The Fallacy of Placing Confidence in Confidence Intervals

Richard D. Morey

Cardiff University

Rink Hoekstra

University of Groningen

Jeffrey N. Rouder

University of Missouri

Michael D. Lee

University of California-Irvine

Eric-Jan Wagenmakers

University of Amsterdam

Author Note

Address correspondence to Richard D. Morey (richarddmorey@gmail.com). We thank Mijke Rhemtulla for helpful discussion during the drafting of the manuscript. Supplementary material, all code, and the source for this document are available at https://github.com/richarddmorey/ConfidenceIntervalsFallacy. Draft date: April 21, 2015.

Abstract

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 (e.g. 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 can lead to unjustified or arbitrary inferences. For this reason, we caution against the sole use of confidence interval theory to justify interval estimates.

KEYWORDS: statistics; confidence intervals; Bayesian methods

The Fallacy of Placing Confidence in Confidence Intervals

"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 allow 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 comprises interval estimates. There are a variety of approaches to interval estimation, differing in their philosophical foundation and computation, but informally are all supposed 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 methodological reformers' proposed programs (G. Cumming, 2014; Loftus, 1996). In the current atmosphere of methodological reform, a firm understanding of what sorts of inferences confidence interval theory does, and does not, allow is critical to decisions about how science is done in the future.

In this paper, we argue that the advocacy of CIs is based on a folk understanding of CI theory, rather than a statistical understanding. We outline three fallacies of CIs, and place these in the philosophical and historical context of CI theory proper. Through an easily-accessible example adapted from the statistical literature, we show how CI theory differs from the folk theory of CIs. Finally, we show the fallacies of confidence in the context of a CI advocated and commonly-used for ANOVA and regression analysis, and discuss the implications of the mismatch between CI theory and the folk theory of CIs.

Our main point is this: confidence intervals may not be used as suggested by

modern proponents because this usage is not justified by confidence interval theory. If used in the way CI proponents suggest, some CIs will provide severely misleading inferences for the given data; other CIs will not. Because such considerations are outside of CI theory, developers of CIs do not test them, and it is therefore often not known whether a given CI yields a reasonable inference or not. For this reason, we believe that appeal to CI theory is redundant in the best cases, when inferences can be justified outside CI theory, and unwise in the worst cases, when they cannot.

The folk theory of confidence intervals

If there is one thing that everyone who writes about confidence intervals agrees on, it is the basic definition: A confidence interval for a parameter — which we call θ, but might be a mean, median, variance, probability, or any number of other quantities — is an interval generated by a procedure that, on repeated sampling, has a fixed probability of containing the parameter. If the probability that the process generates an interval including θ is .5, it is a 50% CI; likewise, the probability is .95 for a 95% CI.

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

The confidence coefficient of a confidence interval derives from the procedure which generated it. It is therefore helpful to differentiate a *procedure* (CP) from a confidence *interval*: an X% confidence procedure is any procedure that generates intervals containing θ on X% of repeated samples, and a confidence interval is a specific interval generated by such a process. A confidence procedure is a random process; a confidence interval is a set of two, fixed numbers.

It seems clear how to interpret a confidence *procedure*: it is any procedure that generates intervals that will contain the true value in a fixed proportion of samples. However, when we compute 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 some observed confidence interval.

Textbook authors and proponents of confidence intervals bridge the gap seamlessly by claiming that confidence intervals have three desirable properties: first, that the confidence coefficient can be read as a measure of the uncertainty one should have that the interval contains the parameter; second, that the CI width is a measure of estimation precision; and third, that the interval contains the "likely" or "reasonable" values for the parameter. These all involve reasoning about the parameter from the observed data: that is, they are "post-data" inferences.

For instance, with respect to 95% confidence intervals, 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." G. 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 understanding of confidence intervals is not correct. We call this mistake the "Fundamental Confidence Fallacy" (FCF) because it seems to flow naturally from the definition of the confidence interval:

Fallacy 1 (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%," or "We have X% confidence that the observed interval contains the true value."*

The reasoning behind the Fundamental Confidence Fallacy seems plausible: on a given sample, we could get any one of the possible confidence intervals. If 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 we obtained one of the confidence intervals that contain the true value. 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" — rather than, for

instance, the more accurate "coverage procedure" — encourages the Fundamental Confidence Fallacy.

We will show in two examples how the reasoning fails. We note here that it has never been an advertised feature of CI theory; the error has been understood since CI theory was born. Neyman said "Consider now the case when a sample...is already drawn and the [confidence interval] given...Can we say that in this particular case the probability of the true value of [the parameter] falling between [the limits] is equal to [X%]? The answer is obviously in the negative" (Neyman, 1937, p. 349). According to frequentist philosopher Mayo (1981) "[the misunderstanding] seems rooted in a (not uncommon) desire for [] confidence intervals to provide something which they cannot legitimately provide; namely, a measure of the degree of probability, belief, or support that an unknown parameter value lies in a specific interval." Recent work has shown that this misunderstanding is pervasive among researchers, who likely learned it from textbooks, instructors, and the confidence interval literature (Hoekstra, Morey, Rouder, & Wagenmakers, 2014).

If confidence intervals cannot be used to assess the certainty with which a parameter is in a particular range, what can they be used for? Proponents of confidence intervals often claim that confidence intervals are useful for assessing the precision with which a parameter can be estimated. This is cited as one of the primary reasons confidence procedures should be used over null hypothesis significance tests (e.g., G. Cumming & S. Finch, 2005; G. Cumming, 2014; Fidler & Loftus, 2009; Loftus, 1993, 1996). For instance, G. 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" (p. 10). Young and Lewis (1997) state that "[i]t is important to know how precisely the point estimate represents the true difference between the groups. The width of the CI gives us information on the precision of the point estimate" (p. 309). This is the second fallacy of confidence intervals, the "precision fallacy":

Fallacy 2 (The Precision fallacy) *"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. To see this most easily, consider that the width of a confidence interval is dependent on the confidence coefficient. Changing from 50% to a 95% confidence will make the confidence interval wider; however, nothing about the precision of the estimate has changed. What has changed is the average frequency at which the dichotomous claim "the parameter is in the interval" is correct for the underlying confidence procedure. Later, we will provide several examples where the width of a CI and the uncertainty with which a parameter is estimated are are inversely related for a fixed confidence coefficient, or not related at all.

We cannot interpret confidence intervals as containing the true value with some probability; we also cannot interpret confidence intervals as indicating the precision of our estimate. There is a third common interpretation of confidence intervals: 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). This is a another fallacy of confidence interval that we call the "likelihood fallacy".

Fallacy 3 (The Likelihood fallacy) *"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.*

A confidence procedure may have a fixed *average* probability of including the true value, but whether on any given sample it includes the "reasonable" values is a different question. As we will show, confidence intervals — even good confidence intervals, as we will see — can exclude almost all reasonable values, and can be empty or infinitely narrow, excluding all possible values (Blaker & Spjøtvoll, 2000; Dufour, 1997; Steiger, 2004; Steiger & Fouladi, 1997; Stock & Wright, 2000). But Neyman (1941) writes,

> "it is not suggested that we can 'conclude' that [the interval contains θ], nor that we should 'believe' that [the interval contains θ]...[we] *decide* to behave as if we actually knew that the true value [is in the interval]. This is done as a result of our decision and has nothing to do with 'reasoning' or 'conclusion'. The reasoning ended when the [CI procedure was derived]. The above process [of using CIs] is also devoid of any 'belief ' concerning the value [] of [θ]." (Neyman, 1941, pp. 133-134)

It may seem strange to the modern user of CIs, but Neyman is quite clear that CIs do not support any sort of reasonable belief about the parameter. Even from a frequentist testing perspective where one accepts and rejects specific parameter values, Mayo and Spanos (2006) note that just because a specific value is in an interval does not mean it is warranted to accept it; they call this the "fallacy of acceptance." If a confidence procedure does not allow an assessment of the probability that an interval contains the true value, if it is not a measure of precision, and if it is not a way of assessing the likelihood or plausibility of the values inside the interval, what is it?

The theory of confidence intervals

In a classic paper, Neyman (1937) laid the formal foundation for confidence intervals. It is easy to describe the practical problem that Neyman saw CIs 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 U – forming an interval (L, U) according to a specified procedure.

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

This recommendation is justified by choosing an procedure 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.

We first focus on the meaning of the statement that θ is in the interval, in step (c). As we have seen, according to CI theory, what happens in step (c) is not a belief, a conclusion, or any sort of reasoning from the data. Furthermore, it is not associated with any level of uncertainty about whether θ is, actually, in the interval. It is merely a dichotomous statement that is meant to have a specified probability of being true, in the long-run.

Evaluation of confidence procedures is based on what can be called the "power" of the procedures, which is the frequency with which false values of a parameter are excluded. Better intervals are shorter on average, excluding false values more often (Lehmann, 1959; Neyman, 1937, 1941; Welch, 1939). Although a "best" confidence procedure does not always exist, we can always compare one procedure to another to decide whether one is better than the other in this way (Neyman, 1952). Confidence procedures are therefore closely related to hypothesis tests: confidence procedures control the rate of including the true value, and better confidence procedures have more power to exclude false values.

Skepticism about the usefulness of confidence intervals arose as soon as Neyman first articulated the theory (Neyman, 1934).[1] In the discussion of Neyman (1934), Bowley — pointing out what we call the fundamental confidence fallacy, expresses skepticism that the confidence interval answers the right question:

> "I am not at all sure that the 'confidence' is not a 'confidence trick.' Does it 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." (p. 609)

[1] Neyman first articulated the theory in another paper before his major theoretical paper in 1937.

In the same discussion, Fisher critiqued the theory for possibly leading to mutually contradictory inferences: "The [theory of confidence intervals] was a wide and very handsome one, but it had been erected at considerable expense, and it was perhaps as well to count the cost. The first item to which he [Fisher] would call attention was the loss of uniqueness in the result, and the consequent danger of apparently contradictory inferences." (p. 618; see also Fisher, 1935). Though, as we will see, the critiques are accurate, in a broader sense they missed the mark. Like modern proponents of confidence intervals, the critics failed to understand that Neyman's goal was different from theirs: Neyman had developed a behavioral theory designed to control error rates, not a theory for reasoning from data (Neyman, 1941).

In spite of the critiques, confidence intervals have grown in popularity to be the most widely used interval estimators. The alternatives — such as Bayesian credible intervals and Fisher's fiducial intervals — are not commonly used. We suggest that this is largely because people believe that confidence intervals, Bayesian intervals, and fiducial intervals are the same thing. In the next section, we will demonstrate the logic of confidence interval theory by building several confidence intervals and comparing them to one another. We will also show how the three fallacies affect inferences with these intervals.

Example 1: The lost submarine

In this section, we present an example taken from the confidence interval literature (J. O. Berger & Wolpert, 1988; Lehmann, 1959; Pratt, 1961; Welch, 1939) designed to bring into focus the how CI theory works. This example is intentionally simple; unlike many demonstrations of CIs, no simulations are needed, and almost all results can be derived by readers with some training in probability.

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

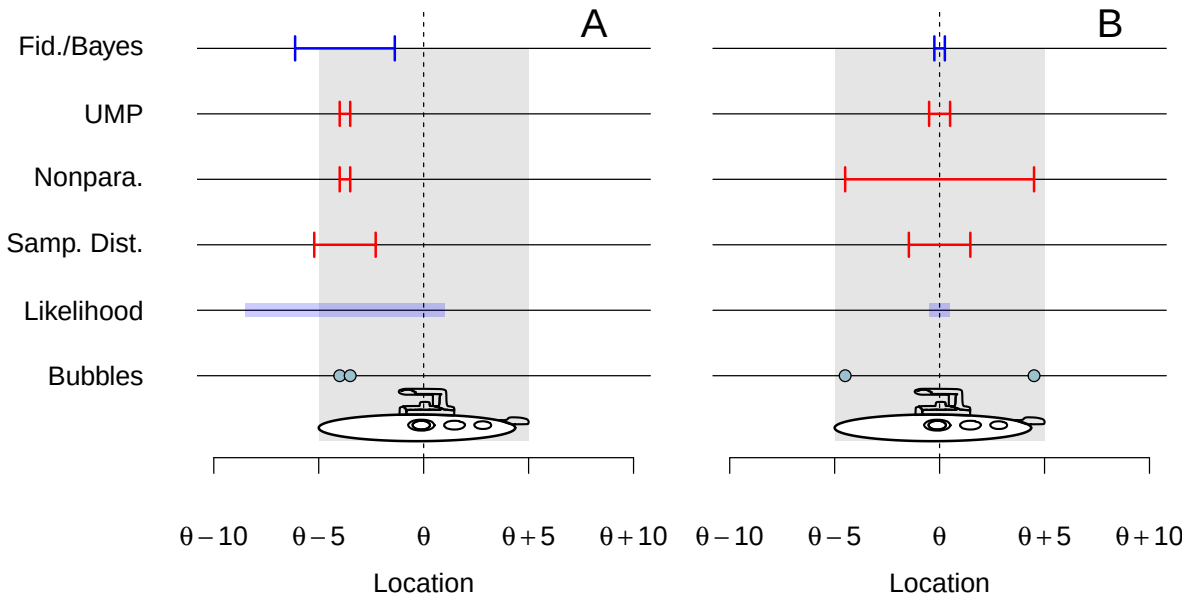

Figure 1. Submersible rescue attempts. Note that likelihood and CIs are depicted from bottom to top in the order in which they are described in the text. See text for details.

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 two 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.

The situation is shown in Figure 1A. The rescue hatch is the unknown location θ, and the bubbles can rise from anywhere with uniform probability between $\theta - 5$ meters (the bow of the submersible) to $\theta + 5$ meters (the stern of the submersible). The rescuers want to use these bubbles — which we will denote x_1 and x_2, x_1 always denoting the smaller of the two for convenience — to infer where the hatch is located.

The rescuers first note that from observing two bubbles, it is easy to rule out all values except those within five meters of both bubbles because no bubble can occur further than 5 meters from the hatch. If the two bubble locations were $x_1 = 4$ and $x_2 = 6$, then the possible locations of the hatch are between 1 and 9, because only these locations are within 5 meters of both bubbles. The function that describes the probability density of the observed bubbles for a particular value of θ is called the "likelihood," and it indexes the information provided by the data about the parameter.

In this case, it is positive only when a value θ is possible given the observed bubbles (see Figures 1 and 4).

Five confidence procedures

A group of four statisticians happen to be on board, and the rescuers decide to ask them for help improving their judgments using statistics. The four statisticians suggest four different 50% confidence procedures. We will outline the four confidence procedures; first, we describe a trivial procedure that no one would ever suggest.

0. A trivial procedure. A trivial 50% confidence procedure can be constructed by flipping a fair coin. If the coin shows "heads", the interval contains the whole ocean, $(-\infty, \infty)$. If the coin shows "tails", the interval contains only the single, *exact* point directly under the middle of the rescue boat. This procedure is obviously a 50% confidence procedure; exactly 50% of the time — when the coin shows "heads" — the rescue hatch will be within the interval. We describe this interval merely to show that *by itself*, a procedure including the true value X% of the time means nothing (see also Basu, 1981). We must obviously consider something more than the confidence property, which we discuss subsequently. We call this procedure the "trivial" procedure.

1. A procedure based on the sampling distribution of the mean. The first statistician suggests building a confidence procedure using the sampling distribution of the mean, $\bar{x} = (x_1 + x_2)/2$. The sampling distribution of $\bar{x}$ has a known triangular distribution with θ as the mean. We can thus use $\bar{x} - \theta$ as a so-called "pivotal quantity" (Casella & R. L. Berger, 2002, see the supplement to this manuscript for more details) by noting that there is a 50% probability that θ is within this same distance of $\bar{x}$ in repeated samples. There is a 50% probability that $\bar{x}$ will differ from θ by less than $5 - 5/\sqrt{2}$, or about 1.46m. This leads to the confidence procedure

$$\bar{x} \pm (5 - 5/\sqrt{2}),$$

which we call the "sampling distribution" procedure.

2. A nonparametric procedure. The second statistician notes that θ is both the mean and median bubble location. Olive (2008) and Rusu and Dobra (2008)

suggested a confidence procedure for the median that in this case is simply the interval between the two observations:

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

It is easy to see that this must be a 50% confidence procedure; the probability that both observations fall below θ is 25%, and likewise for both falling above. There is thus a 50% chance that the two observations encompass θ. Coincidentally, this is the same as 50% Student's t procedure for $n = 2$. We call this the "nonparametric" procedure.

3. The uniformly most-powerful procedure. The third statistician, citing Welch (1939), describes a procedure that can be thought of as a slight modification of the nonparametric procedure. Suppose we obtain a particular confidence interval using the nonparametric procedure. If the nonparametric interval is more than 5 meters wide, then it *must* contain the hatch, because the only possible values are within 5 meters from both bubbles. Moreover, in this case the interval will contain impossible values, because it will be wider than the likelihood. We can exclude these impossible values by truncating the interval to the likelihood whenever the width of the interval is greater than 5 meters:

$$\bar{x} \pm \begin{cases} \frac{x_2 - x_1}{2} & \text{if} \quad x_2 - x_1 < 5 \quad \text{(Nonparametric procedure)} \\ 5 - \frac{x_2 - x_2}{2} & \text{if} \quad x_2 - x_1 \geq 5 \qquad\qquad \text{(Likelihood)} \end{cases}$$

This will not change the probability that the interval contains the hatch, because it is simply replacing one interval that is sure to contain it with another. Pratt (1961) noted that this interval can be justified as an inversion of the uniformly most-powerful (UMP) test. We thus call this procedure the "UMP procedure".

4. An objective Bayesian procedure. The fourth statistician suggests an objective Bayesian procedure. Using this procedure, we simply take the central 50% of the likelihood as our interval:

$$\bar{x} \pm \frac{1}{2}\left(5 - \frac{x_2 - x_1}{2}\right).$$

From the objective Bayesian viewpoint, this can be justified by assuming a prior distribution that assigns equal probability to each possible hatch location. In Bayesian

terms, this procedure generates "credible intervals" for this prior. It can also be justified under Fisher's fiducial theory; see Welch (1939).

Properties of the procedures

The four statisticians report their four confidence procedures to the rescue team, who are understandably bewildered by the fact that there appear to be at least four ways to summarize the data about the hatch location from two bubbles. Just after the statisticians present their confidence procedures to the rescuers, two bubbles appear at locations $x_1 = 1$ and $x_2 = 1.5$. The resulting likelihood and the four confidence intervals are shown in Figure 1A.

After using the observed bubbles to compute the four confidence intervals, the rescuers wonder how to interpret them. It is clear, first of all, why the fundamental confidence fallacy is a fallacy. As Fisher pointed out, for any given problem, as for this one, there are many possible confidence procedures. These confidence procedures will lead to different confidence intervals. In the case of our submersible confidence procedures, all confidence intervals are centered around $\bar{x}$, and so the intervals will be nested within one another. If all intervals had a 50% probability of containing the true value, then all the probability must be contained in the shortest of the intervals. Because each procedure is by itself a 50% procedure, the procedure which chooses the shortest of 50% intervals will contain the true value less than 50% of the time. Hence, believing the FCF results in logical contradiction.

This is not, by itself, a critique of confidence interval theory *proper*. Neyman was very clear that this interpretation was not permissible, using similarly nested confidence intervals to demonstrate the fallacy (Neyman, 1941, pp. 213-215). It is a warning, however, that the improper interpretation of confidence intervals can lead to mutually contradictory inferences, just as Fisher warned.

Even without nested confidence procedures, one can see that the FCF must be a fallacy. Consider Figure 1B, which shows the resulting likelihood and confidence intervals when $x_1 = 0.5$ and $x_2 = 9.5$. When the bubbles are far apart, as in this case,

the hatch can be localized very precisely: the bubbles are far enough apart that they must have come from the bow and stern of the submersible. The sampling distribution, nonparametric, and UMP confidence intervals all encompass the likelihood, meaning that there is 100% certainty that these 50% confidence intervals contain the hatch. Reporting 50% certainty in an interval that surely contains the parameter would clearly be a mistake.

This set of confidence procedures also makes clear the precision fallacy. Consider Figure 2, which shows how the width of each of the intervals produced by the four confidence procedures changes as a function of the width of the likelihood. Because the likelihood represents the possible locations for the hatch, the likelihood's width is a natural measure of the uncertainty in the data. When it is 0, the location of the hatch is known with certainty; when it is wider, there is greater uncertainty.

Intervals from the sampling distribution procedure have a fixed width, and so cannot reveal any information about the precision of the estimate. The nonparametric procedure generates intervals whose widths are *inversely* related to the width of the likelihood. Even more strangely, intervals from the UMP procedure initially increase in width with the uncertainty in the data, but when the width of the likelihood is greater than 5 meters, the width of the UMP interval is inversely related to the uncertainty in the data, like the nonparametric interval. Only the Bayes procedure tracks the uncertainty in the data. This is not a coincidence; we will discuss why subsequently.

To see how the likelihood fallacy manifests in this example, consider again Figure 2. When the uncertainty is high, the likelihood is wide; yet the nonparametric and UMP intervals are extremely narrow, indicating both false precision and excluding almost all likely values. Furthermore, the sampling distribution procedure and the nonparametric procedure can contain impossible values.[2]

[2]In order to construct a better interval, a frequentist would typically truncate the interval to only the possible values, as was done with in generating the UMP procedure from the nonparametric procedure. This is guaranteed to lead to a better procedure. Our point here is that naively assuming that a procedure has good properties on the basis that it is a confidence procedure is a mistake. However, see Velicer et al. (2008) for an example of CI proponents including impossible values in confidence intervals, and Fidler

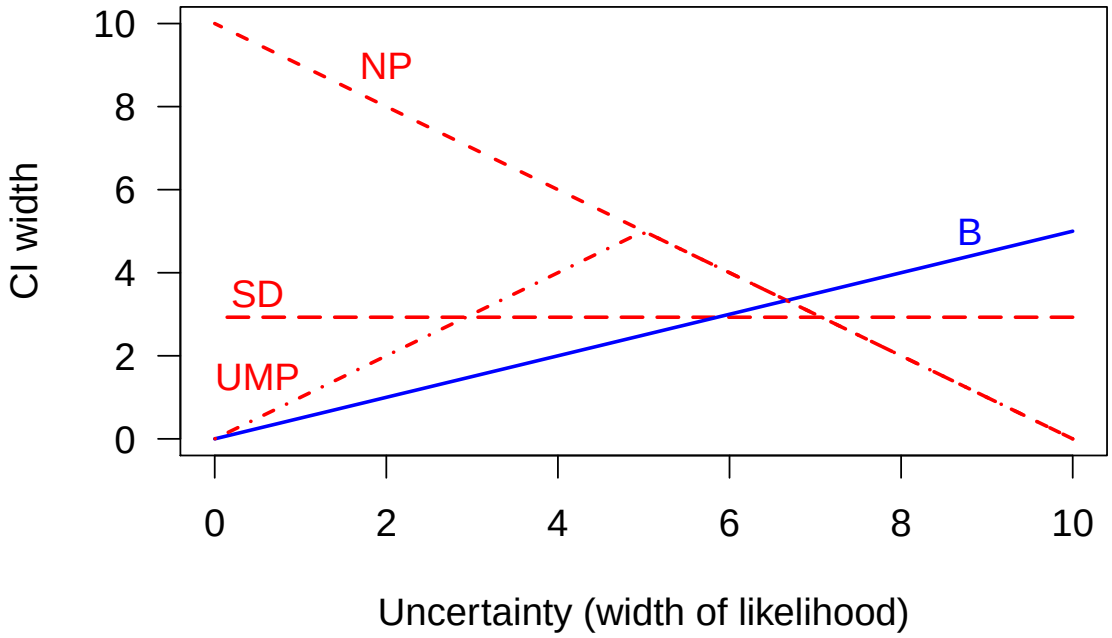

Figure 2. The relationship between CI width and the uncertainty in the estimation of the hatch location for the four confidence procedures. SD: Sampling distribution procedure; NP: Nonparametric procedure; UMP: UMP procedure; B: Bayes procedure. The diagonal gray line represents the width of the likelihood against itself.

Evaluating the confidence procedures

The rescuers who have been offered the four intervals above have a choice to make: which confidence procedure to choose? We have shown that several of the confidence procedures have counter-intuitive properties, but thus far, we have not made any firm commitments about which confidence procedures would be preferred to the others. For the sake of our rescue team, who have a decision to make about which interval to use, we now compare the four procedures directly. We begin with the evaluation of the procedures from the perspective of confidence interval theory, then evaluate them according to Bayesian theory.

---

and Thompson (2001) for a defense of this practice.

As previously mentioned, confidence interval theory specifies that better intervals will include false values less often. Figure 3 shows the probability that each of the procedures include a value θ' at a specified distance from the hatch θ. All procedures are 50% confidence procedures, and so they include the true value θ 50% of the time. Importantly, however, the procedures include false values $\theta' \neq \theta$ at different rates.

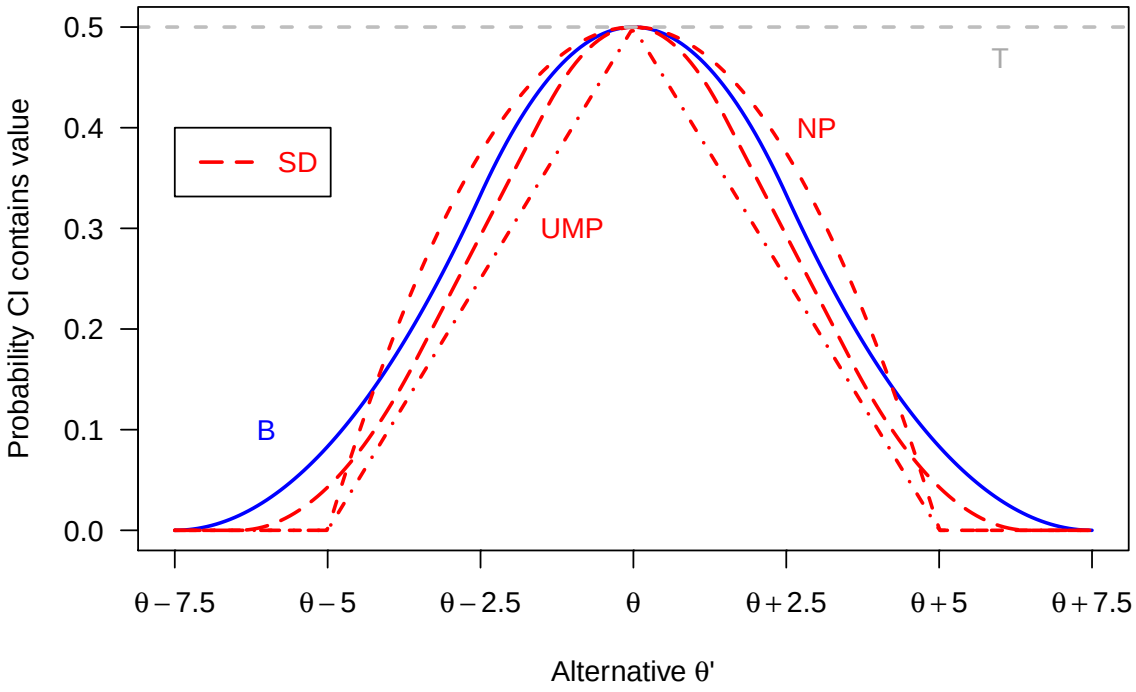

Figure 3. The probability that each confidence procedure includes false values θ'. T: Trivial procedure; SD: Sampling distribution procedure NP: Nonparametric procedure; UMP: UMP procedure; B: Bayes procedure. The line for the sampling distribution procedure (dashed line) is between the lines for the Bayes procedure and the UMP procedure.

The trivial procedure (T; gray horizontal line) is obviously a bad interval because it includes every false value with the same frequency as the true value. The trivial procedure will be worse than any other procedure, unless the procedure is specifically constructed to be pathological. The UMP procedure (UMP), on the other hand, is

better than every other procedure for every value of θ'. This is due to the fact that it was created by inverting a most-powerful test. No other confidence procedure can be better.

The ordering among the three remaining procedures can be seen by comparing their curves. The sampling distribution procedure (SD) is always superior to the Bayes procedure (B), but not to the nonparametric procedure (NP). The nonparametric procedure and the Bayes procedure curves overlap, so one is not preferred to the other. Welch (1939) remarked that the Bayes procedure is "not the best way of constructing confidence limits" using precisely the frequentist comparison shown in Figure 3 with the UMP interval.[3]

The frequentist comparison between the procedures is instructive, because we have arrived at an ordering of the procedures employing the criteria suggested by Neyman and used by the modern developers of new confidence procedures: coverage and power. The UMP procedure is the best, followed by the sampling distribution procedure. The sampling distribution procedure is better than the Bayes procedure. The nonparametric procedure is not preferred to any interval, but neither is it the worst.

We can also examine the procedures from a Bayesian perspective, which primarily concerned with whether the inferences are reasonable in light of the data and what was known before the data were observed (Howson & Urbach, 2006). We have already seen that interpreting the non-Bayesian procedures in this way leads to trouble, and that the Bayesian procedure, unsurprisingly, has better properties in this regard. We will show how the Bayesian interval was derived in order to provide insight into why it has good properties.

Consider the left column of Figure 4, which shows Bayesian reasoning from prior and likelihood to posterior and so-called credible interval. The prior distribution in the top panel shows that prior to observing the data, all the locations are equally probable.

[3]Several readers of a previous draft of this manuscript have noted that frequentists use the likelihood as well, and so may prefer the Bayesian procedure in this example. However, as Neyman (1977) points out, the likelihood has no special status to a frequentist; what matters is the frequentist properties of the procedure, not how it was constructed.

Upon observing the bubbles shown in Figure 1A — also shown in the top of the "likelihood panel" — the likelihood is a function that is 1 for all possible locations for the hatch, and 0 otherwise. To combine our prior knowledge with the new information from the two bubbles, we multiply the prior by the likelihood, which results in the posterior distribution in the bottom row. The central 50% credible interval contains all values in the central 50% of the area of the posterior, shown as the shaded region. The right column of Figure 4 shows a similar computation using a prior distribution that does not assume that all locations are equally likely, as might occur if some other information about the location of the submersible were available.

It is now obvious why the Bayesian credible interval has the properties typically ascribed to confidence intervals. The credible interval can be interpreted as having a 50% probability of containing the true value, because the values within it account for 50% of the posterior probability. It reveals the precision of our knowledge of the parameter, in light of the data and prior, through its relationship with the posterior and likelihood.

Of the five procedures considered, intervals from the Bayesian procedure are the *only ones* that can be said to have 50% probability of containing the true value, upon observing the data. Importantly, the ability to interpret the interval in this way arises from Bayesian theory and not from confidence interval theory. Also importantly, it was necessary to stipulate a prior to obtain the desired interval; the interval should be interpreted in light of the stipulated prior. Of the other four intervals, none can be justified as providing a "reasonable" inference or conclusion from the data, because of their strange properties and that there is no possible prior distribution that could lead to these procedures. In this light, it is clear why Neyman's rejection of "conclusions" and "reasoning" from data naturally flowed from his theory. It is also clear why scientists seeking a method to draw inferences from data might want want to reject confidence interval theory as a basis for evaluating intervals, if they care about the making reasonable inferences from data.

We can now review what we know concerning the four procedures procedures.

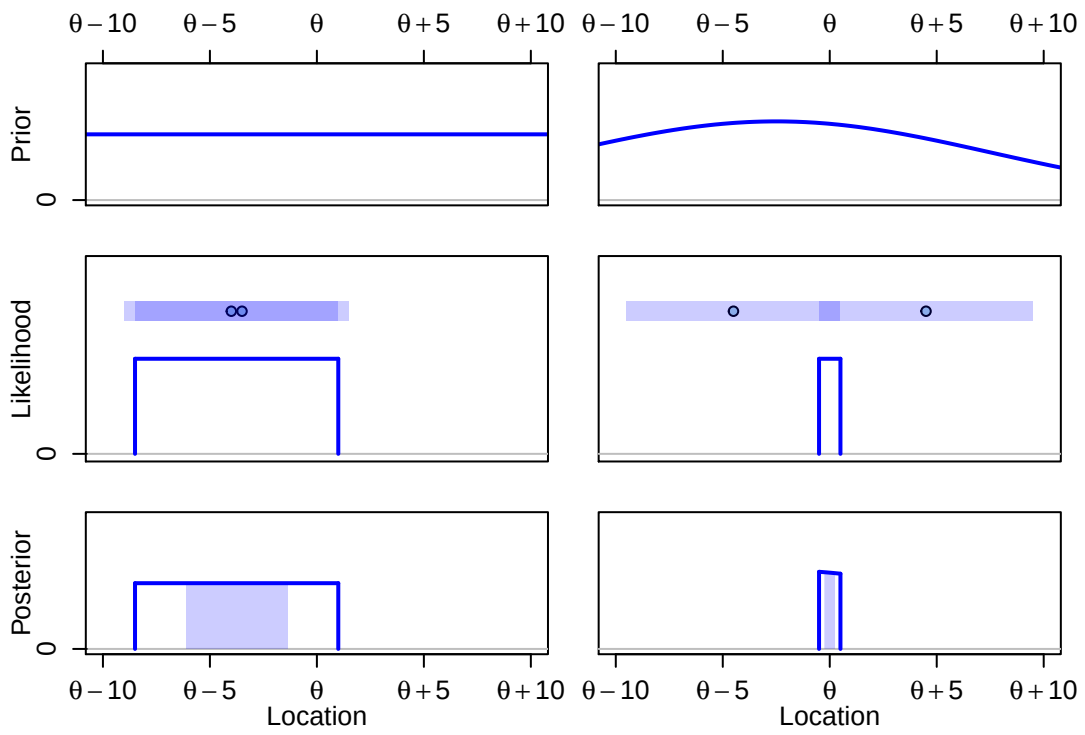

Figure 4. Forming Bayesian credible intervals. Prior information (top) is combined with the likelihood information from the data (middle) to yield a posterior distribution (bottom). In the likelihood plots, the shaded regions show the locations within 5 meters of each bubble; the dark shaded regions show where these overlap, indicating the possible location of the hatch θ. In the posterior plots, the central 50% region (shaded region within posterior) shows one possible 50% credible interval, the central credible interval.

Only the Bayesian procedure — when its intervals are interpreted as credible intervals — allows the interpretation that there is a 50% probability that the hatch is located in the interval. Only the Bayesian procedure properly tracks the precision of the estimate. Only the Bayesian procedure covers the plausible values in the expected way: the other procedures produce intervals that are *known* with certainty — by simple logic — to contain the true value, but still are "50%" intervals. Yet the Bayesian procedure is rejected by frequentist confidence interval theory as inferior.

The disconnect between frequentist theory and Bayesian theory arises from the different goals of the two theories. Frequentist theory is a "pre-data" theory. It looks

forward, devising procedures that will have particular average properties in repeated sampling (Jaynes, 2003; Mayo, 1981, 1982) in the future (see also Neyman, 1937, p. 349). This thinking can be clearly seen in Neyman (1942) as quoted above: reasoning ends once the procedure is derived. Confidence interval theory is vested in the average frequency of including or excluding true and false parameter values, respectively. Any given inference may — or may not — be reasonable in light of the observed data, but this is not Neyman's concern; he disclaims any conclusions or beliefs on the basis of data. Bayesian theory, on the other hand, is a post-data theory: a Bayesian analyst uses the information in the data to determine what is reasonable to believe, in light of the model assumptions and prior information.

Using an interval justified by a pre-data theory to make post-data inferences can lead to unjustified, and possibly arbitrary, inferences. This problem is not limited to the pedagogical pedagogical submersible example (J. O. Berger & Wolpert, 1988; Wagenmakers et al., 2014), but having a simple example is instructive when identifying these issues. In the next section we show how a commonly-used confidence interval leads to similarly flawed post-data inferences.

Example 2: A confidence interval in the wild

The previous example was designed to show, in a simple, accessible example, the logic of confidence interval theory. Further, it shows that confidence procedures cannot be *assumed* to have the properties that analysts desire.

When presenting the confidence intervals, CI proponents almost always focus on estimation of the mean of a normal distribution. In this simple case, frequentist, fiducial, and objective Bayesian answers coincide.[4] However, the proponents of confidence intervals suggest the use of confidence intervals for many other quantities: for instance, standardized effect size Cohen's d (G. Cumming & S. Finch, 2001), medians, (Bonett & Price, 2002; Olive, 2008), correlations (Zou, 2007), ordinal association (Woods, 2007), and many others. Quite often authors of such articles

[4]This should not be taken to mean that inference by confidence intervals is not problematic even in this simple case; see e.g., Brown (1967), Buehler and Feddersen (1963).

provide no analysis of the properties of the proposed confidence intervals beyond showing that contains the true value in the correct proportion of samples: that is, that it is a confidence interval. Sometimes the authors provide an analysis of the frequentist properties of the interval, such as average width. The developers of new confidence procedures do not, however, examine whether their procedures allow for quality post-data reasoning.

As the first example showed, a sole focus on frequentist properties of procedures is potentially disastrous for users of these confidence procedures because a confidence procedure has no guarantee of supporting reasonable inferences about the parameter of interest. Casella (1992) underscores this point with confidence intervals, saying that "we must remember that practitioners are going to make conditional (post-data) inferences. Thus, we must be able to assure the user that any inference made, either pre-data or post-data, possesses some definite measure of validity" (p. 10). Any development of an interval procedure that does not, at least in part, focus on its post-data properties is incomplete at best and extremely misleading at worst: *caveat emptor*.

Can such misleading inferences occur using procedures suggested by proponents of confidence intervals, and in use by researchers? The answer is yes, which we will show by examining a confidence interval for ω^2, the proportion of variance accounted for in ANOVA designs. The parameter ω^2 serves as a measure of effect size when there are more than two levels in a one-way design. This interval was suggested by Steiger (2004, see also Steiger & Fouladi, 1997), cited approvingly by G. Cumming (2014), implemented in software for social scientists (e.g., Kelley, 2007a, 2007b), and evaluated, solely for its frequentist properties, by W. H. Finch and French (2012). The problems we discuss here are shared by other related confidence intervals, such as confidence intervals for η^2, partial η^2, the noncentrality parameter of the F distribution, the signal-to-noise ratio f, RMSSE Ψ, and others discussed by Steiger (2004).

Steiger (2004) introduces confidence intervals by emphasizing a desire to avoid significance tests, and to focus more on the precision of estimates. Steiger says that "the scientist is more interested in knowing how large the difference between the two

groups is (and how precisely it has been determined) than whether the difference between the groups is 0" (pp. 164-165). Steiger and Fouladi (1997) say that "[t]he advantage of a confidence interval is that the width of the interval provides a ready indication of the precision of measurement..." (p. 231). Given our knowledge of the precision fallacy this should raise a red flag.

Steiger then gives a confidence interval on the ω^2 by inverting a significance test. Given the strange behavior of the UMP procedure in the submersible example, this too should raise a red flag. A confidence procedure based on a test, even a good, high-powered test, will not in general yield a procedure that provides for reasonable inferences. We will outline the logic of building a confidence interval by inverting a significance test before showing how Steiger's confidence interval behaves with data.

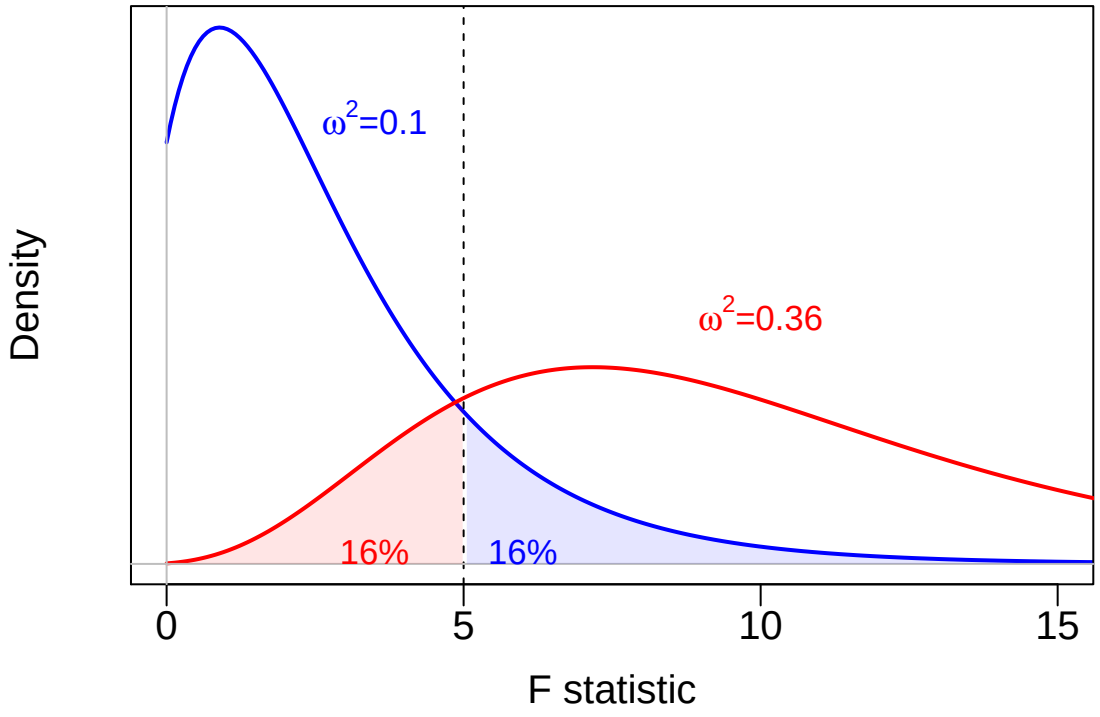

Figure 5. Building a confidence interval by inverting a significance test. The two noncentral F distributions with 16% of F values less than, or greater than, $F = 5$ respectively. When $F = 5$, the 100 - 16 - 16 = 68% confidence interval is [.1, .36].

In order to understand how a confidence interval can be built by inverting a significance test, consider that a two-sided significance test of size α can be thought of

as a combination of two one-sided tests at size $\alpha/2$: one for each tail. The two-sided test rejects when one of the one-tailed tests rejects. To build a 68% confidence interval (e.g., to approximate a standard error), we can use two one-sided tests of size $(1 - .68)/2 = .16$. Suppose we have a one-way design with three groups and $N = 10$ participants in each group. The effect size ω^2 in such a design indexes how large we expect F to be. The distribution of F given the effect size ω^2 is called the noncentral F distribution. When $\omega^2 = 0$ — that is, there is no effect — the familiar central F distribution is obtained.

Figure 5 shows that the value $F(2, 27) = 5$ yields $p = .16$ for two one-sided tests: the one-sided test of $\omega^2 = .1$ that rejects when F is large, and the one-sided test of $\omega^2 = .36$ that rejects when F is small. For any ω^2 value in $[.1, .36]$, the p value for both one-sided tests will be greater than $p > .16$, and hence will not be rejected. A 68% confidence interval for when $F = 5$ can be defined as all ω^2 values that are not rejected by one of the two-tailed tests, and so $[.1, .36]$ is taken as a 68% confidence interval. Because ω^2 cannot be less than 0, sometimes there is no one-sided test that yields $p = .16$. When the p value from the ANOVA F test is greater than $\alpha/2 = .16$, there is no upper-tailed test that will yield .16, so no lower bound on the CI exists. If the F test p value is greater than $1 - \alpha/2$, neither bound exists. If a bound does not exist, Steiger (2004) arbitrarily sets it at 0.

To see how this CI works in practice, suppose we design a three-group, between-subjects experiment with $N = 10$ participants in each group and obtain an $F(2, 27) = 0.18, p = 0.84$. Following recommendations for good analysis practices (e.g. Psychonomics Society, 2012; Wilkinson & the Task Force on Statistical Inference, 1999), we would like to compute a confidence interval on the standardized effects size ω^2. Using software to compute Steiger's CI, we obtain the 68% confidence interval $[0, 0.01]$.

Figure 6A (top interval) shows the resulting 68% interval. If we were not aware of the fallacies of confidence intervals, we might publish this confidence interval thinking provides a good measure of the precision of the estimate of ω^2. Note that the lower limit of the confidence interval is exactly 0, because the lower bound did not exist. In

discussing this situation Steiger and Fouladi (1997) say

> "[Arbitrarily setting the confidence limit at 0] maintains the correct coverage probability for the confidence interval, but the width of the confidence interval may be suspect as an index of the precision of measurement when either or both ends of the confidence interval are at 0. In such cases, one might consider obtaining alternative indications of precision of measurement, such as an estimate of the standard error of the statistic." (Steiger and Fouladi, 1997, p. 255)

Steiger (2004) further notes that "relationship [between CI width and precision] is less than perfect and is seriously compromised in some situations for several reasons" (p. 177). This is rather startling; a major part of the justification for confidence intervals, including the one computed here, is that confidence intervals supposedly allow an assessment of the precision with which the parameter is estimated. The confidence interval fails to meet the purpose for which it was advocated in the first place.

We can confirm the need for Steiger's caution — essentially, a warning about the precision fallacy — by looking at the likelihood, which is probability density of the observed F statistic computed for all true values of ω^2. Notice how narrow the confidence interval is compared to the likelihood of ω^2. The likelihood falls much more slowly as ω^2 gets larger than the confidence interval would appear to imply, if we believed the precision fallacy. We can also compare the confidence interval to a 68% Bayesian credible interval, computed assuming standard "noninformative" priors on the means and the error variance[5]. The Bayesian credible interval is substantially wider, revealing the imprecision with which ω^2 is estimated.

Figure 6B shows the same case, but for a slightly smaller F value. The precision with which ω^2 is estimated has not changed to any measurable degree; yet now the

[5]See supplement for details. We do not generally advocate non-informative priors on parameters of interest (Rouder, Morey, Speckman, & Province, 2012; Wetzels, Grasman, & Wagenmakers, 2012); in this instance we use them as a comparison because many people believe, incorrectly, that confidence intervals numerically correspond to Bayesian credible intervals with noninformative priors.

confidence interval contains only the value $\omega^2 = 0$: or, more accurately, the confidence interval is empty because this F value would always be rejected by one of the pairs of one-sided tests that led to the construction of the confidence interval. As Steiger points out, a "zero-width confidence interval obviously does not imply that effect size was determined with perfect precision," (p. 177), nor can it imply that there is a 68% probability that ω^2 is exactly 0. This can be clearly seen by examining the likelihood and Bayesian credible interval.

Some authors (e.g., Dufour, 1997) interpret empty confidence intervals as indicative of model misfit. In the case of this one sample design, if the confidence interval is empty then the means are more similar than would be expected even under the null hypothesis $\alpha/2$ of the time; that is, $p > 1 - \alpha/2$, and hence F is small. If this model rejection significance test logic is used, the confidence interval itself becomes uninterpretable as the model gets close to rejection, because it appears to indicate false precision (Gelman, 2011). Moreover, in this case the p value is certainly more informative than the CI; the p value provides graded information, while the CI is simply empty for all values of $p > 1 - \alpha/2$.

Panel C shows what happens when we increase the confidence coefficient slightly to 70%. Again, the precision with which the parameter is estimated has not changed, yet the confidence interval now again has nonzero width.

Figure 6D shows the results of an analysis with $F(2, 27) = 4.24, p = 0.03$, and using a 95% confidence interval. Steiger's interval has now encompassed most of the likelihood, but the lower bound is still "stuck" at 0. In this situation, Steiger and Fouladi advise us that the width of the CI is "suspect" as an indication of precision, and that we should "obtain[] [an] alternative indication[] of precision of measurement." As it turns out, here the confidence interval is not too different from the credible interval, though the confidence interval is longer and is unbalanced. However, we would not know this if we did not examine the likelihood and the Bayesian credible interval; the only reason we know the confidence interval has a reasonable width in this particular case is its agreement with the actual measures of precision offered by the

likelihood and the credible interval.

How often will Steiger's confidence procedure yield a "suspect" confidence interval? This will occur whenever the p value for the corresponding F test is $p > \alpha/2$; for a 95% confidence interval, this means that whenever $p > 0.025$, Steiger and Fouladi recommend against using the confidence interval for precisely the purpose that they — and other proponents of confidence intervals — recommend it for. This is not a mere theoretical issue; moderately-sized p values often occur. In a cursory review of papers using Steiger's confidence intervals, we found many that obtained and reported, without note, suspect confidence intervals bounded at 0 (e.g., S. P. Cumming, Sherar, Gammon, Standage, & Malina, 2012; Gilroy & Pearce, 2014; Hamerman & Morewedge, 2015; Lahiri, Maloney, Rogers, & Ge, 2013; Hamerman & Morewedge, 2015; Todd, Vurbic, & Bouton, 2014; Winter et al., 2014). The others did not use confidence intervals, instead relying on point estimates of effect size and p values (e.g. Hollingdale & Greitemeyer, 2014); but from the p values it could be inferred that if they had followed "good practice" and computed such confidence intervals, they would have obtained intervals that according to Steiger could not be interpreted as anything but an inverted F test.

It makes sense, however, that authors using confidence intervals would not note that the interpetation of their confidence intervals is problematic. If a confidence intervals truly contained the most likely values, or if it were an index of the precision, or if the confidence coefficient indexed the uncertainty we should have that the the parameter is in the interval, then it would seem that a CI is a CI: what you learn from one is the same as what you learn from another. The idea that the p value can determine whether the interpretion of a confidence interval is possible is not intuitive in light of the way CIs are typically presented.

We see no reason, however, why our ability to interpret an interval *should* be compromised simply because we obtained a p value that was not low enough. Certainly, the confidence coefficient is arbitrary; if the width is suspect for one confidence coefficient, it makes little sense that the CI width would become acceptable just because we changed the confidence coefficient so the interval bounds did not include 0. Also, if

the width is too narrow with moderate p values, such that it is not an index of precision, it seems that the interval will be too wide in other circumstances, possibly threatening the interpretation as well. This was evident with the UMP procedure in the submersible example: the UMP interval was too narrow when the data provided little information, and was too wide when the data provided substantial information.

Steiger and Fouladi (1997) summarize the central problem with confidence intervals when they say that in order to maintain the correct coverage probability — a frequentist pre-data concern — they sacrifice the very thing researchers want confidence intervals to be: a post-data index of the precision of measurement. If our goal is to move away from significance testing, we should not use methods which cannot be interpreted except as inversions of significance tests. We agree with Steiger and Fouladi that researchers should consider obtaining alternative indications of precision of measurement; luckily, Bayesian credible intervals fit the bill rather nicely, rendering confidence intervals unnecessary.

Discussion

Using the theory of confidence intervals and the support of two examples, we have shown that CIs do not have the properties that are often claimed on their behalf. Confidence interval theory was developed to solve a very constrained problem: how can one construct an procedure that produces intervals containing the true parameter a fixed proportion of the time? Claims that confidence intervals yield an index 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 and unjustified by confidence interval theory.

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 fallacy is driven by a desire to determine which values of the parameter should be taken seriously; and the precision fallacy is driven by a desire to quantify the precision of the

estimates. We support these goals (Morey, Rouder, Verhagen, & Wagenmakers, 2014), but confidence interval theory is not the way to achieve them.

Guidelines for interpreting and reporting intervals

Frequentist theory can be counter-intuitive at times; as Fisher was fond of pointing out, frequentist theorists often seemed disconnected with the concerns of scientists, developing methods that did not suit their needs (e.g., Fisher, 1955, p. 70). This has lead to confusion where practitioners assume that methods designed for one purpose were really meant for another. In order to help mitigate such confusion, here we would like to offer readers a clear guide to interpreting and reporting confidence intervals.

Once one has collected data and computed a confidence interval, how does one then interpret the interval? The answer is quite straightforward: one does not – at least not within confidence interval theory.[6] As Neyman and others pointed out repeatedly, and as we have shown, confidence limits cannot be interpreted as anything besides the result of a procedure that will contain the true value in a fixed proportion of samples. Unless an interpretation of the interval can be specifically justified by some *other* theory of inference, confidence intervals must remain uninterpreted, lest one make arbitrary inferences or inferences that are contradicted by the data. This applies even to "good" confidence intervals, as these are often built by inverting significance tests and may have strange properties (e.g. Steiger, 2004).

In order to help mitigate confusion in the scientific literature, we suggest the following guidelines for reporting of confidence intervals, all informed by our discussion in this manuscript.

Report credible intervals instead. We believe any author who chooses to use

[6]Some recent writers have suggested replacing Neyman's behavioral view on confidence intervals with a frequentist view focused on tests at various levels of "stringency" (see, e.g., Mayo & Cox, 2006; Mayo & Spanos, 2006). Readers who prefer a frequentist paradigm may wish to explore this approach; however, we are unaware of any comprehensive account of CIs within this paradigm, and regardless, it does not offer the properties desired by CI proponents. This is not to be read as an argument against it, but rather a warning that one must make a choice.

confidence intervals should ensure that the intervals correspond numerically with credible intervals under some reasonable prior. Many confidence intervals cannot be so interpreted, but if the authors know they can be, they should be called "credible intervals". This signals to readers that they can interpret the interval as they have been (incorrectly) told they can interpret confidence intervals. Of course, the corresponding prior must also be reported.

This is not to say that one can't also call them confidence intervals if they are; however, readers are likely more interested in the post-data properties of the procedure — not the long-run coverage — if they are interested arriving at substantive conclusions from the interval.

Do not use procedures whose Bayesian properties are not known. As Casella (1992) pointed out, the post-data properties of a procedure are necessary for understanding what can be inferred from an interval. Any procedure whose Bayesian properties have not been explored can have properties that make it unsuitable for post-data inference. Procedures whose properties have not been adequately studied are inappropriate for general use.

Warn readers if the confidence procedure does not correspond to a Bayesian procedure. If it is known that a confidence interval does not correspond to a Bayesian procedure, warn readers that the confidence interval cannot be interpreted as having a X% probability of containing the parameter, that it cannot be interpreted in terms of the precision of measurement, and that cannot be said to contain the values that should be taken seriously: the interval is merely an interval that, prior to sampling, had a X% probability of containing the true value. Authors choosing to report CIs have a responsibility to keep their readers from invalid inferences, because it is almost sure that readers will misinterpret them without a warning (Hoekstra et al., 2014).

***Never* report a confidence interval without noting the procedure and the corresponding statistics.** As we have described, there are many different ways to construct confidence intervals, and they will have different properties. Some will have better frequentist properties than others; some will correspond to credible intervals, and

others will not. It is unfortunately common for authors to report confidence intervals without noting how they were constructed. As can be seen from the examples we've presented, this is a terrible practice because without knowing which confidence intervals was used, it is unclear what can be inferred. In the submersible example, consider a 50% confidence interval .5 meters wide. This could correspond to very precise information (Bayesian interval) or very imprecise information (UMP and nonparametric interval). Not knowing which procedure was used could lead to very poor inferences. In addition, enough information should be presented so that any reader can compute a different confidence interval or credible interval. In most cases, this is covered by standard reporting practices, but in other cases more information may need to be given.

Consider reporting likelihoods or posteriors instead. An interval provides fairly impoverished information. Just as proponents of confidence intervals argue that CIs provide more information than a significance test (although this is debatable for many CIs), a likelihood or a posterior provides much more information than an interval. Recently, G. Cumming (2014) has proposed so-called "cat's eye" intervals which correspond to either fiducial distributions or Bayesian posteriors under a "non-informative" prior. With modern scientific graphics so easy to create, we see no reason why likelihoods and posteriors cannot augment or even replace intervals in most circumstances (e.g. Kruschke, 2010). With a likelihood or a posterior, the arbitrariness of the confidence or credibility coefficient is avoided altogether.

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, and hence all confidence intervals can simply be interpreted in a Bayesian way. 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 might lead one to suspect that it does not matter whether one uses confidence procedures or Bayesian

procedures. We showed, however, that confidence intervals and credible intervals can disagree markedly. The only way to know that a confidence interval is numerically identical to some credible interval is to *prove* it. The correspondence cannot — and should not — be assumed.

More broadly, 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; if one is committed to the procedure that allows a Bayesian interpretation, then one's time is much better spent simply applying Bayesian theory. If the benefits of Bayesian theory are desired — and they clearly are, by proponents of confidence intervals — then there is no reason why Bayesian inference should not be applied in its full generality.

It is important to emphasize, however, that for many of the confidence procedures presented in the applied statistical literature, no effort has been made to show that the intervals have the properties that proponents of confidence intervals desire. We should expect, as a matter of course, that developers of new confidence intervals show that their intervals have the desired inferential properties, instead of just proper coverage of the true value and "short" width. Because developers of confidence intervals have not done this, the push for confidence intervals rests on uncertain ground. Adopting Bayesian inference, where all inferences arise within a logical, unified framework, would render the problems of assessing the properties of these confidence procedures moot. If desired, coverage can also be assessed, but if one is interested primarily in reasonable post-data inference, then those properties should be the priority, not coverage (c.f. Gelman, 2008; Wasserman, 2008).

For advocates of reasoning by intervals, adopting Bayesian inference would have other benefits. 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 confidence procedure can be constructed, even though such information may be critically important to a researcher, policy maker, or criminal defendant (Pratt, Raiffa, & Schlaifer, 1995).

Even in seemingly simple cases where a fixed interval is nested inside a CI, or *vice versa*, one cannot draw conclusions about the probability of a fixed interval. One might assume that an interval nested within a CI must have lower confidence than the CI; however, as shown in Figure 1B, a 100% confidence interval (the likelihood) is nested within some of the 50% confidence intervals. Likewise, one might believe that if a CI is nested within a fixed interval, then the fixed interval must have greater probability than the interval. But in Figure 1A, one can imagine a fixed interval just larger than the 50% UMP interval; this will have much lower than 50% probability of containing the true value, due to the fact that it occupies a small proportion of the likelihood. Knowledge of the FCF prohibits one from using confidence intervals to assess the probability of fixed intervals. Bayesian procedures, on the other hand, offer the ability to compute the plausibility of any given range of values. Because all such inferences must be made from the posterior distribution, inferences must remain mutually consistent (Lindley, 1985; see also Fisher, 1935, for a similar argument).

Moving to credible intervals from confidence intervals would necessitate a shift in thinking, however, away from a test-centric view (e.g., "is 0 in the interval?"). Although every confidence interval can be interpreted as a test, this is not true of credible intervals. Assessing the Bayesian credibility of a specific parameter value by whether it is included in a credible interval is, as James O. Berger (2006) puts it, "simply wrong." When testing a specific value is of interest (such as a null hypothesis), that specific value must be assigned non-zero probability *a priori*. While not conceptually difficult, it is beyond the scope of this paper; see Rouder, Speckman, Sun, Morey, and Iverson (2009), Wagenmakers, Lee, Lodewyckx, and Iverson (2008), or Dienes (2011) for

accessible accounts.

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, and does not require priors. 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. As they say, there is no such thing as a free lunch; one must choose. We suspect that researchers, given the choice, would rather specify priors and get the benefits that come from Bayesian theory. We should not pretend, however, that the choice need not be made. Confidence interval theory and Bayesian theory are not interchangeable, and should not be treated so.

Conclusion

We have suggested that confidence intervals do not support the inferences that their advocates believe they do. It is an interesting question how the theory of confidence intervals began with Neyman as a method of avoiding the problem of reasoning from data by making dichotomous statements (Neyman, 1937, 1941), eventually becoming a method that many believe is the best way to reason from data (e.g. G. Cumming & S. Finch, 2005; G. Cumming & Fidler, 2009) and a way to avoid dichotomous statements (e.g. G. Cumming, 2014; Hoekstra, Finch, Kiers, & Johnson, 2006; Wilkinson & the Task Force on Statistical Inference, 1999). However this confusion ultimately started, we think it is critical in the current atmosphere of statistical reform that we ensure that claims made on behalf of confidence interval theory have a firm foundation. We show here that they do not.

We do not believe that the theory of confidence intervals provides a viable foundation for the future of psychological methods. Confidence procedures that do not have Bayesian properties have other undesirable properties; confidence procedures that *do* have Bayesian properties can be justified using Bayesian theory. If we were to give

up the use of confidence procedures, what would we lose? Abandoning the use of confidence procedures means abandoning a method that merely allows us to create intervals that include 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

Basu, D. (1981). On ancillary statistics, pivotal quantities, and confidence statements. In Y. P. Chaubey & T. D. Dwivedi (Eds.), *Topics in applied statistics* (pp. 1–29). Montreal: Concordia University.

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

Berger, J. O. [James O.]. (2006). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Second edition, Vol. 1, pp. 378–386). Hoboken, New Jersey: John Wiley & Sons.

Blaker, H. & Spjøtvoll, E. (2000). Paradoxes and improvements in interval estimation. *The American Statistician*, *54*(4), 242–247. Retrieved from http://www.jstor.org/stable/2685774

Bonett, D. G. & Price, R. M. (2002). Statistical inference for a linear function of medians: confidence intervals, hypothesis testing, and sample size requirements. *Psychological Methods*, *7*, 370–383.

Brown, L. (1967). The conditional level of Student's t test. *The Annals of Mathematical Statistics*, *38*(4), 1068–1071. Retrieved from http://www.jstor.org/stable/2238826

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. Retrieved from http://www.jstor.org/stable/2238493

Casella, G. (1992). Conditional inference from confidence sets. *Lecture Notes-Monograph Series*, *17*, 1–12. Retrieved from http://www.jstor.org/stable/4355622

Casella, G. & Berger, R. L. (2002). *Statistical inference*. Pacific Grove, CA: Duxbury.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Cumming, G. & Fidler, F. (2009). Confidence intervals: better answers to better questions. *Zeitschrift für Psychologie*, *217*, 15–26.

Cumming, G. & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532–574.

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

Cumming, S. P., Sherar, L. B., Gammon, C., Standage, M., & Malina, R. M. (2012, December). Physical Activity and Physical Self-Concept in Adolescence: A Comparison of Girls at the Extremes of the Biological Maturation Continuum. *Journal of Research on Adolescence*, *22*(4), 746–757. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1532-7795.2012.00821.x/abstract

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

Dufour, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica*, *65*(6), 1365–1387. Retrieved from http://www.jstor.org/stable/2171740

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.

Fidler, F. & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement*, *61*, 575–604.

Finch, W. H. & French, B. F. (2012, February). A Comparison of Methods for Estimating Confidence Intervals for Omega-Squared Effect Size. *Educational and Psychological Measurement*, *72*(1), 68–77. WOS:000300176100005.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of eugenics*, *6*, 391–398.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological)*, *17*, 69–78. Retrieved from http://www.jstor.org/stable/2983785

Gelman, A. (2008). Rejoinder. *Bayesian analysis*, *3*, 467–478. Retrieved from http://www.stat.columbia.edu/~gelman/research/published/badbayesresponsemain.pdf

Gelman, A. (2011, August). Why it doesn't make sense in general to form confidence intervals by inverting hypothesis tests. [blog post]. Retrieved from http://andrewgelman.com/2011/08/25/why_it_doesnt_m/

Gilroy, K. E. & Pearce, J. M. (2014, April). The Role of Local, Distal, and Global Information in Latent Spatial Learning. *Journal of Experimental Psychology*, *40*(2), 212–224.

Hamerman, E. J. & Morewedge, C. K. (2015, March). Reliance on Luck: Identifying Which Achievement Goals Elicit Superstitious Behavior. *Personality and Social Psychology Bulletin*, *41*(3), 323–335. WOS:000349626400002.

Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review*, *13*, 1033–1037.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164.

Hollingdale, J. & Greitemeyer, T. (2014, November). The Effect of Online Violent Video Games on Levels of Aggression. *PLoS ONE*, *9*(11), e111790. Retrieved from http://dx.doi.org/10.1371/journal.pone.0111790

Howson, C. & Urbach, P. (2006). *Scientific reasoning: the Bayesian approach*. La Salle, Illinois: Open Court.

Jaynes, E. (2003). *Probability theory: the logic of science*. Cambridge, UK: Cambridge University Press.

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

Kelley, K. (2007a). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, *20*(8).

Kelley, K. (2007b). Methods for the behavioral, educational, and social sciences: an R package. *Behavioral Research Methods*, *39*(4), 979–984.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*(7), 293–300.

Lahiri, D. K., Maloney, B., Rogers, J. T., & Ge, Y.-W. (2013, January). PuF, an antimetastatic and developmental signaling protein, interacts with the Alzheimer's amyloid-beta precursor protein via a tissue-specific proximal regulatory element (PRE). *Bmc Genomics*, *14*, 68. WOS:000315328100001.

Lehmann, E. H. (1959). *Testing statistical hypotheses*. New York: John Wiley & Sons.

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

Lindley, D. V. (1985). *Making decisions* (2nd ed.). London: Wiley.

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. Retrieved from http://faculty.washington.edu/gloftus/Downloads/ThousandpValues.pdf

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. Retrieved from http://faculty.washington.edu/gloftus/Downloads/CurrentDirections.pdf

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

Mayo, D. G. (1981). In defense of the Neyman-Pearson theory of confidence intervals. *Philosophy of Science*, *48*(2), 269–280.

Mayo, D. G. (1982). On after-trial criticisms of Neyman-Pearson theory of statistics. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1982*, 145–158.

Mayo, D. G. & Cox, D. R. (2006). Frequentist statistics as a theory of inductive inference. *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, *49*, 77–97.

Mayo, D. G. & Spanos, A. (2006). Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *British Journal for the philosophy of science*, *57*, 323–357.

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: a comment on Cumming. *Psychological Science*, 1289–1290.

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. Retrieved from http://www.jstor.org/stable/2342192

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. Retrieved from http://www.jstor.org/stable/91337

Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika*, *32*(2), 128–150. Retrieved from http://www.jstor.org/stable/2332207

Neyman, J. (1952). *Lectures and conferences on mathematical statistics and probability*. Washington, D.C.: Graduate School, U.S. Department of Agriculture.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, *36*(1), 97–131.

Olive, D. J. (2008). *Applied robust statistics*. online electronic book. Retrieved from http://lagrange.math.siu.edu/Olive/ol-bookp.htm

Pratt, J. W. (1961). Book review: Testing Statistical Hypotheses, by E. L. Lehmann. *Journal of the American Statistical Association*, *56*(293), pages. Retrieved from http://www.jstor.org/stable/2282344

Pratt, J. W., Raiffa, H., & Schlaifer, R. (1995). *Introduction to statistical decision theory*. Cambridge, MA: MIT Press.

Psychonomics Society. (2012). *Psychonomic Society guidelines on statistical issues.* Retrieved from http://www.springer.com/psychology?SGWID=0-10126-6-1390050-0

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., 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. Retrieved from http://dx.doi.org/10.3758/PBR.16.2.225

Rusu, F. & Dobra, A. (2008). Sketches for size of join estimation. *ACM Transactions on Database Systems*, *33*, 15:1–15:46.

Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*(2), 164–182.

Steiger, J. H. & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. Harlow, S. Mulaik, & J. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Mahwah, New Jersey: Erlbaum.

Stock, J. H. & Wright, J. H. (2000). Gmm with weak identification. *Econometrica*, *68*(5), 1055–1096. Retrieved from http://www.jstor.org/stable/2999443

Todd, T. P., Vurbic, D., & Bouton, M. E. (2014, July). Mechanisms of Renewal After the Extinction of Discriminated Operant Behavior. *Journal of Experimental Psychology*, *40*(3), 355–368.

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. Retrieved from http://dx.doi.org/10.1111/j.1464-0597.2008.00348.x

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., Verhagen, J., Ly, A., Bakker, M., Lee, D., M. D. Matzke, Rouder, J. N., & Morey, R. D. (2014). A power fallacy. *Behavioral Research Methods*.

Wasserman, L. (2008). Comment on article by Gelman. *Bayesian Analysis*, *3*, 463–466. Retrieved from http://ba.stat.cmu.edu/journal/2008/vol03/issue03/wasserman.pdf

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. Retrieved from http://www.jstor.org/stable/2235987

Wetzels, R., Grasman, R. P., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for ANOVA designs. *American Statistician*, *66*, 104–111.

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

Winter, C., Van Acker, F., Bonduelle, M., Desmyttere, S., De Schrijver, F., & Nekkebroeck, J. (2014, September). Cognitive and psychomotor development of 5-to 6-year-old singletons born after PGD: a prospective case-controlled matched study. *Human Reproduction*, *29*(9), 1968–1977. WOS:000343417900019.

Woods, C. M. (2007). Confidence intervals for gamma-family measures of ordinal association. *Psychological Methods*, *12*(2), 185–204.

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. Retrieved from http://www.sciencedirect.com/science/article/pii/S0196064497701665

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*(4), 399–413.

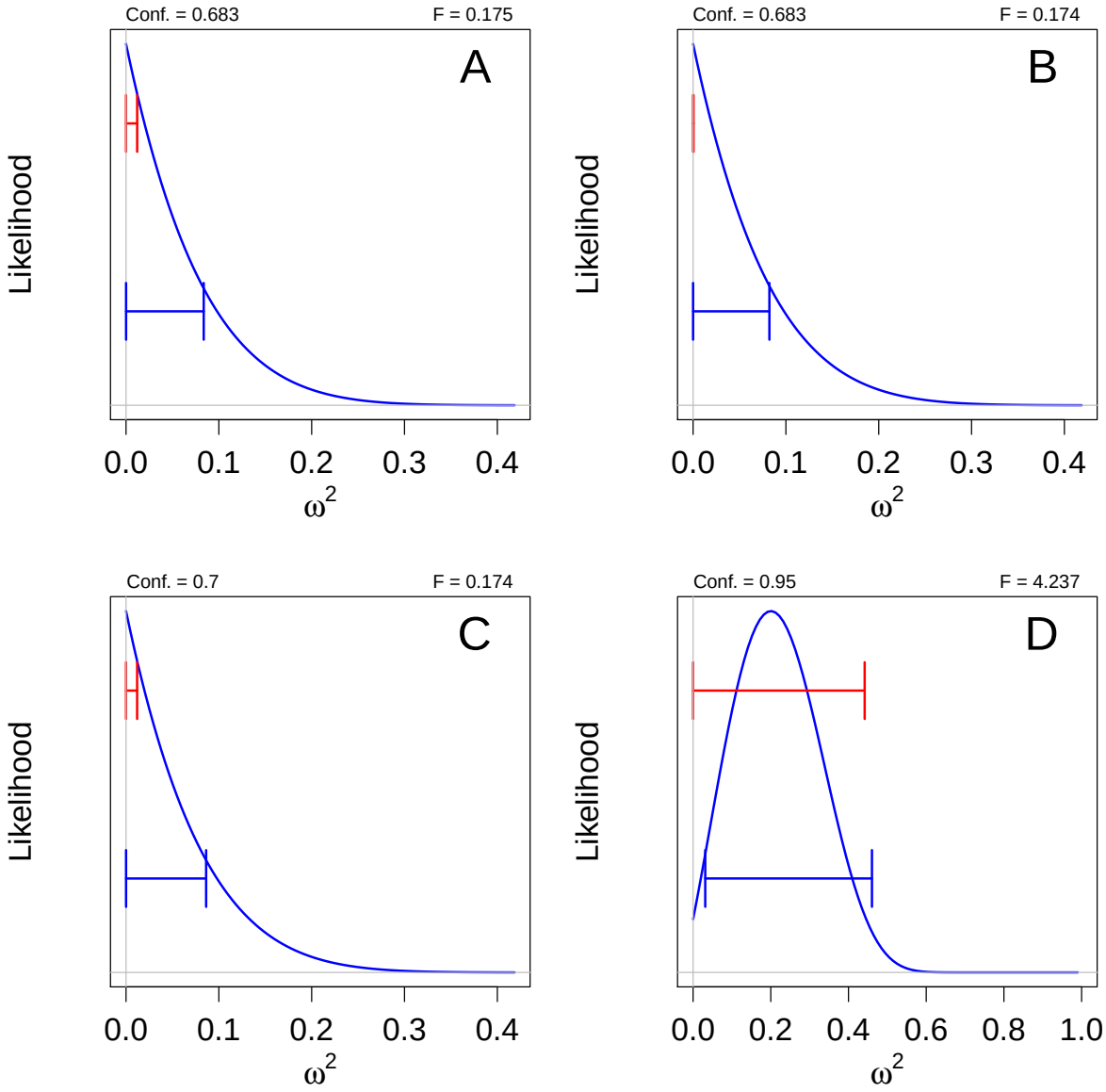

Figure 6. Likelihoods, confidence intervals, and Bayesian credible intervals (highest posterior density, or HPD, intervals) for four hypothetical experimental results. The top interval is Steiger's (2004) confidence interval for ω^2; the bottom interval is the Bayesian HPD. See text for details.