Supplement to "The Fallacy of Placing Confidence in Confidence Intervals"

Richard D. Morey (richarddmorey@gmail.com),
Rink Hoekstra, Jeffrey N. Rouder,
Michael D. Lee, Eric-Jan Wagenmakers

1 The lost submarine: details

In this section we provide more details about the submarine example. We example presented a situation where $N = 2$ observations were distributed uniformly:

$$x_i \overset{iid}{\sim} \text{Uniform}(\theta - 5, \theta + 5),\ i = 1, \ldots, N$$

and the goal is to estimate θ, the location of the submarine hatch. We consider three 50% confidence procedures for θ:

$$\begin{aligned}
&\text{CP1}: \quad \bar{x} \pm \tfrac{|x_1 - x_2|}{2}, \\
&\text{CP2}: \quad \bar{x} \pm \left(5 - \tfrac{5}{\sqrt{2}}\right), \\
&\text{CP3}: \quad \bar{x} \pm \tfrac{1}{4}\left(10 - |x_1 - x_2|\right)
\end{aligned}$$

Proof that these are 50% confidence procedures is straightforward and omitted. CP1 is the same as Student's t interval; CP2 is a fixed-width interval analogous to a z interval used when the population variance is known; CP3 is the Bayesian credible interval based on a uniform prior over θ.

Though we will not prove that CP3 is a 50% confidence procedure, we will show that it arises from Bayes' theorem assuming the use of a uniform prior for θ.The posterior distribution is proportional to the likelihood times the prior. The likelihood is

$$p(x_1, x_2 \mid \theta) \propto \prod_{i=1}^{2} \mathcal{I}(\theta - 5 < x_i < \theta + 5);$$

where $\mathcal{I}$ is an indicator function. Note since this is the product of two indicator functions, it can only be nonzero when both indicator functions' conditions are met; that is, when $x_1 + 5$ and $x_2 + 5$ are both greater than θ, and $x_1 - 5$

and $x_2 - 5$ are both less than θ. If the minimum of $x_1 + 5$ and $x_2 + 5$ is greater than θ, then so to must be the maximum. The likelihood thus can be rewritten

$$p(x_1, x_2 \mid \theta) \propto \mathcal{I}(x_{(2)} - 5 < \theta < x_{(1)} + 5);$$

where $x_{(1)}$ and $x_{(2)}$ are the minimum and maximum observations, respectively. If the prior for θ is proportional to a constant, then the posterior is

$$p(\theta \mid x_1, x_2) \propto \mathcal{I}(x_{(2)} - 5 < \theta < x_{(1)} + 5),$$

This posterior is a uniform distribution over all *a posteriori* possible values of θ (that is, all θ values within 5 meters of all observations), has width

$$10 - |x_1 - x_2|,$$

and is centered around $\bar{x}$. Because the posterior comprises all values of θ the data have not ruled out – and is essentially just the classical likelihood – the width of this posterior can be taken as an indicator of the precision of the estimate of θ. The posterior given the data in the two submarine scenarios is depicted in the top line of Figure 2 in the main text. A 50% credible interval can be constructed by taking half the width of the posterior centered around $\bar{x}$, leading to the formula for CP3.

1.1 Properties of the three confidence procedures

In this section we expand the points in the main text regarding the three myths of confidence intervals.

1.1.1 The precision error

That confidence intervals do not necessarily track the precision of an estimate can be easily seen by noting that the precision of the estimate is inversely proportional to the width of the likelihood, or

$$w(x_1, x_2) = 10 - |x_1 - x_2|.$$

A confidence interval's width should increase as w increases. A brief look at the formulas for CP1, CP2, and CP3 reveal that CP1's width is proportional to $|x_1 - x_2|$, and therefore inversely proportional to w. CP2's width is fixed, and so is also not proportional to w. CP3's width, on the other hand, is proportional to w (which is not surprising, given that it was derived by taking half the posterior's width).

1.1.2 The likelihood error

The easiest way to see that CP1 and CP2 can lead to the likelihood error is to note that they include impossible values. This will occur when the width of

the CI is greater than the width of the likelihood/posterior. For CP1, this will occur when:

$$\begin{aligned} w(x_1, x_2) &< 2 \times |x_1 - x_2|/2 \\ 10 - |x_1 - x_2| &< |x_1 - x_2| \\ |x_1 - x_2| &> 5 \end{aligned}$$

For CP2, this will occur when

$$\begin{aligned} w(x_1, x_2) &< 2 \times \left(5 - \frac{5}{\sqrt{2}}\right) \\ 10 - |x_1 - x_2| &< 10 - \frac{10}{\sqrt{2}} \\ |x_1 - x_2| &> \frac{10}{\sqrt{2}} \end{aligned}$$

Figure 1A shows the proportion of impossible values in confidence intervals from the three confidence procedures as a function of the difference between the two observed data points $|x_1 - x_2|$. Both CP1 and CP2 can include anywhere from 0% to 100% impossible values. If one uses using CP1 or CP2, if $|x_1 - x_2|$ is sufficiently large then one is certain that nearly all the values in the interval are impossible. Bayesian credible intervals, however, can never include impossible values.

1.1.3 Identification of relevant subsets

Demonstrating the precision and likelihood errors using CP1 and CP2 was straightforward. Identification of relevant subsets, on the other hand, is harder. Consider Figure 2A. On the horizontal axis, possible values of the first bubble, denoted x_1, are shown. The vertical axis shows possible values of the second bubble, x_2. The two shaded squares show values for x_1 and x_2 where a confidence interval from CP1 will contain the true location of the hatch, θ. This occurs when $x_1 > \theta$ and $x_2 < \theta$, or when $x_1 < \theta$ and $x_2 > \theta$. The fact that the procedure is a 50% CP can be seen by the fact that the two shaded squares occupy 1/2 of the total possible area.

Instead of considering the two values x_1 and x_2, it is actually more helpful to consider the location and width of the CI. The sample mean $\bar{x}$ is the center of the CI, and the width is $|\omega|$:

$$\begin{aligned} \bar{x} &= \frac{x_1 + x_2}{2} \\ \omega &= x_2 - x_1 \end{aligned}$$

Figure 2B contains the same information as panel A, except appropriately rotated so that the axes show $\bar{x}$ and ω. Possible values of $\bar{x}$ and ω are shown

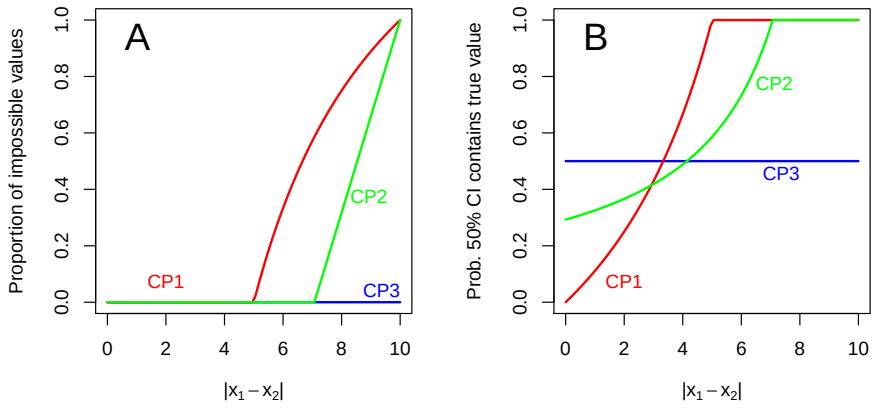

Figure 1: A: The proportion of impossible values in the confidence interval, as a function of the difference between the two observed data points. B: The probability that a 50% CI computed via one of the three confidence procedures includes the true value, as a function of the difference between the observed data points.

within the diamond outlined in black; any values outside these black lines are not possible observations.

Using the Figure 2B, we can work out the probability that CP2 contains the true value, *conditioned on* the observed value of ω. In order to do this, consider a horizontal line at any value $\omega = c$. The proportion of horizontal line that is under the shaded region (within the black diamond) yields the probability. From Figure 2B it is obvious if $\omega > .5$, the entire horizontal line within the diamond is shaded, and thus the probability that the CI contains the true value is 1. These conditional probabilities are also shown as a function of $|\omega|$ in Figure 1B.

Consider the first submarine scenario in light of Panel B. Of all confidence intervals with widths that are the same as the one we observed in the first scenario, how many contain the true value? To determine this, we need only follow the horizontal gray line that passes through point 1. By considering this horizontal slice, we are conditioning on the fact that the observed CI was only 0.5 meters wide. Only a small proportion of this horizontal slice – about 5% – passes over the shaded area. When the two bubbles are 0.5 meters apart, the 50% CI is virtually certain to exclude the true value, showing how wrong the reasoning of the Fundamental Confidence Fallacy is.

Consider now the second submarine scenario. All of this horizontal slice

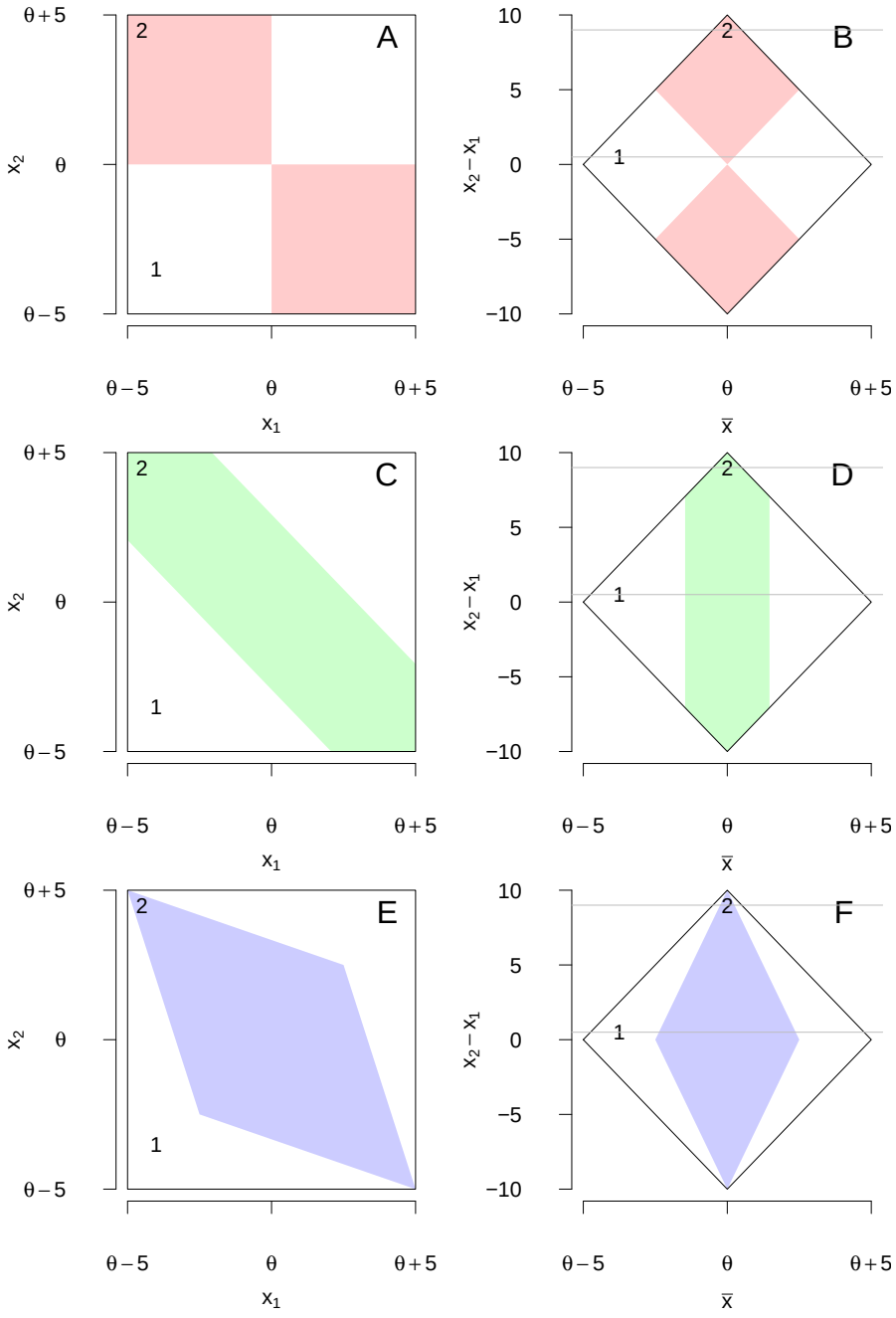

Figure 2: Left: Possible locations of the first (x_1) and second (x_2) bubbles. Right: $x_2 - x_1$ plotted against the mean of x_1 and x_2. Shaded regions show the areas where the respective 50% confidence interval contains the true value (CP1, CP2, and CP3 in the top, middle, and bottom rows, respectively). Points numbered 1 and 2 represent the pairs of bubbles from the first and second scenario, respectively.

through point 2 passes over the shaded area, because when the two bubbles are 9 meters apart, the 50% CI absolutely *must* contain the true value, even though the Fundamental Confidence Fallacy would lead us to believe that we should be only 50% confident that the true value is contained in the interval.

Figures 2C and D show the analogous plot for CP2. Examining panel D it is obvious that the probability that the CI contains the true value is a function of ω. for ω near 0, the probability that the CI contains the true value is about 30%. For larger $|\omega|$, this the probability increases to 100% (see Figure 1B). For the first submarine scenario, the probability that the CI computed from CP2 contains the true value is about 30%; for the second, it is 100%.

The fact that we can condition on $x_2 - x_1 = c$ and obtain a different probability for whether a confidence interval from CP1 or CP2 contains the true value indicates that there are relevant subsets. These relevant subsets are found by simply conditioning on $x_2 - x_1$, which is known because it was observed.

The blue diamond in Figure 2E and F shows where the true value is contained within the CP3, the Bayesian credible procedure. Notice that in Panel B, 50% of the Bayesian credible intervals contain the true value, for *every* value of $x_2 - x_1$: that is, $x_2 - x_1$ is not a recognizable subset for the Bayesian credible procedure. It is important to emphasize, however, that although CP3 is a confidence procedure, the good properties of the interval stem *not* from the fact that the Bayesian credible interval is a confidence interval – that should be clear from the fact that the other two confidence intervals fail – but rather from the fact that it is a Bayesian credible interval.

1.1.4 More about the confidence procedures

At this point, one might suspect that there was something defective about CP1 and CP2. The intervals are easily shown to be a 50% confidence interval, but perhaps there are other valid reasons why a frequentist – that is, someone only concerned with long-run average coverage – might dislike CP1 and CP2. Since CP3 is also a confidence interval, it might be interesting to ask whether a frequentist has any reason to prefer it.

Frequentists typically prefer intervals that are shorter on average, because they have a tendency to *exclude* values of the parameter that are false. Figure 3A shows the distribution of the widths of the three confidence procedures. Interestingly, although the distributions of widths of CP1 and CP3 differ between the two intervals, they have exactly the same average width: exactly 10/3 meters. CP2 has a lower average width than the other two intervals, at a fixed $10(1 - 1/\sqrt{2}) \approx 2.93$ meters.

Shortness, however, isn't by itself the main goal of a confidence procedure. Frequentists prefer intervals that have a lower probability of including all false values of θ, which we can denote θ'. If for every value of θ', one interval has a lower probability of including θ' than the other, a frequentist would prefer

that interval. Figure 3B shows the probability that each interval includes every false value θ'. Of CP1 and CP3, neither interval dominates the other; CP1 has a greater probability of including false values near the true value, but the Bayesian interval can include values farther from the true value. The average probability for CP1 and CP3 in Figure 3B is exactly the same: 2/9. From a frequentist perspective, the CP1 and CP3 are equally justifiable.

CP2, on the other hand, dominates CP3, but does not dominate CP1 (see Welch, 1939, for another example of a CI that dominates CP3 in frequentist terms). On strictly frequentist grounds, we would prefer CP2 to CP3, in spite of the fact that it does not track the precision implied by the data, can include impossible values, and admits relevant subsets. This underscores the fact that to a frequentist, all that matters is the long-run performance of the procedure, and not being able to make reasonable inferences for any given data set.

2 Student's t examples

2.1 Proof of results in t example 1

In the first example, we showed that longer t intervals have a greater probability including the true value, and thus relevant subsets exist. To prove this result requires only basic statistical knowledge, and we prove it here. Following this, we prove a result showing how much we can gain from the use of this knowledge.

There are two facts to know: first, that $\bar{x}$ and s^2 are independent; and second, that $(N-1)s^2/\sigma^2$ has a χ^2_{N-1} distribution. Because the critical t_1 values for a 50% CI are -1 and 1, The probability that the 50% confidence interval contains the true value is

$$Pr\left(\bar{y} - s/\sqrt{2} < \mu < \bar{y} + s/\sqrt{2}\right) = .5$$

This can be rearranged algebraically in a few steps to

$$Pr\left(\bar{y} - s/\sqrt{2} < \mu < \bar{y} + s/\sqrt{2}\right) = Pr\left(\left(\frac{\bar{y}-\mu}{\sigma/\sqrt{2}}\right)^2 < \frac{s^2}{\sigma^2}\right)$$

If we call

$$\begin{aligned} Z &= \left(\frac{\bar{y}-\mu}{\sigma/\sqrt{2}}\right)^2 \\ W &= \frac{s^2}{\sigma^2}, \end{aligned}$$

the condition that the confidence interval includes the true value is $Z < W$. Z is a squared standard normal variate, and hence both Z and W have identical, and independent, χ^2_1 distributions. "Long" confidence intervals occur

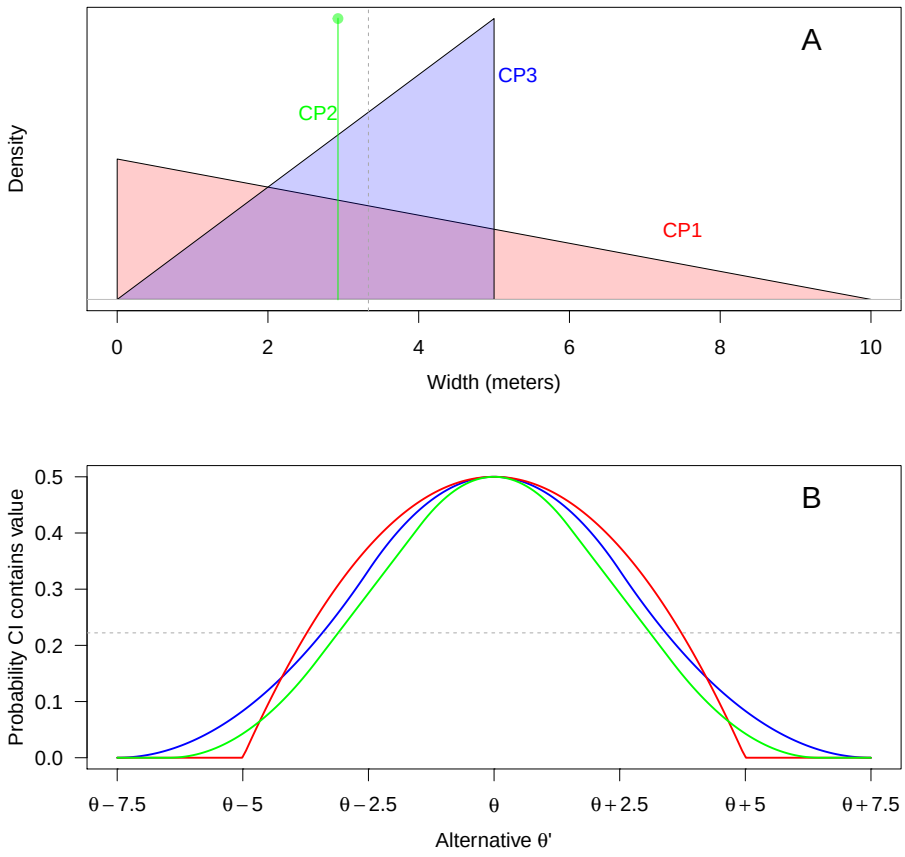

Figure 3: A: Distribution of the widths of the CI presented in the submarine example and of the Bayesian credible interval. The vertical dashed line shows the average width for both intervals. B: Probability that the CI presented in the submarine example and the Bayesian credible interval contain a given alternative false value of θ, denoted θ'. The horizontal dashed line shows the average function value for both intervals.

when W (and thus s) is larger. If we know that W is larger, then this increases the probability that Z is less than W. Because $Pr(Z < W)$ is also the probability that the true value is in the CI, knowing that s is larger increases the probability that the true value is in the CI.

In order to see how we can use this result to make better judgments, it helps to follow Buehler (1959) and consider playing a game. Suppose you are convinced that all 50% confidence intervals have a 50% probability of containing the true value. Paul offers you a bet. You and Paul will $N = 2$ observations from a normal with unknown mean and variance (suppose you ask a third friend to pick the values of μ and σ^2 without your knowledge). Paul will then decide to bet whether the true value is inside or outside the CI. You will be offered the opposite bet.

If you hold the Fundamental Confidence Fallacy, you will be indifferent to the two choices. After all, if there is a 50% probability that any given CI contains the true value, then there is the same 50% probability that it does not. You therefore accept Paul's challenge, thinking that neither of you have any advantage over the other.

As described in the main text, Paul's strategy is simple: pick a number a, look at the observed value of s, and if $s < a$ ("short" intervals), bet that the confidence interval excludes the true value. If $s > a$ ("long" intervals) then Paul bets that the interval includes the true value.

We now prove that Paul will win more than you will (possibly substantially more). Following Buehler, let C be the event $s < a$, C' be the event $s > a$, and let A be the event that the confidence interval contains μ. Also let

$$a_0 = \frac{a^2}{\sigma^2}.$$

Then $s > a$ implies that $s^2/\sigma^2 > a_0$. We seek two probabilities: $Pr(A \mid C)$ (the probability that the confidence interval contains μ given that the interval is "short") and $Pr(A \mid C')$ (the probability that the confidence interval contains μ given that the interval is "long"). Using the same results as above, these can be written:

$$\begin{aligned} Pr(A \mid C) &= Pr(Z < W \mid W < a_0) \\ Pr(A \mid C') &= Pr(Z < W \mid W > a_0). \end{aligned}$$

where Z and W are independent χ^2_1 variates. Let f and F be the PDF and CDF of a χ^2_1 random variable. Then

$$\begin{aligned} Pr(Z < W \text{ and } W < a_0) &= \int_0^{a_0} \int_0^w f(w)f(z)\,dz\,dw \\ &= \int_0^{a_0} f(w) \int_0^w f(z)\,dz\,dw \\ &= \int_0^{a_0} f(w)F(w)\,dw \end{aligned}$$

Make the substitution $q = F(w)$. Then

$$\begin{aligned}\int_0^{a_0} f(w)F(w)\,dw &= \int_0^{F(a_0)} q\,dq \\ &= \frac{1}{2}F(a_0)^2.\end{aligned}$$

This implies that

$$\begin{aligned}Pr(Z < W \mid W < a_0) &= \frac{\frac{1}{2}F(a_0)^2}{Pr(W < a_0)} \\ &= \frac{\frac{1}{2}F(a_0)^2}{F(a_0)} \\ &= \frac{1}{2}F(a_0) \\ &= Pr(A \mid C)\end{aligned}$$

In the same way, one can prove that

$$Pr(A \mid C') = \frac{1}{2} + \frac{1}{2}F(a_0).$$

Let V be the probability that Paul wins. Under this scheme, $Pr(V)$ is:

$$\begin{aligned}Pr(V) &= Pr(\sim A \mid C)P(C) + Pr(A \mid C')P(C') &(1)\\ &= \left(1 - \frac{1}{2}F(a_0)\right)F(a_0) + \frac{1}{2}\left(1 + F(a_0)\right)\left(1 - F(a_0)\right) &(2)\end{aligned}$$

A bit of algebra simplifies this to

$$Pr(V) = \frac{1}{2} + F(a_0)(1 - F(a_0)). \quad (3)$$

Since $0 < a_0 < \infty$, the probability that Paul wins $P(V) > 1/2$. By using the information in s, Paul will, in the long run, outperform you in judgments of whether the confidence interval contains the true value.

How often will Paul win? This depends on the choice of the criterion a. If a is very large or very small relative to the true population standard deviation, then Paul will win more often than not, but only just. If Paul picks his criterion well, then Paul can win up to 3/4 of the time. Figure 4 shows the probability that Paul wins as a function of the criterion.

It is obvious from Eq. 3 that the optimal winning probability is .75, and holds when $F(a_0) = 1 - F(a_0) = .5$. Since F is the CDF of the χ^2_1 distribution, this implies that the optimal a_0 occurs when $\Phi(\sqrt{a_0}) = 0.75$. Because $a_0 = (a/\sigma)^2$, this implies that the optimal value for a is

$$a = \sigma\Phi^{-1}(.75) \approx .67\sigma.$$

But even if one does not have enough knowledge to pick the optimal value for a, Figure 4 shows that for a very wide range of possible values, the probability that Paul wins is substantially above 0.5, and *always* above 0.5.

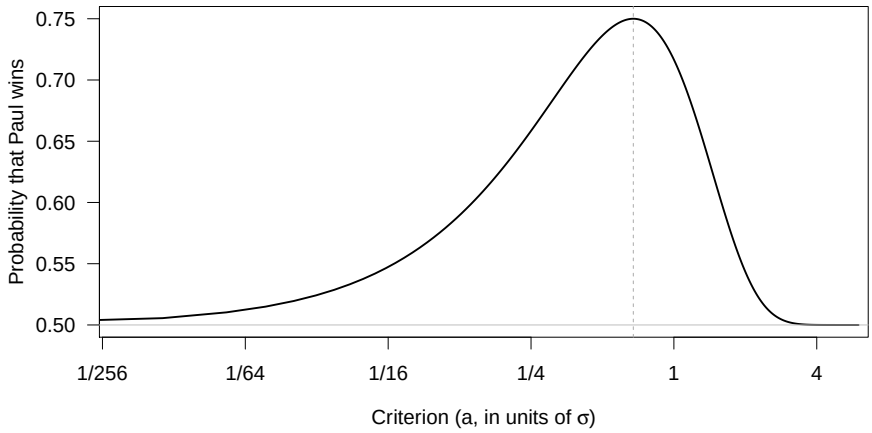

Figure 4: Probability that Paul wins when betting against someone making the Fundamental Confidence Fallacy, as a function of his criterion a. See text for details.

2.2 t example 2

In this section we provide R code to perform the simulation for confirming the claim made in the main text in t example 2, after Pierce (1973). We also try to provide some intuition about why the phenomenon occurs. We first recapitulate the example.

Suppose we sample $N = 2$ observations from a normal with unknown mean μ and unknown variance σ^2, and construct a 50% Student's t confidence interval. Suppose also that we compute the p value from a two-sided Student's t test with $H_0 : \mu = 0$. If $p > .25$ from the hypothesis test, then the probability that the confidence interval contains the true mean is $\geq 2/3$, regardless of the true values of μ and σ^2.

When we first came across this result, we were astounded and had to simulate the result for ourselves. We provide R code so that any reader can simulate the result for themselves.

```
## Set the true values
mu = 1
sigma = 1

## Find critical values for the significance test
critt = qt(.25/2,1)
```

```
## Number of simulations to perform
M = 100000
```

```
## Sample data, and put it in a matrix
y = rnorm(M*2,mu,sigma)
dim(y) = c(M,2)

## Compute sample statistics
ybar = (y[,1] + y[,2])/2
s = abs(y[,1] - y[,2])/sqrt(2)

## Compute test statistic
tval = ybar/s*sqrt(2)

## Is the p value greater than .25?
pvalGtCrit = (tval>critt & tval< -critt)

## Does the CI contain the true value?
contains = apply(y,1,function(v) min(v)<mu & max(v)>mu)

## Compute estimates of conditional probabilities
counts = table(pvalGtCrit,contains)
condProbs = counts / rowSums(counts)
```

The table below shows the probabilities, conditioned on the results of the t test, that the confidence interval contains the true value. The relevant row is the bottom one; it shows that when $\mu = 1$ and $\sigma = 1$ (as defined above), the probability that the CI contains the true value, given that $p > .25$, is approximately 0.69. The result of a (possibly uninteresting) hypothesis test therefore bears on whether we should believe that the confidence interval contains the true value.

	Prob. CI excludes μ	Prob. CI includes μ
$p < .25$	0.75	0.25
$p > .25$	0.31	0.69

Why does this occur? We will not give a formal proof; however, when the problem is demonstrated graphically the reason is clear. Consider Figure 5 (left). The contours show the joint density of $\bar{x}$ and s when $\mu = 0$ and $\sigma = 1$. Inside the dotted "V", the resulting confidence interval will include the true mean μ; outside the dotted "V", the resulting confidence interval will exclude the true mean. The relevant subsets from the previous example are obvious; for large s values, more of the distribution is between the dotted "V"

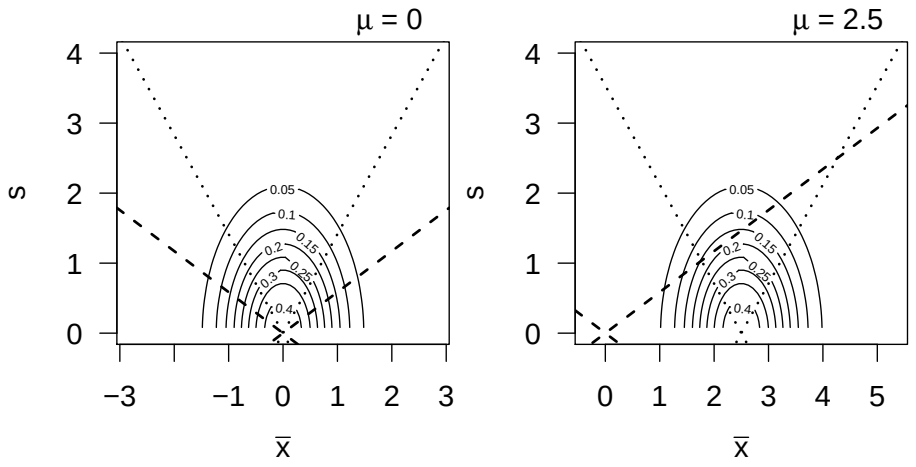

Figure 5: Bivariate contour density of the sample mean $\bar{x}$ and standard deviation s for $\mu = 0$ (left) and $\mu = 2.5$ (right). Within the dotted 'V', the CI contains the true value; within the dashed 'V', the significance test $p > 0.25$. In both panels, $\sigma^2 = 1$ and $N = 2$.

The dashed lines show pairs of $(\bar{x}, s)$ for which $p = .25$. Inside the dashed lines, $p > .25$. The probability that the confidence interval contains the true value, given that $p > .25$, is simply the proportion of area inside the dashed "V" that is inside the dotted "V". This must be greater than 50%, since the unconditional area inside the dotted "V" is 50%. The conditional probability can be proven to be $2/3$.

If we move the true mean μ away from 0, then this probability will get larger. To see this, consider Figure 5 (right), where $\mu = 2.5$. Because the null hypothesis $\mu = 0$ is far away from the true μ, the majority of the region corresponding to $p > .25$ (within the dashed "V") is within the dotted "V". Put another way, μ is very far from 0, yet we obtain $p > .25$ from our significance test against $H_0 : \mu = 0$, this must mean that s is large. If s is large, then it has a greater probability of including the true value.

References

Buehler, R. J. (1959). Some validity criteria for statistical inferences. *The Annals of Mathematical Statistics*, *30*(4), 845–863. Retrieved from http://www.jstor.org/stable/2237430

Pierce, D. A. (1973). On some difficulties in a frequency theory of inference. *The Annals of Statistics*, *1*(2), 241–250. Retrieved from http://www.jstor.org/stable/2958010

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