Stat 341 – Homework 2

Gloria Grace

January 25, 2023

```
p_grid <- seq( from=0 , to=1 , length.out=1000 )
prior <- rep( 1 , 1000 )
likelihood <- dbinom( 6 , size=9 , prob=p_grid )
posterior <- likelihood * prior
posterior <- posterior / sum(posterior)
set.seed(100)
samples <- sample( p_grid , prob=posterior , size=1e4 , replace=TRUE )
```

SR 3E1

```
mean(samples < 0.2)
```

```
## [1] 4e-04
```

SR 3E2

```
mean(samples > 0.8)
```

```
## [1] 0.1116
```

SR 3E3

```
mean(between(samples, 0.2, 0.8))
```

```
## [1] 0.888
```

SR 3E4

```
quantile(samples, probs = 0.2)
```

```
##       20%
## 0.5185185
```

SR 3E5

```
quantile(samples, probs = 0.8)
```

```
##       80%
## 0.7557558
```

SR 3E6

```
HPDI( samples , prob=0.66)
```

```
##     |0.66     0.66|
## 0.5085085 0.7737738
```

SR 3E7

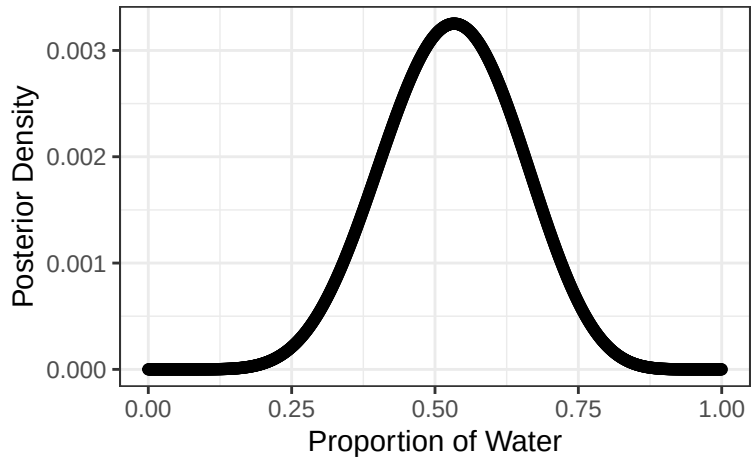
```
PI( samples , prob=0.66)
```

```
##       17%       83%
## 0.5025025 0.7697698
```

SR 3M1

```
p_grid <- seq( from=0 , to=1 , length.out=1000 )
prior <- rep( 1 , 1000 )
likelihood <- dbinom( 8 , size=15 , prob=p_grid )
posterior <- likelihood * prior
posterior <- posterior / sum(posterior)
```

```
tibble(p = p_grid, posterior = posterior) %>%
  ggplot(aes(x = p, y = posterior)) +
  geom_point() +
  geom_line() +
  labs(x = "Proportion of Water", y = "Posterior Density")
```



SR 3M2

```
set.seed(101)
samples <- sample(p_grid, prob = posterior, size = 1e4, replace = TRUE)
HPDI(samples, prob = 0.9)
```

```
##      |0.9      0.9|
## 0.3343343 0.7217217
```

N1

```
movielens <- read_csv('https://sldr.netlify.app/data/movielens.csv')
```

```
## Rows: 1945 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): title, genres, animation, for_kids, comedy, sci_fi, horror, romance
## dbl (2): movieId, rating
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Part A

The quantity that I am working on to estimate is the proportion of the movie that is comedy.

```
table(movielens$comedy)
```

```
##
##     Comedy Not Comedy
##        773       1172
```

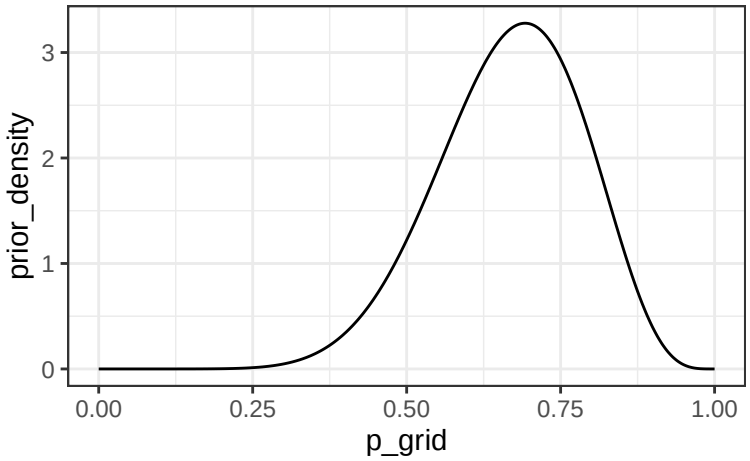
```
total_comedy <- 773
total_Ncomedy <- 1172
```

Part B

Choose a prior that adequately expresses your prior knowledge about the situation, before peeking at the data. Explain your choice. Make a sketch (by hand) or a graph (in R) of the prior, with appropriate axis titles (y axis need not be perfectly to scale)

I don't fully understand what I should input as my shape 1 and 2 for my beta distribution.

```
my_beta_prior <- tibble(p_grid = seq(from = 0, to = 1, length.out = 1000),
                        prior_density = dbeta(p_grid, shape1 = 10, shape2 = 5))
gf_line(prior_density ~ p_grid, data = my_beta_prior)
```



Part C

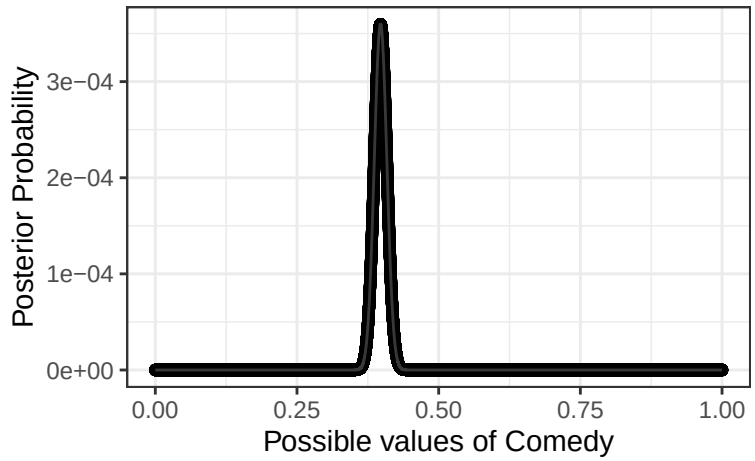
```
grid_movie <-
  tibble(p_grid = seq(from = 0, to = 1, length.out = 100000),      # define grid
         prior  = 1) |>                                            # define prior
  mutate(likelihood = dbinom(total_comedy,
                             size = total_comedy  + total_Ncomedy,
                             prob = p_grid)) |>  # compute likelihood at each value in grid
  mutate(unstd_posterior = likelihood * prior) |>              # compute product of likelihood and prior
  mutate(posterior = unstd_posterior / sum(unstd_posterior))   # standardize the posterior, so it sums


# to peek at the results table
glimpse(grid_movie)
```

```
## Rows: 100,000
## Columns: 5
## $ p_grid          <dbl> 0.0000000000, 0.0000100001, 0.0000200002, 0.0000300003~
## $ prior           <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ likelihood      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ unstd_posterior <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ posterior       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

Part D

Interpret and present your posterior in at least 2 ways. One of them should be a graph (not a number or interval). At least one of them should require sampling from the posterior as a preparatory step. Include a paragraph explaining your choice (why did you choose these 2 particular ways to show/present the posterior)?

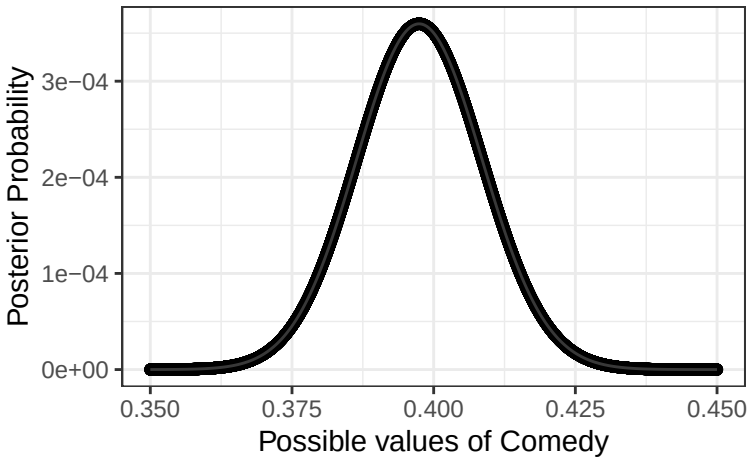
```
grid_plot <- gf_point(posterior ~ p_grid,
                      data = grid_movie) |>
  # this is optional -- adds a line in addition to  the dots
  gf_line(color = 'grey44', alpha = 0.5) |>
  gf_labs(x = 'Possible values of Comedy',
          y = 'Posterior Probability')

grid_plot
```



```
grid_plot |>
  gf_lims(x = c(0.35, 0.45))
```

```
## Warning: Removed 90000 rows containing missing values (`geom_point()`).
```

```
## Warning: Removed 90000 rows containing missing values (`geom_line()`).
```



Part E

Based on my work, I learned that the probability of the movies given that it is a comedy genre is 3e-04 where the possible value of comedy is below 0.4.

I am still confused about what does it mean, since it is binary, I understand there is probability of comedy and not comedy but why is the probability 3e-04? what does the possible values of comedy represent or what is P grid? I think I know what prior and posterior based on what I learn in class but I feel like classes exercise where we have to do our own does not help my understanding that much since a lot of people ask question to.

Part F

I do wonder how the data is obtained, out of all the movies that were picked, how were they pick. I do think that the data or the question that I answer would appear differently if the data were obtained differently too.