distribution when evaluating a new checkout flow. Even with a small sample, the Bayesian model integrates what is already known with what has just been observed, producing more stable and interpretable estimates. This ability to integrate prior knowledge with new data is a major advantage, especially in UX research in which rapid iteration and small sample sizes are common (Lee & Wagenmakers, 2014). By expressing uncertainty through probabilities, Bayesian methods align well with the realities of iterative UX research, while NHST remains useful when studies involve large samples and repeated replications.

The advantages of Bayesian statistics in user research are numerous. In addition to providing a natural and principled way to combine prior information with observed data (Kruschke, 2018), Bayesian inference conditions rely on the observed data and a specified model plus prior in order to form a posterior distribution. The posterior refers to the updated probability of a hypothesis or parameter after observing new data (Gelman et al., 2013). For example, if you

test a new design with a small group of users, Bayesian methods let you update your conclusions directly based on their responses, without requiring a large sample or relying on repeated replications (thus avoiding reliance on *p*-values or confidence intervals in which interpretation depends on long-run repeated sampling). In addition, the results from Bayesian statistics are more interpretable. For example, Bayesian inference allows researchers to say that "the true parameter, such as time on task, has a 95% probability of falling within a given credible interval"—a direct probability statement about the parameter, unlike frequentist confidence intervals. In a frequentist confidence interval, it would be interpreted as "if we were to repeat the study many times, about 95% of those intervals would likely contain the true parameter" (Hazra, 2017).

Bayesian methods are especially useful for handling complex data structures and hierarchical data (for example, users nested within groups or teams, or repeated measures from the same users across multiple tasks or time points) and for handling any missing data. For instance, when analyzing satisfaction scores across multiple groups, if some groups have fewer responses, a Bayesian model can borrow strength from groups with more data to improve estimates for smaller groups and still account for uncertainty. Similarly, Bayesian approaches can handle missing responses without necessarily needing to drop participants or fill in missing values with the mean or last observation, which are methods that can distort variability and introduce bias. One limitation, however, is the need to choose priors carefully, especially when data are limited. Moreover, Bayesian analysis also tends to require more computational power and statistical expertise than traditional frequentist methods (Dienes, 2016).

Bayes Factors: A More Informative Alternative to *P*-Values

The Bayes factor (BF) is a cornerstone of Bayesian hypothesis testing and offers conceptual and practical advantages over traditional *p*-values. It offers a way to compare two hypotheses using Bayesian statistics. It indicates how much more likely the data are under one hypothesis than another. BF is the ratio of the likelihood of the data under two competing hypotheses (usually a null hypothesis versus alternative hypothesis); BF quantifies the strength of evidence in favor of one model over another (Kass & Raftery, 1995). Unlike *p*-values, which only indicate whether to reject or not reject the null hypothesis, the BF allows researchers to directly compare models (that is, competing hypotheses). For example, one model might state that "users in group A and group B complete tasks in the same amount of time," whereas another model might state that "users in group A complete tasks faster than users in group B." The BF indicates how much more (or less) the data support one model over the other, providing evidence for, or against, both hypotheses.

The essential difference between BFs and *p*-values is in hypothesis assessment. A *p*-value indicates the probability of observing data as extreme or more extreme than what was observed, assuming the null hypothesis is true (Wasserstein & Lazar, 2016). In contrast, a BF assesses how well each hypothesis predicts the existing data, offering a fair comparison between competing explanations, rather than just trying to reject one (Rouder et al., 2009). This distinction is particularly important in user research, in which distinguishing between "lack of evidence," meaning there is insufficient data to draw conclusions, and "evidence for a lack of effect," which indicates that the data actively support the absence of an impact, is crucial (Dienes, 2016).

BFs are helpful because they don't depend on repeated sampling. They are robust and allow early interpretation, providing clarity regarding how strongly the data supports both the null and alternative hypotheses (Wagenmakers, 2007). Recent reviews also suggest that using BFs to monitor data collection can lower research costs and minimize participants' unnecessary exposure to study tasks (by allowing studies to stop early once sufficient evidence has been gathered), while maintaining the validity and reliability of statistical inferences (Heck et al., 2023).

Credible Intervals and Their Interpretation

In Bayesian statistics, a 95% credible interval means there is a 95% posterior probability that the parameter lies in the interval, conditional on the model and prior (Kruschke, 2018). This is meaningfully different from the frequentist confidence interval, which reflects the range of intervals that would likely contain the true population parameter, with some uncertainty, if the

study were repeated many times (Hazra, 2017). Therefore, Bayesian credible intervals align better with how researchers naturally think about uncertainty by focusing on what values are most plausible given the data (Morey & Rouder, 2018).

Credible intervals come from the posterior distribution, which represents the updated beliefs about a parameter after combining prior knowledge with the data we have collected (Gelman et al., 2013). This makes them more informative, especially in behavioral research situations where prior information or theoretical expectations are strong (Lee & Wagenmakers, 2014). Bayesian methods, when using informative priors, can yield credible intervals that are narrower and more stable than frequentist confidence intervals, particularly in small-sample settings (Ly et al., 2016). In UX research, in which prior information is often limited, knowledge gained across iterative studies can be incorporated to refine subsequent analyses. As a result, Bayesian credible intervals can provide clearer, more decision-ready estimates of outcomes such as task success rates or satisfaction scores, even when studies involve relatively few participants.

Bayesian credible intervals are now used in many areas, such as individualized therapy analysis (such as how a specific person responds to therapy over time) and cognitive modeling (such as creating mathematical models that simulate how people think, learn, and make decisions) (Wagenmakers et al., 2011). Their intuitive interpretability and the capacity to directly quantify the probability of parameters promote transparent and robust reporting. With the rising accessibility of tools such as JASP and Bayes Factor for R (van Doorn et al., 2021; Morey & Rouder, 2018), Bayesian credible intervals are expected to also become even more prevalent for informed decision-making in UX contexts.

Conclusion

Although NHST can highlight whether results are unlikely under a null hypothesis, it often leaves UX researchers with ambiguities by pointing out what not to believe, without offering clear guidance on what the collected data do support. Bayesian methods shift the focus from statistical significance toward decision-readiness, that is, clearer guidance for making informed design decisions. Bayes factors allow researchers to directly compare competing hypotheses, and credible intervals provide intuitive probability statements about key UX metrics, such as satisfaction ratings. Because Bayesian analysis can borrow strength across groups, handle missing data gracefully, and incorporate prior knowledge from past studies, it aligns naturally with the iterative and resource-constrained nature of UX research. For practitioners, this means more interpretable results, clearer communication with stakeholders, and the ability to make design choices confidently even with modest sample sizes.

References

Bolstad, W. M., & Curran, J. M. (2016). *Introduction to Bayesian statistics* (3rd ed.). Wiley.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology, 72*, 78–89. https://doi.org/10.1016/j.jmp.2015.10.003

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC.

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology, 45*(3), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Hazra, A. (2017). Using the confidence interval confidently. *Journal of Thoracic Disease, 9*(10), 4125–4130. https://doi.org/10.21037/jtd.2017.09.14

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P. C., Derks, K., Dienes, Z., … Hoijtink, H. (2023). A review of applications of the Bayes factor in psychological research. *Psychological Methods, 28*(4), 558–579. https://doi.org/10.1037/met0000483

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

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

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge University Press.

Ly, A., Verhagen, J., & Wagenmakers, E. J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology, 72*, 19–32. https://doi.org/10.1016/j.jmp.2015.06.004

McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan.* Chapman & Hall/CRC.

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes factors for common designs* (Version 0.9.12-4.2) [R package]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=BayesFactor

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

van Doorn, J., van den Bergh, D., Boehm, U., Dablander, F., Derks, K., Draws, T., … Wagenmakers, E. J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review, 28*, 813–826. https://doi.org/10.3758/s13423-020-01798-5

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & Van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology, 100*(3), 426–432. https://doi.org/10.1037/a0022790

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

About the Authors

Mohsen Rafiei, PhD
Dr. Rafiei leads Quantitative UX Research at the Perceptual User Experience (PUX) Lab, working closely with industry partners to bring scientific rigor to real-world design challenges. He is also the director of the Experience Lab at the University of Arkansas at Little Rock, where he serves as an Assistant Professor of Psychological Science.

Iman Tahamtan, PhD
Dr. Tahamtan is a UX lecturer at the University of Tennessee, Knoxville (UTK). His research focuses on analyzing user behavior, needs, and challenges in interactions with digital technologies to enhance their design, usability, and accessibility. He also serves on the editorial board of the *Journal of User Experience* (*JUX*).