Example for large sample credible intervals

Example 1: Prevalence of Recessive Gene

If gene frequencies are in equilibrium, the genotypes AA, Aa, and aa occur with probabilities $(1-\theta)^2$, $2\theta(1-\theta)$, and θ^2 respectively, where θ represents the overall prevalence of the recessive a gene in the population. Plato et al. (1964) published the following data on a haptoglobin type in a sample of 190 people:

Haptoglobin Type	AA	Aa	aa
Count	112	68	10

Let's regard the vector $\mathbf{x} = (x_1, x_2, x_3) = (112, 68, 10)$ as a realization of the random variable $\mathbf{X} \sim \text{Multinomial}\big((1-\theta)^2, 2\theta(1-\theta), \theta^2\big)$.

To save some time and allow us to focus on the results of interest here, I'll give you the likelihood function, its first and second derivatives with respect to θ, and the form of the posterior:

Preliminary Results

General form of Multinomial pmf

$f(\mathbf{x}|\mathbf{p}) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$

Likelihood function

In our example, $p_1 = (1-\theta)^2$, $p_2 = 2\theta(1-\theta)$, and $p_3 = \theta^2$.

$$\begin{aligned}\mathcal{L}(\theta|\mathbf{x}) &= f(\mathbf{x}|\theta) \\ &= \{(1-\theta)^2\}^{x_1}\{2\theta(1-\theta)\}^{x_2}\{\theta^2\}^{x_3}\end{aligned}$$

I'm going to leave this in terms of $x_1, x_2,$ and x_3 for now.

Log-likelihood function

$$\begin{aligned}\ell(\theta|\mathbf{x}) &= \log[\mathcal{L}(\theta|\mathbf{x})] \\ &= \log\left[\{(1-\theta)^2\}^{x_1}\{2\theta(1-\theta)\}^{x_2}\{\theta^2\}^{x_3}\right] \\ &= x_1 \log\{(1-\theta)^2\} + x_2 \log\{2\theta(1-\theta)\} + x_3 \log\{\theta^2\}\end{aligned}$$

First and second derivatives of log-likelihood function

The first derivative of the log-likelihood is:

$$\frac{d}{d\theta}\ell(\theta|\mathbf{x}) = \cdots = \frac{-2x_1\theta}{\theta(1-\theta)} + \frac{x_2(1-2\theta)}{\theta(1-\theta)} + \frac{2x_3(1-\theta)}{\theta(1-\theta)}$$

The second derivative of the log-likelihood is:

$$\frac{d^2}{d\theta^2}\ell(\theta|\mathbf{x}) = \cdots = -\frac{2x_1 + x_2}{(1-\theta)^2} - \frac{2x_3 + x_2}{\theta^2}$$

Maximum likelihood estimator

Setting the first derivative equal to 0, we obtain a maximum likelihood estimator of $\hat{\theta}^{MLE} = \frac{X_2+2X_3}{2n}$.

It can be verified that this gives a global maximum of the likelihood function.

Posterior Distribution

Suppose we adopt a prior of $\Theta \sim \text{Uniform}(0,1)$

The prior distribution for Θ has density $f_\Theta(\theta) = \begin{cases} 1 \text{ if } \theta \in [0,1] \\ 0 \text{ otherwise} \end{cases}$.

Additionally, in part (a) we showed that $f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) = \{(1-\theta)^2\}^{x_1}\{2\theta(1-\theta)\}^{x_2}\{\theta^2\}^{x_3}$.

Applying Bayes' Rule, we find that

$$f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = \cdots = \begin{cases} c\{(1-\theta)^2\}^{x_1}\{2\theta(1-\theta)\}^{x_2}\{\theta^2\}^{x_3} \text{ if } \theta \in [0,1] \\ 0 \text{ otherwise} \end{cases}$$

The integral is kindof annoying, but can be done.

Problems for you

There is just one written problem here. Do this problem before continuing to lab 11.

1. Find a large-sample normal approximation to the posterior distribution for θ.

Use the approximation centered at the maximum likelihood estimate $\hat{\theta}^{MLE}$. Your answer will be in terms of x_1, x_2, x_3, and n.

Solution:

The observed Fisher information at the maximum likelihood estimate is:

$$\begin{aligned} J(\hat{\theta}^{MLE}) &= -\frac{d^2}{d\theta^2}\ell(\theta|\mathbf{x})|_{\theta=\hat{\theta}^{MLE}} \\ &= \frac{2x_1 + x_2}{(1-\hat{\theta}^{MLE})^2} + \frac{2x_3 + x_2}{\left(\hat{\theta}^{MLE}\right)^2} \\ &= \frac{2x_1 + x_2}{(1-\frac{x_2+2x_3}{2n})^2} + \frac{2x_3 + x_2}{\left(\frac{x_2+2x_3}{2n}\right)^2} \end{aligned}$$

The approximate posterior distribution for Θ is therefore

$$\Theta|X_1 = x_1, X_2 = x_2, X_3 = x_3 \sim \text{Normal}\left[\frac{x_2+2x_3}{2n}, \left\{\frac{2x_1+x_2}{(1-\frac{x_2+2x_3}{2n})^2} + \frac{2x_3+x_2}{\left(\frac{x_2+2x_3}{2n}\right)^2}\right\}^{-1}\right]$$