Assignment_11.1_HillZach

Zach Hill

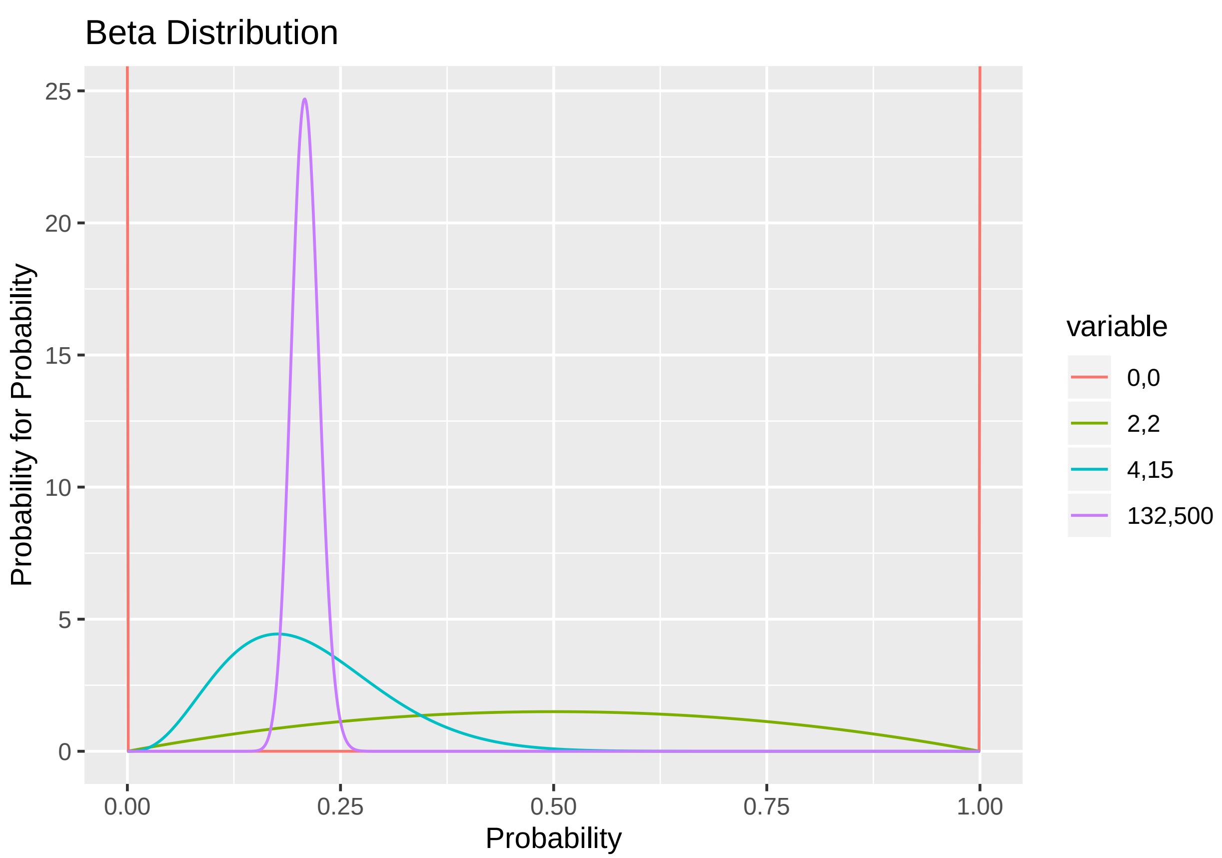
May 26, 2019

a. Using the Beta distribution and the probability function P(p) = Beta(k + 1,n - k + 1), plot the probability distributions for the following values.

No data collected. k = 0, n = 0. k = 2, n = 2 k = 4, n = 15 k = 132, n = 500

```
x <- seq(0, 1, length=1000)

beta_dist <- data.frame(cbind(x, dbeta(x, 0, 0), dbeta(x, 2, 2), dbeta(x, 4, 15), dbeta(x, 132, 500)))
colnames(beta_dist) <- c("x", "0,0", "2,2", "4,15", "132,500")
beta_dist <- melt(beta_dist, x)
ggplot(beta_dist, aes(x, value, colour = variable)) +
  geom_line() +
  labs(title = "Beta Distribution") +
  labs(x = "Probability", y = "Probability for Probability")
```



b. In the previous part of this problem, you plotted the probability distribution for different values of k (number of successes) and n (number of trials). Based on these plots, provide your best estimate of the success rate.

No data collected. k = 0, n = 0

With no data, you can't calculate probability. As expected the probability is zero across the interval

k = 2, n = 2

Probability of finding a 100% probability is low and likely caused by the low number of trials

k = 4, n = 15

With a greater number of trials we see an increased likelihood of the probability being correct

k = 132, n = 500

With the number of trials being greater than 30, this sample size shows a high probability of being correct

c. Load the data from ab_test.csv. Using all the data, plot probability distribution for the test case A and B on the same plot. Based on these plots, which one has the higher conversion rates?

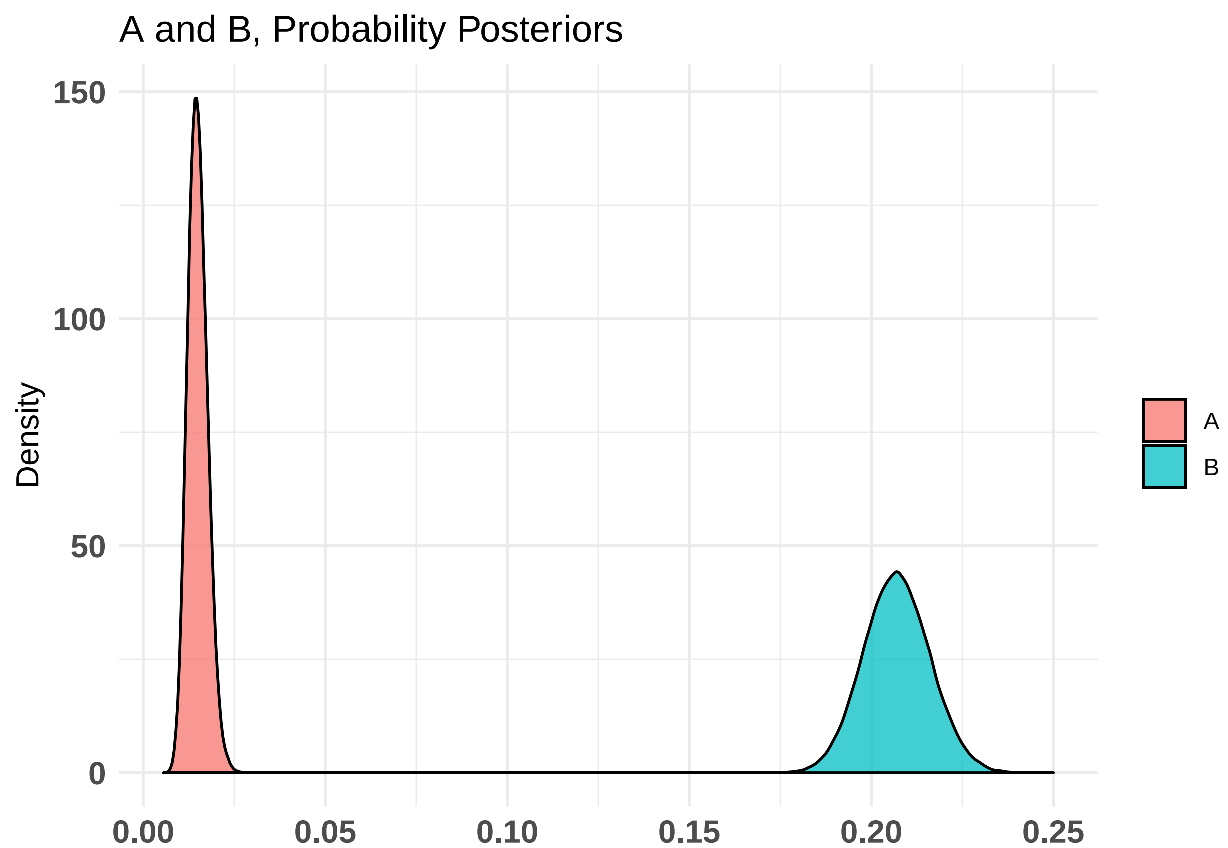
```
input_file <- './ab_test.csv'
data <- read_csv(input_file)

dfA <- data[which(data$label=="A"),]
dfB <- data[which(data$label=="B"),]

sA <- nrow(data[which(data$label=="A" & data$is_success==1),])
sB <- nrow(data[which(data$label=="B" & data$is_success==1),])
nA <- nrow(data[which(data$label=="A"),])
nB <- nrow(data[which(data$label=="B"),])
pA <- sA/nA
pB <- sB/nB

test_A <- bayesTest(dfA$is_success, dfB$is_success, distribution = "bernoulli", priors = c("alpha" = pA
plot_A <- plot(test_A)
plot_A[2]
```

```
## $posteriors
## $posteriors$Probability
```



d. Using the qbeta function (quantile function of the Beta function) calculate the 95% confidence interval (i.e., quantiles between 2.5% and 97.5%) for A and B.

Test A

```
qbeta(c(.025, .975), sA+1, nA-sA+1)
```

```
## [1] 0.01054987 0.02133403
```

Test B

```
qbeta(c(.025, .975), sB+1, nB-sB+1)
```

```
## [1] 0.1898208 0.2253185
```

e. Finally, you will examine what the distributions would like at different points during the data collection process. Repeat steps c and d, but only include data on or before the date provided.

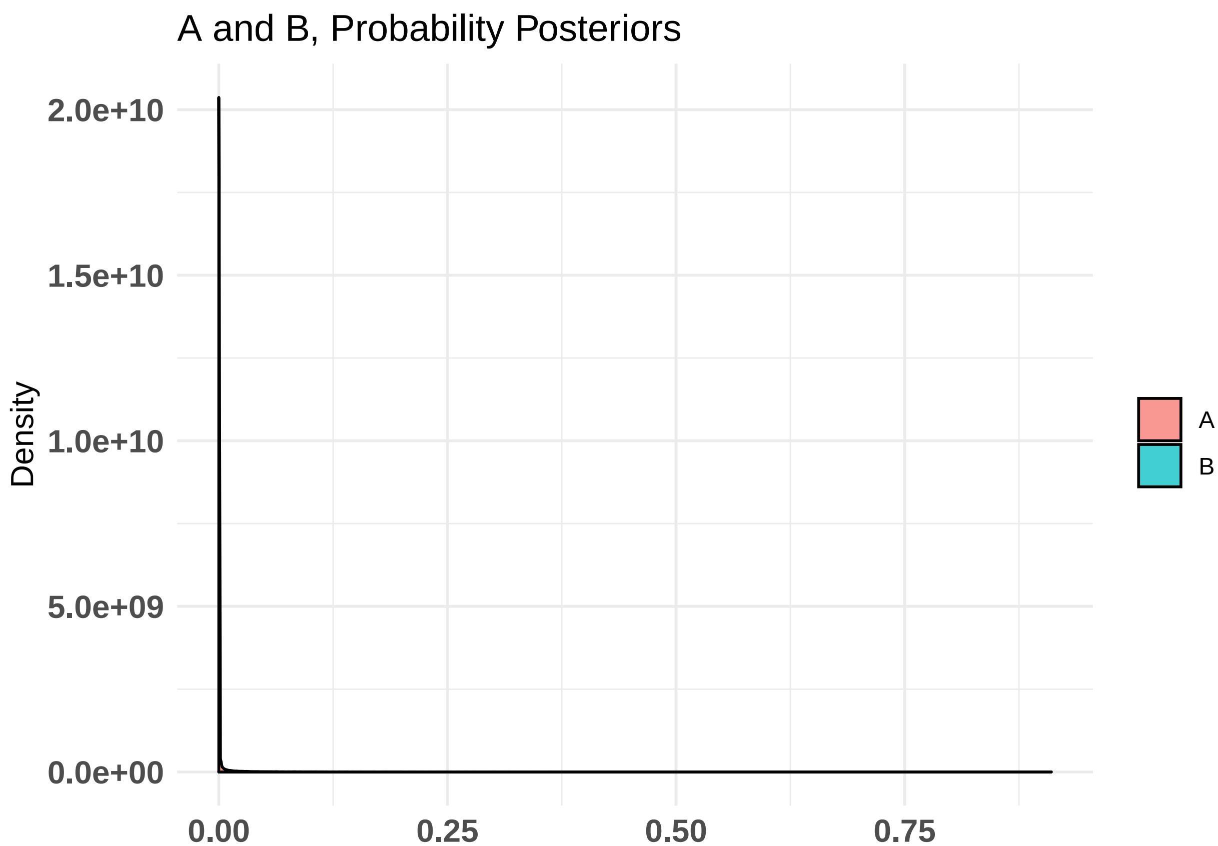
On or before 2009-09-01

```
dfA_date1 <- dfA[which(dfA$timestamp < "2009-09-02"),]
dfB_date1 <- dfB[which(dfB$timestamp < "2009-09-02"),]

sA_date1 <- nrow(dfA_date1[which(dfA_date1$label=="A" & dfA_date1$is_success==1),])
sB_date1 <- nrow(dfB_date1[which(dfB_date1$label=="B" & dfB_date1$is_success==1),])
nA_date1 <- nrow(dfA_date1[which(dfA_date1$label=="A"),])
nB_date1 <- nrow(dfB_date1[which(dfB_date1$label=="B"),])
```

```
test_e1 <- bayesTest(dfA_date1$is_success, dfB_date1$is_success, distribution = "bernoulli", priors = c
plot_e1 <- plot(test_e1)
plot_e1$posteriors
```

```
## $Probability
```



Test A

```
qbeta(c(.025, .975), sA_date1+1, nA_date1-sA_date1+1)
```

```
## [1] 0.002298972 0.284914153
```

Test B

```
qbeta(c(.025, .975), sB_date1+1, nB_date1-sB_date1+1)
```

```
## [1] 0.07485463 0.60009357
```

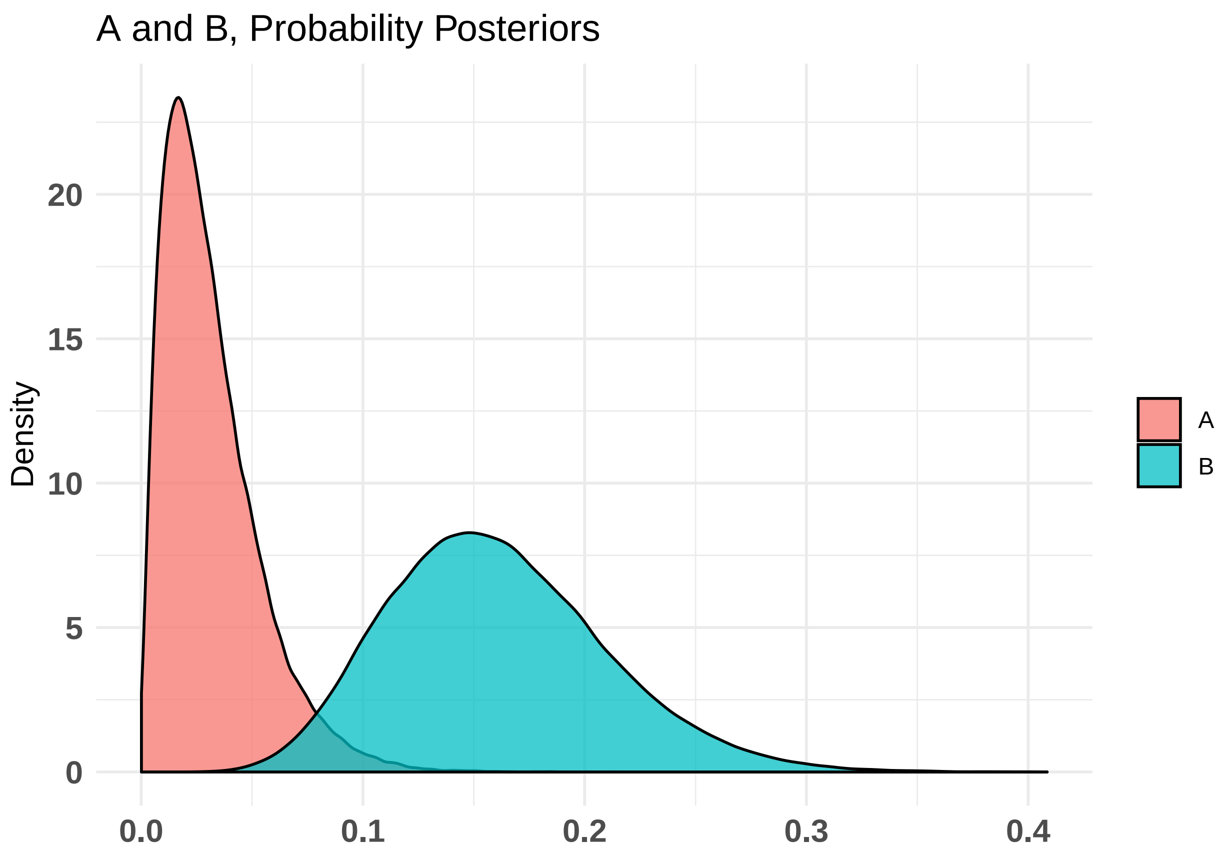
On or before 2009-10-15

```
dfA_date2 <- dfA[which(dfA$timestamp < "2009-10-16"),]
dfB_date2 <- dfB[which(dfB$timestamp < "2009-10-16"),]

sA_date2 <- nrow(dfA_date2[which(dfA_date2$label=="A" & dfA_date2$is_success==1),])
sB_date2 <- nrow(dfB_date2[which(dfB_date2$label=="B" & dfB_date2$is_success==1),])
nA_date2 <- nrow(dfA_date2[which(dfA_date2$label=="A"),])
nB_date2 <- nrow(dfB_date2[which(dfB_date2$label=="B"),])
```

```
test_e2 <- bayesTest(dfA_date2$is_success, dfB_date2$is_success, distribution = "bernoulli", priors = c
plot_e2 <- plot(test_e2)
plot_e2$posteriors
```

```
## $Probability
```



Test A

```
qbeta(c(.025, .975), sA_date2+1, nA_date2-sA_date2+1)
```

```
## [1] 0.009621079 0.106770321
```

Test B

```
qbeta(c(.025, .975), sB_date2+1, nB_date2-sB_date2+1)
```

```
## [1] 0.08747323 0.27868263
```

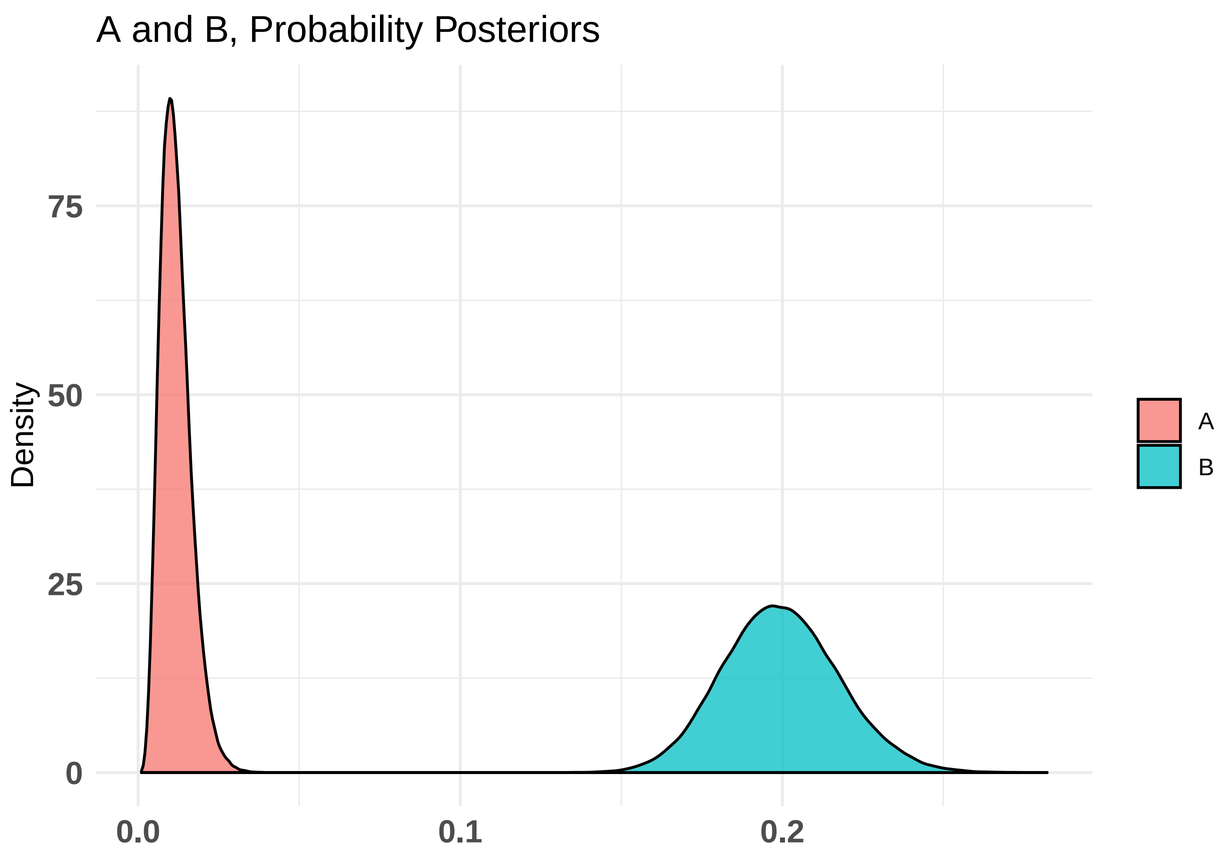
On or before 2009-12-24

```
dfA_date3 <- dfA[which(dfA$timestamp < "2009-12-25"),]
dfB_date3 <- dfB[which(dfB$timestamp < "2009-12-25"),]

sA_date3 <- nrow(dfA_date3[which(dfA_date3$label=="A" & dfA_date3$is_success==1),])
sB_date3 <- nrow(dfB_date3[which(dfB_date3$label=="B" & dfB_date3$is_success==1),])
nA_date3 <- nrow(dfA_date3[which(dfA_date3$label=="A"),])
nB_date3 <- nrow(dfB_date3[which(dfB_date3$label=="B"),])
```

```
test_e3 <- bayesTest(dfA_date3$is_success, dfB_date3$is_success, distribution = "bernoulli", priors = c
plot_e3 <- plot(test_e3)
plot_e3$posteriors
```

```
## $Probability
```



Test A

```
qbeta(c(.025, .975), sA_date3+1, nA_date3-sA_date3+1)
```

```
## [1] 0.005481779 0.025184735
```

Test B

```
qbeta(c(.025, .975), sB_date3+1, nB_date3-sB_date3+1)
```

```
## [1] 0.1666631 0.2372583
```