Computational Bayesian Statistics: Credible Intervals

A **credible interval** is a Bayesian version of a frequentist's confidence interval. We begin by reviewing frequentist confidence intervals.

1 Frequentist Confidence Intervals

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a distribution that depends on a one-dimensional parameter θ. To construct a confidence interval for θ we must

1. Come up with an estimator $\widehat{\theta}$ for θ.
2. Come up with some function of both $\widehat{\theta}$ and θ whose distribution is known and, in particular, not dependent on the unknown θ. This function is known as a **pivotal quantity**.
3. Use the distribution from Step 2 to come up with "critical values" that the pivotal quantity is between with a prespecified probability.
4. Using the inequalities from Step 3, solve for bounds for the unknown θ.

Example 1: Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the $N(\mu, 1)$ distribution.

Derive a 90% confidence interval for μ

Step One: An estimator of μ is the sample mean $\widehat{\mu} = \overline{X}$.

Step Two: The distribution of $\overline{X}$ is $N(\mu, 1/n)$. Thus, we know that

$$\frac{\overline{X} - \mu}{\sqrt{1/n}} = \frac{\overline{X} - \mu}{1/\sqrt{n}} \sim N(0, 1).$$

Since this distribution is completely known (not only is it known to be normal but it does not depend on any unknown parameters), $\frac{\overline{X}-\mu}{1/\sqrt{n}}$ is a pivotal quantity.

Step Three: Suppose that $Z \sim N(0, 1)$, we can find (using a $z-$table or computer that

$$P(-1.645 < Z < 1.645) = 0.90.$$

Since $\frac{\overline{X}-\mu}{1/\sqrt{n}}$ has the same distribution, we also have that

$$P\left(-1.645 < \frac{\overline{X} - \mu}{1/\sqrt{n}} < 1.645\right) = 0.90.$$

Step Four: Solving for μ "in the middle" gives

$$P\left(\overline{X} - 1.645\frac{1}{\sqrt{n}} < \mu < \overline{X} + 1.645\frac{1}{\sqrt{n}}\right) = 0.90.$$

The 90% confidence interval for μ is given by

$$\boxed{\left(\overline{X} - 1.645\frac{1}{\sqrt{n}}, \overline{X} - 1.645\frac{1}{\sqrt{n}}\right).}$$

Interpretation: Note that μ is not a random variable in the frequentist paradigm. Here, it is the endpoints of the interval that are random. Once data is collected for some fixed sample size n and the sample mean is computed. The confidence interval becomes numerical. Suppose the interval is then, for example, $(-3.421, 4.718)$. The true mean μ is fixed and is either in this interval or not in this interval. It is **not** in there with probability 0.90. The randomness came from the random sampling from the normal distribution. Different samples will result in different sample means which will result in different confidence intervals. In the long run, with repreated sampling, the true mean μ will be correctly captured by the interval 90% of the time. $\square$

Example 2: Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the exponential distribution with rate λ. Derive a 90% confidence interval for λ

Step One: Since the mean of this distribution is $1/\lambda$, one decent idea for an estimator of λ is $\widehat{\lambda} = 1/\overline{X}$.

Step Two: The distribution of $\overline{X}$ can be found using moment generating functions. It is $\overline{X} \sim \Gamma(n, n\lambda)$.

$1/\overline{X}$ has an inverse gamma distribution but we will see that is is sufficient to work with the distribution of $\overline{X}$.

We need to find a function of $1/\overline{X}$ and λ whose distribution is known and free of the unknown λ. Any function of $\overline{X}$ and λ can be rewritten as a function of $1/\overline{X}$ and λ. This is why we can work with $\overline{X}$ instead of $1/\overline{X}$.

If $Y \sim \Gamma(\alpha, \beta)$ and $c > 0$ is a constant, we can show that $cY \sim \Gamma(\alpha, \beta/c)$. Applying this to $\overline{X} \sim \Gamma(n, n\lambda)$, we see that $\lambda\overline{X} \sim \Gamma(n, n)$. Since this is a completely known distribution, $\lambda\overline{X}$ is a pivotal quantity.

Step Three: Suppose that $Y \sim \Gamma(n, n)$. We can find, using a computer, critical values a and b that solve

$$0.90 = P(a < Y < b) = \int_a^b \frac{1}{\Gamma(n)} n^n y^{n-1} e^{-ny}\, dy.$$

There are many pairs of values for a and b that will solve this problem. Ideally, one would choose a and b as a solution to this integral equation with the restriction that $b - a$ is minimized so that the resulting confidence interval will be shortest.

(In reality though, if you are up on your Mathematical Statistics, you would transform the pivotal quantity $\lambda \overline{X}$ into something having a χ^2 distribution and use a χ^2-table. A χ^2 distribution is a special case of the gamma distribution that depends on only one parameter as opposed to two paramters for the gamma distribution. There are many tables floating around that give that give critical values for χ^2 distributions that have been obtained through numerical integration of the χ^2 pdf!)

Step Four: Assuming that you have solved for a and b (they will depend on n and on the 0.90), you would write

$$a < Y < b$$

$$\Downarrow$$

$$a < \lambda \overline{X} < b$$

and solve for λ "in the middle" to get

$$\frac{a}{\overline{\overline{X}}} < \lambda < \frac{b}{\overline{\overline{X}}}.$$

The 90% confidence interval for λ is then given by

$$\boxed{\left(\frac{a}{\overline{\overline{X}}}, \frac{a}{\overline{\overline{X}}} \right)}$$

where, again, a and b will depend on n and 0.90. (And ideally would be written as chi-squared critical values!)

Interpretation: Once again, we have an interval with random endpoints that will contain the true value of λ with probability 0.90. Once we collect the sample and compute the numerical value of $\overline{X}$ and hence the numerical confidence interval, we will have a fixed interval that either contains λ or doesn't contain λ. In the long run, with repreated sampling, the true mean λ will be correctly captured by the interval 90% of the time. □

2 Bayesian Credible Intervals

As Bayesians, we are thinking of parameters, such as μ for the normal distribution, as random variables. Thus, it now makes sense to write statements like

$$P(-3.421 < \mu < 4.718) = 0.90.$$

This probability can be computed by integrating a prior pdf for μ but, if we want to let the data speak, we'd better use the posterior pdf for μ given the data!

Like all Bayesian results, the credible interval will be affected by the choice of the prior distribution for μ. In the following examples, we will compare results for different priors.

Example 1: Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the $N(\mu, 1)$ distribution.

Compute a 90% credible interval for μ under the assumptions

(a) μ has a flat prior

(b) μ has a $N(\mu_0, \sigma_0^2)$ prior for known hyperparameters μ_0 and σ_0^2

Solution to (a):

The likelihood is

$$f(\vec{x}|\mu) = (2\pi)^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

The prior is

$$f(\mu) \propto 1, \quad -\infty < \mu < \infty.$$

The posterior is

$$\begin{aligned} f(\mu|\vec{x}) &\propto f(\vec{x}|\mu)\cdot f(\mu) \\ &\propto e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}\cdot 1 \\ &= e^{-\frac{1}{2}\sum x_i^2+\mu\sum x_i-\frac{n}{2}\mu^2} \\ &\propto e^{-\frac{1}{2}\sum x_i^2+\mu\sum x_i-\frac{n}{2}\mu^2} \\ &= \vdots \\ &\propto e^{-\frac{n}{2}(\mu-\overline{x})^2}. \end{aligned}$$

That is, the posterior distribution for μ given $\vec{X} = \vec{x}$ is $N(\overline{x}, 1/n)$.

We wish to find critical values a and b such that

$$P(a < \mu < b|\vec{X} = \vec{x}) = 90\%$$

In this conditional world, we have that $\frac{\mu-\overline{x}}{1/\sqrt{n}} \sim N(0,1)$.

We know that for a $N(0,1)$ random variable Z,

$$P(-1.645 < Z < 1.645) = 0.90.$$

(Just as in the frequentist case, there are other non-symmetric values as well!)

So,

$$\begin{aligned}
0.90 &= P(-1.645 < Z < 1.645) \\
&= P\left(-1.645 < \tfrac{\mu-\overline{x}}{1/\sqrt{n}} < 1.645 \middle| \vec{X} = \vec{x}\right) \\
&= \vdots \\
&= P\left(\overline{x} - 1.645\tfrac{1}{\sqrt{n}} < \mu < \overline{x} + 1.645\tfrac{1}{\sqrt{n}}\right)
\end{aligned}$$

The 90% credible interval for μ, using a flat prior, is

$$\boxed{\left(\overline{x} - 1.645\frac{1}{\sqrt{n}}, \overline{x} + 1.645\frac{1}{\sqrt{n}}\right).}$$

Interpretation: This looks very similar to what we had in the frequentist case but the interpretation is very different. The lowercase notation for $\overline{x}$ indicates that the sample mean, and hence the endpoints of the interval, have been fixed and computed. As a frequentist, the parameter μ would be fixed and would be either in the interval or not. It would not be in there "with some probability".

In the Bayesian case though, we are thinking of μ as random, and we can say, using a flat prior and observing the data as $x_1, x_2, \ldots, x_n$, that μ is in this interval with probability 0.90!

Solution to (b):

The likelihood is still

$$f(\vec{x}|\mu) = (2\pi)^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

The prior is now

$$f(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}.$$

The posterior is then

$$\begin{aligned}
f(\mu|\vec{x}) &\propto f(\vec{x}|\mu) \cdot f(\mu) \\
&\propto e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2} \cdot e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2} \\
&= \vdots \\
&\propto e^{-\frac{1}{2(\sigma^2)^*}(\mu=\mu^*)^2}
\end{aligned}$$

where

$$\begin{aligned} \mu^* &= \tfrac{\mu_0+\sigma_0^2\sum x_i}{1+n\sigma_0^2} \\ (\sigma^2)^* &= \tfrac{\sigma_0^2}{1+n\sigma_0^2} \end{aligned}$$

Thus, we have that

$$\mu|\vec{x} \sim N\left(\mu^*, (\sigma^2)^*\right).$$

We wish to find critical values a and b such that

$$P(a < \mu < b|\vec{X} = \vec{x}) = 90\%$$

In this conditional world, we have that $\frac{\mu-\mu^*}{\sqrt{(\sigma^2)^*}} \sim N(0,1)$.

We know that for a $N(0,1)$ random variable Z,

$$P(-1.645 < Z < 1.645) = 0.90.$$

So,

$$\begin{aligned} 0.90 &= P(-1.645 < Z < 1.645) \\ &= P\left(-1.645 < \tfrac{\mu-\mu^*}{\sqrt{(\sigma^2)^*}} < 1.645 \middle| \vec{X} = \vec{x}\right) \\ &= \vdots \\ &= P\left(\overline{x} - 1.645\tfrac{1}{\sqrt{n}} < \mu < \overline{x} + 1.645\tfrac{1}{\sqrt{n}}\right) \end{aligned}$$

The 90% credible interval for μ, using the conjugate $N(\mu_0, \sigma_0^2)$, is

$$\boxed{\left(\mu^* - 1.645\sqrt{(\sigma^2)^*},\ \mu^* + 1.645\sqrt{(\sigma^2)^*}\right).}$$

Note that, as $\sigma_0^2 \to \infty$, the $N(\mu_0, \sigma_0^2)$ conjugate prior is an ever flattening bell curve that is "squishing down" to a flat line. So, the flat uninformative prior can be thought of as a limiting case of the conjugate prior. Indeed, in this case and for fixed n,

$$\mu^* = \frac{\mu_0 + \sigma_0^2\sum x_i}{1 + n\sigma_0^2} \to \frac{\sum x_i}{n} = \overline{x}$$

and

$$(\sigma^2)^* = \frac{\sigma_0^2}{1+n\sigma_0^2} \to \frac{1}{n}$$

as $\sigma_0^2 \to \infty$. These limiting parameters match the parameters in the posterior distribution for μ under the assumption of a flat prior.

Still, our resulting interval is kind of convoluted. In my personal opinion, too many people fall back on conjugate priors for computational convenience. I would go with the flat prior here unless I really knew something a priori about μ. I wouldn't worry so much about the fact that it is a normal distribution used for the prior as much as the parameters. Perhaps prior experimentation/data suggests that the mean is around 3. I would then use a normal prior with mean 3 and a variance that expresses how confident I am about that prior estimate. (Small prior variance for high confidence and larger prior variance for less confidence.) $\square$

Example 2: Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the exponential distribution with rate λ.

Compute an 85% credible interval for λ under the assumptions

(a) λ has a flat prior

(b) λ has a $\Gamma(\alpha, \beta)$ prior for known hyperparameters α and β

Solution to (a):

The likelihood is

$$f(\vec{x}|\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \prod_{i=1}^n I_{(0\infty)}(x_i).$$

The prior is

$$f(\lambda) \propto 1, \quad \lambda > 0.$$

The posterior is

$$\begin{aligned} f(\lambda|\vec{x}) &\propto f(\vec{x}|\lambda) \\ &\propto \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \cdot 1 \\ &= \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \end{aligned}$$

As a function of λ. this looks like a gamma distribution. In particular,

$$\lambda|\vec{x} \sim \Gamma(n+1, \sum_{i=1}^n x_i).$$

We wish to find critical values a and b such that

$$P(a < \lambda < b|\vec{X} = \vec{x}) = 0.85.$$

If you've had MathStat, the best approach would be to multiply $a < \lambda < b$ through by appropriate constants in order to move from working with a gamma distribution to a chi-squared distribution. Then, give your cutoffs in terms of symbolic chi-squared critical values of numerical ones from a χ^2-table.

Otherwise, you need to numerically solve

$$\int_a^b \frac{1}{\Gamma(n+1)} \left(\sum_{i=1}^{n} x_i \right)^{n+1} \lambda^n e^{-\lambda \sum x_i} \, d\lambda.$$

There are many sets of values for a and b that will solve this. For simplicity, you could take $a = 0$ and just solve for b. Alternatively, you could take $b = \infty$ and try to solve for a.

In frequentist statistics, taking $a = 0$ is often done for simplicity but sometimes people will use Calculus (or numerics) to find the shortest possible confidence interval. After all, we are giving an interval estimate for λ and we would like to be as precise as possible.

In Bayesian statistics, if you wanted to optimize your credible interval, the goal would be to find an interval of values for λ that have the **highest posterior density**.

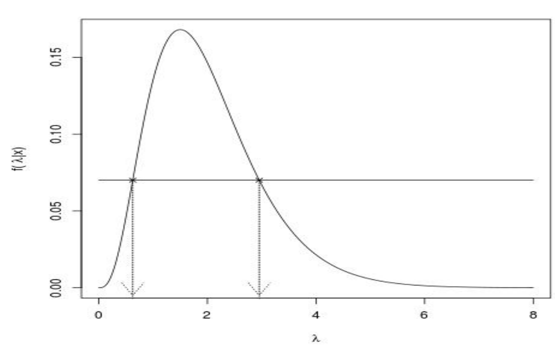
Definition:
In Bayesian statistics, a $100(1-\alpha)\%$ **highest posterior density** region for a parameter θ is a subset $\mathcal{C}$ of the parameter space that is defined by

$$\mathcal{C} = \{\theta : f(\theta|\vec{x}) \geq k\}$$

where k is the largest value such that

$$\int_{\mathcal{C}} f(\theta|\vec{x}) \, d\theta = 1 - \alpha.$$

Basically this means that you want to find the highest horizontal line that, when intersected with the posterior pdf, defines θ values that, when integrating between, will give you $1-\alpha$. For the exponential example, this will give two values as depicted below.



These values are the endpoints of the highest posterior density region for λ. Note that, in the case of multimodal posterior densities, the highest posterior density region may consist of a collection of disjoint intervals. (This is why it's called a "region" and not a highest posterior density "interval".)

Solution to (b): Similar because the prior is a conjugate prior– we will just get a different gamma posterior.