Machine Learning Problem Assignment 2

Swapnil Asawa, swa12@pitt.edu

January 30, 2019

1 Problem 1

In this problem we study the influence of the sample size on the estimate of the mean.The data for this experiment are in file mean study data.txt in the homework assignment folder. The data were generated from the normal distribution with mean=15 and standard deviation=5.

1.1

Load the data in the mean study data.txt Calculate and report the mean and standard deviation of the data. Compare them to the true mean and std above.

main=importdata('main_study_data.txt');
standarddev=std(main)
standarddev = 5.0279
mean=mean(main)
mean = 15.0415
The actual mean of the dataset as calculated in part 1 is 15.0415 but the dataset was sampled from a normal distribution of mean 15. As the dataset is made up of finite random samples of the the source distribution, it might not fit 100percent the model from which it cam from with slight irregularities, it could have a bit difference in means from the actual distribution, which is 0.0415.

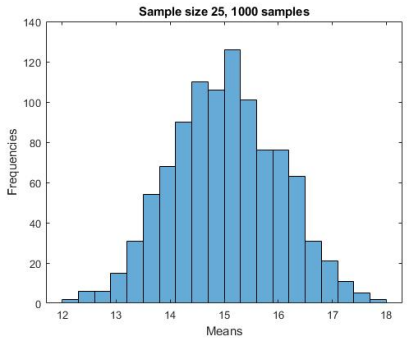
1.2

Write (and submit) a function [newdata] = subsample(data, k) that randomly selects k instances from the data in the mean study data.txt.

function [newdata]=subsample(data,k)
d=length(data);
indices = randperm(d);
indices=indices(1:k);
y= data(indices)
end

1.3

Use the above function to randomly generate 1000 subsamples of the data of size 25. For each subsample calculate its mean and save the results in the vector of 1000 means. Plot a histogram of 1000 mean values using 20 bins.




```
for i=(1:1000)
Mean25(i)=mean(subsample(main,25));
end
histogram(Mean25,20)
mean(Mean25)
ans = 15.0341
std(Mean25)
ans = 1.0041
```

1.4

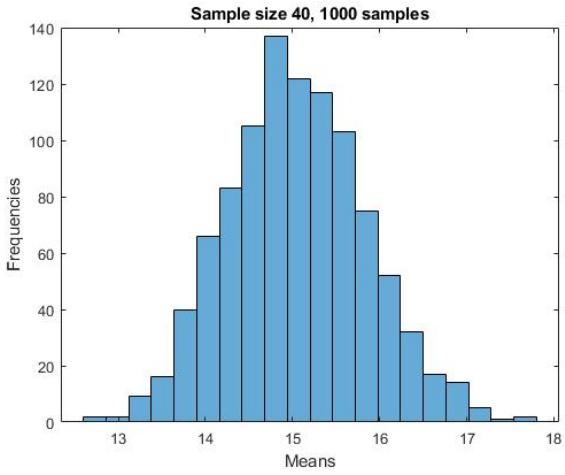
Include the histogram in your report. Analyze and the means calculated on 1000 subsamples of size 25 and compare them to the true mean and the mean that was calculated in step 1 on all examples in the dataset. Report your observations.

True mean= 15
Mean of dataset(Part 1)= 15.0415
Mean of means of samples= 15.0341
Figure can be seen at the top of the page. As we can see, the means of different samples vary from 12 to 18 and centered towards 15.0341. The distribution of the means is normal and has center at 15.0341 and std= 1.0041. So approx, 68.8% means are from range 15.0341+-1.0041. This mean is greater than the true mean and lesser than the dataset mean. In other words, sampled mean is more accurate than the dataset mean in predicting true estimate of the mean of the underlying generating function.

1.5

Repeat steps form part 3 but now generate 1000 subsamples of size 40. Include the histogram in the report and compare it to the histogram generated in part 4 for subsamples of size 25, and to the mean of the original data. What are the differences? What conclusions can you make by comparing the means for subsamples of size 25 and 40.

```
for i=(1:1000)
Mean40(i)=mean(subsample(main,40));
end
histogram(Mean25,20)
hold on
```

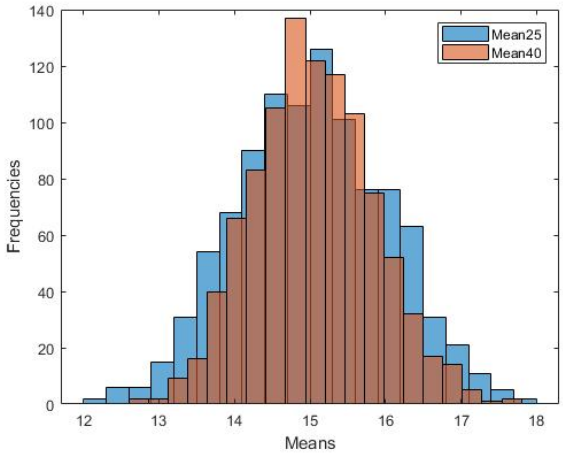


histogram(Mean40,20)
hold off
mean(Mean25)
ans = 15.0341
mean(Mean40)
ans =15.0561
std(Mean25)
ans = 1.0041
std(Mean40)
ans =0.7923
The distribution of means with sample size 40 is less spread out as compared to sample size 25. This is because, the more the samples size, the better chances of is their average estimating the actual mean value with lesser error. So, when the sample size is 40, sample means are more near the actual mean, and hence the distribution is less spread out with variance =0.7923 as compared to 1.0041 of sample size 25. Comparing the mean is giving me non-intuitive result. Mean of 40sized 1000 samples is not as good as mean of 25sized 1000 samples in this random sampling. But every time I repeat the for loop, I get different mean values for both sizes. Sampling mean of size 40 is expected to be better than of 25. Variance is much lesser with sample size of 40.

1.6

Take first 25 examples from the original data in the mean study data.txt and calculate their mean. Use the function t-test to calculate and report the 0.95 confidence interval for the mean estimates. Does the true mean value fall into the 0.95 confidence interval?

mean1st25=mean(main(1:25))
mean1st25 = 14.5652
[h,p,ci,stats] = ttest(main(1:25), 15, 'Alpha',0.05)
h =0



p = 0.6465
ci = 12.6329 16.4974
stats = struct with fields:
tstat: -0.4645
df: 24
sd: 4.6811

With confidence level 0.95, The true mean is between the range 12.6329 to 16.4974, which it is.

2

2.1

Please write and submit the function: [train test] = kfold crossvalidation(data, k, m) that takes the data, k (the number of folds) and m (the target fold) as inputs, and returns the training and testing data sets, such that the testing set corresponds to m-th fold under the k-th fold cross-validation scheme. To implement the procedure please place the folds over indexes of the data, by assuring that each fold has equal number of entries that do not overlap. If this is not possible, the fold sizes (number of instances in each fold) should differ by at most one. The file should be named kfold_crossvalidation.m.

```
function [train, test]=kfold_crossvalidation(data, k, m)
    subSize=length(data)/k;
    modSize=mod(length(data),k);
    subSets=[];
    for i=(1:modSize)
    subSets=[subSets;((i-1)*subSize+i:i*subSize+i)];
    end
    for j=(modSize+1:k)
    subSets=[subSets;((j-1)*subSize+modSize+1 :j*subSize+modSize)];
    end
```

```
    trainSets=subSets;
    trainSets(m,:)=[];
    trainind=trainSets ;
    trainind=reshape(trainind,1,[]);
    testind=subSets(m,:);
    train=data(trainind);
    test=data(testind);
end
```

2.2

Run/test your function on data in the file resampling data.txt. More specifically, run your kfold crossvalidation function on all data in the file by setting k (number of folds) to 10 and by varying the test fold index (parameter m) from 1, to 10. For each test data (generated for the different value of m) that were returned by your function calculate the mean and std and report them

```
resampling=importdata('resampling_data.txt')
for i=(1:10)
    [train, test]= kfold_crossvalidation(resampling,10,i);
    MeanSampling(i)=mean(test);
    StdSampling(i)=std(test);
end
MeanSampling
MeanSampling =
    Columns 1 through 7
    3.9938 1.8277 2.1446 1.7954 2.0849 1.7627 2.1046
    Columns 8 through 10
    1.0343 1.5837 2.4246
StdSampling
StdSampling =
    Columns 1 through 7
    4.4391 3.6272 2.3504 3.1598 3.3794 3.2640 3.4622
    Columns 8 through 10
    2.5800 3.4184 2.2831
```

3

Part a. Assume you have 2 fair dice. What are the probabilities associated with the different outcomes that are obtained by summing together the numbers on the two dice?
Part b. Calculate the expected value of the outcome for the 2 fair dice roll experiment.
Part c. Assume you play the two dice game from part a. 5 times. What is the probability, we never see the outcome of 4? What is the probability we see odd-sum outcomes in all 5 trials.

Problem 3

Part a 2 fair dice

$S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

Total variations of outcomes = 36

① $P(2) = \frac{1}{36} \times 1$ because $\{1,1\}$ is the only way sum could be 1

$P(2) = \frac{1}{36}$

② $E(\text{Sum} = 3) = \{\{2,1\}, \{1,2\}\}$

$P(3) = \frac{1}{6} \times \frac{1}{6} + \frac{1}{6} \times \frac{1}{6} = \frac{2}{36} = \frac{1}{18}$

③ $E(\text{Sum} = 4) = \{\{1,3\}, \{2,2\}, \{3,1\}\}$

$P(4) = \frac{3}{36}$.

④ $E\{\text{Sum} = 5\} = \{\{1,4\}, \{2,3\}, \{3,2\}, \{4,1\}\} = \frac{4}{36}$

$P(5) = 4/36$

⑤ $E(\text{Sum} = 6) = \{(1,5)(2,4)(3,3)(4,2)(5,1)\}$

$P(6) = 5/36$

⑥ $E(\text{Sum} = 7) = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$

$P(7) = 6/36$

⑦ E(Sum = 8) = {(2,6) (3,5) (4,4) (5,3) (6,2)}

P(8) = 5/36

⑧ E(Sum = 9) = {(3,6), (4,5), (5,4), (6,3)}

P(9) = 4/36

⑨ E(Sum = 10) = {(4,6), (5,5), (6,4)}

P(10) = 3/36

⑩ E(Sum = 11) = {(5,6) (6,5)}

P(11) = 2/36

⑪ E(Sum = 12) = {(6,6)} = 1/36

→ P(12) = 1/36.

Part b. $E(X) = \sum_{x=2}^{12} x\,p(x)$

$$= \left[2\times\frac{1}{36} + 3\times\frac{2}{36} + 4\times\frac{3}{36} + 5\times\frac{4}{36} + 6\times\frac{5}{36} + 6\times\frac{7}{36} + 8\times\frac{5}{36} + 9\times\frac{4}{36} + 10\times\frac{3}{36} + 11\times\frac{2}{36} + 12\times\frac{1}{36}\right]$$

$$= \frac{1}{18}\left[1 + 3 + 6 + 10 + 15 + 21 + 20 + 18 + 15 + 11 + 6\right]$$

$$= \frac{1}{18}\left[126\right] = 7$$

$$E(\text{Sum of 2 fair dice}) = 7$$

Part c) $P(\text{not seeing outcome 4 in 5 rolls})$

$= ?$

$= P(\bar{4} \text{ in game } 1) \times P(\bar{4} \text{ in } g_2)\, P(\bar{4} \text{ in } G_3)\, P(\bar{4} \text{ in } G_4) \times P(\bar{4} \text{ in } G_5)$

$\because$ all games are independent

$$P(\bar{4} \text{ in 1 game}) = 1 - \frac{3}{36}$$

$$= \frac{33}{36} = \frac{11}{12}.$$

$$\Rightarrow P(\bar{4} \text{ in all 5 games}) = \left(\frac{11}{12}\right)^5 = 0.6472$$

P(odd sum in all)
= P(odd sum in 1)^5
∵ independent trials
⇒ P(odd sum in 1 trial)
= P(3, 5, 7, 9, 11) ~~3, 5~~

~~= P(3)~~
= P(3) + P(5) + P(7) + P(9) + P(11)
(mutually exclusive)

$= \frac{2}{36} + \frac{4}{36} + \frac{6}{36} + \frac{4}{36} + \frac{2}{36}$

$= \frac{1}{36}[18] = \boxed{\frac{1}{2}}$

⇒ P(odd sum in all)

$= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$

$= 2^{-5} = \boxed{0.03125}$ Ans

.jpg

4

A pharmaceutical company has developed a nearly accurate test for the disease A. The accuracy of the test is 99%, that is, with probability 0.99 it gives the correct result (the same probability for disease-positive-test and no-disease-negative-testcombinations are assumed) and only in 1% of tested cases (probability 0.01) the result is wrong. The incidence of the disease in the population is 0.01% (probability 0.0001). Compute the probability that somebody from wide population who has tested positive indeed suffers from the disease. Would you recommend the test to be widely adopted?

Prob 4) $P(\text{Disease}) = 0.01\%$

$= 0.0001 \Rightarrow P(\bar{D}) = 0.9999$

$P(\text{Disease} / +ve) = ?$

$P(+ve / \text{Disease}) = P(+/D)$

$= 0.99$

$P(-ve / \text{Not Disease}) = P(-/\bar{D})$

$= 0.99$

$P(D/+ve) = \frac{P(D \cap +)}{P(+)}$

$= \frac{P(+/D)\,P(D)}{P(+)}$

$= \frac{P(+/D)\,P(D)}{P(+/D)P(D) + P(+/\bar{D})P(\bar{D})}$

lets say D^γ

$P(+/\bar{D}) = 1 - P(-/\bar{D})$

$\Rightarrow P(\bar{D}) = 0.9999$

$\Rightarrow D^\gamma = 0.99 \times 0.0001 + 0.01 \times 0.9999$

$D^{\gamma} = 9.9 \times 10^{-5} + 9999 \times 10^{-5}$

$= 1009.8 \times 10^{-5}$

$N^{\gamma} = 0.99 \times 0.0001$

$= 9.9 \times 10^{-5}$

$\Rightarrow P(D/+) = \frac{9.9 \times 10^{-5}}{1009.8 \times 10^{-5}}$

$\Rightarrow P(D/+) = 0.0098$

No, I would not recommend the test to be widely adopted because even when it gives +ve result for the disease, the chance of having that disease is as low as 0.01 approximately which is not reasonable evidence of the disease.

5

Assume we have conducted a coin toss experiment with 100 coin flips. The results of the experiment are in file 'coin.txt' where 1 means a head and 0 means a tail. Assume that represents the probability of observing a head.

5.1

What is the ML estimate of thetha?

```
tbl=tabulate(coint)
N1=tbl(2,2)
N2=tbl(1,2)
MLthetha=N1/(N1+N2)
tbl =
0 35 35
1 65 65
N1 = 65
N2 =35
MLthetha =0.6500
```

5.2

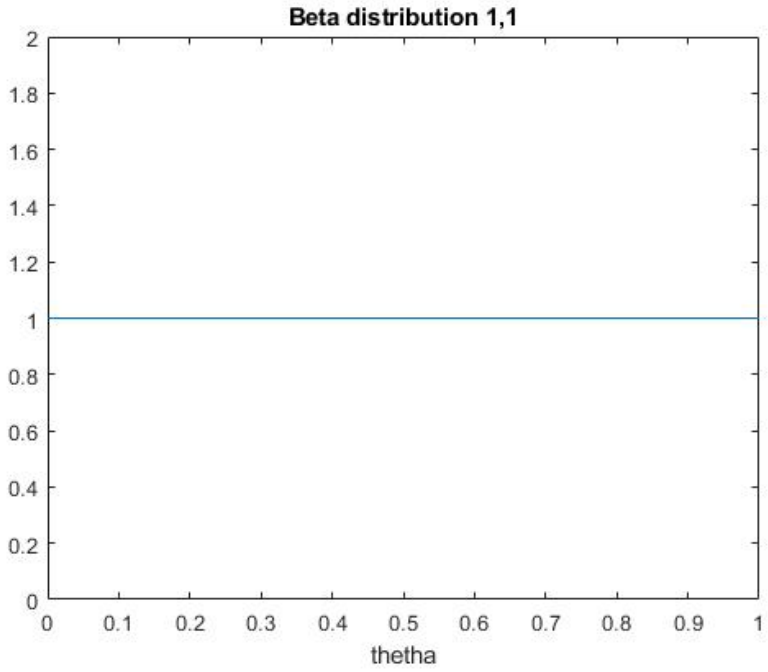
Assume the prior on is defined by a Beta distribution Beta(—1, 1). Plot and report both the prior and the posterior distributions on .

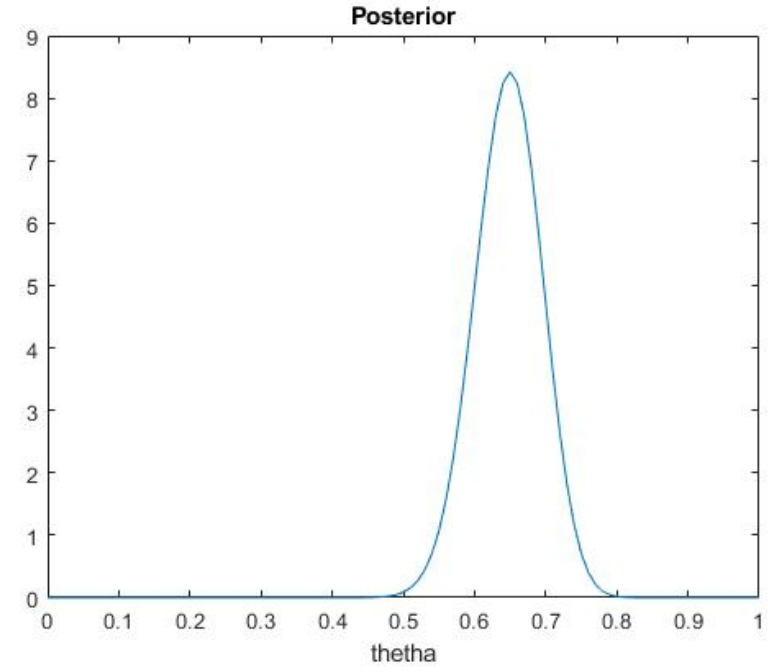
```
Xaxis=0:0.01:1
y=betapdf(Xaxis,1,1)
figure
plot(Xaxis,y)
posty=betapdf(Xaxis,1+N1,1+N2)
figure
plot(Xaxis,posty)
```

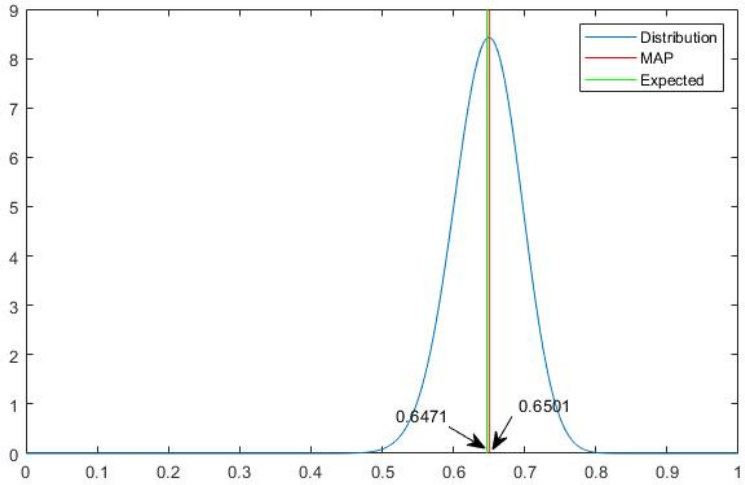
5.3

Calculate and report: – the MAP estimate of based on the posterior from part b. – Expected value of based on the posterior from part b. Show (plot) both the MAP estimate, and the expected value of on the plot of the posterior of you have generated in part b

```
thethaRange=0:0.0001:1
posty=betapdf(thethaRange,1+N1,1+N2);
[M,thethaMAP]=max(posty)
M = 8.4239
thethaMAP = 6501
estimateThetha= trapz(thethaRange,thethaRange.*posty)
estimateThetha = 0.6471
figure
plot(thethaRange,posty)
hold on
x = [0.6501 .6501];
```


Beta distribution 1,1
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
thetha


Posterior
9
8
7
6
5
4
3
2
1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
thetha

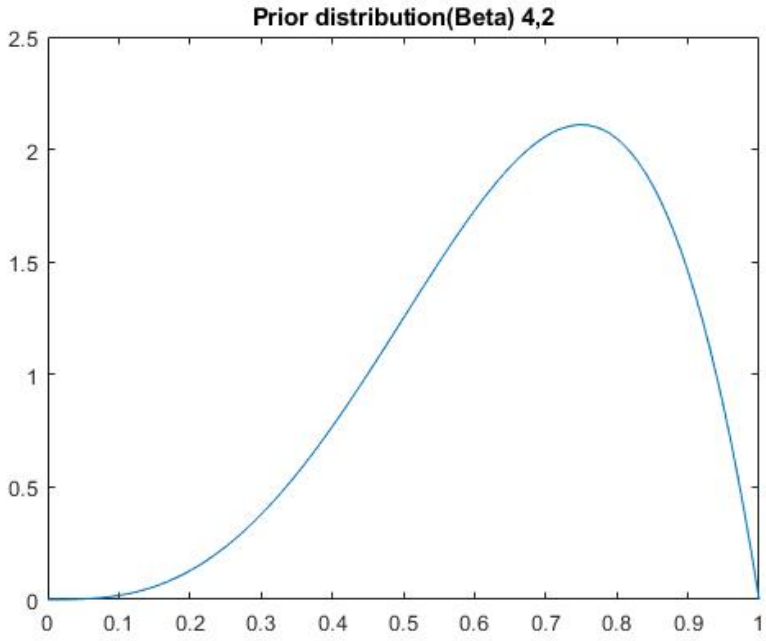


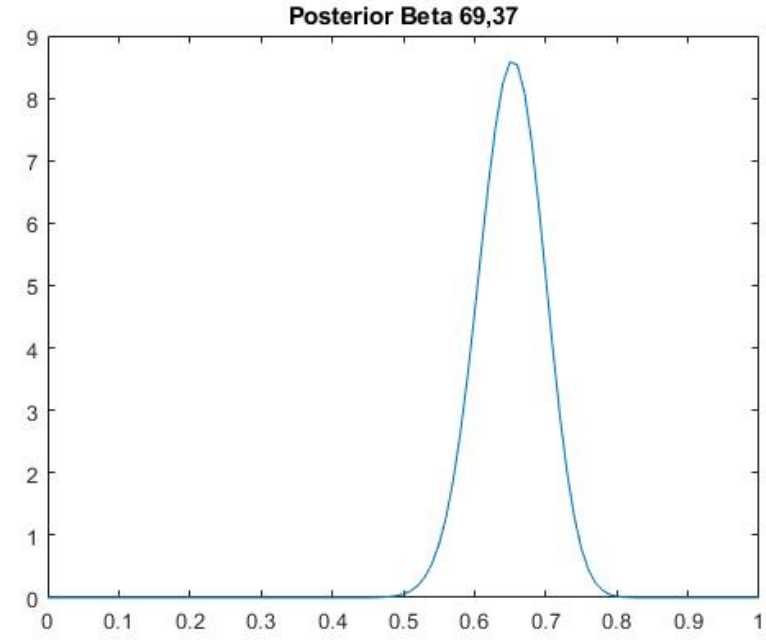

```
y = [0 9];
pl = line(x,y);
pl.Color = 'r';
x2 = [0.6471 .6471];
p2=line(x2,y);
p2.Color = 'g';
```

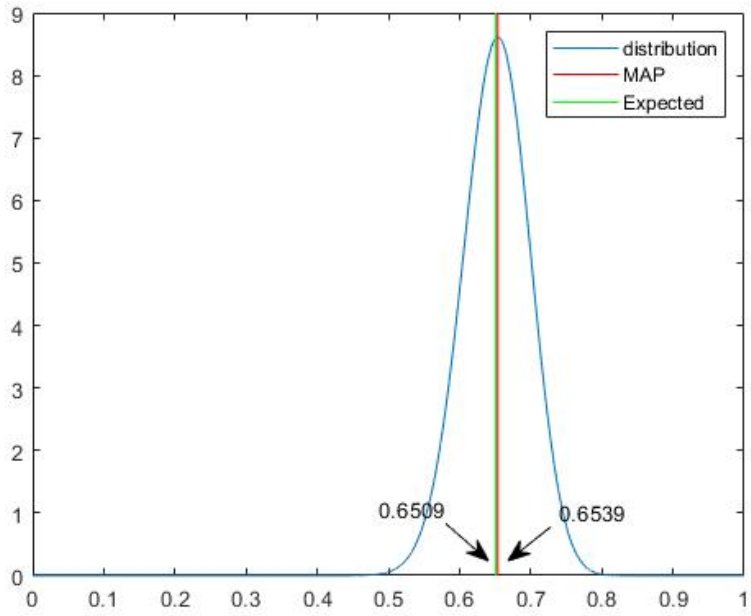
5.4

Repeat part b and c by assuming that the prior on follows Beta(—4, 2).

```
posty=betapdf(thethaRange,4+N1,2+N2);
estimateThetha= trapz(thethaRange,thethaRange.*posty)
estimateThetha =0.6509
[M,thethaMAP]=max(posty)
M = 8.6103
thethaMAP = 6539
figure
plot(thethaRange,posty)
hold on
x = [0.6539 .6539];
y = [0 9];
pl = line(x,y);
pl.Color = 'r';
x2 = [0.6509 .6509];
p2=line(x2,y);
p2.Color = 'g';
```


Prior distribution(Beta) 4,2
2.5
2
1.5
1
0.5
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Posterior Beta 69,37
9
8
7
6
5
4
3
2
1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


distribution
MAP
Expected
0.6509
0.6539
0
1
2
3
4
5
6
7
8
9
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1