

# What's the difference between a confidence interval and a credible interval?

Asked 9 years, 8 months ago   Active 1 year, 2 months ago   Viewed 130k times

▲ Joris and Srikant's exchange [here](#) got me wondering (again) if my internal explanations for the difference between confidence intervals and credible intervals were the correct ones. How you would explain the difference?

248



bayesian

confidence-interval

frequentist

credible-interval

fiducial



277



edited Apr 13 '17 at 12:44



Community ♦

1

asked Sep 1 '10 at 13:53



Matt Parker

5,077

5

24

36

## 9 Answers

Active

Oldest

Votes



342



I agree completely with Srikant's explanation. To give a more heuristic spin on it:

Classical approaches generally posit that the world is one way (e.g., a parameter has one particular true value), and try to conduct experiments whose resulting conclusion -- no matter the true value of the parameter -- will be correct with at least some minimum probability.



+50



As a result, to express uncertainty in our knowledge after an experiment, the frequentist approach uses a "confidence interval" -- a range of values designed to include the true value of the parameter with some minimum probability, say 95%. A frequentist will design the experiment and 95% confidence interval procedure so that out of every 100 experiments run start to finish, at least 95 of the resulting confidence intervals will be expected to include the true value of the parameter. The other 5 might be slightly wrong, or they might be complete nonsense -- formally speaking that's ok as far as the approach is concerned, as long as 95 out of 100 inferences are correct. (Of course we would prefer them to be slightly wrong, not total nonsense.)

Bayesian approaches formulate the problem differently. Instead of saying the parameter simply has one (unknown) true value, a Bayesian method says the parameter's value is fixed but has been chosen from some probability distribution -- known as the prior probability distribution. (Another way to say that is that before taking any measurements, the Bayesian assigns a probability distribution, which they call a belief state, on what the true value of the parameter happens to be.) This "prior" might be known (imagine trying to estimate the size of a truck, if we know the overall distribution of truck sizes from the DMV) or it might be an assumption drawn out of thin air. The Bayesian inference is simpler -- we collect some data, and then calculate the probability of different values of the parameter GIVEN the data. This new probability distribution is called the "a posteriori probability" or simply the "posterior." Bayesian approaches can summarize

their uncertainty by giving a range of values on the posterior probability distribution that includes 95% of the probability -- this is called a "95% credibility interval."

A Bayesian partisan might criticize the frequentist confidence interval like this: "So what if 95 out of 100 experiments yield a confidence interval that includes the true value? I don't care about 99 experiments I DIDN'T DO; I care about this experiment I DID DO. Your rule allows 5 out of the 100 to be complete nonsense [negative values, impossible values] as long as the other 95 are correct; that's ridiculous."

A frequentist die-hard might criticize the Bayesian credibility interval like this: "So what if 95% of the posterior probability is included in this range? What if the true value is, say, 0.37? If it is, then your method, run start to finish, will be WRONG 75% of the time. Your response is, 'Oh well, that's ok because according to the prior it's very rare that the value is 0.37,' and that may be so, but I want a method that works for ANY possible value of the parameter. I don't care about 99 values of the parameter that IT DOESN'T HAVE; I care about the one true value IT DOES HAVE. Oh also, by the way, your answers are only correct if the prior is correct. If you just pull it out of thin air because it feels right, you can be way off."

In a sense both of these partisans are correct in their criticisms of each others' methods, but I would urge you to think mathematically about the distinction -- as Srikant explains.

Here's an extended example from that talk that shows the difference precisely in a discrete example.

When I was a child my mother used to occasionally surprise me by ordering a jar of chocolate-chip cookies to be delivered by mail. The delivery company stocked four different kinds of cookie jars -- type A, type B, type C, and type D, and they were all on the same truck and you were never sure what type you would get. Each jar had exactly 100 cookies, but the feature that distinguished the different cookie jars was their respective distributions of chocolate chips per cookie. If you reached into a jar and took out a single cookie uniformly at random, these are the probability distributions you would get on the number of chips:

Cookie jars **A, B, C, D** have the following distribution of cookies with chocolate chips:

$P(\text{ chips }   \text{ jar } )$	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>
<b>0</b>	1	12	13	27
<b>1</b>	1	19	20	70
<b>2</b>	70	24	0	1
<b>3</b>	28	20	0	1
<b>4</b>	0	25	67	1
total	100%	100%	100%	100%

A type-A cookie jar, for example, has 70 cookies with two chips each, and no cookies with four chips or more! A type-D cookie jar has 70 cookies with one chip each. Notice how each vertical column is a probability mass function -- the conditional probability of the number of chips you'd get, given that the jar = A, or B, or C, or D, and each column sums to 100.

I used to love to play a game as soon as the deliveryman dropped off my new cookie jar. I'd pull one single cookie at random from the jar, count the chips on the cookie, and try to express my

uncertainty -- at the 70% level -- of which jars it could be. Thus it's the identity of the jar (A, B, C or D) that is the **value of the parameter** being estimated. The number of chips (0, 1, 2, 3 or 4) is the **outcome** or the observation or the sample.

Originally I played this game using a frequentist, 70% confidence interval. Such an interval needs to make sure that **no matter** the true value of the parameter, meaning no matter which cookie jar I got, the interval would cover that true value with at least 70% probability.

An interval, of course, is a function that relates an outcome (a row) to a set of values of the parameter (a set of columns). But to *construct* the confidence interval and guarantee 70% coverage, we need to work "vertically" -- looking at each column in turn, and making sure that 70% of the probability mass function is covered so that 70% of the time, that column's identity will be part of the interval that results. Remember that it's the vertical columns that form a p.m.f.

So after doing that procedure, I ended up with these intervals:

$P(\text{chips} \mid \text{jar})$	A	B	C	D
0	1	12	13	27
1	1	[19	20	70]
2	[70	24]	0	1
3	28	[20]	0	1
4	0	[25	67]	1
coverage	70%	88%	87%	70%

For example, if the number of chips on the cookie I draw is 1, my confidence interval will be {B,C,D}. If the number is 4, my confidence interval will be {B,C}. Notice that since each column sums to 70% or greater, then no matter which column we are truly in (no matter which jar the deliveryman dropped off), the interval resulting from this procedure will include the correct jar with at least 70% probability.

Notice also that the procedure I followed in constructing the intervals had some discretion. In the column for type-B, I could have just as easily made sure that the intervals that included B would be 0,1,2,3 instead of 1,2,3,4. That would have resulted in 75% coverage for type-B jars (12+19+24+20), still meeting the lower bound of 70%.

My sister Bayesia thought this approach was crazy, though. "You have to consider the deliverman as part of the system," she said. "Let's treat the identity of the jar as a random variable itself, and let's *assume* that the deliverman chooses among them uniformly -- meaning he has all four on his truck, and when he gets to our house he picks one at random, each with uniform probability."

"With that assumption, now let's look at the joint probabilities of the whole event -- the jar type **and** the number of chips you draw from your first cookie," she said, drawing the following table:

$P(\text{jar} \cap \text{chips})$	A	B	C	D	total
0	1/4	12/4	13/4	27/4	13.25%
1	1/4	19/4	20/4	70/4	27.50%
2	70/4	24/4	0/4	1/4	23.75%
3	28/4	20/4	0/4	1/4	12.25%
4	0/4	25/4	67/4	1/4	23.25%
total	25%	25%	25%	25%	

Notice that the whole table is now a probability mass function -- meaning the whole table sums to 100%.

"Ok," I said, "where are you headed with this?"

"You've been looking at the conditional probability of the number of chips, given the jar," said Bayesia. "That's all wrong! What you really care about is the conditional probability of which jar it is, given the number of chips on the cookie! Your 70% interval should simply include the list jars that, in total, have 70% probability of being the true jar. Isn't that a lot simpler and more intuitive?"

"Sure, but how do we calculate that?" I asked.

"Let's say we **know** that you got 3 chips. Then we can ignore all the other rows in the table, and simply treat that row as a probability mass function. We'll need to scale up the probabilities proportionately so each row sums to 100, though." She did:

$P(\text{jar}   \text{chips})$	A	B	C	D	total
0	1.9	22.6	24.5	50.9	100%
1	0.9	17.3	18.2	63.6	100%
2	73.7	25.3	0.0	1.1	100%
3	57.1	40.8	0.0	2.0	100%
4	0.0	26.9	72.0	1.1	100%

"Notice how each row is now a p.m.f., and sums to 100%. We've flipped the conditional probability from what you started with -- now it's the probability of the man having dropped off a certain jar, given the number of chips on the first cookie."

"Interesting," I said. "So now we just circle enough jars in each row to get up to 70% probability?" We did just that, making these credibility intervals:

$P(\text{jar}   \text{chips})$	A	B	C	D	credibility
0	1.9	22.6	<b>[24.5 50.9]</b>		75%
1	0.9	17.3	<b>[18.2 63.6]</b>		82%
2	<b>[73.7]</b>	25.3	0.0	1.1	74%
3	<b>[57.1 40.8]</b>		0.0	2.0	98%
4	0.0	26.9	<b>[72.0]</b>	1.1	72%

Each interval includes a set of jars that, *a posteriori*, sum to 70% probability of being the true jar.

"Well, hang on," I said. "I'm not convinced. Let's put the two kinds of intervals side-by-side and compare them for coverage and, assuming that the deliveryman picks each kind of jar with equal probability, credibility."

Here they are:

#### Confidence intervals:

confidence	A	B	C	D	credibility
0	1	12	13	27	0%
1	1	[19	20	70]	99%
2	[70	24]	0	1	99%
3	28	[20]	0	1	41%
4	0	[25	67]	1	99%
coverage	70%	88%	87%	70%	

#### Credibility intervals:

credibility	A	B	C	D	credibility
0	1	12	[13	27]	75%
1	1	19	[20	70]	82%
2	[70]	24	0	1	74%
3	[28	20]	0	1	98%
4	0	25	[67]	1	72%
coverage	98%	20%	100%	97%	

"See how crazy your confidence intervals are?" said Bayesia. "You don't even have a sensible answer when you draw a cookie with zero chips! You just say it's the empty interval. But that's obviously wrong -- it has to be one of the four types of jars. How can you live with yourself, stating an interval at the end of the day when you **know the interval is wrong**? And ditto when you pull a cookie with 3 chips -- your interval is only correct 41% of the time. Calling this a '70%' confidence interval is bullshit."

"Well, hey," I replied. "It's correct 70% of the time, no matter which jar the deliveryman dropped off. That's a lot more than you can say about your credibility intervals. What if the jar is type B? Then your interval will be wrong 80% of the time, and only correct 20% of the time!"

"This seems like a big problem," I continued, "because your mistakes will be correlated with the type of jar. If you send out 100 'Bayesian' robots to assess what type of jar you have, each robot sampling one cookie, you're telling me that on type-B days, you will expect 80 of the robots to get the wrong answer, each having >73% belief in its incorrect conclusion! That's troublesome, especially if you want most of the robots to agree on the right answer."

"PLUS we had to make this assumption that the deliveryman behaves uniformly and selects each type of jar at random," I said. "Where did that come from? What if it's wrong? You haven't talked to him; you haven't interviewed him. Yet all your statements of *a posteriori* probability rest on this statement about his behavior. I didn't have to make any such assumptions, and my interval meets its criterion even in the worst case."

"It's true that my credibility interval does perform poorly on type-B jars," Bayesia said. "But so what? Type B jars happen only 25% of the time. It's balanced out by my good coverage of type A, C, and D jars. And I never publish nonsense."

"It's true that my confidence interval does perform poorly when I've drawn a cookie with zero chips," I said. "But so what? Chipless cookies happen, at most, 27% of the time in the worst case (a type-D jar). I can afford to give nonsense for this outcome because NO jar will result in a wrong answer more than 30% of the time."

"The column sums matter," I said.

"The row sums matter," Bayesia said.

"I can see we're at an impasse," I said. "We're both correct in the mathematical statements we're making, but we disagree about the appropriate way to quantify uncertainty."

"That's true," said my sister. "Want a cookie?"

edited Nov 18 '12 at 21:29

answered Sep 1 '10 at 18:46



Keith Winstein

4,230 1 13 8

- 
- 19 Good answer - just one minor point, you say "...Instead of saying the parameter has one true value, a Bayesian method says the value is chosen from some probability distribution...." This is not true. A Bayesian fits the probability distribution to express the uncertainty about the true, unknown, fixed value. This says which values are plausible, given what was known before observing the data. The actual probability statement is  $Pr[\theta_0 \in (\theta, \theta + d\theta) | I]$ , where  $\theta_0$  is the true value, and  $\theta$  the hypothesised one, based on information  $I$ . – [probabilityislogic](#) Feb 5 '11 at 11:34
- 
- 1 ...cont'd... but it is much more convenient to just write  $p(\theta)$ , with the understanding of what it means "in the background". Clearly this can cause much confusion. – [probabilityislogic](#) Feb 5 '11 at 11:38
- 
- 17 sorry to revive this super old post but a quick question, in your post in the section where the frequentist criticizes the Bayesian approach you say: "What if the true value is, say, 0.37? If it is, then your method, run start to finish, will be WRONG 75% of the time." How did you get those numbers? how does 0.37 correspond to 75% wrong? Is this off of some type of probability curve? Thanks – [BYS2](#) Jul 6 '12 at 11:18
- 
- 5 @BYS2, when the author says that "What if the true value is, say, 0.37? If it is, then your method, run start to finish, will be WRONG 75% of the time", they are just giving example numbers they made up. In this particular case, they would be referring to some prior distribution that had a very low value at 0.37, with most of its probability density elsewhere. And we assume that our example distribution would perform very poorly when the true value of the parameter happens to be 0.37, similarly to how Bayesia's credibility intervals failed miserably when the jar happened to be type-B. – [Garrett](#) Sep 23 '14 at 7:03
- 
- 2 The author says "you will expect 80 of the robots to get the wrong answer, each having >73% belief in its incorrect conclusion!", but this should have been >72% belief, since 72% is the minimum credibility in the credibility intervals table. – [Garrett](#) Sep 23 '14 at 7:07
- 

My understanding is as follows:



## Background





Suppose that you have some data  $x$  and you are trying to estimate  $\theta$ . You have a data generating process that describes how  $x$  is generated conditional on  $\theta$ . In other words you know the distribution of  $x$  (say,  $f(x|\theta)$ ).

## Inference Problem

Your inference problem is: What values of  $\theta$  are reasonable given the observed data  $x$ ?

## Confidence Intervals

Confidence intervals are a classical answer to the above problem. In this approach, you assume that there is *true, fixed* value of  $\theta$ . Given this assumption, you use the data  $x$  to get to an estimate of  $\theta$  (say,  $\hat{\theta}$ ). Once you have your estimate you want to assess where the true value is in relation to your estimate.

Notice that under this approach the true value is *not* a random variable. It is a fixed but unknown quantity. In contrast, your estimate *is* a random variable as it depends on your data  $x$  which was generated from your data generating process. Thus, you realize that you get different estimates each time you repeat your study.

The above understanding leads to the following methodology to assess where the true parameter is in relation to your estimate. Define an interval,  $I \equiv [lb(x), ub(x)]$  with the following property:

$$P(\theta \in I) = 0.95$$

An interval constructed like the above is what is called a confidence interval. Since, the true value is unknown but fixed, the true value is either in the interval or outside the interval. The confidence interval then is a statement about the likelihood that the interval we obtain actually has the true parameter value. Thus, the probability statement is about the interval (i.e., the chances that interval which has the true value or not) rather than about the location of the true parameter value.

In this paradigm, it is meaningless to speak about the probability that a true value is less than or greater than some value as the true value is *not* a random variable.

## Credible Intervals

In contrast to the classical approach, in the bayesian approach we assume that the true value is a random variable. Thus, we capture the our uncertainty about the true parameter value by a imposing a prior distribution on the true parameter vector (say  $f(\theta)$ ).

Using bayes theorem, we construct the posterior distribution for the parameter vector by blending the prior and the data we have (briefly the posterior is  $f(\theta|-) \propto f(\theta)f(x|\theta)$ ).

We then arrive at a point estimate using the posterior distribution (e.g., use the mean of the posterior distribution). However, since under this paradigm, the true parameter vector is a random variable, we also want to know the extent of uncertainty we have in our point estimate. Thus, we construct an interval such that the following holds:

$$P(l(\theta) \leq \theta \leq ub(\theta)) = 0.95$$

The above is a credible interval.



## Summary

Credible intervals capture our current uncertainty in the location of the parameter values and thus can be interpreted as probabilistic statement about the parameter.

In contrast, confidence intervals capture the uncertainty about the interval we have obtained (i.e., whether it contains the true value or not). Thus, they cannot be interpreted as a probabilistic statement about the true parameter values.

answered Sep 1 '10 at 16:01

user28

- 
- 2 A 95% confidence interval by definition covers the true parameter value in 95% of the cases, as you indicated correctly. Thus, the chance that your interval covers the true parameter value is 95%. You can sometimes say something about the chance that the parameter is larger or smaller than any of the boundaries, based on the assumptions you make when constructing the interval (pretty often the normal distribution of your estimate). You can calculate  $P(\theta > ub)$ , or  $P(ub < \theta)$ . The statement is about the boundary, indeed, but you can make it. – Joris Meys Sep 2 '10 at 11:44
- 
- 9 Joris, I can't agree. Yes, for any value of the parameter, there will be >95% probability that the resulting interval will cover the true value. That doesn't mean that after taking a particular observation and calculating the interval, there still is 95% conditional probability given the data that THAT interval covers the true value. As I said below, formally it would be perfectly acceptable for a confidence interval to spit out [0, 1] 95% of the time and the empty set the other 5%. The occasions you got the empty set as the interval, there ain't 95% probability the true value is within! – Keith Winstein Sep 2 '10 at 14:21 
- 
- @ Keith : I see your point, although an empty set is not an interval by definition. The probability of a confidence interval is also not conditional on the data, in contrary. Every confidence interval comes from a different random sample, so the chance that your sample is drawn such that the 95%CI on which it is based does not cover the true parameter value, is only 5%, regardless of the data. – Joris Meys Sep 2 '10 at 15:02 
- 
- 1 Joris, I was using "data" as a synonym for "sample," so I think we agree. My point is that it's possible to be in situations, after you take the sample, where you can prove with absolute certainty that your interval is wrong -- that it does not cover the true value. This does not mean that it is not a valid 95% confidence interval. So you can't say that the confidence parameter (the 95%) tells you anything about the probability of coverage of a particular interval after you've done the experiment and got the interval. Only an a posteriori probability, informed by a prior, can speak to that. – Keith Winstein Sep 2 '10 at 17:46
- 
- 4 @svadalli - the Bayesian approach does not take the view that  $\theta$  is random. It is not  $\theta$  that is distributed ( $\theta$  is fixed but unknown), it is the *uncertainty about  $\theta$  which is distributed, conditional on a state of knowledge about  $\theta$* . The actual probability statement that  $f(\theta)$  is capturing is  $Pr(\theta \text{ is in the interval } (\theta, \theta + d\theta) | I) = f(\theta)d\theta$ . In fact, the exact same argument applies to  $X$ , it too can be considered fixed, but unknown. – probabilityislogic Jan 27 '11 at 16:14
- 

I disagree with Srikant's answer on one fundamental point. Srikant stated this:

13

"Inference Problem: Your inference problem is: What values of  $\theta$  are reasonable given the observed data  $x$ ?"

In fact this is the BAYESIAN INFERENCE PROBLEM. In Bayesian statistics we seek to calculate  $P(\theta | x)$  i.e the probability of the parameter value given the observed data (sample). The CREDIBLE INTERVAL is an interval of  $\theta$  that has a 95% chance (or other) of containing the true value of  $\theta$  given the several assumptions underlying the problem.

The FREQUENTIST INFERENCE PROBLEM is this:



Are the observed data  $x$  reasonable given the hypothesised values of  $\theta$ ?

In frequentist statistics we seek to calculate  $P(x|\theta)$  i.e the probability of observing the data (sample) given the hypothesised parameter value(s). The CONFIDENCE INTERVAL (perhaps a misnomer) is interpreted as: if the experiment that generated the random sample  $x$  were repeated many times, 95% (or other) of such intervals constructed from those random samples would contain the true value of the parameter.

Mess with your head? That's the problem with frequentist statistics and the main thing Bayesian statistics has going for it.

As Sikrant points out,  $P(\theta|x)$  and  $P(x|\theta)$  are related as follows:

$$P(\theta|x) = P(\theta)P(x|\theta)$$

Where  $P(\theta)$  is our prior probability;  $P(x|\theta)$  is the probability of the data conditional on that prior and  $P(\theta|x)$  is the posterior probability. The prior  $P(\theta)$  is inherently subjective, but that is the price of knowledge about the Universe - in a very profound sense.

The other parts of both Sikrant's and Keith's answers are excellent.

edited Sep 4 '10 at 10:28

answered Sep 4 '10 at 10:22



Thylacoleo

4,321 5 20 30

Technically, you are correct but do note that the confidence interval gives the set of parameter values for which the null hypothesis is true. Thus, "are the observed data  $x$  reasonable given our hypothesis about  $\theta$ ?" can be re-phrased as "What true values of  $\theta$  would be a compatible hypothesis given the observed data  $x$ ?" Note that the re-phrased question does not necessarily imply that  $\theta$  is being assumed to be a random variable. The re-phrased question exploits the fact that we perform null hypothesis tests by inspecting if the hypothesized value falls in the confidence interval. – user28 Sep 4 '10 at 15:32

@svadali - confidence intervals evaluate *data* for a fixed hypothesis. Thus when changing the "fixed" part of the equation, if you fail to take account of the probability of the hypothesis prior to observing your data, then you are bound to come up with inconsistencies and incoherent results. Conditional probability is not "constrained" when changing the conditions (e.g. by changing the conditions you can change a conditional probability from 0 to 1). The prior probability takes account of this arbitrariness. Conditioning on  $X$  is done because we are certain  $X$  has occurred-we did observe  $X$ ! – probabilityislogic Jan 27 '11 at 17:11

The answers provided before are very helpful and detailed. Here is my \$0.25.

13

Confidence interval (CI) is a concept based on the classical definition of probability (also called the "Frequentist definition") that probability is like proportion and is based on the axiomatic system of Kolmogorov (and others).



Credible intervals (Highest Posterior Density, HPD) can be considered to have its roots in decision theory, based on the works of Wald and de Finetti (and extended a lot by others).

As people in this thread have done a great job in giving examples and the difference of hypotheses in the Bayesian and frequentist case, I will just stress on a few important points.

1. CIs are based on the fact that inference MUST be made on all possible repetitions of an experiment that can be seen and NOT only on the observed data where as HPDs are based ENTIRELY on the observed data (and obv. our prior assumptions).
2. In general CIs are NOT coherent (will be explained later) where as HPDs are coherent (due to their roots in decision theory). Coherence (as I would explain to my grand mom) means: given a betting problem on a parameter value, if a classical statistician (frequentist) bets on CI and a bayesian bets on HPDs, the frequentist IS BOUND to lose (excluding the trivial case when HPD = CI). In short, if you want to summarize the findings of your experiment as a probability based on the data, the probability HAS to be a posterior probability (based on a prior). There is a theorem (cf Heath and Sudderth, Annals of Statistics, 1978) which (roughly) states: Assignment of probability to  $\theta$  based on data will not make a sure loser if and only if it is obtained in a bayesian way.
3. As CIs don't condition on the observed data (also called "Conditionality Principle" CP), there can be paradoxical examples. Fisher was a big supporter of CP and also found a lot of paradoxical examples when this was NOT followed (as in the case of CI). This is the reason why he used p-values for inference, as opposed to CI. In his view p-values were based on the observed data (much can be said about p-values, but that is not the focus here). Two of the very famous paradoxical examples are: (4 and 5)
4. Cox's example (Annals of Math. Stat., 1958):  $X_i \sim \mathcal{N}(\mu, \sigma^2)$  (iid) for  $i \in \{1, \dots, n\}$  and we want to estimate  $\mu$ .  $n$  is NOT fixed and is chosen by tossing a coin. If coin toss results in H, 2 is chosen, otherwise 1000 is chosen. The "common sense" estimate - sample mean is an unbiased estimate with a variance of  $0.5\sigma^2 + 0.0005\sigma^2$ . What do we use as the variance of sample mean when  $n = 1000$ ? Isn't it better (or sensible) to use the variance of sample mean estimator as  $0.001\sigma^2$  (conditional variance) instead of the actual variance of the estimator, which is HUGE!! ( $0.5\sigma^2 + 0.0005\sigma^2$ ). This is a simple illustration of CP when we use the variance as  $0.001\sigma^2$  when  $n = 1000$ .  $n$  stand alone has no importance or no information for  $\mu$  and  $\sigma$  (ie  $n$  is ancillary for them) but GIVEN its value, you know a lot about the "quality of data". This directly relates to CI as they involve the variance which should not be conditioned on  $n$ , ie we will end up using the larger variance, hence over conservative.
5. Welch's example: This example works for any  $n$ , but we will take  $n = 2$  for simplicity.  $X_1, X_2 \sim \mathcal{U}(\theta - 1/2, \theta + 1/2)$  (iid),  $\theta$  belongs to the Real line. This implies  $X_1 - \theta \sim \mathcal{U}(-1/2, 1/2)$  (iid).  $\frac{1}{2}(X_1 + X_2) - \theta$  (note that this is NOT a statistic) has a distribution independent of  $\theta$ . We can choose  $c > 0$  s.t.  $\text{Prob}_\theta(-c \leq \bar{x} - \theta \leq c) = 1 - \alpha (\approx 99\%)$ , implying  $(\bar{x} - c, \bar{x} + c)$  is the 99% CI of  $\theta$ . The interpretation of this CI is: if we sample repeatedly, we will get different  $\bar{x}$  and 99% (at least) times it will contain true  $\theta$ , BUT (the elephant in the room) for a GIVEN data, we DON'T know the probability that CI will contain true  $\theta$ . Now, consider the following data:  $X_1 = 0$  and  $X_2 = 1$ , as  $|X_1 - X_2| = 1$ , we know FOR SURE that the interval  $(X_1, X_2)$  contains  $\theta$  (one possible criticism,  $\text{Prob}(|X_1 - X_2| = 1) = 0$ , but we can handle it mathematically and I won't discuss it). This example also illustrates the concept of coherence beautifully. If you are a classical statistician, you will definitely bet on the 99% CI without looking at the value of  $|X_1 - X_2|$  (assuming you are true to your profession). However, a bayesian will bet on the CI only if the value of  $|X_1 - X_2|$  is close to 1. If we condition on  $|X_1 - X_2|$ , the interval is coherent and the player won't be a sure loser any longer (similar to the theorem by Heath and Sudderth).
6. Fisher had a recommendation for such problems - use CP. For the Welch's example, Fisher suggested to condition of  $X_2 - X_1$ . As we see,  $X_2 - X_1$  is ancillary for  $\theta$ , but it provides information about theta. If  $X_2 - X_1$  is SMALL, there is not a lot of information about  $\theta$  in the

data. If  $X_2 - X_1$  is LARGE, there is a lot of information about  $\theta$  in the data. Fisher extended the strategy of conditioning on the ancillary statistic to a general theory called *Fiducial Inference* (also called his greatest failure, cf Zabell, Stat. Sci. 1992), but it didn't become popular due to lack of generality and flexibility. Fisher was trying to find a way different from both the classical statistics (of Neyman School) and the bayesian school (hence the famous adage from Savage: "Fisher wanted to make a Bayesian omelette (ie using CP) without breaking the Bayesian eggs"). Folklore (no proof) says: Fisher in his debates attacked Neyman (for Type I and Type II error and CI) by calling him a *Quality Control guy* rather than a *Scientist*, as Neyman's methods didn't condition on the observed data, instead looked at all possible repetitions.

7. Statisticians also want to use Sufficiency Principle (SP) in addition to the CP. But SP and CP together imply the Likelihood Principle (LP) (cf Birnbaum, JASA, 1962) ie given CP and SP, one must ignore the sample space and look at the likelihood function only. Thus, we only need to look at the given data and *NOT* at the whole sample space (looking at whole sample space is in a way similar to repeated sampling). This has led to concept like Observed Fisher Information (cf. Efron and Hinkley, AS, 1978) which measure the information about the data from a frequentist perspective. The amount of information in the data is a bayesian concept (and hence related to HPD), instead of CI.
8. Kiefer did some foundational work on CI in the late 1970s, but his extensions haven't become popular. A good source of reference is Berger ("Could Fisher, Neyman and Jeffreys agree about testing of hypotheses", Stat Sci, 2003).

## Summary:

(As pointed out by Srikant and others)

CIs can't be interpreted as probability and they don't tell anything about the unknown parameter GIVEN the observed data. CIs are statements about repeated experiments.

HPDs are probabilistic intervals based on the posterior distribution of the unknown parameter and have a probability based interpretation based on the given data.

Frequentist property (repeated sampling) property is a desirable property and HPDs (with appropriate priors) and CI both have them. HPDs condition on the given data also in answering the questions about the unknown parameter

(Objective NOT Subjective) Bayesians agree with the classical statisticians that there is a single TRUE value of the parameter. However, they both differ in the way they make inference about this true parameter.

Bayesian HPDs give us a good way of conditioning on data, but if they fail to agree with the frequentist properties of CI they are not very useful (analogy: a person who uses HPDs (with some prior) without a good frequentist property, is bound to be doomed like a carpenter who only cares about the hammer and forgets the screw driver)

At last, I have seen people in this thread (comments by Dr. Joris: "...assumptions involved imply a diffuse prior, i.e. a complete lack of knowledge about the true parameter.") talking about lack of

knowledge about the true parameter being equivalent to using a diffuse prior. I DON'T know if I can agree with the statement (Dr. Keith agrees with me). For example, in the basic linear models case, some distributions can be obtained by using a uniform prior (which some people called diffuse), BUT it DOESN'T mean that uniform distribution can be regarded as a LOW INFORMATION PRIOR. In general, NON-INFORMATIVE(Objective) prior doesn't mean it has low information about the parameter.

*Note:* A lot of these points are based on the lectures by one of the prominent bayesians. I am still a student and could have misunderstood him in some way. Please accept my apologies in advance.

edited Nov 18 '18 at 23:03



Giorfindel

580 1 7 13

answered Sep 16 '10 at 9:35



suncoolsu

6,002 27 46

"the frequentist IS BOUND to lose " Looking at the most-voted answer, I'd assume this depends on the utility function (e.g. not if regret optimization is going on). Intuitively, it might also depend on the ability to determine the prior function... – Abel Molina Aug 27 '14 at 19:08

4 "the frequentist IS BOUND to lose"...\*conditional on having the appropriate prior\* (which, in general, is not so easy). Perfect example: gambling addicts are 99% certain their luck will change this time. Those who incorporate this prior into their decision analysis tend not to do so well in the long run. – Cliff AB Sep 3 '16 at 0:55

1 I don't think you should abbreviate confidence intervals as *C*'s in an answer about the distinction between credible intervals and confidence intervals. – Hugh Apr 21 '17 at 15:06

10

Always fun to engage in a bit of philosophy. I quite like Keith's response, however I would say that he is taking the position of "Mr forgetful Bayesia". The bad coverage when type B and type C can only come about if (s)he applies the same probability distribution at every trial, and refuses to update his(her) prior.



You can see this quite clearly, for the type A and type D jars make "definite predictions" so to speak (for 0-1 and 2-3 chips respectively), whereas type B and C jars basically give a uniform distribution of chips. So, on repetitions of the experiment with some fixed "true jar" (or if we sampled another biscuit), a uniform distribution of chips will provide evidence for type B or C jars.

And from the "practical" viewpoint, type B and C would require an enormous sample to be able to distinguish between them. The KL divergences between the two distributions are  $KL(B||C) \approx 0.006 \approx KL(C||B)$ . This is a divergence equivalent to two normal distributions both with variance 1 and a difference in means of  $\sqrt{2 \times 0.006} = 0.11$ . So we can't possibly be expected to be able to discriminate on the basis of one sample (for the normal case, we would require about 320 sample size to detect this difference at 5% significance level). So we can justifiably collapse type B and type C together, until such time as we have a big enough sample.

Now what happens to those credible intervals? We actually now have 100% coverage of "B or C"! What about the frequentist intervals? The coverage is unchanged as all intervals contained both B and C or neither, so it is still subject to the criticisms in Keith's response - 59% and 0% for 3 and 0 chips observed.

But let's be pragmatic here. If you optimise something with respect to one function, it can't be expected to work well for a different function. However, both the frequentist and bayesian intervals do achieve the desired credibility/confidence level on the average. We have  $(0 + 99 + 99 + 59 + 99)/5 = 71.2$  - so the frequentist has appropriate average credibility. We also have  $(98 + 60 + 66 + 97)/4 = 80.3$  - the bayesian has appropriate average coverage.

Another point I would like to stress is that the Bayesian is not saying that "the parameter is random" by assigning a probability distribution. For the Bayesian (well, at least for me anyways) a probability distribution is a description of what is known about that parameter. The notion of "randomness" does not really exist in Bayesian theory, only the notions of "knowing" and "not knowing". The "knowns" go into the conditions, and the "unknowns" are what we calculate the probabilities for, if of interest, and marginalise over if a nuisance. So a credible interval describes what is known about a fixed parameter, averaging over what is not known about it. So if we were to take the position of the person who packed the cookie jar and knew that it was type A, their credibility interval would just be [A], regardless of the sample, and no matter how many samples were taken. And they would be 100% accurate!

A confidence interval is based on the "randomness" or variation which exists in the different possible samples. As such the only variation that they take into account is that in a sample. So the confidence interval is unchanged for the person who packed the cookie jar and knew that it was type A. So if you drew the biscuit with 1 chip out of the type A jar, the frequentist would assert with 70% confidence that the type was not A, even though they know the jar is type A! (if they maintained their ideology and ignored their common sense). To see that this is the case, note that nothing in this situation has changed the sampling distribution - we have simply taken the perspective of a different person with "non-data" based information about a parameter.

Confidence intervals will change only when the data changes or the model/sampling distribution changes. credibility intervals can change if other relevant information is taken into account.

Note that this crazy behavior is certainly not what a proponent of confidence intervals would actually do; but it does demonstrate a weakness in the philosophy underlying the method in a particular case. Confidence intervals work their best when you don't know much about a parameter beyond the information contained in a data set. And further, credibility intervals won't be able to improve much on confidence intervals unless there is prior information which the confidence interval can't take into account, or finding the sufficient and ancillary statistics is hard.

answered Jun 14 '11 at 16:37



**probabilityislogic**

**21.1k** 3 68 90

I can't say I understood Keith's explanation of the jar example, a quick question: I repeat the experiment  $m$  times, collected  $m$  different samples, so now I've computed  $m$  different CIs (each with 95% confidence level), now what is CI? Does it mean 95% of  $m$  CIs should cover the true value? – **avocado** Jan 2 '14 at 13:11

@loganecolss - this is indeed true, but only in the limit as  $m \rightarrow \infty$ . This accords with standard "probability" = "long run frequency" interpretation underlying CIs. – **probabilityislogic** Jan 2 '14 at 13:30

Yes, in the limit. Then for one or just a couple of samples, the CIs doesn't mean anything, right? Then what's the point of calculating the CI, if I don't have tons of samples? – **avocado** Jan 2 '14 at 13:42

3 @loganecolss - that's why I'm a Bayesian. – **probabilityislogic** Jan 2 '14 at 13:53



- 2 @nazka - sort of. I would say it is always best to use a Bayesian approach regardless of how much data you have. If this can be well approximated by a frequentist procedure, then use that. Bayesian is not a synonym for slow. — [probabilityislogic](#) Dec 3 '14 at 9:00



7



As I understand it: A credible interval is a statement of the range of values for the statistic of interest that remain plausible given the particular sample of data that we have actually observed. A confidence interval is a statement of the frequency with which the true value lies in the confidence interval when the experiment is repeated a large number of times, each time with a different sample of data from the same underlying population.



Normally the question we want to answer is "what values of the statistic are consistent with the observed data", and the credible interval gives a direct answer to that question - the true value of the statistic lies in a 95% credible interval with probability 95%. The confidence interval does not give a direct answer to this question; it is not correct to assert that the probability that the true value of the statistic lies within the 95% confidence interval is 95% (unless it happens to coincide with the credible interval). However this is a very common misinterpretation of a frequentist confidence interval as it the interpretation that would be a direct answer to the question.

The paper by Jayne's that I discuss in another question gives a good example of this (example #5), where a perfectly correct confidence interval is constructed, where the particular sample of data on which it is based rules out any possibility of the true value of the statistic being in the 95% confidence interval! This is only a problem if the confidence interval is incorrectly interpreted as a statement of plausible values of the statistic on the basis of the particular sample we have observed.

At the end of the day, it is a matter of "horses for courses", and which interval is best depends on the question you want answered - just choose the method that directly answers that question.

I suspect confidence intervals are more useful when analysing [designed] repeatable experiments (as that is just the assumption underlying the confidence interval), and credible intervals better when analysing observational data, but that is just an opinion (I use both sorts of intervals in my own work, but wouldn't describe myself as an expert in either).

answered Sep 4 '10 at 11:07



**Dikran Marsupial**

**39.5k** 5 111 155

- 6 The issue with confidence intervals in repeated experiments, is that in order for them to work, the conditions of the repeatable experiment need to stay the same (and who would believe that?), whereas the Bayesian interval (if used properly) conditions on the data observed, and thus provides allowances for changes which occur in the real world (via data). I think it is the *conditioning rules* of Bayesian statistics which make it so hard to outperform (I think it is impossible: only equivalence can be achieved), and the automatic machinery which it achieves this that make it seem so slick. — [probabilityislogic](#) Jan 27 '11 at 16:40



3



Generic and consistent confidence and credible regions.  
<http://dx.doi.org/10.6084/m9.figshare.1528163> with code at  
<http://dx.doi.org/10.6084/m9.figshare.1528187>



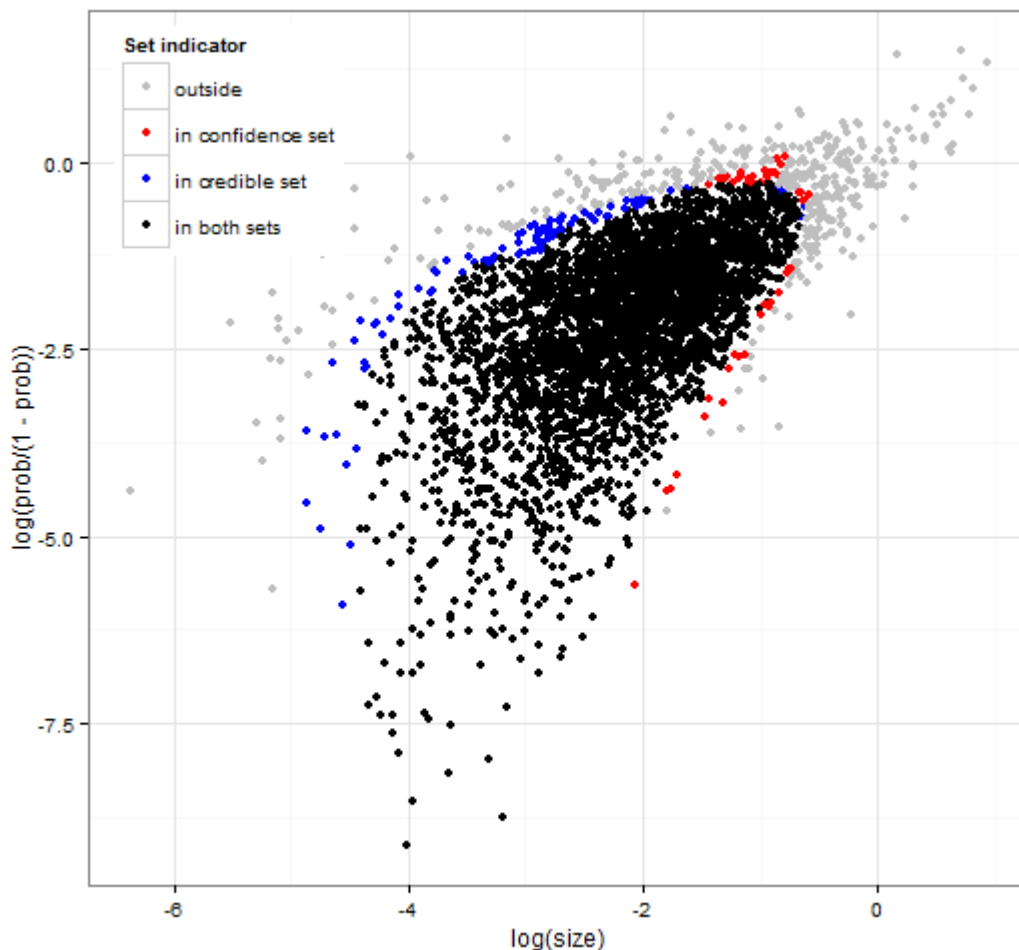


Provides a description of credible intervals and confidence intervals for set selection together with generic R code to calculate both given the likelihood function and some observed data. Further it proposes a test statistics that gives credible and confidence intervals of optimal size that are consistent with each other.

In short and avoiding formulas. The Bayesian **credible interval** is based on the **probability of the parameters given the data**. It collects the parameters that have a high probability into the credible set/interval. The 95% credible interval contains parameters that together have a probability of 0.95 given the data.

The frequentist **confidence interval** is based on the **probability of the data given some parameters**. For each (possibly infinitely many) parameter, It first generates the set of data that is likely to be observed given the parameter. It then checks for each parameter, whether the selected high probability data contains the observed data. If the high probability data contains the observed data, the corresponding parameter is added to the confidence interval. Thus, the confidence interval is the collection of parameters for which we cannot rule out the possibility that the parameter has generated the data. This gives a rule such that, if applied repeatedly to similar problems, the 95% confidence interval will contain the true parameter value in 95% of the cases.

95% credible set and 95% confidence set for an example from a negative binomial distribution



edited Oct 1 '15 at 5:36

answered Sep 3 '15 at 16:20



user36160

316 3 7

The description of the confidence intervals is not correct. The "95%" comes from the probability that a sample from the population will product an interval that contains the true value of the parameter. –

[jlimahaverford](#) Sep 29 '15 at 20:58

@jlimahaverford - The description is correct as is yours. To make the link to what you describe, I added "This gives a rule such that, if applied repeatedly to similar problems, the 95% credible interval will contain the true parameter value in 95% of the cases." – [user36160](#) Sep 30 '15 at 19:22 ✎

- 1 I was not talking about your description of credible intervals I was talking about confidence intervals. I'm now noticing that in the middle of your paragraph on confidence intervals you start talking about credible again, and I think this is a mistake. The important idea is this "If this were the true value of the parameter, what is the probability that I would draw a sample this extreme or more. If the answer is greater than 5% it's in the confidence interval." – [jlimahaverford](#) Sep 30 '15 at 19:33

@jlimahaverford - agree and corrected - Thanks. – [user36160](#) Sep 30 '15 at 19:37

hmm, I am not seeing it corrected. – [jlimahaverford](#) Sep 30 '15 at 19:39

3

I found a lot of interpretations about confidence interval and credible set are wrong. For example, confidence interval cannot be expressed in this format  $P(\theta \in CI)$ . If you look closely on the 'distributions' in the inference of frequentist and Bayesian, you will see Frequentist works on Sampling Distribution on the data while Bayesian works on (posterior) distribution of the parameter. They are defined on totally different Sample Space and Sigma Algebra.



So yes you can say 'If you repeat the experiment a lot of times, approximately 95% of the 95% CIs will cover the true parameter'. Although in Bayesian you get to say 'the true value of the statistic lies in a 95% credible interval with probability 95%', however, this 95% probability (in Bayesian) itself is only an estimate. (Remember it is based on the condition distribution given this specific data, not the sampling distribution). This estimator should come with a random error due to random sample.

Bayesian try to avoid the type I error issue. Bayesian always say it does not make sense to talk about type I error in Bayesian. This is not entirely true. Statisticians always want to measure the possibility or error that 'Your data suggests you to make a decision but the population suggests otherwise'. This is something Bayesian cannot answer (details omitted here). Unfortunately, this may be the most important thing statistician should answer. Statisticians do not just suggest a decision. Statisticians should also be able to address how much the decision can possibly go wrong.

I have to invent the following table and terms to explain the concept. Hope this can help explain the difference of Confidence Interval and Credible Set.

Please note that the posterior distribution is  $P(\theta_0 | Data_n)$ , where  $\theta_0$  is defined from the prior  $P(\theta_0)$ . In frequentist the sampling distribution is  $P(Data_n; \theta)$ . The sampling distribution of  $\hat{\theta}$  is  $P(\hat{\theta}_n; \theta)$ . The subscript  $n$  is the sample size. Please do not use the notation  $P(Data_n | \theta)$  to present the sampling distribution in frequentist. You can talk about random data in  $P(Data_n; \theta)$  and  $P(\hat{\theta}_n; \theta)$  but you cannot talk about random data in  $P(\theta_0 | Data_n)$ .

Method	Estimate	Error Range due to Sampling Error
Frequentist	$\hat{\theta}$	Confidence Interval
Bayesian	Posterior Distribution. Credible Set is just a feature of Posterior Distribution	???????

The '???????' explains why we are not able to evaluate type I error (or anything similar) in Bayesian.

Please also note that credible sets can be used to approximate confidence intervals under some circumstances. However this is only mathematical approximation. The interpretation should go with frequentist. The Bayesian interpretation in this case does not work anymore.

[Thylacoleo](#)'s notation in  $P(x|\theta)$  is not frequentist. This is still Bayesian. This notation causes a fundamental problem in measure theory when talking about frequentist.

I agree with the conclusion made by [Dikran Marsupial](#). If you are the FDA reviewer, you always want to know the possibility that you approve a drug application but the drug is actually not efficacious. This is the answer that Bayesian cannot provide, at least in classic/typical Bayesian.

edited Apr 13 '17 at 12:44



Community ♦

1

answered Jul 3 '13 at 6:14



Chester Lin

31 2

This is more of a comment but too long. In the following paper:

<http://www.stat.uchicago.edu/~lekheng/courses/191f09/mumford-AMS.pdf> Mumford have the following interesting comment:

While all these really exciting uses were being made of statistics, the majority of statisticians themselves, led by Sir R.A. Fisher, were tying their hands behind their backs, insisting that statistics couldn't be used in any but totally reproducible situations and then only using the empirical data. This is the so-called 'frequentist' school which fought with the Bayesian school which believed that priors could be used and the use of statistical inference greatly extended. This approach denies that statistical inference can have anything to do with real thought because real-life situations are always buried in contextual variables and cannot be repeated. Fortunately, the Bayesian school did not totally die, being continued by DeFinetti, E.T. Jaynes, and others.

edited Feb 9 '19 at 22:51



Ben - Reinstate Monica

51.8k 2 64 211

answered Dec 24 '16 at 1:13



kjetil b halvorsen

44.2k 9 100 332



**Highly active question.** Earn 10 reputation in order to answer this question. The reputation requirement helps protect this question from spam and non-answer activity.

