Paper 3020-2015

Jeffreys Interval for One-Sample Proportion with SAS/STAT® Software

Wu Gong, The Children's Hospital of Philadelphia

ABSTRACT

This paper introduces Jeffreys interval for one-sample proportion using SAS® software. It compares the credible interval from a Bayesian approach with the confidence interval from a frequentist approach. Different ways to calculate the Jeffreys interval are presented using PROC FREQ, the QUANTILE function, a SAS program of random walk Metropolis sampler, and PROC MCMC.

INTRODUCTION

Statisticians give estimates to decision makers. For example, given a fact that 81 events occurred after 263 trials (Newcombe, 1998), what is the probability of success? A simple answer is 30.8%. If given another fact that 8 events occurred after 26 trials, a simple point estimate would be almost same, 30.8%. Most people may agree intuitively that the first fact gives more accurate estimate than the second fact. Anyway, how to express the accuracy, the belief, the uncertainty, or the confidence of the estimate has been a big question for statisticians. Historically, there are two approaches to measure the estimate regarding its variation. One is called frequentist approach, the other is called Bayesian's approach.

This paper serves an introduction to the Jeffreys interval for one-sample proportion using SAS/STAT® software. It explains some basic concepts of Bayesian statistics, and how these concepts could be implemented in different SAS procedures. Related SAS codes and outputs are also presented.

The section "Interval Estimate" compares two methods of Ward confidence interval and Jeffreys credible interval. The next section "Jeffreys Interval" explains how Jeffreys interval been derived and how to use QUANTILE function to get values of the interval. The following section "Random Walk Metropolis Algorithm" will explain the detailed steps of the algorithm and how to get posterior sample with SAS codes. Finally, the section "PROC MCMC" will show how to use PROC MCMC procedure to get numbers.

INTERVAL ESTIMATE

Interval estimate uses sample to calculate an interval of values for an unknown population parameter. Interval estimate and point estimate both are methods of statistical inference. They make prediction about the universe population by sample. While point estimate is a single value of the best guess, interval estimate deals with random component of the sampling, and tells audience how precise the estimate is. Two popular forms of interval estimation are confidence intervals from a frequentist approach, and credible intervals from Bayesian approach.

A confidence interval is an interval bounded by an upper limit and a lower limit. Frequentists treat the unknown parameter as a fixed value. Assuming sampling repeatedly from the universe with the fixed parameter, each sampling will generate a different interval. The uncertainty of the estimate is expressed as a probability of how often the interval will contain the true population parameter (Lesaffre & Lawson, 2012). This probability is called nominal confidence level. Often a 95% level is used in practice. The confidence interval for a binomial proportion is derived through a normal approximation to the binomial distribution. It is also called Wald interval.

$$Binomial\ Confidence\ Interval: \hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

In Bayesian statistics, the estimated parameter is treated as a random variable. A credible interval or credible region covers the probability that the population parameter would fall in the interval. It expresses the degree of belief that the parameter falls in this interval. Two methods could be used for defining a credible interval. One is called equal-tailed interval where the probability of being below the interval is equal to the probability of being above. The other one is called highest probability density (HPD) interval which is the narrowest interval that covers the nominal credible level. All values within the HPD interval are more probable than any other values outside of the HPD interval (Woodworth, 2004).

SAS procedure PROC FREQ provides Wald confidence interval and Jeffreys equal-tailed credible interval for the binomial proportion with binomial-options. The SAS codes to calculate Wald and Jeffreys interval, brief explanation of the codes, and SAS output are followed.

```
data d1; ❶
    do i=1 to 263;
           if i <= 81 then x=1;
           else x=0;
           output;
    end;
    drop i;
run;

title "Table 1. Wald and Jeffreys Interval";
ods noproctitle;
ods select  binomialcls; ❷
proc freq data=d1 order=data;
    tables x/binomial(Wald Jeffreys); ❸
run;
title;
```

❶ Create a table with 263 records, and 81 of them are successes.
❷ Select ODS output of Binomial Confidence Limits table.
❸ Binomial option for Wald confidence interval and Jefferys credible interval.

Table 1. Wald and Jeffreys Interval

Binomial Proportion for x = 1		
Type	95% Confidence Limits	
Wald	0.2522	0.3638
Jeffreys	0.2545	0.3656

JEFFREYS INTERVAL

A Bayesian approach starts with a prior distribution of the parameter. The prior distribution represents estimator's belief about the parameter before any observation, and the posterior distribution is the updated belief about the parameter after observation. By multiplying the prior distribution with the likelihood of observed data, the Bayesian theorem gives you posterior distribution of the parameter. "You use the posterior distribution to carry out all inference" (SAS/STAT(R) 9.3 User's Guide, 2011) including Jeffreys credible interval.

It is often required that a prior is objective and has minimal influence on the posterior. A flat prior is obviously a candidate. Anyway, a flat prior becomes non-flat after transformation. For example, a distribution of parameter becomes non-flat after a log transformation. It had been "unacceptable that the conclusions of a Bayesian analysis depends on what scale the flat prior was taken on" (Lesaffre & Lawson, 2012) until Harold Jeffreys proposed a rule to construct priors. The Jeffreys prior is a non-informative prior invariant under transformation or called re-parameterization. The Jeffreys prior for the binomial proportion is a Beta distribution with parameters $(1/2, 1/2)$. After observing r successes in n trials, the posterior distribution could be derived and has a closed form formula of Beta distribution with parameters $(r + 1/2, n - r + 1/2)$ (Lee, 2012).

The Jeffreys interval is a Bayesian credible interval using the Jeffreys prior. Since the posterior distribution is known, the equal tailed 95% credible interval is simply an interval bounded by 2.5% percentile and 97.5% percentile. The SAS codes to calculate Jeffreys interval using QUANTILE function, brief explanation of codes, and SAS output are followed.

```
title "Table 2. Jeffreys Interval by Calculating Quantiles of Beta Distribution";
data jeff;
    n=263;
    r=81;
    p=r/n;
    u=quantile('beta',0.975,r+1/2,n-r+1/2,); ❶
    l=quantile('beta',0.025,r+1/2,n-r+1/2,); ❷
run;
proc print data=jeff noobs;
    var n r p l u;
    format p l u 7.4;
run;
title;
```

❶ Upper limit equals quantile at .975 of distribution $Beta(r + 1/2, n - r + 1/2)$.
❷ Lower limit equals quantile at .025 of distribution $Beta(r + 1/2, n - r + 1/2)$.

Table 2. Jeffreys Interval by Calculating Quantiles of Beta Distribution

N	R	P	L	U
263	81	.3080	.2545	.3656

RANDOM WALK METROPOLIS ALGORITHM

The posterior distribution of a one-sample proportion has a closed form solution, $Beta(r + 1/2, n - r + 1/2)$. That means a mathematical formula for the probability density function of the distribution is derivable through a set of equations. This method is called an analytic solution. Sometimes, a closed form distribution doesn't exist or it might be difficult to be derived or calculated. Then comes numerical method. Numerical method solves equation by guessing or approximating. A numerical algorithm repeats guessing again and again, produces result closer and closer to the real answer. Although each run of approximating could produce a little different result than another, an appropriate degree of accuracy can be achieved after enough number of iterations.

The Bayesian statistics became popular after a numerical method called Metropolis Algorithm was developed. The concept of the Metropolis Algorithm comes from Von Neumann. If "you want to sample from some specific probability distribution, simply sample from any distribution you have on hand, but keep only the good samples" (Beichl & Sullivan, 2000). The Metropolis Algorithm is a Markov chain Monte Carlo (MCMC) method for generating random samples for a probability distribution. The generated sample is called posterior sample and could be used for statistical inference. For example, the posterior sample could be used to derive upper and lower limits of credible interval.

Markov train is a sequence of numbers, of which every successive number is generated only depending on its preceding number. Russian mathematician Andrei Andreyevich Markov introduced this concept (Gamerman & Lopes, 2006). In this one-sample proportion example, the statistic of interest is to estimate the population probability for a success event. The unknown parameter is modelled as a random variable with value from 0 to 1. To build a Markov train, the algorithm starts with an arbitrary value, for example 0.5. Then, it generates the next value through a stochastic process based on this current value. The stochastic process could be a random walk. It mimics a particle moves randomly from one position to left or right on the line. The distances the particle moves for each step could be a normal distribution with zero mean and a variance, for example, $N(0, 0.08)$. So the next value on the sequence is 0.5 plus a random distance. After multiple iterations, a sequence of numbers is generated to construct a Markov train.

$$Random\ walk\ Markov\ train\colon\ \theta^{n} = \theta^{n-1} + \omega_n, \omega_n \in N(0, \delta)$$

In Metropolis Algorithm, not all values on Markov train are accepted in the posterior sample in order to construct a specific distribution. Good samples are accepted and bad samples are rejected. The new generated candidate value is compared with the current value. If the observations (81 events over 263 trials) are more likely to happen under the candidate value, this candidate value is more likely to be accepted in the posterior sample. Then the ratio of likelihoods for observed data under the current and candidate sample population parameter is evaluated. The ratio is used to determine the probability whether the candidate is accepted or rejected.

Steps of the Metropolis algorithm:

1, Set an arbitrary initial value of the sample, $p_0 = 0.5$, which is the population parameter.
2, Set a normal distribution as the proposal distribution to generate the next candidate sample, $p_1 \sim N(p_0, \delta)$. The variance δ defines how far the step is for a new generated sample value goes from the current value.
3, Generate p_1 based on p_0 (random walk from p_0 to p_1).
4, Calculate likelihood for observed sample by prior distribution and binomial likelihood.

$$Density\ of\ the\ Prior\ Distribution\colon \begin{cases} D\left(p_0; \frac{1}{2}, \frac{1}{2}\right) = p_0^{\frac{1}{2}-1}(1-p_0)^{\frac{1}{2}-1} \Big/ B\left(\frac{1}{2}, \frac{1}{2}\right) \\ D\left(p_1; \frac{1}{2}, \frac{1}{2}\right) = p_1^{\frac{1}{2}-1}(1-p_1)^{\frac{1}{2}-1} \Big/ B\left(\frac{1}{2}, \frac{1}{2}\right) \end{cases}$$

$$Binomial\ Likelihood\colon \begin{cases} L(X|p_0) = p_0^{r}(1-p_0)^{n-r} \\ L(X|p_1) = p_1^{r}(1-p_1)^{n-r} \end{cases}$$

5, If the candidate value has higher likelihood than the current value, accept the candidate.
6, If the candidate value has lower likelihood, accept the candidate in a probability equals ratio.

$$Ratio = \frac{L(X|p_1) \cdot D\left(p_1; \frac{1}{2}, \frac{1}{2}\right)}{L(X|p_0) \cdot D\left(p_0; \frac{1}{2}, \frac{1}{2}\right)}$$

7, If accept, set $p_0 = p_1$. If reject, discard the candidate and keep the current value again in the posterior sample.
8, Go to step 3.

The SAS DATA step could be considered as a process of numerical iteration. And it is very easy to implement a Random Walk Metropolis Algorithm through a do while loop. The RETAIN statement helps to generate values sequentially. Once the posterior sample is generated, empirical percentiles are calculated through PROC UNIVARIATE procedure to construct the equal tailed credible interval. The SAS codes for the Random Walk Metropolis Algorithm, brief explanation of codes, and SAS output are followed.

```
%let nmc=1010000;
%let c=0.08;

data PosteriorSample;
    call streaminit(123);
    n=263;
    r=81;
    p0=0.5; ❶
    retain p0;
    do i = 1 to &nmc.; ❷
           p1=p0+rand('normal',0,&c.); ❸
           do while(p1 lt 0 or p1 gt 1);
                  p1=p0+rand('normal',0,&c.); ❹
           end;
           logratio=(1/2-1)*log(p1)+(1/2-1)*log(1-p1)+r*log(p1)+(n-r)*log(1-p1)
                  - (1/2-1)*log(p0)-(1/2-1)*log(1-p0)-r*log(p0)-(n-r)*log(1-p0); ❺
           if log(rand('uniform')) <= logratio then do; ❻
                  p0=p1;
           end;
           if i gt 10000 and floor(i/20)*20 eq i then do; ❼
                  output;
           end;

    end;
    keep i p0;
run;
title "Table 3. Jeffreys Interval by Random Walk Metropolis Algorithm";
proc univariate data=PosteriorSample noprint;
    var p0;
    output out=PP pctlpts  = 2.5 97.5 pctlpre  = pct;
run;
proc print data=pp noobs;
    format pct2_5 pct97_5 6.4;
run;
title;
```

❶ Arbitrary initial value of the current candidate, p_0.
❷ Number of total iteration. It will control total number of records on posterior sample after consideration of burn-in and thinning.
❸ Set proposal distribution of random walk as normal distribution. Variance defines step size.
❹ Regenerate p_1 in case the probability less than zero or greater than 1.
❺ Calculate the log ratio of likelihood of the sample under two proposals. See step 6.
❻ Reject or accept the proposal.
❼ Drop the first 10000 records (burn-in) and keep one record every 20 records (thinning).

Table 3. Jeffreys Interval by Random Walk Metropolis Algorithm

pct2_5	*pct97_5*
0.2548	0.3657

PROC MCMC

SAS procedure PROC MCMC uses a self-tuning random walk Metropolis algorithm to obtain posterior sample (Chen, 2009). The MCMC procedure produces Jeffreys equal-tailed interval and highest probability density interval. It also provides diagnostics for the convergence of the posterior sample. PROC MCMC has strong flexibility to perform a wide range of Bayesian statistical analysis.

The SAS codes to calculate Jefreys interval using PROC MCMC procedure, brief explanation of the codes, and related SAS output are followed.

```
title "Table 4. Jeffreys Credible Interval by PROC MCMC";
ods select PostIntervals;
proc mcmc data=jeff
      seed=123
      outpost=PosteriorSample ❶
      nbi=10000 ❷
      nthin=20 ❸
      nmc=1000000 ❹
      statistics=(summary interval) diagnostics=none plots=none;
   parms prb 0.5; ❺
   prior prb ~ beta(1/2,1/2); ❻
   model r ~ binomial(n,prb); ❼
run;
title;
```

❶ Output posterior sample into a dataset called PosteriorSample. There will be 1000000/20=50000 records in the posterior sample.
❷ Number of burn-in iterations that be discharged before saved into posterior sample.
❸ Number of thinning controls the thinning of the Markov chain and keep one of every 20 samples.
❹ Number of iteration.
❺ The parameter of posterior probability is named as “prb” with arbitrary initial value of 0.5.
❻ Prior distribution of beta(1/2,1/2).
❼ It specifies the likelihood function as binomial. There are r events occurred after n trials under binomial likelihood with probability prb.

Table 4. Jeffreys Credible Interval by PROC MCMC

Posterior Intervals					
Parameter	Alpha	Equal-Tail Interval		HPD Interval	
PRB	0.050	0.2547	0.3660	0.2543	0.3652

CONCLUSION

SAS provides various methods in Bayesian analysis and MCMC procedure is a great tool of them. In the scenario of one-sample proportion, the Jeffreys credible interval could be calculated using several different ways. Credible interval serves a summary of posterior information. It has more meaningful interpretation than the confidence interval. Also, once the posterior sample has been generated, it has advantages to derive all other statistics such as mean, median, variance and all quantiles. The Bayesian posterior could be used to answer decision makers’ questions more directly and intuitively.

REFERENCES

Beichl, I., & Sullivan, F. (2000). The Metropolis Algorithm. *Computing in Science and Engineering*, 65-69.

Chen, F. (2009). Bayesian Modeling Using the MCMC Procedure. *SAS Global Forum 2009.* SAS Institute Inc.

Gamerman, D., & Lopes, H. F. (2006). *Markov Chain Monte Carlo : Stochastic Simulation for Bayesian Inference.* Chapman & Hall/CRC.

Lee, P. M. (2012). *Bayesian Statistics : An Introduction (4th Edition).* John Wiley & Sons.

Lesaffre, E., & Lawson, A. B. (2012). *Bayesian Biostatistics.* John Wiley & Sons, Ltd.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*.

SAS/STAT(R) 9.3 User's Guide. (2011). SAS Institute Inc.

Woodworth, G. G. (2004). *Biostatistics: A Bayesian Introduction.* John Wiley & Sons, Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Wu Gong
Healthcare Analytic Unit/CPCE/PolicyLab, The Children’s Hospital of Philadelphia
3535 Market Street,
Philadelphia, PA 19104
Work Phone: (267) 426-9445
gongw1@email.chop.edu