Name: Hany Hamed,
Group Number: BS18-Robotics
Course: Statistical Techniques for Data Science and Robotics (STDSR) - Spring 2022,
Assignment 2 Code

Contents

1 Preliminaries

The assignment is based on Bayesian inference.

$P(\theta|X) = \frac{P(X|\theta)\cdot P(\theta)}{P(X)}$

$P(\theta|X)$ is the posterior probability

$P(X|\theta)$ is the likelihood probability

$P(\theta)$ is the prior probability

$\frac{P(X|\theta)}{P(X)}$ is the normalized likelihood

2 Question 1

Find the conditional distibution of X (the number of samples containing Giardia cysts) given θ

In order to compute the conditional probability $P(X|\theta)$, that means that given θ what is the probability of the number of samples containing Giardia cysts.

By choosing the binomial distribution as following $Bin(n, \theta)$ such that θ is the probability that the sample contains Giardia cysts. Thus, binomial distribution is the best selection for modeling the conditional probability

$P(X = x|\theta, N = n) \sim Bin(n, \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$

3 Question 2

Find α and β for the Beta distribution for the prior distribution of θ

Source: Beta distribution wiki

$\mathbb{E}[\theta] = 0.2$

$\sigma(\theta) = 0.16$

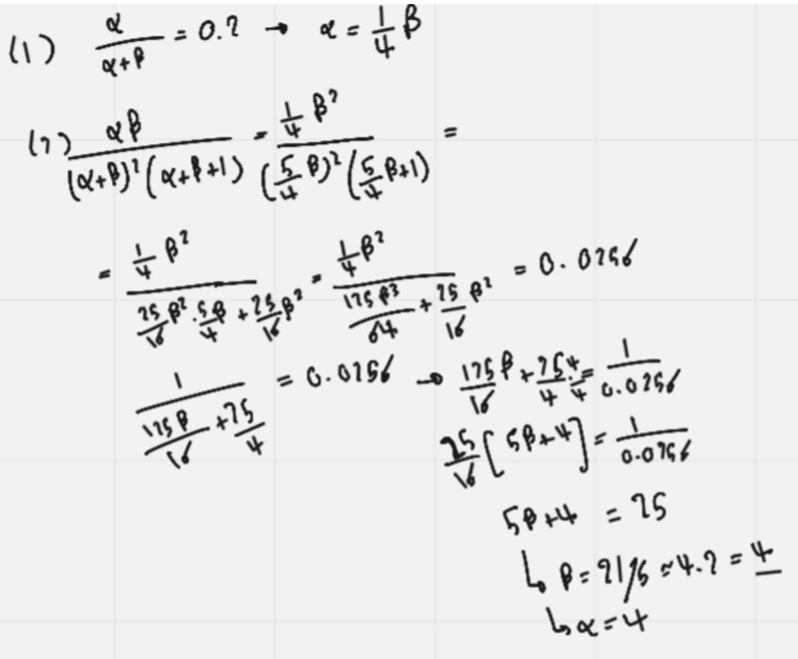
$Var[\theta] = \sigma^2 = 0.0256$

$\theta \sim Beta(\alpha, \beta)$ $P(\theta) = f(\theta|\alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$ such that $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is constant regardless the theta

$\mathbb{E}[Beta(\alpha, \beta)] = \frac{\alpha}{\alpha+\beta} = 0.2$

$Var[Beta(\alpha, \beta)] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = 0.0256$

Solve two equations with two unknowns and get the parameters



$$(1)\ \frac{\alpha}{\alpha+\beta} = 0.2 \rightarrow \alpha = \frac{1}{4}\beta$$

$$(2)\ \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\frac{1}{4}\beta^2}{(\frac{5}{4}\beta)^2(\frac{5}{4}\beta+1)} =$$

$$= \frac{\frac{1}{4}\beta^2}{\frac{25}{16}\beta^2 \cdot \frac{5\beta}{4} + \frac{25}{16}\beta^2} = \frac{\frac{1}{4}\beta^2}{\frac{125\beta^3}{64} + \frac{25}{16}\beta^2} = 0.0256$$

$$\frac{1}{\frac{125\beta}{16} + \frac{25}{4}} = 0.0256 \rightarrow \frac{125\beta}{16} + \frac{25}{4}\cdot\frac{4}{4} = \frac{1}{0.0256}$$

$$\frac{25}{16}\left[5\beta + 4\right] = \frac{1}{0.0256}$$

$$5\beta + 4 = 25$$

$$\hookrightarrow \beta = 21/5 \approx 4.2 = \underline{4}$$

$$\hookrightarrow \alpha = 4$$

Figure 1: Solving the two equations

$\alpha = 1, \beta = 4$

4 Question 3

Find the posterior distribution of θ and its mean and standard deviation

Source: Bayes' theorem with beta distribution as prior

We are going to use the Bayes' theorem, such that $n = 116,\ x = 17$

$P(\theta|X = x, N = n) = \frac{P(X=x|\theta)\cdot P(\theta)}{P(X=x)} \sim \frac{f_{Bin(n,\theta)}(x)\cdot f_{Beta(1,4)}(\theta)}{P(X=x)}$

$P(\theta|X = x, N = n) = \frac{(\binom{n}{x}\theta^x(1-\theta)^{n-x})\cdot(\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)})}{P(X=x)}$

$P(\theta|X = 17, N = 116) = \frac{(\binom{116}{17}\theta^{17}(1-\theta)^{99})\cdot(\frac{(1-\theta)^3}{B(1,4)})}{P(X=17)}$

Such that $P(X = x)$ is the marginal probability which is formalized as following given that θ is continuous $P(X = 17) = \int_0^1 P(X = 17|\theta, 116)P(\theta)d\theta$ which is a constant

Next step is to combine all the constants togther and the same for the variable terms as following:

$P(\theta|X = x, N = n) = \frac{\binom{116}{17}}{B(1,4)P(X=17)}[\theta^{17}(1-\theta)^{102}]$

As the first term is just a constant, $P(\theta|X = x, N = n) \propto \theta^{17}(1-\theta)^{102}$

$\theta^{17}(1-\theta)^{102} = \theta^{18-1}(1-\theta)^{103-1}$

And we are able using the normalized factor in the beta function for the posterior distribution as following: $P(\theta|X = x, N = n) = \frac{1}{B(18,103)}\theta^{18-1}(1-\theta)^{103-1} = Beta(18, 103)$

Therefore, posterior distribution of $\theta = \theta|X \sim Beta(18, 103)$

$\mathbb{E}[\theta|X] = \frac{18}{18+103} = \frac{18}{121} = 0.148760331$

$Var(\theta|X) = \frac{18*103}{121^2*122} = 0.00103795651 = \sigma^2$

Therefore, the standard deviation $\sigma = \sqrt{Var} = 0.0322173324$

5 Question 4

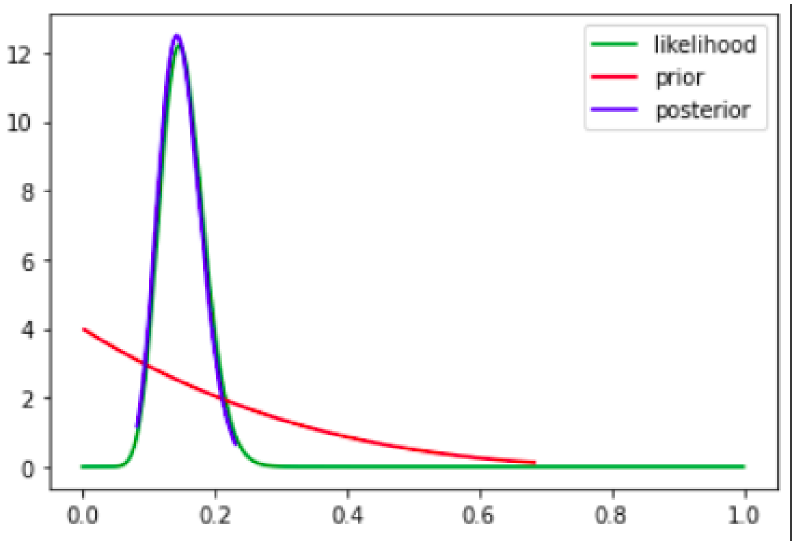


Figure 2: Prior, posterior and normalized likelihood

6 Question 5

CDF of posterior

$P(\theta < 0.1|x, n) = CDF(Beta(18, 103), \theta < 0.1) = 0.05309437699304309$

```
from scipy.stats import beta
rv = beta(18, 103)
rv.cdf(0.1)
```

7 Question 6

Source: Stackoverflow

Central confident interval using ppf from python is [0.09138957252823003, 0.21710689824337648]

```
from scipy.stats import beta
[beta.ppf(0.025, 18, 103), beta.ppf(0.975, 18, 103)]
```

8 Question 7

Source: Stackexchange Using the hints:

Construct a density function of beta distribution

$n^* = 50$, then the posterior predictive probability that x = 5 = y of these contain Giardia cysts from the source $P(x = 5, previousexperiment)$ =original pdf times the posterior $P(X = y)P(\theta|x) = Bin(n^*, \theta)P(\theta|x)$

Take an integral substituting needed values:

$$P = \int_0^1 \binom{n^*}{y} \theta^y (1-\theta)^{(n^*-y)} \frac{\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}}{\frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)}} d\theta =$$

$$P = \int_0^1 \binom{n^*}{y} \theta^y (1-\theta)^{(n^*-y)} \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)} \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} d\theta =$$

$$P = \binom{n^*}{y} \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)} \int_0^1 \theta^y (1-\theta)^{(n^*-y)} \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} d\theta =$$

$$P = \binom{n^*}{y} \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)} \int_0^1 \theta^{y+\alpha+x-1}(1-\theta)^{(n-y+n^*-x+\beta-1)} d\theta$$

Using python, such that $\alpha = 1, \beta = 4, n^* = 50, x = 17, y = 5, n = 116$

Thus, $P = 0.11087022918190788$

```
import operator as op
from functools import reduce
from scipy.special import gamma, factorial
import scipy.integrate as integrate


def ncr(n, r):
    # Source: https://stackoverflow.com/questions/4941753/is-there-a-math-ncr-function-in-pyt
    r = min(r, n-r)
    numer = reduce(op.mul, range(n, n-r, -1), 1)
    denom = reduce(op.mul, range(1, r+1), 1)
    return numer // denom  # or / in Python 2

x3 = np.linspace(0, 1, 1000)
alpha = 1
beta = 4
y = 5
n_ = 50
x = 17
n = 116
int_val = integrate.quad(lambda theta: theta**(y+alpha-1+x)*(1-theta)**(n-y+n_-x+beta-1), 0, 1
val = int_val * gamma(alpha+beta+n)/(gamma(alpha+x)*gamma(beta+n-x)) * ncr(n_,y)
print(val)
```

9 Question 8

Hypothesis testing

Assuming that the null hypothesis is true, we need to determine the p-value that represents observing the experiment (finding at least 17 sample out of 116 contains the cysts) such that $\theta = 0.2$ (minimum) if θ increased, the number of samples will increase as well

Therefore, we need to compute $p = P(X \leq 17|\theta = 0.2) = CDF(Bin(17, 116|\theta = 0.2) = 0.08947619410317142$

As $p = 0.089 > 0.05$ such that 0.05 is the significance level. Therefore, we fail to reject the null hypothesis (H_0)

```
from scipy.stats import binom
n = 116
p =  0.2
x = np.arange(binom.ppf(0.01, n, p),
              binom.ppf(0.99, n, p))
rv = binom(n,p)
p = rv.cdf(17)
print(p)
```

10 Question 9

Hypothesis testing (Bayesian point of view)

Here, we are dealing with it differently we transform it into probability, we calculate the probability of the null hypothesis given that the number of samples = 17

$P(H_0|X = 17) = P(\theta \geq 0.2|X = 17) = 1 - P(\theta \leq |X = 17) = 1 - CDF(Beta(18, 103|\theta = 0.2) = 0.06472685723749172$

```
from scipy.stats import binom
a,b = 18,103
rv = beta(a, b)
p = rv.cdf(0.2)
print(p)
print(1-p)
```

Thus, the true probability that a one-liter water sample from this type of site contains Giardia cysts (θ) is larger than or equal to 0.2 given the observation of 17 samples (The null hypothesis) has a probability 0.06473, and the alternative hypothesis has a probability = 0.9353