Problem 1: In a research program on human health risk from recreational contact with water contaminated with pathogenic microbiological material, the National Institute of Water and Atmosphere (NIWA) instituted a study to determine the quality of New Zealand stream water at a variety of catchment types. Out of n=116 one-litre water samples from sites identified as having heavy environmental impact from birds (seagulls) and water fowl, y=17 samples contained Giardia cysts.

(a) Let p denote the true probability that a one-litre sample from this type of site contains Giardia cysts. What is the probability distribution of y?
(b) What is the classical estimate for p obtained by maximizing the likelihood for observed data and 95% confidence interval? (This procedure is called "Maximum Likelihood Estimation")
(c) Under (Non-informative) Jeffrey's prior (i.e. Beta (.5,.5)), what is the estimate and 95% credible interval?
(d) Assume that based on scientific knowledge, the prior distribution for p is determined to be Beta (1,4). What is the estimate and 95% credible interval under this prior?.
(e) (**Predictive Distribution**) If 5 additional 1-litre samples are taken, (based on (d)) what is the probability that atleast 2 are contaminated?

Problem 2: Recall, our problem where we computed

$$\theta := P(HIV|Positive) = \frac{P(Positive|HIV) * P(HIV)}{P(Positive|HIV) * P(HIV) + P(Positive|No\ HIV) * P(No\ HIV)}$$

Which we will write as:

$$\theta = \frac{p_1 p_0}{p_1 p_0 + p_2(1 - p_0)}$$

Where, $p_0 = P(HIV)$, $p_1 = P(Positive|HIV)$ and $p_2 = \left(Positive\middle|No\ HIV\right)$

Suppose p_0 was estimated from a sample of 100000, p_1 from a sample of 5000 and p_2 from a sample of 10000. Can we compute a confidence interval for θ ?

This was a problem that was presented to James Berger (a renowned Bayesian statistician) by a doctor Dr. Mossman. While there are some classical statistical approaches which can be used here, James Berger proposed the following approach which is computationally simple and which he demonstrated performs better than other classical methods.

Step 1: Assume Jeffrey's prior for each p_i and obtain the posterior distribution of each p_i

Step 2: Simulate a value for p_0, p_1, p_2 from the posterior and compute θ. This is one possible realization of θ

Step 3: Repeat step 2 , 10000 (a large number) of times to obtain a distribution for θ

Reference: Berger (2006), "The case for Objective Bayesian Analysis", Bayesian Analysis, No. 3, pages 385-402.

Problem 3: Calibration of vendor models for internal experience

A firm has developed a new account management product for corporations. Prior to developing the product, in a market research survey of 100 randomly chosen companies, 30 companies had expressed interest in adapting such a product. Since the firm started selling the product in the market, they have approached 20 companies. So far, 4 of them have bought the product. So the success rate has been 20%, way below the prior idea for success rate based on market survey.

Meeting the customer, carrying out multiple demos, following up, and convincing them to buy etc takes enormous effort and cost. Hence, they approached a consultant to help them identify and target customers who are more likely to buy such a product. The consultant provided a model that uses various company characteristics such as industry, size, product mix etc.. to predict such a likelihood. The consultant however cautioned them that this is a generic model and may need to be calibrated to the company's own experience.

The firm applied this model on the 4 companies which have bought the product and the 16 companies which had decided not to buy. Of the four that had bought the product, the model was able to correctly predict 3 companies as "likely to buy" and it predicted 1 as "not likely to buy". Of the 16 companies that did not buy, the model correctly predicted 10 companies as "not likely to buy" and the other 6 as "likely to buy".

Based on this information the firm wants to determine the expected success rate if it were to adapt this model and also the uncertainty around that estimate.