The shortcomings of p-values and confidence intervals and how Bayesian alternatives save the day

Ksenia Lepikhina
University of Colorado, Boulder
STAT 5700: Philosophy of Statistics
Professor: Brian Zaharatos

November 13, 2019

The goal of this paper is to take a stance against the proposal that confidence intervals should be used instead of p-values and argue that the Bayesian counterparts should be used instead. The Philosophy of Statistics course at CU Boulder has raised numerous issues pertaining to p-values that do not seem redeemable. Many [1] discuss the shortcomings of p-values and suggest confidence intervals as alternatives but fail to elaborate on the issues with confidence intervals with respect to comparable statistics in Bayesian inference. This paper will start by defining p-values and confidence intervals and then describe why statistics has historically frequently used both. It will then counter by describing the issues with the two frequentist concepts. Finally, it will propose why the Bayesian alternatives should be preferred.

To begin, a p-value is a probability derived from a hypothesis test. Given a null hypothesis and an alternative hypothesis, a test statistic can be computed which describes the difference between the characteristics of the observed data and the null hypothesis. This test statistic can be represented as a p-value, a probability, using the probability distribution of the test statistic under the null hypothesis [2]. A p-value can be compared to a pre-decided significance level, typically represented as α (usually $\alpha = 0.05$), in order to make conclusions about the null hypothesis. If a p-value is less than or equal to the significance level, the null hypothesis can be rejected and the result is statistically significant. If a p-value is greater than the significance level, then we fail to reject the null hypothesis. At this point, it is important to note that failing to reject the null hypothesis is not the same as accepting the null hypothesis. Distinguishing these terms is important because a hypothesis test does not determine which hypothesis is correct but rather tests to see if the available evidence and data serve to reject the null hypothesis.

A confidence interval is an interval estimation of a population parameter computed with a test statistic. The true interpretation of a confidence interval is that it is an interval such that over repeated samples from the population, the true parameter will lie within the confidence intervals $x\%$ of the time. It describes the uncertainty associated with a certain

sampling method. Confidence intervals are derived by adding and subtracting some sort of error (critical value multiplied with standard deviation of the sample statistic) from a sample statistic. Here, it is important to draw attention to the meaning of a confidence interval. In frequentist statistics, the population parameter stays constant and thus when a confidence interval is calculated, the population parameter is fixed (but unknown), and the probabilistic statement is made about the interval.

P-values have historically been the most common statistical tool for drawing conclusions in scientific research. In August 2019, Google Scholar returned up to 2.85 million citations that included the phrase "statistically significant" [3]. Part of the reason for the frequent use of p-values is because they are easy to understand (if interpreted properly). P-values also allow the statistician to assign a strength for the evidence against the null hypothesis. The purpose of a p-value it to indicate whether or not randomness has been eliminated as an explanation for a certain result. P-values certainly serve their purpose in statistics for that reason, however issues arise when their role in science and statistics is misinterpreted and inflated.

P-values are frequently misused and misinterpreted which causes them to often be subjected to "scientific misconduct" [1]. One issue is that p-values are misinterpreted when a hypothesis is not defined beforehand and a "significant" result is found by data dredging (finding patterns in data that find a significant result and then defining a null hypothesis). Another problem with p-values is the issue of multiple comparisons — this occurs when multiple statistical tests are performed simultaneously which reduces the value of a statistical finding [1]. When multiple tests are performed, some will have p-values less than the predefined significance levels and some will be greater causing an increase in false positives. P-values are also affected by the number of observations in a sample meaning that a phenomena present in the population might not be observed as significant in a small sample and a phenomena observed in a large enough sample may appear significant when it isn't in the population. [1]

Given the shortcomings of p-values, confidence intervals have frequently been suggested as alternatives [1][4][5]. The primary argument for confidence intervals as a superior tool is that "confidence intervals offer more information than significance tests." [4] One of the ways they provide information is by presenting the results (the actual intervals) at the level of the data [5]. The intervals are on the same scale as the data meaning that we can directly compare our data with the interval. While p-values provide info about the statistical significance, confidence intervals provide the same as well as information about how well our statistical test worked. For example, a wide confidence interval could suggest a poor test. Additionally, compared to a point estimate (like a p-value), confidence intervals provide a range of values for the population and also provide a probability of the interval covering the true value [5].

Similar to p-values, confidence intervals also have their deficiencies. For one, they are also quite easy to misinterpret. A common misinterpretation of a confidence interval is: "There is an $x\%$ chance that the true population parameter falls within the confidence interval." The problem with this interpretation is that it tries to assign a probability to the location of the true parameter even though the true parameter is not a random variable. The true parameter does not change and thus we can not make any probabilistic statements about where it will lie within the interval. To reiterate, the true interpretation of a confidence interval is that over repeated samples, the true parameter will lie within the confidence intervals $x\%$ of the time. Given that the two interpretations appear quite similar, it is easy to make mistakes.

In addition to the misinterpretations of confidence intervals, there are a couple of arguments presented by Morey et al. [6] on other deficiencies. The fallacies brought up in the paper include that knowing the proper definition of a confidence interval does not tell us about the inferences that can be made from a specific interval, and that confidence intervals do not necessarily contain likely values of the parameter [6]. The first raises the issue that knowing a confidence interval does not mean that moving beyond statistical inference and into practical inference is clear. The second argument raises the issue that often times, for a small interval, we get high precision with poor accuracy meaning that the confidence interval

actually contains some impossible values [6]. With a wide confidence interval, we will likely have a high accuracy (imagine using the whole number line as your confidence interval), but our precision will be incredibly low [6].

A potential solution to the problems with p-values and confidence intervals is Bayesian inference and its alternatives to those frequentist statistical tools. The Bayesian alternative to p-values are called Bayes factors [7] and the alternative to confidence intervals are called credible intervals. Bayes factors are the ratio of likelihoods between two hypotheses. The argument in favor of Bayes factors is that the ratio of likelihoods adds a method for *accepting* a null hypothesis (as opposed to hypothesis testing — rejecting or failing to reject the null). A Bayes factor can be any positive number; a large number suggests high evidence for the alternative hypothesis, a small number suggests high evidence for the null hypothesis, and 1 suggests that there is no evidence for one hypothesis over the other. Unlike p-values, with Bayes factors, a statement of the type "the null hypothesis is more strongly supported by the data than the alternative hypothesis" is valid. While Bayes factors do not avoid the issues of data dredging and multiple comparisons [8], they at least resolve the interpretation issue as well as allow for the acceptance of a null hypothesis.

A credible interval is an interval in which a parameter falls with a specified probability. The definition of a credible interval is often how confidence intervals are misinterpreted. Arguably, it is more intuitive to specify a probability for a parameter (credible) than an interval (confidence). The difference is that Bayesian inference fixes the interval and guarantees that $x\%$ of the time the parameter will fall within the credible interval (the parameter is a random variable from a distribution) and frequentist inference fixes the parameter and guarantees that $x\%$ of the time, the confidence interval will cover the parameter [9]. Like p-values and Bayes factors, credible intervals are easier to interpret than confidence intervals and are therefore easier to use practically. Credible intervals also resolve the issue (Morey et al. [6]) that it is not clear how to make inference from confidence intervals. From a credible interval, a specific claim about the parameter can be made. The other issue Morey et al. [6]

raise is that confidence intervals do not necessarily contain likely values of the parameter. Yet again, credible intervals solve this issue. The credible interval strictly contains the most likely values of the parameter.

Given the analysis in this paper, it is clear that both p-values and confidence intervals have their drawbacks. The most significant flaw with these frequentist concepts is that they are frequently misinterpreted. While Bayes factors and credible intervals share some of the shortcomings of p-values and credible intervals, they largely resolve the interpretation issue. Bayes factors allow for the acceptance of a null hypothesis rather than failing to reject it which is much clearer as it avoids issues of double negatives. Credible intervals allow a statistician to make probabilistic statements about a parameter within an interval rather than a probabilistic statements about the interval. Bayesian alternative are, perhaps, the best "cultural alternative" to p-values.

References

[1] Jonas Ranstam. *Why the P-value culture is bad and confidence intervals a better alternative*. 2012.

[2] David Jean Biau, Brigitte M Jolles, and Raphaël Porcher. "P value and the theory of hypothesis testing: an explanation for new researchers". In: *Clinical Orthopaedics and Related Research®* 468.3 (2010), pp. 885–892.

[3] Aubrey Clayton. *The Flawed Reasoning Behind the Replication Crisis - Issue 74: Networks*. Aug. 2019. URL: http://nautil.us/issue/74/networks/the-flawed-reasoning-behind-the-replication-crisis.

[4] Eduard Brandstätter and Johannes Kepler. "Confidence intervals as an alternative to significance testing". In: *Methods of Psychological Research Online* 4.2 (1999), pp. 33–46.

[5] Jean-Baptist du Prel, Gerhard Hommel, Bernd Röhrig, et al. "Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications". In: *Deutsches Ärzteblatt International* 106.19 (2009), p. 335.

[6] Richard D Morey, Rink Hoekstra, Jeffrey N Rouder, et al. "The fallacy of placing confidence in confidence intervals". In: *Psychonomic bulletin & review* 23.1 (2016), pp. 103–123.

[7] Robert E Kass and Adrian E Raftery. "Bayes factors". In: *Journal of the american statistical association* 90.430 (1995), pp. 773–795.

[8] Christian P Robert. "The expected demise of the Bayes factor". In: *Journal of Mathematical Psychology* 72 (2016), pp. 33–37.

[9] Jake VanderPlas. *Frequentism and Bayesianism III: Confidence, Credibility, and why Frequentism and Science do not Mix*. Wrote multiple statistics and python textbooks. URL: http://jakevdp.github.io/blog/2014/06/12/frequentism-and-bayesianism-3-confidence-credibility/.