Stat 474/574 Homework 1 - Due Wed., February 8

Kaushiki Singh[1]

Unless instructed otherwise, please return a printed version or email a pdf version of your assignment during the class period.

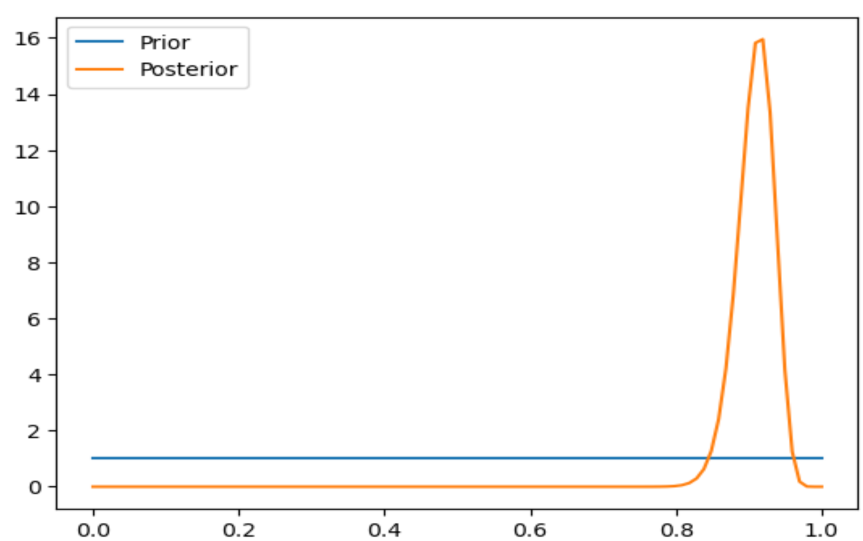
The codes of the following assignment have been written in Python

Problem 1: In the General Social Survey, n = 129 females aged 65 and older were asked whether they were generally happy. Let $Y_i = 1$ if the i^{th} woman reported being generally happy and let $Y_i = 0$ otherwise. One hundred and eighteen women (91%) reported being generally happy and 11 women (9%) said that they were not generally happy.

1.1 Consider first a uniform prior distribution for θ, the probability that a woman will report being generally happy. Simulate 1000 values from the posterior distribution of θ and summarize results.

Summary Statistics is as below:

```
Mean: 0.9091
Median: 0.9109
Mode: 0.808
95% Credible Interval: [0.8536 0.9514]
```

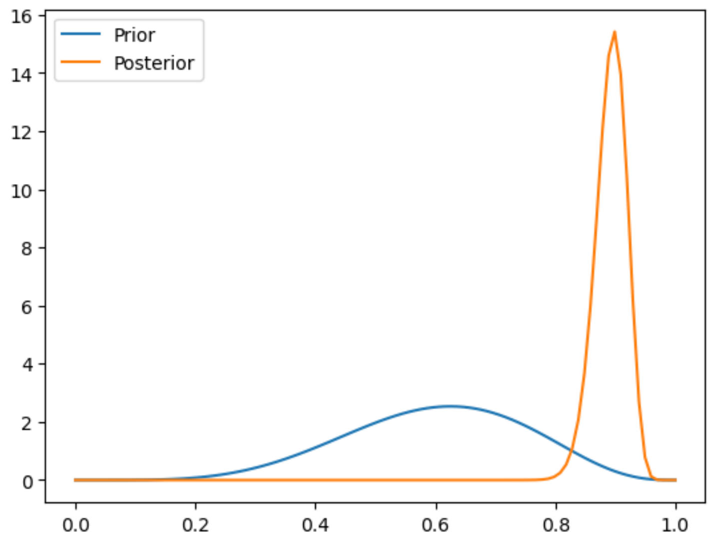


[1] All codes and figures have been generated in python. The python jupyter notebook file has been submitted separately.

1.2 Consider now a conjugate beta prior, such that $E(\theta) \approx 0.6$. You are not very confident that 0.6 is a good guess for θ. Again simulate 1000 values from the posterior distribution of θ. Summarize your results.

Summary Statistics is as below:

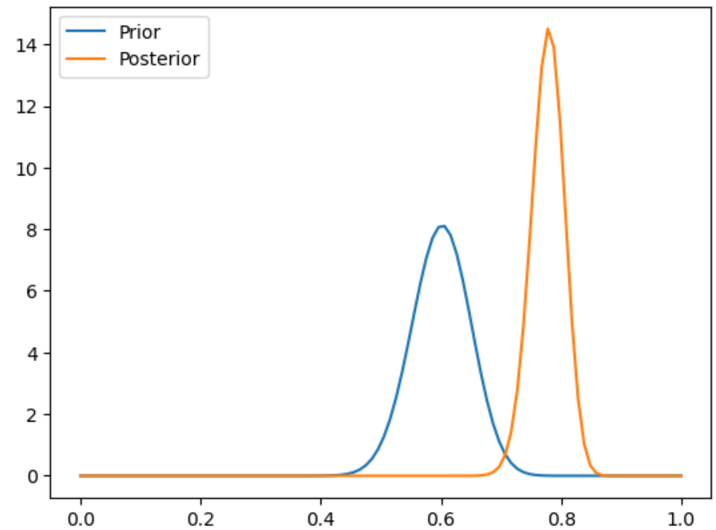
```
Mean: 0.8922
Median: 0.894
Mode: 0.7862
95% Credible Interval: [0.8356 0.9379]
```



1.3 Suppose that some additional information has become available, and you are now quite comfortable about 0.6 as a guess for θ. Modify your prior accordingly and repeat the analysis. Comment on what you observe.

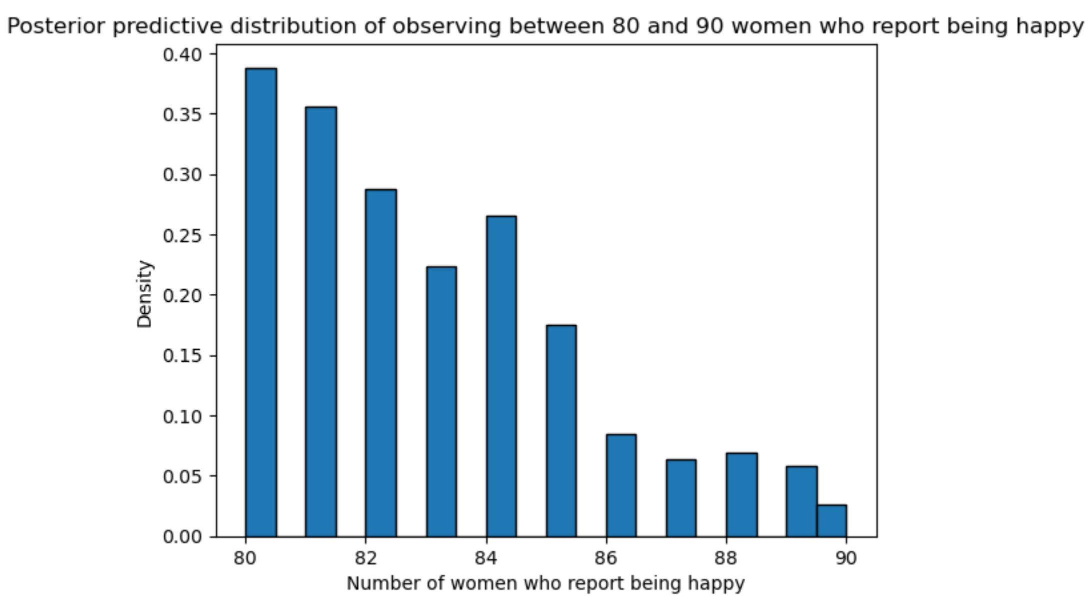
Modifying the prior as Beta(60,40):

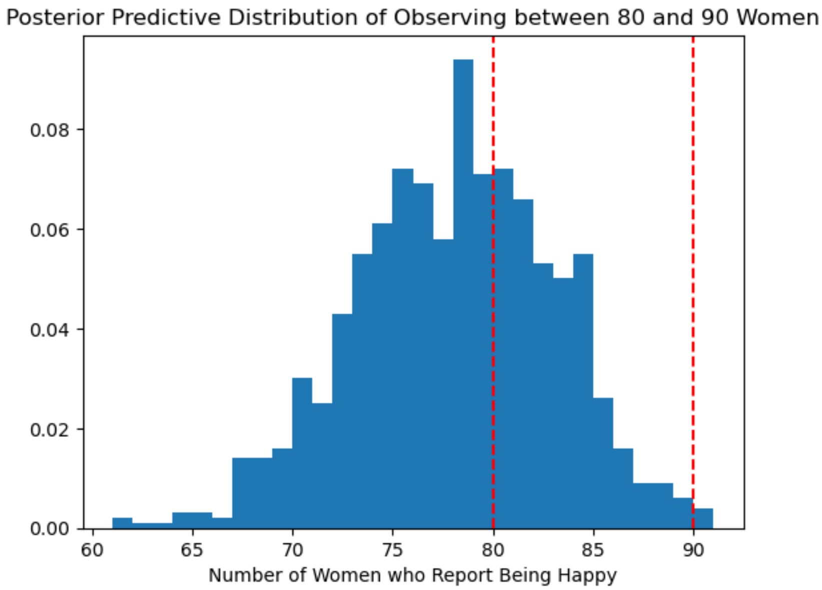
```
Mean: 0.7772
Median: 0.7783
Mode: 0.683
95% Credible Interval: [0.7213 0.8287]
```



1.4 Using the posterior distribution in 1.3, compute the posterior predictive distribution of observing between 80 and 90 women who report being happy if an additional 100 women were to be interviewed.

```
Probability of observing between 80 and 90 women who report being ha
ppy: 0.371
```



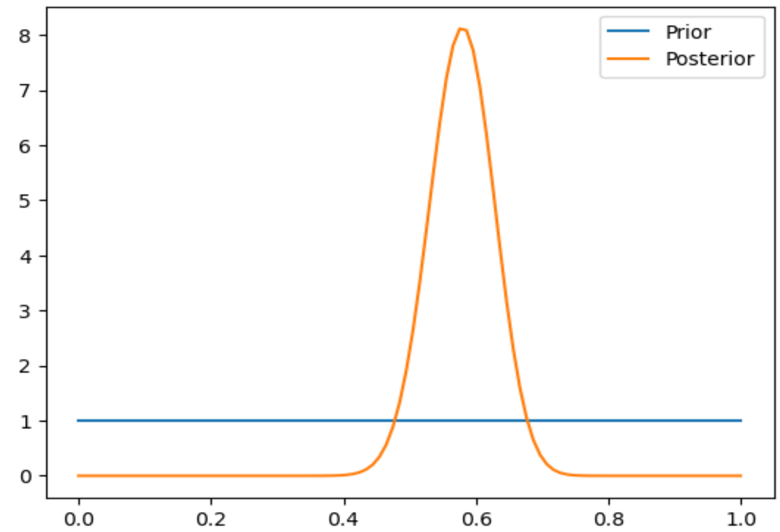


The red vertical lines in the above graph shows the distribution of number of women between 80 and 90 who report being happy.

Problem 2: (To do this problem, you can use the function BernBeta in the LearnBayes package). An election is approaching and the latest poll of 100 people shows that 58 prefer candidate A and 42 prefer candidate B. You are curious about the chances that candidate A will prevail.

2.1 Before you read the results of the poll, you had not thought much about the chances that candidate A would win (or a priori, you had no information about the probability of success for candidate A). What is the 95% Highest Posterior Density set (HPD set) on the probability that A will win after you have read the results of the poll?

Starting with beta(1,1), the prior and posterior distributions are as below:




```
Having calculated the posterior, the HPD set is as below:
95% HPD Set: [0.4818  0.6721]
```

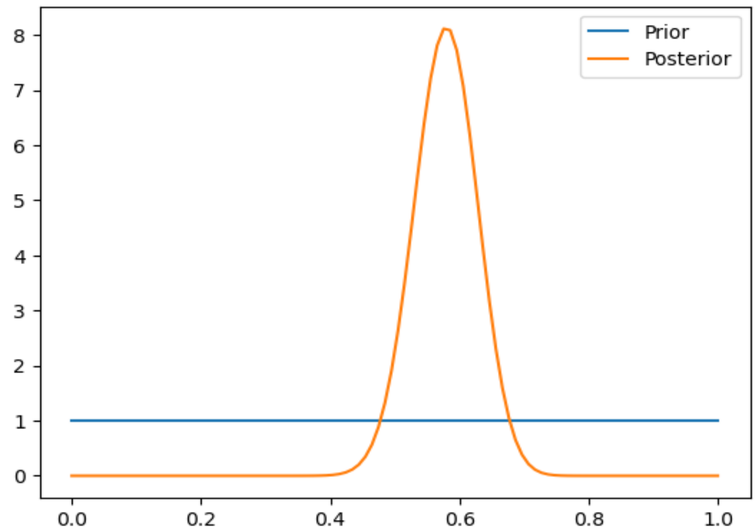
2.2 Using simulation, compute a 95% credible set for the probability that A will win. How do the two intervals compare?

```
The credible set is as below:
95% Credible Interval: [0.4781 0.6678]
```

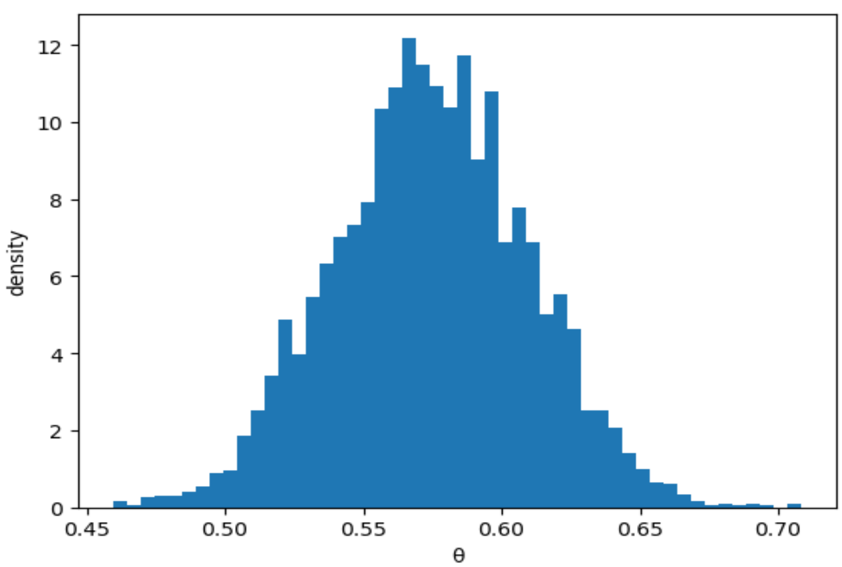
On comparing the 95% Highest Posterior Density set (HPD set) and the 95% credible set for the probability that A will win, we find that both the sets are pretty much similar.

2.3 Just to confirm, you poll a second set of 100 randomly chosen persons from the same population. You find that in this second set, 57 prefer candidate A and the rest prefer B. How does the 95% HPD set you computed in part a change?

The prior and posterior distributions are as below:




```
Having calculated the posterior in this case, the HPD set is as belo
w:
```
`95% HPD Set: [0.5056 0.6415]`
`95% Credible Interval: [0.5096 0.6413]`

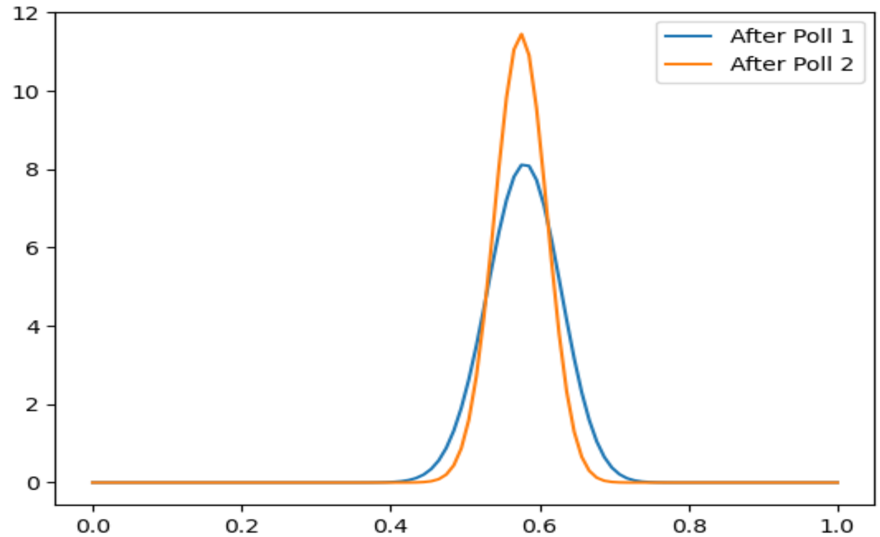


The HPD set now becomes narrower than before i.e. in comparison to part (a).

2.4 Based on the two surveys, is it reasonable to conclude that the population is evenly divided in terms of preferences for candidates A and B?

Looking at HPD and the credible sets for the two surveys, in case the sets for both surveys overlap, we can say that the population is evenly divided in terms of preferences for candidates A and B. Therefore, on comparing 2.1-2.2 with 2.3 we can see that there is substantial overlap between the HPD and the credible sets and therefore, one can say that the population is more or less evenly divided in terms of preferences for candidates A and

B. The posteriors after poll 1 and after poll 2 are plotted as below:



The overlap of the two intervals suggests that there is a significant overlap in the probability of candidate A winning.

2.5 Based on the two surveys, what is the probability that candidate A will win by at least a 10% margin?

Based on the two surveys, what is the probability that candidate A will win by at least a 10% margin is **0.7582.**

Problem 3: Suppose that two different persons (Joe and Sam) are interested in estimating the proportion θ of students at a college who commute to school. Joe uses the discrete prior shown in Table 1, while Sam decides instead to use a Beta(3,2) prior for the proportion of commuters θ.

θ	0.1	0.2	0.3	0.4	0.5
$p(\theta)$	0.5	0.2	0.2	0.05	0.05

Table 1: *Joe's discrete prior for* θ

3.a Use R to compute the mean and standard deviation of θ for Joe's prior and for Sam's prior. Do Joe and Sam have similar prior beliefs about the proportion of commuters?

Like for other questions, Python has been used for this question as well:

Summarizing priors of Joe and Sam as below:

```
Joe's prior mean:  0.195
Joe's prior standard deviation:  0.1161
Sam's prior mean:  0.6
Sam's prior standard deviation:  0.04
```

Thus, the priors are different for both Joe and Sam.

3.b Suppose that you carry out a survey of the students in the college and find that out of 100 student surveyed, 30 of them commute. Use the function pdisc in LearnBayes to find the posterior probability under **each of the priors** that no more than 20% of the students in the college are commuters.

The posterior probabilities under Joe and Sam's priors are as below:

```
P(θ <= 0.2 | Data) under Joe's prior:  0.0549
P(θ <= 0.2 | Data) under Sam's prior:  0.0252
```

3.c Find a 90% credible set for θ under each of the two models.

```
90% credible set for Joe's prior:: [0.2, 0.3]
90% credible set for Sam's prior: [0.2422, 0.3904]
```