The Fallacy of Placing Confidence in Confidence Intervals

Richard D. Morey and Rink Hoekstra
University of Groningen

Jeffrey N. Rouder
University of Missouri

Michael D. Lee
University of California-Irvine

Eric-Jan Wagenmakers
University of Amsterdam

Interval estimates – estimates of parameters that include an allowance for sampling uncertainty – have long been touted as a key component of statistical analyses. There are several kinds of interval estimates, but the most popular are confidence intervals (CIs): intervals that contain the true parameter value in some known proportion of repeated samples, on average. The width of confidence intervals is thought to index the precision of an estimate; the parameter values contained within a CI are thought to be more plausible than those outside the interval; and the confidence coefficient of the interval (typically 95%) is thought to index the plausibility that the true parameter is included in the interval. We show in a number of examples that CIs do not necessarily have any of these properties, and generally lead to incoherent inferences. For this reason, we recommend against the use of the method of CIs for inference.

DRAFT

"You keep using that word. I do not think it means what you think it means."

Inigo Montoya, *The Princess Bride* (1987)

The development of statistics over the past century has seen the proliferation of methods designed to make inferences from data. Methods vary widely in their philosophical foundations, the questions they are supposed to address, and their frequency of use in practice. One popular and widely-promoted class of methods are interval estimates, which include frequentist confidence intervals, Bayesian credible intervals and highest posterior density (HPD) intervals, fiducial intervals, and likelihood intervals. These procedures differ in their philosophical foundation and computation, but informally are all designed to be estimates of a parameter that account for measurement or sampling uncertainty by yielding a range of values for the parameter instead of a single value.

Of the many kinds of interval estimates, the most popular is the confidence interval (CI). Confidence intervals are introduced in almost all introductory statistics texts; they are recommended or required by the methodological guidelines of many prominent journals (e.g., Psychonomics Society, 2012; Wilkinson & the Task Force on Statistical Inference, 1999); and they form the foundation of proposed methodological reformers' programs (Cumming, 2014; Loftus, 1996). However, there is a tremendous amount of confusion in among researchers, methodologists, and textbook authors about what exactly a confidence interval is and how it may be interpreted. We believe that the confusion about what CIs are drives their promotion; if researchers understood what CIs actually are, and what inferences that can or cannot support, they would not be promoted as commonly as they are. Our goal is to alleviate confusion about CIs and to call into question whether they can be used for sound inference.

We begin by precisely defining confidence intervals. We then outline three common myths about confidence interval that have been perpetuated by proponents of confidence intervals. Using several examples, we show how it is not necessary that confidence intervals have any of the properties commonly ascribed to them; that is, confidence intervals, as general inference tools, have been misrepresented. Finally, we discuss methods that – under certain assumptions – do have the properties that researchers desire.

Confidence Intervals

In a classic paper, Neyman (1937) laid the formal foundation for confidence intervals. Before defining confidence intervals, we describe the practical problem that Neyman saw confidence intervals as solving. Suppose a researcher is interested in estimating a parameter, which we may call θ. This parameter could be a population mean, an effect size, a variance, or any other quantity of interest. Neyman suggests that researchers perform the following three steps:

a. Perform an experiment, collecting the relevant data.

b. Compute two numbers – the smaller of which we can call L, the greater of which U – forming an interval (L, U) according to an algorithm.

7698 words. Address correspondence to Richard D. Morey (richarddmorey@gmail.com). We thank Mijke Rhemtulla for helpful discussion during the drafting of the manuscript.

c. State that $L < \theta < U$ – that is, that θ is in the interval.

This recommendation is justified by choosing an algorithm for step (b) such that in the long run, the researcher's claim in step (c) will be correct, on average, $X\%$ of the time. A confidence interval is any interval computed using such a procedure.

Definition 1 (Confidence interval) *A $X\%$ confidence interval for a parameter θ is an interval (L, U) generated by an algorithm that in repeated sampling has an $X\%$ probability of containing the true value of θ (Neyman, 1937).*

Although skepticism about the usefulness of confidence intervals began as soon as Neyman laid out the theory (e.g., the discussion of Neyman, 1934)[1], confidence intervals have grown in popularity to be the most widely used interval estimators. Perhaps the most commonly used CI is the CI for the mean of the normal distribution with unknown variance.

$$\bar{x} \quad \pm \quad t^* \frac{s}{\sqrt{N}} \tag{1}$$

where $\bar{x}$ and s are the sample mean and standard deviation, N is the sample size, and t^* the quantile from Student's t_{N-1} distribution chosen such that

$$Pr(|t_{N-1}| < |t^*|) = X\%$$

for an $X\%$ CI. This confidence interval is taught to first-year statistics students all over the world and used in papers throughout the scientific literature.

Figure 1 shows 200 random 50% confidence intervals, all constructed from draws of sample size $N = 2$ from the same Normal(100,15^2) population. Some of the confidence intervals, denoted by dark lines, include the true mean $\theta = 100$; others do not and are denoted with light lines. Of this sample of 200 confidence intervals, 107 (53.5%) contain the true value; if we were to continue sampling more CIs, this proportion would approach the confidence coefficient $X = 50\%$. Note that with two observations $t^* = 1$ and $s = |x_1 - x_2|/\sqrt{2}$, meaning that the 50% Student's t CI is simply the interval between the two observations. We make use of the simplicity of the 50% CI with two observations throughout.[2]

The definition of a confidence interval seems, on its face, straightforward: a CI is an interval generated by some procedure that creates intervals containing the true value of a parameter in some fixed proportion of repeated samples, on average. Put another way, if one were to always make the dichotomous claim that the true value *is* in a specific interval computed from the procedure, one would be correct in that same proportion of repeated samples. "Confidence" is thus an average property of a procedure. It is conceptually helpful, therefore, to distinguish between a confidence *procedure* (CP) and a confidence *interval* (CI). For the purposes of this paper we will use the term confidence procedure to denote the algorithm, and confidence interval to denote specific realizations from the algorithm.

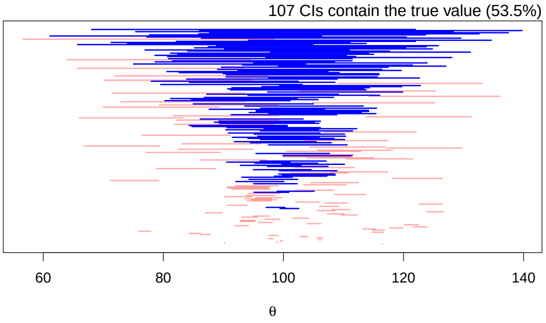

Figure 1. 200 random 50% CIs for a normal mean with unknown variance, based on 2 draws from a Normal(100, 15^2) distribution. Dark (blue) lines denote CIs that contain the true mean; light (red) lines denote those that do not. CIs are sorted by length.

Myths of Confidence

Confidence intervals are described broadly as tools for extracting the necessary information about the parameter from the data. However, the relationship between the definition of the confidence interval and anything a researcher would want to know is unclear: we might know the average properties of the procedure, but what implications does this have for inference from a specific interval? Various heuristic explanations are used by textbook authors and proponents of confidence intervals in the literature to help bridge the gap between the theoretical definition of the confidence interval and properties that are important to analysts, such as the plausibility of specific parameter values or the precision of an estimate. In this section, we explain how the various heuristic explanations of confidence intervals are actually myths: they are not true of confidence intervals in general. We present two examples that show how the logic of inference by CI fails.

Example 1: The lost submarine

[1] For instance, in this discussion Bowley states "Does [the confidence interval] really lead us towards what we need – the chance that in the universe which we are sampling the proportion is within these certain limits? I think it does not. I think we are in the position of knowing that either an improbable event has occurred or the proportion in the population is within the limits. To balance these things we must make an estimate and form a judgment as to the likelihood of the proportion in the universe [that is, a prior probability] – the very thing that is supposed to be eliminated."

[2] The more typical choice of 95% confidence is, of course, arbitrary. All of the points we make in this paper extend to 95% CIs and $N > 2$, with the drawback that the arguments would be more mathematically involved and less transparent. Because our goal is to demonstrate that the *logic* of using confidence intervals for inference is flawed, we opt to use the simplest case.

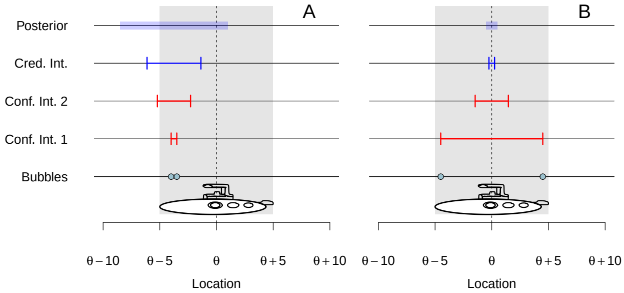

Figure 2. Submersible rescue attempts. See text.

A 10-meter-long research submersible with several people on board has lost contact with its surface support vessel. The submersible has a rescue hatch exactly halfway along its length, to which the support vessel will drop a rescue line. Because the rescuers only get one rescue attempt, it is crucial that when the line is dropped to the craft in the deep water that the line be as close as possible to this hatch. The researchers on the support vessel do not know where the submersible is, but they do know that it forms distinctive bubbles. These bubbles could form anywhere along the craft's length, independently, with equal probability, and float to the surface where they can be seen by the support vessel.[3]

The situation is shown in Figure 2A. The rescue hatch is the unknown location θ, and the bubbles can rise anywhere from $\theta-5$ meters (the bow of the submersible) to $\theta+5$ meters (the stern of the submersible). The rescuers want to use these bubbles to learn where the hatch is located, so they consult the frequentist statistician on board. The statistician, being an advocate of the use of confidence intervals, tells the rescuers how confidence intervals can be used to estimate the location of the hatch. He notes that the location of these first two bubbles form a 50% confidence interval for the θ, because there is a 50% probability that two bubbles will be on opposite sides of the hatch. In more familiar terms, the statistician has chosen the 50% confidence procedure

$$\bar{x} \pm \frac{|x_1 - x_2|}{2},$$

where x_1 and x_2 are the locations of the first and second bubbles, respectively, and $\bar{x}$ is the mean location. The statistician justifies this confidence procedure on the grounds that it is the same as the 50% Student's t procedure with $N = 2$. We denote this procedure Confidence Procedure 1 (CP1).

The rescuers see the first two bubbles, shown as circles in Figure 2A. The two bubbles are very close together, yielding the narrow 50% confidence interval shown above the bubbles. The statistician excitedly reports to the rescue workers that this narrow CI indicates that the knowledge of the hatch location is quite precise, and that he is 50% certain that the hatch is in the confidence interval. He advises them to drop the rescue line.

The statistician's opinion has been guided by the claims made by the advocates of CIs. The relationship between CIs and precision, or power, is often cited as one of the primary reasons they should be used over null hypothesis significance tests (e.g., Cumming & Finch, 2005; Cumming, 2014; Fidler & Loftus, 2009; Loftus, 1993, 1996). For instance, Cumming (2014) writes that "[l]ong confidence intervals (CIs) will soon let us know if our experiment is weak and can give only imprecise estimates," and Young and Lewis (1997) state that "[t]he width of the CI gives us information on the precision of the point estimate."

One of the rescue workers, however, is skeptical. She notes that there are a wide range of locations that are possible for the location of the hatch. Because the craft is 10 meters long, no bubble can originate more than 5 meters from hatch. Given two bubbles, the only possible locations for the hatch are within 5 meters of both bubbles. These values are shown as the shaded region in Figure 2A labeled "posterior". The bubbles themselves give no reason to prefer any of these locations over another. The skeptical rescue worker says that because the second bubble was so close to the first, collectively the bubbles have actually provided very *imprecise* information. She suggests waiting for more bubbles before dropping the line.

The statistician, on the authority of the many advocates of confidence intervals, convinces the rescuers to drop the line inside the 50% CI. The rescue line misses the hatch by 3.75 meters. By the time the rescue workers realize they have missed the hatch, there is no more time for another attempt. All crew on board the submersible are lost.

The statistician has fallen victim to a myth about confidence intervals that we dub the "precision error":

Myth 1 (The Precision Error) *"The width of a confidence interval indicates the precision of our knowledge about the parameter. Narrow confidence intervals show precise knowledge, while wide confidence errors show imprecise knowledge."*

There is no necessary connection between the precision of an estimate and the size of a confidence interval. In the case of the submarine, the narrow confidence interval yielded imprecise imprecise information about the location of the submarine hatch. In fact, as we shall see, the narrowness of an interval from CP1 and the precision are actually *inversely* related, as the next situation will make clear.

Oddly enough, a second boat is in a similar situation to the first half a world away. The researchers on the boat have contact with their submersible, and like the first boat are planning to mount a rescue using the distinctive bubbles. They consult their statistician, who advises them to wait for

[3] This example was adapted from Welch (1939) and Berger and Wolpert (1988).

the first two bubbles and use a 50% CI from Confidence Procedure 1. While the rescue team waits for the bubbles to appear, the statistician notes that CP1 does not directly make use of the known variance of the bubble locations. She quickly computes an alternative 50% confidence procedure, which we denote Confidence Procedure 2 (CP2):

$$\bar{x} \pm \left(5 - \frac{5}{\sqrt{2}}\right) \qquad (2)$$

This confidence procedure makes use of the fact that there is a 50% probability that $\bar{x}$ falls within $5 - 5/\sqrt{2} = 1.46$ meters of the hatch. The rescuers alert the statistician that the first two bubbles have been observed, as shown in Figure 2B. The bubbles are almost 10 meters apart.

The statistician computes the confidence intervals using CP1 and CP2. The interval from CP1 is nearly as wide as it can possibly be. At first, the statistician despairs, thinking that the first two bubbles have led to extremely imprecise information. But then the statistician notes that the width of the interval from CP2 never changes, confusingly suggesting that the width of the CI need not be function of precision in the data.

Faced with seemingly contradictory information from the two CIs about the precision of the estimate $\bar{x}$, the statistician decides to follow a different line of reasoning. Thinking her CP2 superior due to the fact that it directly used information about the width of the submersible, she reasons about the likely values for the location of the hatch. She believes that by virtue of being contained in her interval, the values inside the confidence interval should all be taken seriously as estimates of the hatch location. This second, widely-mistaken interpretation of intervals – that specific parameters in the interval are "likely" or "plausible" – we dub the likelihood error:

Myth 2 (The Likelihood Error) *"A confidence interval contains the likely values for the parameter. Values inside the confidence interval are more likely than those outside." This error exists in several varieties, sometimes involving plausibility, credibility, or reasonableness of beliefs about the parameter.*

Loftus (1996), for instance, says that the CI gives an "indication of how seriously the observed pattern of means should be taken as a reflection of the underlying pattern of population means." This logic is used when when confidence intervals are used to test theory (Velicer et al., 2008) or to argue for the null (or practically null) hypothesis (Loftus, 1996). The problem with the logic of the likelihood error is that a confidence interval may contain impossible values, or may exclude values that are just as plausible as ones inside the interval. For instance, the CI for the first submarine failed to contain most of the likely values, leading the statistician to believe his estimate was very precise.

The second statistician notes that the interval from CP2 is 2.93 meters wide. Thinking that all values within this interval are "likely", she advises risking waiting for a few more bubbles to narrow the confidence interval. However, a perceptive rescue worker challenges the statistician. He notes that they know very precisely where the hatch must be: halfway between the two bubbles! The bubbles were almost as far apart as they could be, which means they must have come from opposite ends of the craft. If they came from opposite ends of the craft, then the hatch must be almost exactly in the middle. The only possible locations for the hatch are shown in the top line of Figure 2B, labeled "posterior." In fact, nearly all of the values contained in the 50% CIs computed from CP1 and CP2 are impossible; the hatch could not possibly be located there.

The rescue worker advises to ignore the statistician and to drop the rescue line halfway between the bubbles. The argument of the rescue worker convinces the other rescue workers. The rescue line meets the hatch, and all crew members aboard the submersible are saved.

In the rescue attempts of both submarines, the statisticians used the same logic recommended by the advocates of confidence intervals in an effort to determine the location of the submarine hatch. In both cases, the judgment of the statistician was flawed. We now consider how the confidence interval could be so misleading. In the example, the likelihood and precision errors are easily seen; CIs can contain mostly impossible values, and the precision of an estimate available in the data can be inversely related to the narrowness of the CI, or not related at all. The logic of the likelihood and precision interpretations of CIs simply do not follow from the definition of a confidence procedure; some confidence intervals may have these properties, others may not.

If the precision and likelihood interpretations of confidence intervals are incorrect, what can we say about CIs? The definition of a confidence interval makes clear how to interpret a confidence procedure. However, when we compute a specific interval from the data and must interpret it, we are faced with difficulty. It is not obvious how to move from our knowledge of the properties of the confidence procedure to the interpretation of the confidence interval.

Textbook authors and proponents of confidence intervals bridge the gap seamlessly by claiming that the properties of confidence procedures can be applied to individual confidence intervals. For instance, Masson and Loftus (2003) state that "[t]he interpretation of the confidence interval constructed around that specific mean would be that there is a 95% probability that the interval is one of the 95% of all possible confidence intervals that includes the population mean. Put more simply, in the absence of any other information, there is a 95% probability that the obtained confidence interval includes the population mean." Cumming (2014) writes that "[w]e can be 95% confident that our interval includes [the parameter] and can think of the lower and upper limits as likely lower and upper bounds for [the parameter].

This interpretation, although seemingly natural, is incorrect. We dub this incorrect reasoning the "Fundamental Confidence Fallacy" (FCF). The FCF is fundamental because it seems to flow naturally from the definition of the confidence interval:

Myth 3 (The Fundamental Confidence Fallacy) *"If the probability that a random interval contains the true value is X%, then the plausibility (or probability) that a particular observed interval contains the true value is also X%."*

The reasoning behind the Fundamental Confidence Fallacy seems plausible: on a given sample, we could get any one of the possible confidence intervals. Because 95% of the possible confidence intervals contain the true value, without any other information it seems reasonable to say that we have 95% certainty that the true value is in our calculated confidence interval. This interpretation is suggested by the name "confidence interval" itself: the word "confident", in lay use, is closely related to concepts of plausibility and belief. The name "confidence interval" therefore invites the FCF as an interpretation.

The first hint that the FCF is a fallacy is the fact that, as we have seen, for any given problem there could be more than one confidence procedure. If the mere fact that a confidence procedure has a confidence coefficient of X% implied that any interval computed from that procedure has a X% probability of containing the true parameter value, then two intervals computed with the same data from two different 50% confidence procedures will both have a 50% probability of containing the true value.

Consider CP1 and CP2; both are centered around $\bar{x}$. If intervals from CP1 and CP2 both have a 50% probability of containing the true value, then the laws of probability require that all of the 50% probability be concentrated in the shorter of the two intervals – otherwise, the longer of the two intervals would have $>$ 50% probability of containing the true value. This seems to imply that if we have two procedures that are always nested within one another, as CP1 and CP2 are, then we can simply take the shorter of the two — whichever that is — as our 50% CI. But since both CP1 and CP2 are 50% CIs, then the procedure taking the shorter of the two must contain the true value less than 50% of the time, meaning it cannot be a 50% confidence procedure. The FCF leads to contradiction.

We do not need two confidence procedures to see why the FCF is a fallacy. In the second submarine scenario, a statistician using CP1 under the FCF would believe, on the basis that she has computed a 50% CI, that she is 50% certain that the true value is between the two bubbles. Considering the fact that bubbles cannot be more than 5 meters from the hatch, if the bubbles are more than 5 meters apart a CI computed from CP1 *must* contain the hatch location. All CIs from CP1 that are wider than 5 meters contain θ, with certainty. Under the FCF, the data would yield *both* 100% certainty and 50% certainty. Likewise, in the first scenario, the bubbles were only 0.05 meters apart. Of confidence intervals that are 0.05 meters wide, only 5% contain the true value (see the supplement for an explanation). On the basis of this fact, one could state 5% certainty that the hatch is in the interval, and yet the FCF would lead one to also claim 50% certainty. The FCF leads to multiple, contradictory inferences from the same data.

The Fundamental Confidence Fallacy seems to follow directly from the definition of a confidence interval. How can one use a confidence procedure that yields 50% CIs, yet *not* be 50% certain that they contain the true value? We explore the roots of this seeming paradox.

Relevant subsets

You may have heard the old joke about the statistician who drowned wading across a river. He knew that the river was one meter deep...on average. The joke works (as far as it goes) because anyone can see that the statistician was foolish for considering only the average depth of the river. Surely at some point he could see that the water was dangerously deep and yet he pressed on, presumably reassured by the sophistication of his research into the depth of the river.

The statistician in the joke failed to use relevant depth of the river *at his location* to inform his actions. The Fundamental Confidence Fallacy entails a similar confusion: the confusion of the average properties of a confidence procedure with what is known about a particular confidence interval. Consider the submarine example: if we were to know that the width of the confidence interval from CP1 had a width of 9 meters, should we report 50% confidence in the interval, or 100% confidence? It is a 50% CI in the sense that it was generated from a 50% confidence procedure. If we restrict our attention to intervals of width 9 meters, however, we find that 100% of these intervals contain the true value.

Fisher (1959b) argued that probability statements like the ones above were critically flawed. In particular, upon observing the data, one can see that the observed data are part of a subset of data space (that is, possible data) for which the CI has a different probability of containing the true value than the average probability. In the submarine example, these subsets can be identified by how far apart the bubbles were. In cases like these, the subset into which the data fall is relevant to judging the probability that the CI contains the true value; hence, these special subsets of the data are called *relevant subsets* (Buehler, 1959; Buehler & Feddersen, 1963; Casella, 1992; Robinson, 1979).

The existence of relevant subsets indicates a failure of properly conditioning on the data: there is the information in data that is not being used. By not using this relevant information, researchers who believe the Fundamental Confidence Fallacy are blinding themselves to important aspects of the data. Someone who wanted to make good use of their data would surely never want to use an interval that admitted relevant subsets, because that would involve ignoring what could be plainly seen in the data. Furthermore, if one believes the Fundamental Confidence Fallacy, the existence of relevant subsets can cause mutually contradictory confidence statements.

Although there is nothing about confidence procedures themselves that prevent relevant subsets – their goal is a long-run average performance, not to support reasonable, specific inferences – we can create a procedure that eliminates the relevant subsets in the submersible example. Bayesian intervals called credible intervals take into account all the data through conditioning. If we take the central 50% of the intervals labeled "posterior" in Figure 2, we obtain the following credible interval[4]:

$$\bar{x} \pm \left(5 - \frac{|x_1 - x_2|}{2}\right)$$

The credible intervals for the first and second submarine rescue scenarios are shown in Figure 2 along the line labeled "Cred. Int." Notice that the Bayesian credible interval properly tracks the precision given by the data (as shown by the posterior). Incidentally, this Bayesian credance procedure yields another 50% confidence procedure.

It is clear that to a scientist caring primarily about what the data at hand say about the parameter, the Bayesian interval is preferable. The Bayesian interval does not ignore the information in the data by leaving relevant subsets; it cannot include impossible values; and its narrowness is directly proportional to the precision implied by the data. It is derived from the Bayesian posterior distribution, which is a unique statement about the uncertainty about the parameter, given the data and a prior. In contrast, confidence intervals are not unique; for any given problem, there may be several mutually contradictory confidence procedures. Finally, the Bayesian interval, through its use of a prior, has the interpretation that the advocates of desire: that the plausibility that the true value is within the interval is 50%.

We will explore the properties of Bayesian procedures in the Discussion; for now, we present a scenario that will be more familiar to psychological researchers: Normal data with the Student's *t* confidence interval.

Example 2: Student's *t* interval

The submarine example was specifically tailored to show how each of the three broad claims about confidence intervals fail. Many users of CIs, however, will never encounter such a problem. Their data is roughly normal, and so they use Student's *t* confidence interval almost exclusively. As we will argue in the discussion, this does not make the argument against CIs less powerful: CIs are advocated as a general inferential tool, so they should work generally. If CIs only appear to make good inferences in a small number of circumscribed situations, their logical basis is suspect. Moreover, the propenents of CIs often directly state or indirectly imply that *all* CPs have the properties erroneously ascribed to them, by virtue of them being CIs (see, for example, Cumming, 2014, who states that these properties "generally appl[y] to any CI"). It turns out, however, that the contradictory conclusions caused by relevant subsets occur even with popular Student's *t* confidence interval.

Consider a situation in which $N = 2$ observations are to be drawn from a normal distribution with unknown mean μ and standard deviation σ. As previously discussed, the typical 50% Student's *t* confidence procedure for these data is

$$\bar{x} \pm \frac{s}{\sqrt{2}} = \bar{x} \pm \frac{|x_1 - x_2|}{2},$$

or simply the minimum observation to the maximum observation. The fact that this is a 50% confidence procedure implies that if we observe the data, compute the CI, and then make the dichotomous claim "the CI contains the true value," – or, alternatively, "the CI does not contain the true value" our claim will be correct 50% of the time, on average. If relevant subsets exist, then we can find a way to improve our accuracy above 50% correct by only using information in the data.

Although we did not previously point it out, we have already presented the evidence of relevant subsets using Student's *t* intervals. If we re-examine Figure 1, it is apparent that the intervals near the bottom of the figure — the short intervals — almost never contain the true value. The longer intervals near the top, in contrast, almost always contain the true value. In general, the probability that a Student's *t* CI contains the true value is an increasing function of the standard error.

In order to exploit this fact, we require a way of separating short intervals from long ones. Buehler (1959) suggested a simple strategy for improving on the 50% accuracy of the CI: pick a positive number — any positive number will do — and if the sample standard deviation s is larger than this number, claim that the interval contains the true value. If the standard deviation is smaller the selected number, claim that the interval excludes the true value. No matter what the true mean, true standard deviation, and positive number one selects, this strategy will lead to a greater than 50% success rate for selecting intervals that do or do not contain the true value.

How far can you raise your accuracy above the nominal 50% correct rate for the confidence interval? The accuracy

[4] This credible interval is formed by assuming a prior distribution that assigns equal plausibility to all possible values of θ. A derivation is provided in the supplement to this article.

of the strategy depends on the number you selected to separate the short intervals from the long ones. If the number you selected is near $.67\sigma$, you can win about 3/4 of the time. Your accuracy will drop as your criterion moves further from this optimal value. The proof of this fact is not provided by Buehler (1959), but we have provided it in the supplement because it will help understand why the relevant subsets arise.

Given that your accuracy depends on how well you can guess the true population standard deviation σ, it might seem that increasing your accuracy above 50% requires prior information about σ. It is important to emphasize that the probability that you are correct with Buehler's strategy is *always* greater than 50%, no matter what the criterion. It is true, however, that it is bounded at .5. If you are very, very far off in your guess of σ, then your accuracy will be negligibly larger than 50%. This boundedness has led statisticians to call this relevant subset a *semirelevant* subset (Buehler, 1959; Robinson, 1979), which is typically considered less problematic than fully relevant subsets, where the probability is bounded away from 50%. This might cause one to suspect that the problem is not as bad as it might first seem; certainly, the problem with Student's t intervals does not seem as dire as it was with the CI in the submarine rescue example.

There are two responses to this defense of the Student's t confidence interval. First, as Buehler (1959) points out, information about the error in one's measurements – whatever these measurements might be – is precisely the kind of information that is known to a competent experimenter, and that would garner rough agreement across experimenters. One does not need very specialized knowledge of σ to increase one's accuracy substantially above 50%. If, for instance, you underestimate σ by a factor of 10 – a very large underestimation that we suspect most experimentalists would be able to beat – your accuracy will still about 55%. A statistician may find this semi-relevant subset less interesting, but a scientist concerned with making reasonable, accurate inferences should find it disconcerting.

The second response to the defense against the Student's t CI's semirelevant subsets is that in fact, the Student's t CI admits fully relevant subsets. We presented the semirelevant subset first for conceptual clarity. Consider the following scheme, suggested by Pierce (1973). Suppose we sample $N = 2$ observations from a normal population with unknown mean and variance, then compute a 50% Student's t confidence interval. We then *simultaneously* perform a two-sided t test against the null hypothesis that $\mu = 0$ (it is irrelevant whether or not we are interested in the hypothesis that $\mu = 0$). If the p value for the t test is greater than .25, then the probability that the 50% confidence interval contains the true value is greater than or equal to 2/3, regardless of the true mean and standard deviation of the population. In a similar but more dramatic fashion to Buehler's (1959) example, the procedure selects for large CIs by requiring a large p value.[5]

Pierce's (1973) relevant subset shows that even with the Student's t confidence interval, one could claim with 50% certainty that the true mean is within the CI, and simultaneously know that the data are part of a relevant subset for which CIs contain the true mean with a probability of at least 2/3. Confidence procedures fail because they can provide contradictory advice on how confident we should be that the true value is inside the interval. Indeed, confidence procedures were not designed to offer such advice. Confidence procedures were merely designed to allow the analyst to make certain kinds of dichotomous statements about whether an interval contains the true value, in such a way that the statements are true a fixed proportion of the time *on average* (Neyman, 1937). Expecting them to do anything else is expecting too much.

Inference without Relevant Subsets

Relevant subsets threaten the coherence and uniqueness of frequentist inferences with confidence intervals and can be thought of as an example of a more general problem, the reference class problem (Venn, 1888). Frequentist probability is defined as a long-run proportion, or the number of events that occur with a specified property (e.g., CIs that contain the true value) out of some reference class of events. If there are relevant subsets, then there are multiple probability statements we could make (Fisher, 1959a), depending on the reference class we choose.

The submarine example makes the implications clear. Is our reference class all CIs, in which case 50% of CIs will contain the true value? Or is our reference class all CIs with a particular width, in which case anywhere from 0% (for very narrow CIs) to 100% (for CIs wider than 5 meters)? Either confidence statement is valid, from a frequentist perspective, but they contradict one another.

It seems clear that restricted reference classes are the preferable: when possible, our inferences should be based on the most specific descriptions of the data possible. One might be tempted to try to solve the problem simply: if our data are part of a known relevant subset, then we should condition our inference on that fact. In Pierce's (1973) example above, if $p > .25$, then we should report a confidence of greater than or equal to 2/3. However, this is no solution; an observation may be a member of multiple overlapping relevant subsets, and then the reference class problem rears its ugly head again.

The desire to base inference only on the data that were observed, and not on the average properties across all possible data sets, suggests a simple solution to the relevant subset

[5] If, like the authors of this paper, you find this result unbelievable and counterintuitive, R code is provided in the supplement to show the result by simulation.

problem. There is one reference class that is as specific as possible that makes the relevant subsets problem disappear: the data itself. Bayesian inference, for instance, makes use of Bayes' theorem, which states that

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta})}{p(\mathbf{y})} p(\boldsymbol{\theta}),$$

where $\mathbf{y}$ is the data and $\boldsymbol{\theta}$ is a vector of unknown parameters. The posterior $p(\boldsymbol{\theta} \mid \mathbf{y})$ yields a probability distribution that represents the uncertainty about the parameters, given the observed data. The prior, $p(\boldsymbol{\theta})$, represents our uncertainty about the parameters before observing the data. If the prior probability distribution is *proper* – that is, it represents a valid probability distribution – then there can be no relevant subsets (Casella, 1992; Robinson, 1979). The probability statements that arise from Bayesian inference with proper priors must be unique and consistent. Furthermore, the interpretation of probability statements that arise from Bayesian inference are interpretable as statements of plausibility, unlike frequentist probability statements which have no such interpretation.

Notice, though, that in order to rid ourselves the relevant subsets that make can make reasoning from confidence intervals problematic, we require a prior distribution, $p(\boldsymbol{\theta})$. As we have argued elsewhere (Rouder, Morey, Verhagen, Province, & Wagenmakers, submitted), reasonable inference requires bringing information to the table. Only if we bring information to the table, in the form of a reasonable prior distribution, can we take all the information off the table. The belief that one can make inferences without committing to using prior information strikes us like a gambler who tries to win without paying the ante. The rules of the table, unfortunately, do not allow this.

The argument presented here – that frequentist inference, in this case with CIs – leads to incoherent inferences, is but one of a number of arguments for moving away from frequentist inferential methods. Elsewhere, we have argued for Bayesian inference as a viable replacement (de Vries & Morey, 2013; M. Lee & Wagenmakers, 2005; Morey, Rouder, Verhagen, & Wagenmakers, in press; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Rouder & Morey, 2011; Rouder, Morey, Speckman, & Province, 2012; Wagenmakers, 2007; Wagenmakers, M. D. Lee, Lodewyckx, & Iverson, 2008; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). A full accounting of Bayesian inference is beyond the scope of this article; for the interested reader, in addition to the articles just mentioned we also recommend Dienes (2011), P. M. Lee (2004), and Edwards, Lindman, and Savage (1963).

Discussion

Using two examples, we have shown that confidence intervals do not have the properties that are often claimed on their behalf. Confidence intervals were developed to solve a very constrained problem: how can one construct an interval that contains the true mean a fixed proportion of the time? This definition, which concerns only average performance, does not support reasonable inference from specific data. Claims that confidence intervals yield an impression of precision, that the values within them are plausible, and that the confidence coefficient can be read as a measure of certainty that the interval contains the true value, are all errors.

Good intentions underlie the advocacy of confidence intervals: it would be excellent to have procedures with the properties claimed. The FCF is driven by a desire to assess the plausibility that an interval contains the true value; the likelihood error is driven by a desire to determine which values of the parameter are likely; and the precision error is driven by a desire to quantify the precision of the estimates. We support these goals (Morey et al., in press), but CIs are not the way to achieve them.

Confidence intervals versus credible intervals

One of the misconceptions regarding the relationship between Bayesian inference and frequentist inference is that they will lead to the same inferences. In the case where data are normally distributed, for instance, there is a particular prior that will lead to a confidence interval that is numerically identical to Bayesian credible intervals computed using the Bayesian posterior (Jeffreys, 1961; Lindley, 1965). This occurs, for instance, in the Student's t scenario described above.[6] This might lead one to suspect that it does not matter whether one uses confidence procedures or Bayesian procedures.

If researchers were only expected to study phenomena that were normally distributed, and researchers were only expected to make a single inference from the data – the confidence interval – then inference by confidence procedures might seem indistinguishable from inference by Bayesian procedures. The defense of confidence procedures by noting that, in some restricted cases, they numerically correspond to Bayesian procedures is actually no defense at all. One must first choose *which* confidence procedure, of many, to use, and if one is committed to the procedure that corresponds to Bayesian inference, then this is an admission that it was the Bayesian procedure that was desired all along. More broadly, if psychologists are to be sophisticated statistical thinkers, they should not be limited to a single inferential statement under restrictive assumptions.

[6] The fact that the confidence interval and objective Bayesian credible interval are numerically the same might lead one to believe that Bayesian intervals are susceptible to relevant subsets as well. However, the objective Bayesian interval is not a proper probability distribution. Bayesian inference with proper priors will be immune to relevant subsets.

Loftus (1993) argued in the context of recommending abandoning significance testing, that limiting ourselves to common designs and assumptions (e.g., normal populations) severely limits the inferences we can make. To prevent arbitrary limitations on statistical inferences, if Bayesian interpretations are desired, Bayesian inference should be applied in its full generality – not just when it numerically corresponds with frequentist inference. The correspondence between confidence procedures and Bayesian procedures is not a general rule. In some cases, for instance with count data, there are many different confidence intervals (among others the Wald, the Agresti-Coull, the Clopper-Pearson, the arcsine, and the logit; see Brown, Cai, & DasGupta, 2001, for a review). These confidence procedures all yield different inferences among themselves, not to mention differences with Bayesian credible intervals.

In some cases confidence procedures do not even allow an inference. The end-points of a confidence interval are always set by the data. Suppose, however, we are interested in determining the plausibility that a parameter is in a particular range; for instance, in the United States, it is against the law to execute criminals who are intellectually disabled. The criterion used for intellectual disability in the US state of Florida is having an IQ lower than 70. Since IQ is measured with error, one might ask what confidence we have that a particular criminal's IQ is between 0 and 70. In this case, the interval is no longer a function of the sample. The long-run probability that the true value is inside a fixed interval is unknown and is either 0 or 1, and hence no CP can be constructed, even though such information may be critically important to a researcher, policy maker, or criminal defendant.

Even in seemingly simple cases where a fixed interval is nested inside a CI, or *vice versa*, one cannot draw conclusions about the confidence of a fixed interval. One might assume that an interval nested within a CI must have lower confidence than the CI; however, in the second submersible rescue scenario, a 100% confidence interval (all the possible values of θ) was nested within both CI1 and CI2, which were 50% CIs. Likewise, one might believe that if a CI is nested within a fixed interval, then the fixed interval must have greater confidence than the interval. In the first submersible rescue scenario, intervals within which the 50% CI1 were nested had low plausibility, due to their narrowness. In contrast, Bayesian procedures offer the ability to compute the plausibility of any given range of values, and are guaranteed to yield statements that are mutually coherent. Because all inferences must be made through Bayes theorem, inferences must remain internally consistent (c.f. Stone & Dawid, 1972).

Finally, we believe that in science, the meaning of our inferences are important. Bayesian credible intervals support an interpretation of probability in terms of plausibility, thanks to the explicit use of a prior. Confidence intervals, on the other hand, are based on a philosophy that does not allow inferences about plausibility, but do not require a prior. Using confidence intervals as if they were credible intervals is an attempt to smuggle Bayesian meaning into frequentist statistics, without proper consideration of a prior. Priors have consequences, and must be carefully considered. There is no free lunch; to get reasonable inference, one must pay a price (Rouder, Morey, Verhagen, et al., submitted).

Conclusion

We have suggested that confidence intervals do not support the inferences that their advocates believe they do. The problems with confidence intervals – particularly the fact that they admit can relevant subsets – shows a fatal flaw with their logic. They cannot be used to draw reasonable inferences. We recommend that their use be abandoned.

We therefore take stock of what we would be giving up, if we were to give up the use of confidence procedures. Abandoning the use of confidence procedures means abandoning a method that merely allows us to create intervals that contain the true value with a fixed long-run probability. We suspect that if researchers understand that this is the only thing they will be losing, they will not consider it a great loss. By adopting Bayesian inference, they will gain a way of making principled statements about precision and plausibility. Ultimately, this is exactly what the advocates of CIs have wanted all along.

References

Berger, J. O. & Wolpert, R. L. (1988). *The likelihood principle (2nd ed.)* Hayward, CA: Institute of Mathematical Statistics.

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*(2), 101–133.

Buehler, R. J. (1959). Some validity criteria for statistical inferences. *The Annals of Mathematical Statistics*, *30*(4), 845–863.

Buehler, R. J. & Feddersen, A. P. (1963). Note on a conditional property of Student's t^1. *The Annals of Mathematical Statistics*, *34*(3), 1098–1100.

Casella, G. (1992). Conditional inference from confidence sets. *Lecture Notes-Monograph Series*, *17*, 1–12.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*.

Cumming, G. & Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170–180.

de Vries, R. M. & Morey, R. D. (2013). Bayesian hypothesis testing for single-subject designs. *Psychological Methods*, *18*(2), 165–185.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Fidler, F. & Loftus, G. R. (2009). Why figures with error bars should replace p values: some conceptual arguments and empirical demonstrations. *Zeitschrift fūr Psychologie*, *217*(1), 27–37.

Fisher, R. A. (1959a). Mathematical probability in the natural sciences. *Metrika*, *2*(1), 1–10.

Fisher, R. A. (1959b). *Statistical Methods and Scientific Inference* (Second). Edinburgh, UK: Oliver and Boyd.

Jeffreys, H. (1961). *Theory of probability (3rd edition)*. New York: Oxford University Press.

Lee, M. & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*, 662–668.

Lee, P. M. (2004). *Bayesian statistics: An introduction (3rd ed.)* New York: Wiley.

Lindley, D. V. (1965). *Introduction to probability and statistics from a Bayesian point of view, part 2: Inference*. Cambridge, England: Cambridge University Press.

Loftus, G. R. (1993). A picture is worth a thousand p-values: On the irrelevance of hypothesis testing in the computer age. *Behavior Research Methods, Instrumentation and Computers*, *25*, 250–256.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current directions in psychological science*, *5*, 161–171.

Masson, M. E. J. & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, *57*, 203–220.

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (in press). Why hypothesis tests are essential for psychological science: a comment on Cumming. *Psychological Science*.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, *97*(4), 558–625.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *236*, 333–380.

Pierce, D. A. (1973). On some difficulties in a frequency theory of inference. *The Annals of Statistics*, *1*(2), 241–250.

Psychonomics Society. (2012). *Psychonomic Society guidelines on statistical issues.*

Robinson, G. K. (1979). Conditional properties of statistical procedures. *The Annals of Statistics*, *7*(4), 742–755.

Rouder, J. N. & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*, 682–689.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374.

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (submitted). The $p < .05$ rule and the hidden costs of the free lunch in inference.

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 and Review*, *16*, 225–237.

Stone, M. & Dawid, A. P. (1972). Un-Bayesian implications of improper Bayes inference in routine statistical problems. *Biometrika*, *59*(2), 369–375.

Velicer, W. F., Cumming, G., Fava, J. L., Rossi, J. S., Prochaska, J. O., & Johnson, J. (2008). Theory testing using quantitative predictions of effect size. *Applied Psychology*, *57*(4), 589–608.

Venn, J. (1888). *The logic of chance* (3rd). London: Macmillan.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p values. *Psychonomic Bulletin and Review*, *14*, 779–804.

Wagenmakers, E.-J., Lee, M. D., Lodewyckx, T., & Iverson, G. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. Boelen (Eds.), *Practical Bayesian approaches to testing behavioral and social science hypotheses* (pp. 181–207). New York: Springer.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi. A comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432.

Welch, B. L. (1939). On confidence limits and sufficiency, with particular reference to parameters of location. *The Annals of Mathematical Statistics*, *10*(1), 58–69.

Wilkinson, L. & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Young, K. D. & Lewis, R. J. (1997). What is confidence? part 1: the use and interpretation of confidence intervals. *Annals of Emergency Medicine*, *30*(3), 307–310.