

We consider a Taylor approximation of the function

$$\ell(\mathbf{y}; \boldsymbol{\theta}) = \log f_Y(\mathbf{y}; \boldsymbol{\theta}) \quad \boldsymbol{\theta} \in \mathbb{R}^d$$

with respect to $\boldsymbol{\theta}$ around $\boldsymbol{\theta}_0$.

$$\ell(\mathbf{y}; \boldsymbol{\theta}) \simeq \ell(\mathbf{y}; \boldsymbol{\theta}_0) + \dot{\ell}(\mathbf{y}; \boldsymbol{\theta}_0)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \ddot{\ell}(\mathbf{y}; \boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

where

$$\dot{\ell}(\mathbf{y}; \boldsymbol{\theta}) = \frac{\partial \ell(\mathbf{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad \ddot{\ell}(\mathbf{y}; \boldsymbol{\theta}) = \frac{\partial^2 \ell(\mathbf{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$$

are $(d \times 1)$ and $(d \times d)$ respectively.

Evaluating for each of $y_1, \dots, y_n$ and summing the result, we have

$$\ell_n(\theta) \simeq \ell_n(\theta_0) + \dot{\ell}_n(\theta_0)^\top (\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^\top \ddot{\ell}_n(\theta_0)(\theta - \theta_0)$$

Assuming that for n large we can replace θ_0 by $\hat{\theta}_n$, we have

$$\begin{aligned}\ell_n(\theta) &\simeq \ell_n(\hat{\theta}_n) + \dot{\ell}_n(\hat{\theta}_n)^\top (\theta - \hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)^\top \ddot{\ell}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n) \\ &= \ell_n(\hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)^\top \ddot{\ell}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n)\end{aligned}$$

as $\ell_n(\hat{\theta}_n) = \mathbf{0}_d$.

This tells us that the likelihood can be approximated in the vicinity of $\hat{\theta}_n$ by a quadratic:

$$\begin{aligned}\ell_n(\hat{\theta}_n) - \ell_n(\theta) &= -\frac{1}{2}(\theta - \hat{\theta}_n)^\top \ddot{\ell}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n) \\ &= \frac{1}{2}(\theta - \hat{\theta}_n)^\top \left\{ -\ddot{\ell}_n(\hat{\theta}_n) \right\} (\theta - \hat{\theta}_n)\end{aligned}$$

where this quantity is non-negative, as $\hat{\theta}_n$ is a maximizer of ℓ_n .

A quadratic approximation for the likelihood at θ around $\hat{\theta}_n$ is

$$\begin{aligned}\exp\{\ell_n(\theta)\} &\doteq \exp\{\ell_n(\hat{\theta}_n)\} \exp\left\{\frac{1}{2}(\hat{\theta}_n - \theta)^\top \ddot{\ell}_n(\hat{\theta}_n)(\hat{\theta}_n - \theta)\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\theta - \hat{\theta}_n)^\top \{-\ddot{\ell}_n(\hat{\theta}_n)\}(\theta - \hat{\theta}_n)\right\}.\end{aligned}$$

Note that we can also write the right-hand side as proportional to

$$\exp\left\{-\frac{1}{2}(\hat{\theta}_n - \theta)^\top \{-\ddot{\ell}_n(\hat{\theta}_n)\}(\hat{\theta}_n - \theta)\right\}.$$

Using the Normal Approximation in Bayesian inference

Treating θ as a proposed true data generating parameter, and assuming that n is large enough so that $\hat{\theta}_n$ is close to θ , this essentially corresponds to the *frequentist* assumption that

$$\hat{\theta}_n | \theta \approx \text{Normal}_d \left(\theta, \left\{ -\ddot{\ell}_n(\theta) \right\}^{-1} \right)$$

or equivalently that

$$\sqrt{n}(\hat{\theta}_n - \theta) \approx \text{Normal}_d \left(\mathbf{0}_d, \left\{ -\frac{1}{n} \ddot{\ell}_n(\theta) \right\}^{-1} \right).$$

which allows the construction of frequentist *confidence intervals*.

Using the Normal Approximation in Bayesian inference

In the frequentist calculation, it is often assumed that as n is large

$$-\frac{1}{n}\ddot{\ell}_n(\theta) \simeq \mathcal{I}_n(\theta)$$

where

$$\mathcal{I}_n(\theta) = \mathbb{E}_{Y|\theta} \left[- \frac{\partial^2 \ell(Y; \mathbf{t})}{\partial \mathbf{t} \partial \mathbf{t}^\top} \bigg|_{\mathbf{t}=\theta} ; \theta \right]$$

is the Fisher information.

Using the Normal Approximation in Bayesian inference

In the *Bayesian* version, the posterior can be approximated by noting that

$$\pi_n(\theta) \propto \exp\{\ell_n(\theta)\}\pi_0(\theta)$$

and assuming that n is large so that $\pi_0(\theta)$ can be treated as constant in the neighbourhood of $\hat{\theta}_n$, so that

$$\pi_n(\theta) \equiv \text{Normal}_d \left(\hat{\theta}_n, \left\{ -\ddot{\ell}_n(\hat{\theta}_n) \right\}^{-1} \right).$$

which allows the construction of Bayesian *credible intervals*.

Example

Suppose that a Bayesian model is built for realizations for an infinitely exchangeable binary sequence.

- ▶ $Y_i|\theta \sim \text{Bernoulli}(\theta)$
- ▶ $f_Y(y; \theta) = \theta^y(1 - \theta)^{1-y}$ for $y = 0, 1$
- ▶ Log-mass function and derivatives:

$$\ell(y; \theta) = y \log \theta + (1 - y) \log(1 - \theta)$$

$$\dot{\ell}(y; \theta) = \frac{y}{\theta} - \frac{1 - y}{1 - \theta}$$

$$\ddot{\ell}(y; \theta) = -\frac{y}{\theta^2} - \frac{1 - y}{(1 - \theta)^2}$$

- Derivative of the log-likelihood

$$\dot{\ell}_n(\theta) = \frac{s_n}{\theta} - \frac{t_n}{(1 - \theta)}$$

where

$$s_n = \sum_{i=1}^n y_i \quad t_n = \sum_{i=1}^n (1 - y_i)$$

yields the maximum likelihood estimate

$$\hat{\theta}_n = \frac{s_n}{s_n + t_n} = \frac{s_n}{n}$$

Example

- This leads to

$$\begin{aligned}\ddot{\ell}_n(\hat{\theta}_n) &= - \left[\frac{s_n}{\hat{\theta}_n^2} + \frac{t_n}{(1 - \hat{\theta}_n)^2} \right] = - \left[\frac{n\hat{\theta}_n}{\hat{\theta}_n^2} + \frac{n(1 - \hat{\theta}_n)}{(1 - \hat{\theta}_n)^2} \right] \\ &= - \frac{n}{\hat{\theta}_n(1 - \hat{\theta}_n)}\end{aligned}$$

- We also have that

$$\begin{aligned}\mathcal{I}_n(\theta) &= \mathbb{E}_{Y|\theta} \left[\frac{Y}{\theta^2} + \frac{1 - Y}{(1 - \theta)^2}; \theta \right] = \frac{\theta}{\theta^2} + \frac{1 - \theta}{(1 - \theta)^2} \\ &= \frac{1}{\theta(1 - \theta)}\end{aligned}$$

Example

- $(1 - \gamma)\%$ confidence interval: recall that a confidence interval is the *random* interval $[L_n, U_n]$ such that

$$P_{Y|\theta}[\theta \in [L_n, U_n]; \theta] = 1 - \gamma.$$

that is, *for each given θ* ,

$$P_{Y|\theta}[L_n \leq \theta \leq R_n; \theta] = 1 - \gamma.$$

To obtain the interval, we usually construct a *pivotal quantity*, that is some quantity

$$T_n(\theta)$$

whose distribution does not depend on θ .

Example

Here we use

$$T_n(\theta) = \frac{\hat{\theta}_n - \theta}{\sqrt{\frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n}}}$$

and deduce from the asymptotic theory that for any θ

$$T_n(\theta) \rightsquigarrow \text{Normal}(0, 1).$$

Hence for an *equal-tailed* interval, we have

$$[L_n, U_n] : \quad \hat{\theta}_n \pm z_{1-\gamma/2} \sqrt{\frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n}}$$

where z_a is the a th quantile of the standard Normal distribution.

Example

Note that this is a *frequentist* statement:

- ▶ Suppose we *know* θ ;
- ▶ *if* the data are generated from $f_Y(y; \theta)$,
- ▶ *then* the interval $[L_n, U_n]$ will contain θ with probability $1 - \gamma$ under the $f_Y(y; \theta)$ model.

- ▶ The quantities L_n and U_n are functions of the data random variables through the *estimator*;
- ▶ we estimate the interval by computing numerical values l_n and u_n which are functions of the *estimate*.

Specifically, this is *not* a probability statement about the unknown θ ; in the frequentist formulation, such a probability statement is not possible, as the unknown θ is *not a random quantity*.

- $(1 - \gamma)\%$ credible interval: by the approximation theory

$$\pi_n(\theta) \simeq \text{Normal} \left(\hat{\theta}_n, \frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n} \right).$$

Hence we can make the *Bayesian* statement for the unknown θ : in light of the data and the model specification that

$$P_{\pi_n}[\theta \in [l_n, u_n]] = 1 - \gamma.$$

where l_n and u_n are quantities computed from the *estimate*.

But if we consider the standardized parameter

$$\phi = \frac{\sqrt{n}(\theta - \hat{\theta}_n)}{\sqrt{\hat{\theta}_n(1 - \hat{\theta}_n)}}$$

then it is clear that

$$\pi_n(\phi) \simeq \text{Normal}(0, 1)$$

revealing that the Bayesian credible interval will be the same as the frequentist interval.

Example

Numerical example: $n = 20$, $\theta_0 = 0.65$.

```
set.seed(39)
n<-20
th0<-0.65
y<-rbinom(n,1,th0)
sn<-sum(y)
tn<-sum(1-y)
c(sn,tn)  #Data

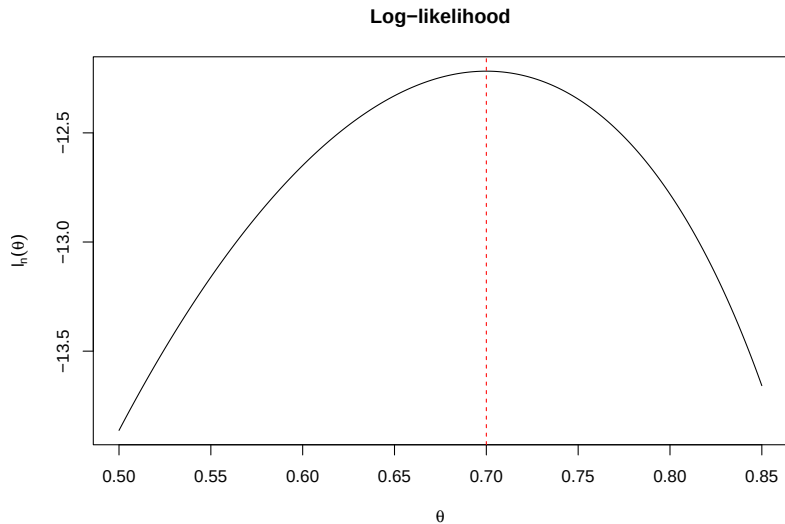
## [1] 14  6

th.n<-sn/(sn+tn)
th.n      #MLE

## [1] 0.7

log.like<-function(xv,sv,tv){
  return(sv*log(xv)+tv*log(1-xv))
}
thv<-seq(0.50,0.85,by=0.001)
ll.n<-log.like(thv,sn,tn)
par(mar=c(4,4,4,2))
plot(thv,ll.n,type='l',
      xlab=expression(theta),ylab=expression(l[n](theta)))
title('Log-likelihood')
abline(v=th.n,lty=2,col='red')
```

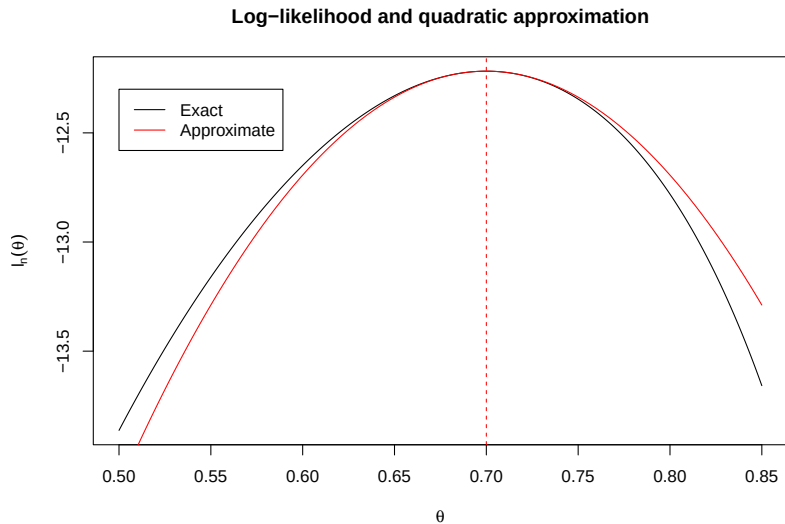
Example



Example

```
#Quadratic approximation
max.val<-log.like(th.n,sn,tn)
qln<-max.val-0.5*(thv-th.n)^2/(th.n*(1-th.n)/n)
par(mar=c(4,4,4,2))
plot(thv,ll.n,type='l',
      xlab=expression(theta),ylab=expression(l[n](theta)))
title('Log-likelihood and quadratic approximation')
lines(thv,qln,col='red')
abline(v=th.n,lty=2,col='red')
legend(0.5,-12.3,c('Exact','Approximate'),lty=1,col=c('black','red'))
```

Example



Assuming a $\pi_0(\theta) \equiv \text{Uniform}(0, 1)$, we have that

$$\pi_n(\theta) \equiv \text{Beta}(s_n + 1, t_n + 1)$$

and using the quadratic approximation we have

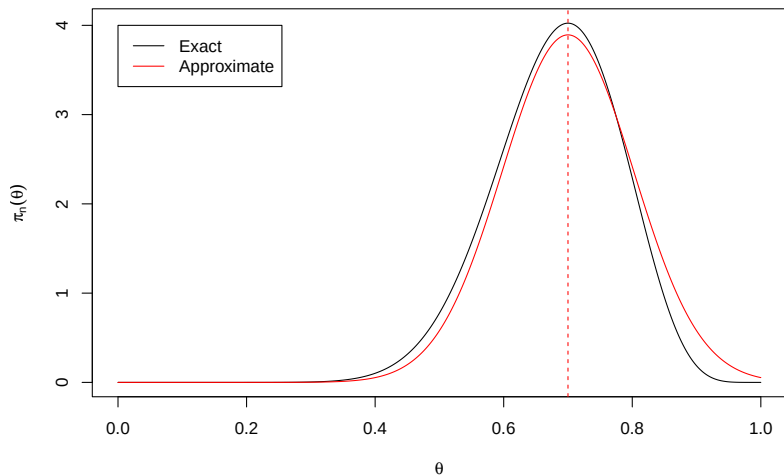
$$\pi_n(\theta) \simeq \text{Normal}(\hat{\theta}_n, \hat{\theta}_n(1 - \hat{\theta})/n)$$

Example

```
#Posterior and approximation
thfull<-seq(0,1,by=0.001)
pi.n<-dbeta(thfull,sn+1,tn+1)
qlnfull<--0.5*(thfull-th.n)^2/(th.n*(1-th.n)/n)
const<-1/sqrt(2*pi*(th.n*(1-th.n)/n))
par(mar=c(4,4,4,2))
plot(thfull,pi.n,type='l',
      xlab=expression(theta),ylab=expression(pi[n](theta)))
title('Posterior and quadratic approximation')
lines(thfull,exp(qlnfull)*const,col='red')
abline(v=th.n,lty=2,col='red')
legend(0,4,c('Exact','Approximate'),lty=1,col=c('black','red'))
```

Example

Posterior and quadratic approximation



Example

Equal-tail intervals: for $\gamma = 0.05$, and a 95% interval

```
#Exact  
qbeta(c(0.025,0.975),sn+1,tn+1)  
  
## [1] 0.4782489 0.8541231  
  
#Approximate  
qnorm(c(0.025,0.975),th.n,sqrt(th.n*(1-th.n)/n))  
  
## [1] 0.4991635 0.9008365
```

that is

Exact : $[0.4782, 0.8541]$

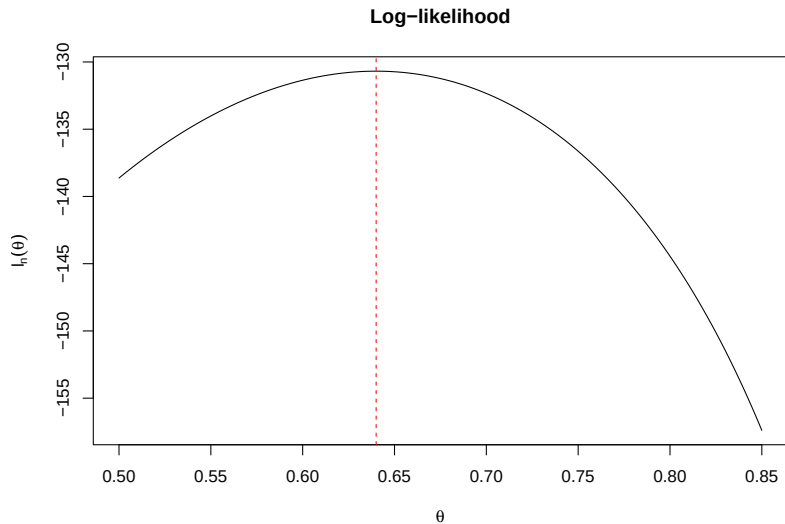
Approximate : $[0.4992, 0.9008]$

Example

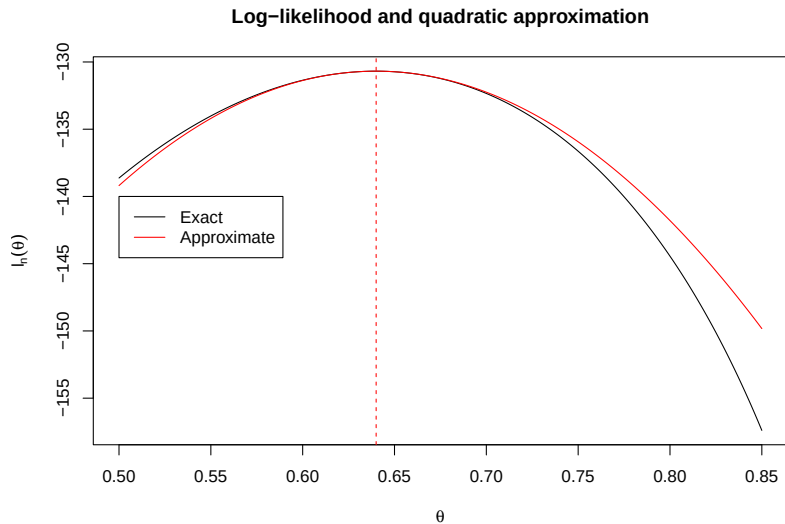
Larger sample size: $n = 200$, $\theta_0 = 0.65$.

```
## [1] 128 72  
## [1] 0.64
```

Example

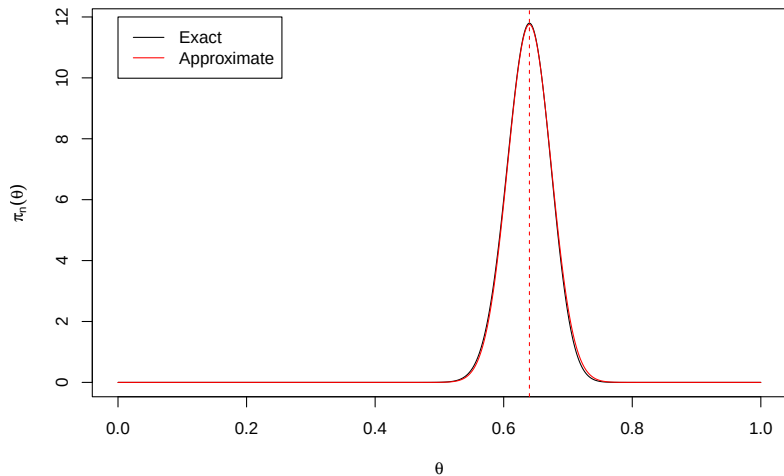


Example



Example

Posterior and quadratic approximation



Example

Equal-tail intervals: for $\gamma = 0.05$, and a 95% interval

```
## [1] 0.5713075 0.7033200  
## [1] 0.5734766 0.7065234
```

that is

Exact : $[0.5713, 0.7033]$

Approximate : $[0.5735, 0.7065]$