Examples for large sample confidence and credible intervals

Example 1: Asbestos (adapted from Rice)

Here's a description of this example, quoted from Rice

> "The following study was done at the National Institute of Science and Technology (Steel et al. 1980). Asbestos fibers on filters were counted as part of a project to develop measurement standard for asbestos concentration. Asbesto dissolved in water was spread on a filter, and 3-mm diameter punches were taken from the filter and mounted on atransmission electron microscope. An operator counted the number of fibers in each of 23 grid squares, yielding the following counts:"

31, 29, 19, 18, 31, 28, 34, 27, 34, 30, 16, 18, 26, 27, 27, 18, 24, 22, 28, 24, 21, 17, 24

We have $\sum_{i=1}^{23} x_i = 573$ and $\frac{1}{23}\sum_{i=1}^{23} x_i = 24.9$

Let's use the model

$X_i \sim \text{Poisson}(\lambda)$

(Why? According to Wikipedia, "the Poisson distribution describes the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event.")

Find and interpret point and interval estimates for λ in the Frequentist and Bayesian paradigms based on the large-sample Normal approximations to the sampling distribution of the MLE and the posterior distribution.

Example 2: Prevalence of Recessive Gene

In the first problem on Problem Set 2, you derived maximum likelihood and Bayesian estimators for the prevalence of a recessive gene in a population. As a reminder, here was the setup:

Overview/Reminder

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

Here's one way to specify a model for these data. Let's give each genotype a number: $AA \Leftrightarrow 1$, $Aa \Leftrightarrow 2$, and $aa \Leftrightarrow 3$. Define the random variables X_{ij} for each individual $i = 1, \ldots, n$ and genotype $j = 1, 2, 3$ by

$$X_{ij} = \begin{cases} 1 \text{ if individual } i \text{ has genotype } j \\ 0 \text{ otherwise} \end{cases}$$

We can then define the random vector $X_i = (X_{i1}, X_{i2}, X_{i3})$ to represent the genotype of individual i; for example, if we observed the vector $x_i = (0, 1, 0)$, this would represent the case that individual i has genotype Aa. Let's model the X_i as independent and identically distributed (i.e., we're thinking of the individuals in our sample as independent of each other) with each having a Categorical$\big((1-\theta)^2, 2\theta(1-\theta), \theta^2\big)$ distribution. The probability mass function of this distribution is as follows:

$f(x_i|\theta) = \{(1-\theta)^2\}^{x_{i1}}\{2\theta(1-\theta)\}^{x_{i2}}\{\theta^2\}^{x_{i3}}$.

This p.m.f. gives the probabilty that $X_{i1} = x_{i1}$, $X_{i2} = x_{i2}$, and $X_{i3} = x_{i3}$.

For each individual i, we have that $E(X_{i1}) = (1-\theta)^2$, $E(X_{i2}) = 2\theta(1-\theta)$, and $E(X_{i3}) = \theta^2$.

The counts in the table above are a summary of these data: $y_1 = \sum_{i=1}^n x_{i1} = 112$, $y_2 = \sum_{i=1}^n x_{i2} = 68$, and $y_3 = \sum_{i=1}^n x_{i3} = 10$. Here, y_1 is the count of the total number of subjects with genotype AA, y_2 is the count of the total number of subjects with genotype Aa, and y_3 is the count of the total number of subjects with genotype aa.

Results from Problem Set 2

You should have obtained the following results in Problem Set 2:

Frequentist

The maximum likelihood estimator is given by $\widehat{\theta}_{MLE} = \frac{\sum_{i=1}^n X_{i2} + 2\sum_{i=1}^n X_{i3}}{2n}$

The maximum likelihood estimate is $\widehat{\theta}_{MLE} = \frac{68+2*10}{2*190} \approx 0.232$

The Fisher information is

$$I(\theta) = \frac{2}{\theta(1-\theta)}$$

For a large enough sample size n, the approximate sampling distribution of $\hat{\theta}_{MLE}$ is therefore approximately $\hat{\theta}_{MLE} \sim \text{Normal}\left(\theta^*, \frac{\theta^*(1-\theta^*)}{2n}\right)$

Bayesian

We found that a large-sample normal approximation to the posterior distribution for θ was $\theta \sim \text{Normal}\left(\hat{\theta}_{MLE}, \frac{-1}{\frac{d^2}{d\theta^2}L(\hat{\theta}_{MLE}|x_1,\ldots,x_n)}\right)$, where the second derivative of the full log-likelihood is $\frac{d^2}{d\theta^2}L(\theta|x_1,\ldots,x_n) = -\frac{2y_1+y_2}{(1-\theta)^2} - \frac{2y_3+y_2}{\theta^2}$

The maximum likelihood estimate was 0.232. Plugging in to the expression above for the second derivative of the log-likelihood, we have that

$\frac{d^2}{d\theta^2}L(\theta|x_1,\ldots,x_n) = -\frac{2y_1+y_2}{(1-\theta)^2} - \frac{2y_3+y_2}{\theta^2} \approx -2130$

Therefore, the posterior distribution for θ is approximately Normal$(0.232, 1/2130)$

New Problems: Find and Interpret Interval Estimates

Find and interpret a frequentist 95% confidence interval and a Bayesian 95% credible interval for θ.