Bayesian Analysis

A Credible Set Problem

Submitted by: Soumarya Basak

Instructor: Prof Gaurangadeb Chattopadhay

UNIVERSITY OF CALCUTTA

Department of Statistics

Contents

Abstract

The main motivation for this document is to understand the different approaches for finding an interval estimation of a population parameter by *Bayesian* method and usual *Frequentest (Classical)* method. At the end of the paper it will be clear enough to understand the view of this two conflicts.

The document is consist of an example of binomial distribution, where $X|\theta \sim bin(n, \theta)$ where θ is the parameter, and we wish to get an interval estimation for θ. Now for this we have two approaches. We go through all the process and discuss the result of their comparison.

1 The Bayesian Approach

The Bayesian approach is much more different from the usual inference procedure what we are used to with. The first and the most important part is that here we consider some *prior information* about θ and based on our data, we want to improve the information of θ for the further purpose , called *posterior information.* So the θ is a random quality here. Hence the estimation procedure becomes different from the usual one and totally based on the *posterior information* about the θ, $\pi(\theta|x)$ which is the conditional distribution of θ given data. If the prior information about θ is given by $\tau(\theta)$, then the posterior distribution of $\theta|x$ is given by,

$$\pi(\theta|x) = \frac{p(x|\theta).\tau(\theta)}{\int_\theta p(x|t).\tau(t)dt}$$

1.1 Credible Interval

For an interval estimation in Bayesian Inference we find a Credible Interval or perhaps sometimes a credible set for the within which the θ lies with certain confidence.

An interval $[l(x), u(x)]$ based on the observed values x is called a credible interval at confidence level $1-\alpha$ if

$$P\Big\{l(x) < \theta < u(x)|x\Big\} \geq 1-\alpha$$

Sometimes, a set $c_x \in \Theta$ is called a credible set at level $1-\alpha$ given the sample point x if

$$P\Big\{\theta \in c_x|x\Big\} \geq 1-\alpha$$

1.1.1 Choices of $l(x)$ & $u(x)$

The most easiest way of considering such l and u are such that,

$$P\Big[\theta < l(x)\Big] \leq \alpha/2 \; ; \; P\Big[\theta > u(x)\Big] \leq \alpha/2$$

As we omitting the 1st and last $\alpha/2^{th}$ quantiles.

1.1.2 HPD Credible Set

Define, $c_{x,k} = \{\theta \in \Theta|\pi(\theta|x) \geq k\}$
And then take, $c_x = sup\{k : P[\theta \in c_{x,k}|x] \geq 1-\alpha\}$
and,

$$P[\theta \in c_x|x] \geq 1-\alpha$$

Note that,

$$\pi(\theta|x) \geq k \equiv (l < \theta < u)|x$$

So it reduced to the previous case[1].

1.2 Our Example (Beta-Binomial Model)

For our example we have data X, from a $bin(n, \theta)$ distribution and try to find an interval estimation of parameter θ based on our observation.

From our knowledge it is meaningful to consider a prior information about such that $0 < \theta < 1$, that is we may consider $\theta \sim Beta(\alpha, \beta)$ where, α and β is known.
Hence the prior distribution of θ is given by,

$$\tau(\theta) \equiv \theta \sim Beta(\alpha, \beta)$$

Then the posterior distribution of $\theta|(x = x^*)$ is given by,

$$\pi(\theta|x = x^*) \equiv \theta|(x = x^*) \sim Beta(x^* + \alpha, n - x^* + \beta)$$

1.2.1 The Data

Suppose for a fix n we have data from $bin(n, \theta)$ distribution where $\theta \sim Beta(\alpha, \beta)$. Note that we have $N = 10^6$ observations of θ and for each θ we have a data x from corresponding binomial distribution.

[1] It is preferable to choose a interval or set where the probability distribution has higher chances to occure, so it is preferable to choose a HPD credible set.

1.2.2 The Algorithm

So, we have 10^6 values of θ and x respectively. Note that for bayesian analysis we need to fix our data x at some value say x^*.

Step 1 First of all we fix our data at $x = x^*$. i,e from the 10^6 data values we are interested in the data values equal to x^*. Say there are N_{x^*} data values equals to x^*. And so we would further work with N_{x^*} much values of θ from $Beta(\alpha, \beta)$

An interval $[l(x^*), u(x^*)]$ is said to be a credible interval with confidence level 95% if,

$$P\Big\{l(x) < \theta < u(x)|x = x^*\Big\} \geq 0.95$$

we need to check whether our data satisfies this or not.

Step 2 Our next task is to find the $l(x^*)$ and $u(x^*)$ From the choices of $l(x)$ and $u(x)$ discussed above we have that,
$l(x^*)$ is the 0.025^{th} quantile of $\pi(\theta|x^*)$, and
$u(x^*)$ is the $(1 - 0.025)^{th}$ quantile of $\pi(\theta|x^*)$
Since n is fixed we can have $l(x^*), u(x^*)$ from $Beta(x^* + \alpha, n - x^* + \beta)$

$$\text{i.e. } l(x^*) = F^{-1}(0.025) \quad \& \quad u(x^*) = F^{-1}(1 - 0.025)$$

Where $F(x)$ is the cdf of $Beta(x^* + \alpha, n - x^* + \beta)$

Step 3 Now from the data we need to find an estimate of confidence coefficient which is given by,

$$\frac{\#\theta^j \in [l(x^*), u(x^*)]}{N_{x^*}} \quad \forall j = 1, 2, N_{x^*}$$

1.2.3 Result

Our discussion in this regard is displayed in the following table 1.2.3.

Table 1: Confidence Coefficient for different values of prior parameters for different n and x^*

(α, β)	$X = 2$				$X = 3$				$X = 13$		
	$n = 5$	$n = 15$	$n = 25$	$n = 50$	$n = 5$	$n = 15$	$n = 25$	$n = 50$	$n = 15$	$n = 25$	$n = 50$
$(2, 2)$	0.950376978	0.95061632	0.9487983	0.94816728	0.95017405	0.94984719	0.9502734	0.94423605	0.94848191	0.9505117	0.9515362
$(2, 5)$	0.95009970	0.9500943	0.9513395	0.94724853	0.9515739	0.95040764	0.9497224	0.95112732	0.94599483	0.94881423	0.9495308
$(2, 0.5)$	0.94874500	0.94633838	0.95359239	0.94573643	0.94941258	0.95111615	0.95508727	0.954621848	0.9493799	0.9496386	0.94911267
$(0.5, 2)$	0.949412586	0.949379903	0.9509927	0.949255618	0.94874500	0.9482979	0.94810062	0.95093159	0.94633838	0.94969777	0.9502670
$(4, 2)$	0.95015763	0.9460966	0.95152603	0.93129770	0.950310324	0.9494186	0.9470752	0.956043	0.95012819	0.948873500	0.9504950
$(0.5, 0.5)$	0.94969318	0.94979411	0.950212631	0.9505704665	0.95048869	0.94855509	0.951532	0.95012459	0.95049030	0.94888633	0.94940350
$(1, 5)$	0.9499355	0.95035757	0.94983348	0.949929260	0.9489569	0.95051419	0.9501159	0.94947432	0.9416216	0.95024482	0.9478797
$(5, 1)$	0.94895692	0.941621621	0.95370370	1	0.94993553	0.94858387	0.9675925	1	0.95035757	0.9495256	0.9358208
$(1, 1)$	0.9492547	0.95103981	0.9490324	0.9504525	0.9500863	0.9501720	0.9501578	0.9508363	0.9498720	0.9502165	0.9507980

Note. The above data is collected based on a random sample with a fix `seed(7700)` it may vary from sample to sample although the conclusion will remain the same.

1.2.4 Observation

From the Table 1 of 1.2.3, it can be noticed that as n increases the confidence coefficient decreases, fixing the others parameters α, β & X^*. It is because as n increase the value of distinct X (data) increase, and as our N is fixed at 10^6, the number of observation for a fix X^* decreases, so the number of θ 's given X^* decreases, so then approximation also decreases. Although it doesn't decreases very much from 0.95, hence we can give a conclusion that Credible set works well for this example.

Perhaps, for fix n and X^* it is difficult to recognize any pattern within the confidence coefficient for different values of hyperparameter (α, β) and also the same for the different choices of X^* for fix n, α, β.

2 The Classical Approach

Now in our usual classical approach the population parameter is treated as a *fixed unknown* quantity, so the inference for θ(population parameter) is done in completely in different way. An interval$[T_1(\underset{\sim}{X}), T_2(\underset{\sim}{X})]$ is called a $(1-\alpha)\%$ confidence interval for θ if

$$P[T_1(\underset{\sim}{X}) < \theta < T_2(\underset{\sim}{X})] \geq 1-\alpha \qquad \forall\theta$$

2.1 Beta-Binomial Example

Our problem is the same, to find an interval estimation of parameter θ, but θ is not random hare.

2.1.1 The Algorithm

For a fix n we have data from $bin(n, \theta)$ distribution where $\theta \sim Beta(\alpha, \beta)$. Note that we have $N = 10^6$ observations of θ and for each θ we have a data x from corresponding binomial distribution. Since θ is fixed in classical approach then we need to fix a θ.

Step 1 So, first fix a θ at θ^*. i.e. out of all values now we are interested in the data which comes from $bin(n, \theta^*)$, Let say there are N_{θ^*} values for which $\theta = \theta^*$, so now we will only interested in the N_{θ^*} data values come from $bin(n, \theta^*)$ distribution.

Step 2 So, we have data,

$$(\theta^*, x^1), (\theta^*, x^2), ..., (\theta^*, x^{N_{\theta^*}})$$

Then the estimate of confidence coefficient is given by,

$$\frac{\#\theta^* \in [T_1(X^i), T_2(X^i)]}{N_{\theta^*}} \qquad i = 1, 2, ..., N_{\theta^*}$$

Step 3 Now we need to find the intervals $[T_1(X^i), T_2(X^i)]$.

Here we will discuss 3 suggested confidence interval for a binomial model.

2.2 Wald's Confidence Interval

The most commonly used confidence interval for a binomial proportion is based of the large sample distribution of the estimate of the binomial proportion. Note that the estimate of θ is given by,

$$\widehat{\theta} = \frac{x}{n} \qquad \text{,where } n \text{ is the sample size}$$

Then by *central limit theorem* $\widehat{\theta}$ follows *Normal distribution.*
Note that,

$$Var(\widehat{\theta}) = Var(\frac{X}{n}) = \frac{\theta(1-\theta)}{n}$$

Since θ is unknown then,

$$\widehat{Var(\widehat{\theta})} = \frac{\widehat{\theta}(1-\widehat{\theta})}{n}$$

Then, for large n, by CLT,

$$\frac{\widehat{\theta} - E[\widehat{\theta}]}{\sqrt{\widehat{Var(\widehat{\theta})}}} = \frac{\widehat{\theta} - \theta}{\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}} \sim N(0,1)$$

Therefore, the 95% confidence interval for θ is given by,

$$\left[\widehat{\theta} - Z_{0.025}\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}} \quad , \quad \widehat{\theta} + Z_{0.025}\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}\right] \tag{1}$$

where, Z_α is the α^{th} quantile of *Standard Normal Distribution.*

2.3 Agresti-Coull Confidence Interval

Another approximation for confidence interval of binomial proportion is suggested by *Agresti* and *Coull*. Although it is quite similar with the previous one unless the estimator of θ. The estimate of θ is given by,

$$\widehat{\widehat{\theta}} = \frac{x+2}{n+4}$$

i.e. we add 2 success and 2 failure to our observation and that estimates θ

Therefore, the 95% confidence interval for θ is given by,

$$\left[\widehat{\widehat{\theta}} - 2\sqrt{\frac{\widehat{\widehat{\theta}}(1-\widehat{\widehat{\theta}})}{n}} \quad , \quad \widehat{\widehat{\theta}} + 2\sqrt{\frac{\widehat{\widehat{\theta}}(1-\widehat{\widehat{\theta}})}{n}}\right] \tag{2}$$

Note that the 0.025^{th} quantile for standard normal is 1.96 which is approximated by 2 in this interval.

2.4 Clopper-Pearson Confidence Interval

Unlike the above two, this confidence interval doesn't assume normality, perhaps it is based on the exact distribution. It is determined from the exact cumulative distribution function of binomial distribution which is a *incomplete beta function*. Note that if $X \sim bin(n, \theta)$, then the *cdf* is given by,

$$F(x) = Incomplete - Beta(x, n-x+1)$$

so,

$$1 - F(x) = Incomplete - Beta(1+x, n-x)$$

Therefore, the $(1-\alpha)\%$ confidence interval for binomial proportion based on the exact distribution is given by,

$$\left[Beta\big(\frac{\alpha}{2}; x, n-x+1\big) \quad , \quad Beta\big(1-\frac{\alpha}{2}; 1+x, n-x\big)\right] \tag{3}$$

2.5 Result

From the data we have, we will calculate the confidence coefficient for these three confidence interval for different values of n. In the following Table 2 we are listing our results.

Table 2: Confidence Coefficient for different values of n for different choice of confidence interval

(α, β)	Wald's Interval					Agresti-Coull					Clopper Pearson				
	$n = 5$	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 5$	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 5$	$n = 25$	$n = 50$	$n = 100$	$n = 200$
$(2, 2)$	0.809	0.9035	0.9532	0.946	0.952	0.996	0.970	0.968	0.959	0.960	0.996	0.970	0.968	0.959	0.960
$(2, 5)$	0.808	0.900	0.951	0.948	0.954	0.997	0.969	0.968	0.961	0.962	0.997	0.969	0.968	0.961	0.962
$(2, 0.5)$	0.820	0.905	0.956	0.947	0.954	0.997	0.970	0.971	0.959	0.963	0.997	0.970	0.971	0.959	0.963
$(0.5, 2)$	0.806	0.901	0.954	0.953	0.953	0.996	0.969	0.969	0.965	0.961	0.996	0.969	0.969	0.965	0.961
$(4, 2)$	0.801	0.899	0.950	0.948	0.954	0.997	0.967	0.963	0.960	0.961	0.997	0.967	0.963	0.960	0.961
$(0.5, 0.5)$	0.809	0.905	0.952	0.950	0.949	0.996	0.971	0.969	0.961	0.958	0.996	0.971	0.969	0.961	0.958
$(1, 5)$	0.805	0.899	0.953	0.948	0.952	0.997	0.969	0.968	0.962	0.961	0.997	0.969	0.968	0.962	0.961
$(5, 1)$	0.822	0.919	0.961	0.953	0.959	0.997	0.977	0.973	0.961	0.965	0.997	0.977	0.973	0.961	0.965
$(1, 1)$	0.807	0.901	0.953	0.950	0.952	0.996	0.968	0.968	0.961	0.961	0.996	0.968	0.968	0.961	0.961

Note. The above data is collected based on a random sample with a fix `seed(7700)` it may vary from sample to sample although the conclusion will remain the same.

2.5.1 Observation

Table 2 depicts that for smaller values of n the *Wald's interval* doesn't works well, but as n increases the interval gets better and for sufficiently large n it shows desire results. It is because, the *Wald's Interval* is based on the large sample distribution of x, which doesn't works well for smaller n.

Whereas the other two intervals show better results even for smaller n as well as larger n. *Clopper Pearson's confidence interval* being based on the exact distribution of x, gives the better estimate every time, and as n increases keeping α, β fix the approximation show a decreasing nature, it is because, as n increases the number of distinct x increases and as our N is fixed, the number of observations for $x|\theta^*$ shows a vast variability, so the confidence level starts dropping.

It becomes a good coincidence that the *Agresti-Coull confidence interval* shows almost same results as the *Clopper Pearson's Interval* shows, and also the confidence intervals become same. So, *Agresti Coull Interval* is a better approximation than the *Wald's Interval*. But it is not necessarily always conincides with *Clopper Pearson Interval*, as the *Clopper Pearson Interval* uses the exact distribution, and the others use an approximation. So it always better to use the *Clopper Pearson confidence interval* for better result.

Keep in mind that it is useless to find any pattern in confidence coefficient for different values of α, β, keeping n fix. As for the *Frequentest* approach the prior information is unnecessary, θ is fixed unkown quantity here, so the computation confidence level is independent of whatever α, β is. Even if anyone got a pattern in that it is nothing but meaningless.