Examples 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 $\mathbf{X} \sim$ Multinomial $\big((1-\theta)^2, 2\theta(1-\theta), \theta^2\big)$.

To save some time/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

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

You should find the numeric values of the mean and variance of this normal approximation based on our data set.

Mean:

```
mle <- (68 + 2 * 10)/(2 * 190)
mle
```

```
## [1] 0.2315789
```

Variance:

$\frac{d^2}{d\theta^2}\ell(\theta|x_1, \ldots, x_n)|_{\theta=\hat{\theta}^{MLE}} = -\frac{2 \cdot 112 + 68}{\left(1-\hat{\theta}^{MLE}\right)^2} - \frac{2 \cdot 10 + 68}{\left(\hat{\theta}^{MLE}\right)^2}$

```
second_deriv_loglik <- -(2 * 112 + 68)/(1 - mle)^2 - (2 * 10 + 68)/(mle^2)
post_approx_var <- -1/second_deriv_loglik
post_approx_var
```

```
## [1] 0.0004682898
```

Therefore, the posterior distribution for θ is approximately Normal$(0.232, 0.000468)$

2. Add a plot of the pdf of the normal approximation to the plot of the actual posterior below.

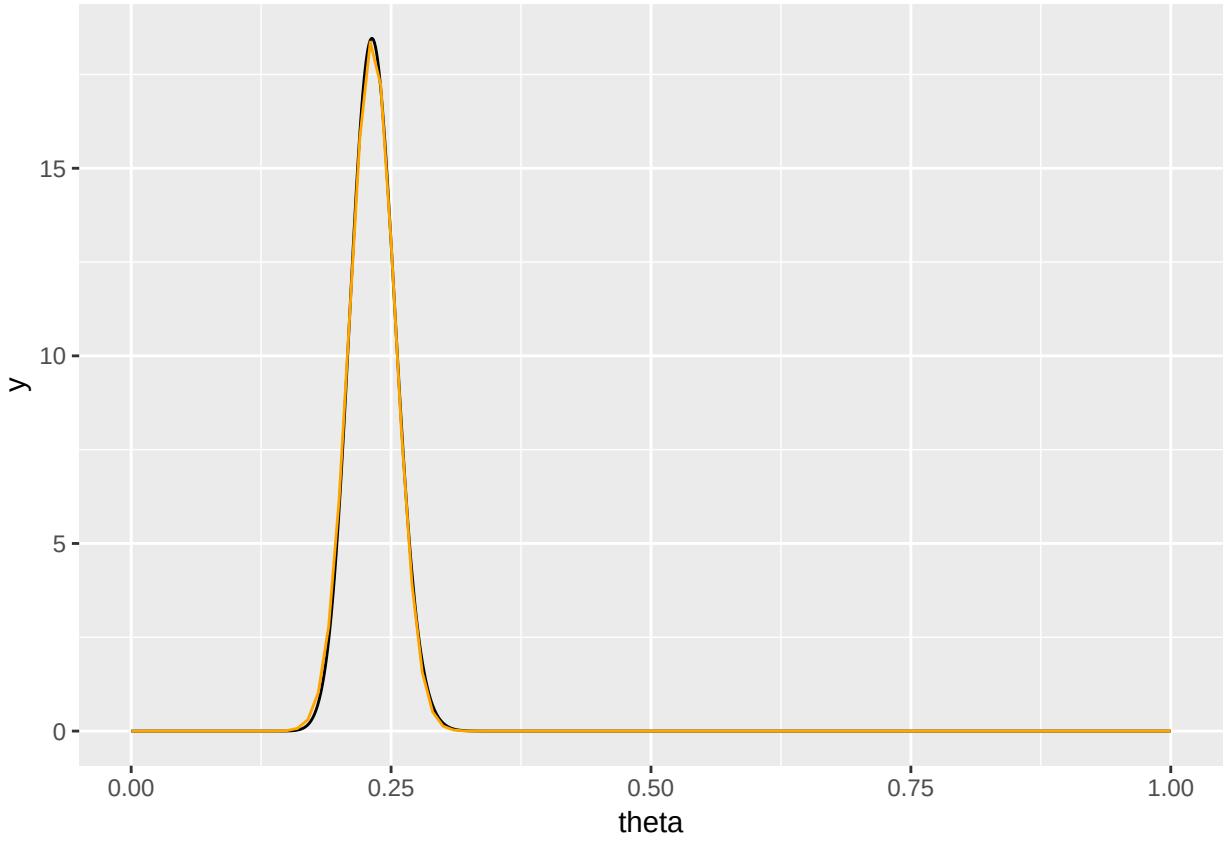
I used Wolfram Alpha to figure out how to calculate the constant c. That's the first 4 lines of the calculation of the log density in the `dposterior` function below:

```
library(ggplot2)

dposterior <- function(theta, x_1, x_2, x_3, log = FALSE) {
    n <- x_1 + x_2 + x_3
    log_d_posterior <- sum(log(seq_len(2 * n + 1))) -
        x_2 * log(2) -
        sum(log(seq_len(2 * x_1 + x_2))) -
        sum(log(seq_len(x_2 + 2 * x_3))) +
        2 * x_1 * log(1 - theta) +
        x_2 * log(2 * theta * (1 - theta)) +
        2 * x_3 * log(theta)

    if(log) {
        return(log_d_posterior)
    } else {
        return(exp(log_d_posterior))
    }
}

ggplot(data = data.frame(theta = c(0, 1)), mapping = aes(x = theta)) +
    stat_function(fun = dposterior, args = list(x_1 = 112, x_2 = 68, x_3 = 10), n = 1001) +
    stat_function(fun = dnorm, args = list(mean = 0.232, sd = sqrt(0.000468)), color = "orange")
```



3. Find and interpret an approximate Bayesian 95% credible interval for θ based on the normal approximation.

```
qnorm(c(0.025, 0.975), mean = 0.232, sd = sqrt(0.000468))
```

```
## [1] 0.1895995 0.2744005
```

There is probability 0.95 that the overall prevalence of the recessive a gene in the population is between about 0.19 and 0.27.