SYST 664: HW Assignment 5

Jake Whalen

February 26, 2018

Table of Contents

[Problem 1 2](#_Toc507410457)

[Part A 2](#_Toc507410458)

[Part B 2](#_Toc507410459)

[Part C 3](#_Toc507410460)

[Part D 5](#_Toc507410461)

[Problem 2 5](#_Toc507410462)

[Part A 6](#_Toc507410463)

[Part B 6](#_Toc507410464)

[Part C 7](#_Toc507410465)

[Part D 7](#_Toc507410466)

[Part E 8](#_Toc507410467)

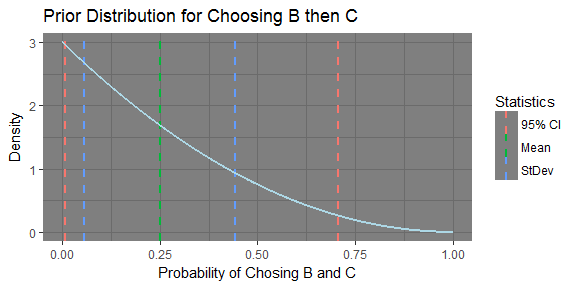
# 

# Problem 1

In previous years, students in this course collected data on people’s preferences in the two Allais gambles from Assignment 2. For this problem, we will assume that responses are independent and identically distributed, and the probability is that a person’s response is B in the first gamble and C in the second gamble.

### Part A

Assume that the prior distribution for is Beta(1, 3). Below is a plot of a Beta(1, 3) distribution.



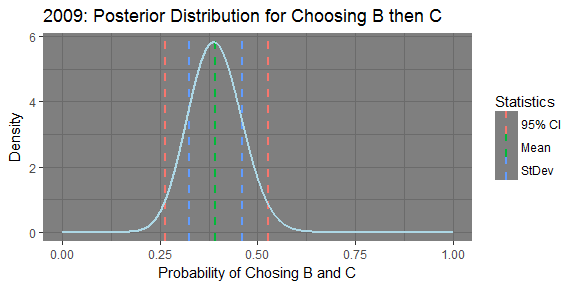
The table below shows the prior mean and standard deviation for as well as a 95% symmetric tail area credible interval for the prior probability that a person would choose B and C.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameters | Mean | StDev | Median | Q.025 | Q.975 |
| alpha: 1 , beta: 3 | 0.25 | 0.1936492 | 0.2062995 | 0.0084038 | 0.7075982 |

I do not think this is a reasonable prior distribution to use for this problem. I think that the prior distribution should have a higher mean. More people would be inclined to take the certainty of B in gamble 1 specifically. In gamble 2 the difference in winning nothing in C and D is 0.01. Therefore most people would be inclined to choose C because of the possibility of winning $100 more then in D.

### Part B

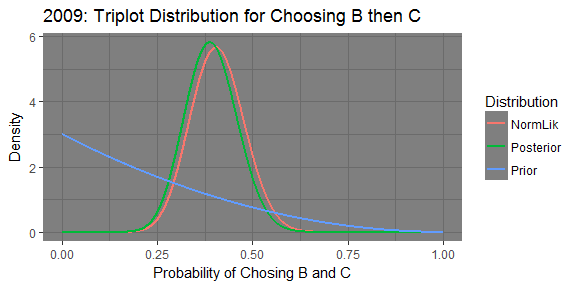
In 2009, 19 out of 47 respondents chose B and C. The plot below shows the posterior distribution for the probability that a person in this population would choose B and C.



The table below shows the posterior mean and standard deviation, and a 95% symmetric tail area credible interval for the posterior probability that a person in this population would choose B and C.

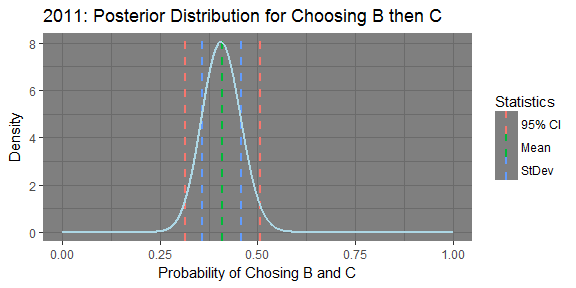
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameters | Mean | StDev | Median | Q.025 | Q.975 |
| alpha: 20 , beta: 31 | 0.3921569 | 0.0677055 | 0.3907365 | 0.2640784 | 0.5282508 |

Below is a triplot of the prior, posterior and normalized likelihood.



### Part C

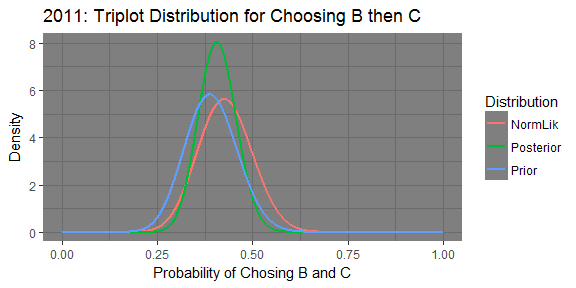
In 2011 another 47 responses were collected. This time, 20 out of 47 people said they preferred B and C. Assume that the Spring 2011 respondents have the same distribution of responses as the Spring 2009 respondents. Repeat part b, using the posterior distribution from the 2009 sample as your prior distribution. The plot below shows the posterior distribution for the probability that a person in this population would choose B and C.



The table below shows the posterior mean and standard deviation, and a 95% symmetric tail area credible interval for the posterior probability that a person in this population would choose B and C.

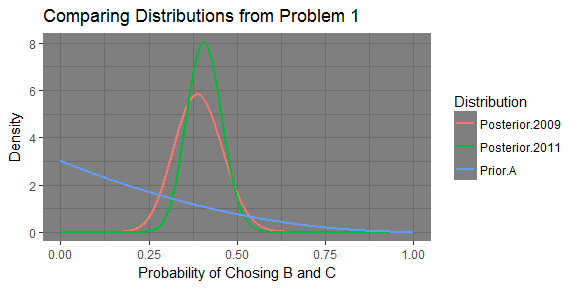
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameters | Mean | StDev | Median | Q.025 | Q.975 |
| alpha: 40 , beta: 58 | 0.4081633 | 0.049397 | 0.407536 | 0.3133473 | 0.506531 |

Below is a triplot of the prior, posterior and normalized likelihood.



### Part D

The plot below shows the distributions generated in the three previous parts of this problem.



We started out with the Beta(1, 3) prior distribution from part A (Prior.A). Then we calculated the posterior after having collected more data on peoples preferences in 2009 (Posterior.2009). Then again in 2011 more data was collected and the posterior distribution was calculated again (Posterior.2011).

The table below shows the means, standard deviation, median, and 95% credible intervals for each of the distributions.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameters | Mean | StDev | Median | Q.025 | Q.975 |
| alpha: 1 , beta: 3 | 0.2500000 | 0.1936492 | 0.2062995 | 0.0084038 | 0.7075982 |
| alpha: 20 , beta: 31 | 0.3921569 | 0.0677055 | 0.3907365 | 0.2640784 | 0.5282508 |
| alpha: 40 , beta: 58 | 0.4081633 | 0.0493970 | 0.4075360 | 0.3133473 | 0.5065310 |

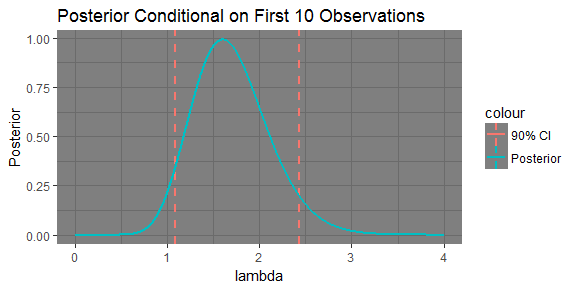
The prior has a very low mean and median when compared to the posteriors. The prior is also very spread out compared to the posteriors. The posteriors ahow the distribution mean centering in on 0.4. The spread of the distributions are also decreasing as more data is collected.

# Problem 2

This problem concerns the automobile data from Assignments 3 and 4.

### Part A

As in Assignment 4, assume that counts of cars per 15-second interval are independent and identically distributed Poisson random variables with unknown mean . Assume a uniform prior distribution for . (As for Assignment 4, you can approximate this prior distribution by using a Gamma distribution with shape 1 and scale 10,000.) The car counts for the first 10 time intervals are 2, 2, 1, 1, 0, 4, 3, 0, 2, 1. Find the posterior distribution for conditional on the first 10 observations. Plot the posterior density function for conditional on the first 10 observations. Find a 90% 2-sided posterior credible interval for .



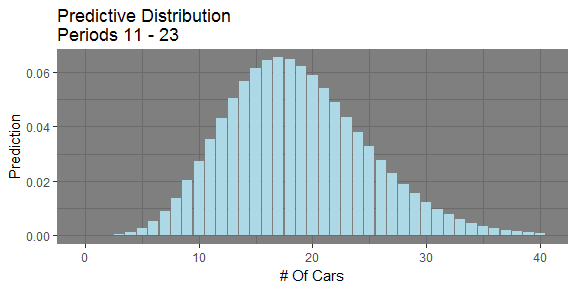
* 90% 2-sided posterior credible interval for
  + Upper: 1.083
  + Lower: 2.43

### Part B

Using the posterior distribution from Part a, find the predictive distribution for the total number of cars in the next 11 time periods. The predictive distribution is a Negative Binomial. Finding its parameters (size and prob) can be accomplished using the parameters from the posterior in part A.

* The resulting Values are:
  + 17
  + 0.476193

Plot the probability mass function for the predictive distribution.



### Part C

Find the posterior predictive probability that between 10 and 30 cars will pass by in the next 11 time periods. This is done by summing the probabilities from 10 cars to 30 cars arriving in the distribution above.

p2.b.df %>%  
 filter(cars >= 10,   
 cars <= 30) %>%  
 summarise(pred = sum(Prediction)) %>%  
 .$pred

[1] 0.9054938

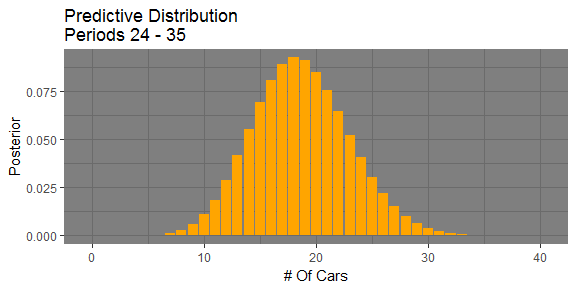
### Part D

If the number of cars passing during a single time period has a Poisson distribution with parameter , then the number of cars passing during 11 time periods has a Poisson distribution with parameter 11. Use the mean of the posterior distribution as a point estimate of , and use this point estimate to find a Poisson distribution for predicting the number of cars passing during the next 11 periods. The point estimate is calculated first using the posterior parameters found in part B.

p2.d.pointEst <- 11 \* p2.a.shape \* p2.a.scale  
p2.d.pointEst

[1] 18.69981

This estimate is used in the R function dpois to create the predictive distribution for the next 11 time periods. The plot below shows the result of this approach.



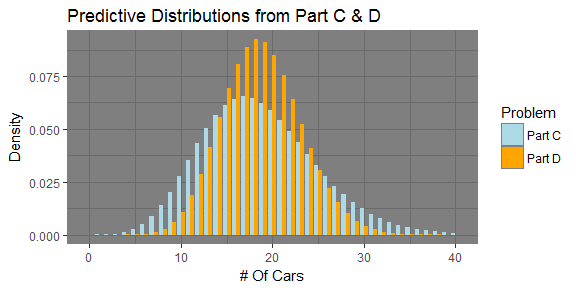
The code below calculates the probability that between 10 and 30 cars pass in these 11 time periods.

p2.d.dist.df %>%  
 filter(cars >= 10,   
 cars <= 30) %>%  
 summarise(post = sum(Posterior)) %>%  
 .$post

[1] 0.9838827

### Part E

The plot below shows the distributions for Parts c and d.



The table below shows the probabilities of 10 to 30 cars passing in the respective period of intervals.

|  |  |
| --- | --- |
| Periods | Probability |
| Periods 11-23 | 0.9054938 |
| Periods 24-35 | 0.9838827 |

The distribution from part C is more spread out and thus has a higher amount of probability below 10 and over 30. This results in the lower probability as compared to the distribution from part D in the table. By using the mean from part C as a point estimate for part D the predictive distribution in D reduced sread and became tighter around the mean.