Name: Truong Nguyen
Group: DS-01

Assignment 2

Task 1:

1)

Assuming that the samples are independent and that the probability of a sample containing Giardia cysts is the same for all samples, X follows a binomial distribution with parameters n and θ. The probability mass function (PMF) of X is given by:

$$P(X = x \vee \theta) = (n\ choose\ x) \times \theta^{x} \times (1-\theta)^{(n-x)}$$

where (n choose x) is the binomial coefficient, which is the number of ways to choose x items from n items without regard to order and is given by:

$$\text{(n choose x)} = \frac{n!}{x! * (n-x)!}$$

where n! denotes the factorial of n, which is the product of all positive integers up to n.
Therefore, the conditional distribution of X given θ is a binomial distribution with parameters n and θ, and its PMF is given by the equation above.

2)

Mean and variance of Beta distribution are as follow:

$$\mu = \frac{\alpha}{\alpha+\beta} = 0.2$$

$$\sigma^{2} = \frac{\alpha\beta}{(\alpha+\beta)^{2}(\alpha+\beta+1)} = 0.16^{2}$$

solve the above system of equations we get:

$$\alpha = 1.05 \approx 1$$

$$\beta = 4.2 \approx 4$$

3)

The posterior distribution of θ given the observed data X can be obtained using Bayes' theorem:

$$P(\theta \vee X) = P(X \vee \theta) \times P(\theta)$$

where P(X | θ) is the likelihood of the data given θ, and P(θ) is the prior distribution of θ.

From part 1, we know that X follows a binomial distribution with parameters n = 116 and θ.
Therefore, the likelihood function is:

$$P(X = 17 \vee \theta) = (116\ choose\ 17) \times \theta^{17} \times (1-\theta)^{(116-17)} = (116\ choose\ 17) \times \theta^{17} \times (1-\theta)^{99}$$

From part 2, we know that the prior distribution of θ is a Beta distribution with parameters α = 1 and β = 4. Therefore, the prior distribution is:

$$P(\theta) \propto \frac{\theta^{(\alpha-1)} \times (1-\theta)^{(\beta-1)}}{Beta(\alpha,\beta)} \propto \frac{\theta^{(1-1)} \times (1-\theta)^{(4-1)}}{Beta(1,4)} \propto \frac{(1-\theta)^3}{Beta(1,4)}$$

Substituting these expressions into Bayes' theorem, we get:

$$P(\theta \vee X) \propto \frac{\theta^{17} * (1-\theta)^{102}}{Beta(18,103)}$$

The posterior parameters α’ and β’ therefore are:

$$\alpha' = 18$$

$$\beta' = 103$$

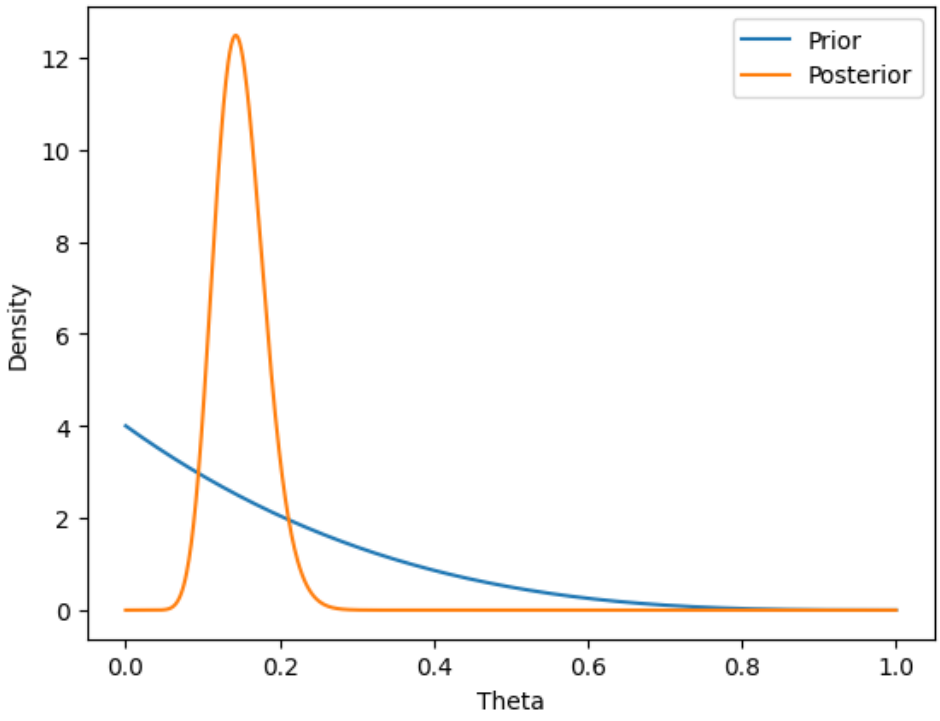
And so the posterior mean and standard deviation are:

$$\mu' = \frac{\alpha'}{\alpha' + \beta'} = 0.148$$

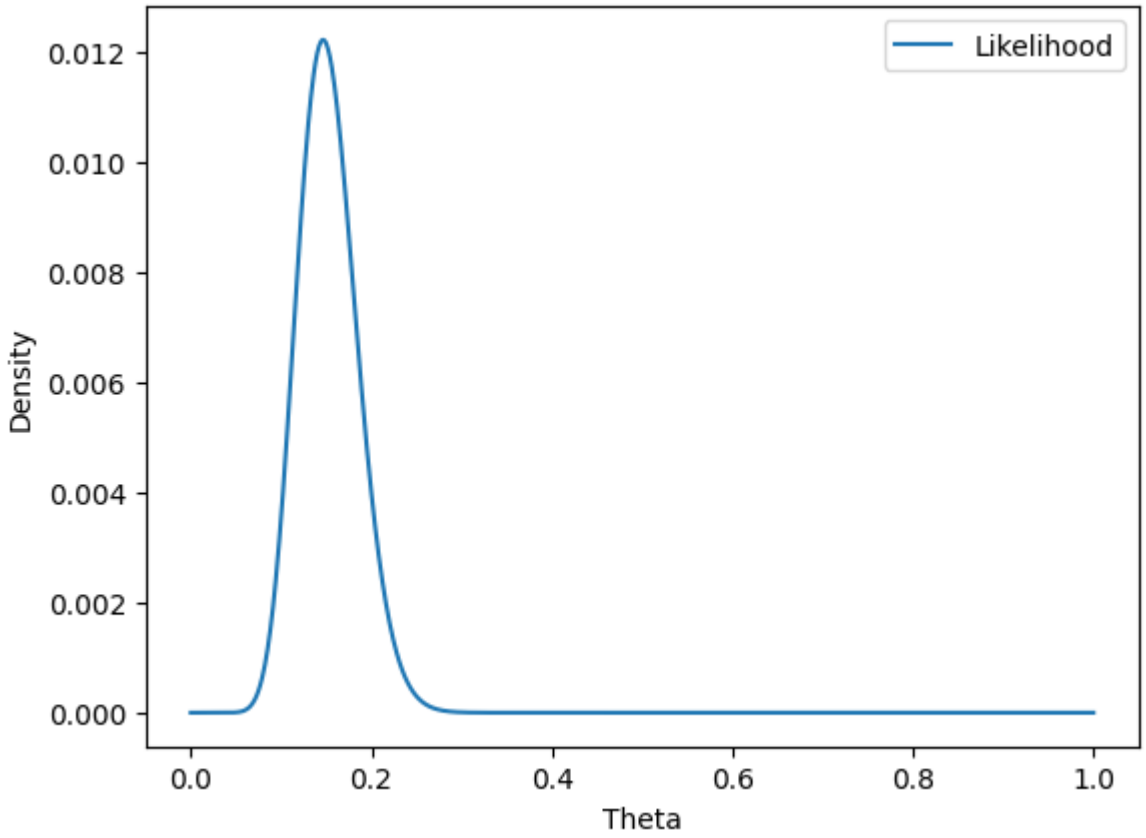
$$\sigma'^2 = \frac{\alpha'\beta'}{(\alpha' + \beta')^2(\alpha' + \beta' + 1)} = 0.00103$$

4) Plot the prior, posterior and normalized likelihood (code is included in Juputer Notebook attached)

```
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt
%matplotlib inline

# Define the parameters of the Beta prior distribution
alpha = 1
beta = 4

# Define the data
x = 17
n = 116

# Calculate the parameters of the posterior distribution
alpha_post = alpha + x
beta_post = beta + n - x

# Define the range of values for theta
theta_range = np.linspace(0, 1, 1000)

# Calculate the prior, posterior, and normalized likelihood
prior = stats.beta.pdf(theta_range, alpha, beta)
posterior = stats.beta.pdf(theta_range, alpha_post, beta_post)

# Plot the prior, posterior, and normalized likelihood
plt.plot(theta_range, prior, label='Prior')
plt.plot(theta_range, likelihood, label='Likelihood')
plt.xlabel('Theta')
plt.ylabel('Density')
plt.legend()
plt.show()
```



```
likelihood = stats.binom.pmf(x, n, theta_range)  # normalized by integrating over theta_range

# Normalize the likelihood to have integral of 1
sum_likelihood = np.sum(likelihood)
likelihood = likelihood / sum_likelihood

plt.plot(theta_range, likelihood, label='Likelihood')
plt.xlabel('Theta')
plt.ylabel('Density')
plt.legend()
plt.show()
```



5) To find the posterior probability that θ < 0.1, we need to integrate the posterior distribution from 0 to 0.1. We can do this using the cumulative distribution function (CDF) of the beta distribution, which is available in many software libraries including SciPy in Python.

```
import scipy.stats as stats

# Calculate the posterior probability that theta < 0.1
post_prob = stats.beta.cdf(0.1, alpha_post, beta_post)

print("The posterior probability that theta < 0.1 is:", post_prob)
```

And the result is: 0.053094376993042654

6) To find a central 95% posterior credible interval for θ, we can use the quantile function (inverse CDF) of the beta distribution. This function tells us the value of θ for a given probability. We can use it to find the lower and upper bounds of the interval that contains the central 95% of the posterior distribution.

```
# Calculate the central 95% posterior credible interval for theta
lower = stats.beta.ppf(0.025, alpha_post, beta_post)
upper = stats.beta.ppf(0.975, alpha_post, beta_post)

print("The central 95% posterior credible interval for theta is: [{:.3f}, {:.3f}]".format(lower, upper))
```

And the result is: [0.091, 0.217]

Task 2:

Refer to notebook attached