# CASE STUDY - FUNDAMENTALS

## Bayesian Data Analysis (MA480)

*Aaradhana Bharill,CM Box 3074*

## Introduction

For the past five years, the math department has been offering calculus 1 classes in fall and winter quarter. In this study, we are trying to predict the expected number of freshmen who will need to take Calculus 1 next fall (fall 2018) at Rose Hulman given that there are 550 incoming freshmen, so that all the freshmen can take the class in the fall. As long as there are no major changes, the obtained model can also help estimate the number of freshmen who need to take calculus 1 in the upcoming years at Rose Hulman.

## Methods

### Data Collection

Data about the number of students who took calculus 1 in fall and winter quarter for the years, 2000-2017, was collected from the Rose Hulman schedule lookup page (https://prodwebxe-hv.rose-hulman.edu/regweb-cgi/reg-sched.pl?&ticket=ST-a8d97648dda742008a4bad5fb5892e29-bxeeis-hv.rose-hulman.edu). No calculus 1 classes were offered in spring quarter.

Then, the number of incoming freshmen for all those years was obtained from the Rose Hulman admissions department. I decided to leave out the data for the years before 2013 since in all those years, calculus 1 was offered only in fall quarter whereas after 2013, it was offered in both, fall and winter quarter. However, last year the capacity of all the classes was at 23 unlike the previous years when it was 26, 27 or 28, despite the fact that the number of incoming freshmen was slightly more than the last two years. So I decided that the best estimator for the number of students who will need caclulus 1 next fall is last fall and winter only.

### Model

The likelihood of freshmen needing to take calculus was modeled by the binomial distribution with parameters, N = number of incoming freshmen in the year 2017 = 557, and $\theta$ = proportion of incoming freshmen who will need to take calculus. The total number of students who took calculus 1 in the year 2017, was x = 305.

The prior distribtuion of $\theta$ was modeled with the uniform distribution, $\theta \sim$ Uniform(0, 1) (=beta(1, 1)) since we have no prior information about $\theta$.

### Posterior Distribution

This is the posterior distribution: $\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)}$ The numerator is simply the likelihood since the prior distributon is unif(0, 1), and the denominator integrates to the beta-binomial(N, 1, 1).

$\pi(\theta|y) = \frac{\binom{n}{x}\theta^x(1-\theta)^{N-x}}{\binom{n}{x}\frac{\Gamma(x+1)\Gamma(N-x+1)}{\Gamma(x+1+N-x+1)}}$ where N = 557, and x = 305.

$\binom{n}{x}$ in the numerator and denominator can be cancelled out, and the denominator can be brought into the numerator and the following expression is obtained:

$\pi(\theta|y) = \frac{\Gamma(x+1+N-x+1)}{\Gamma(x+1)\Gamma(N-x+1)} * (\theta^x(1-\theta)^{N-x})$ where N = 557, and x = 305.

Therefore the posterior distribution is $\theta|x \sim$Beta(x+1, N-x+1), where N = 557, and x = 305.

## Results

The posterior predictive, $\pi(x*|x)$, for the likelihood:

$P(x*|\theta) = Binomial(550, \theta)$

and the posterior distribution:

$\pi(\theta|x) = Beta(306{,}253)$,

for N=550 (incoming freshmen), is Beta-Binomial(550, 306, 253).

The mean of this beta binomial distribution = $\frac{550*306}{306+253}$ = 301 students. Therefore 301 students are expected to want to take calculus 1 next fall.

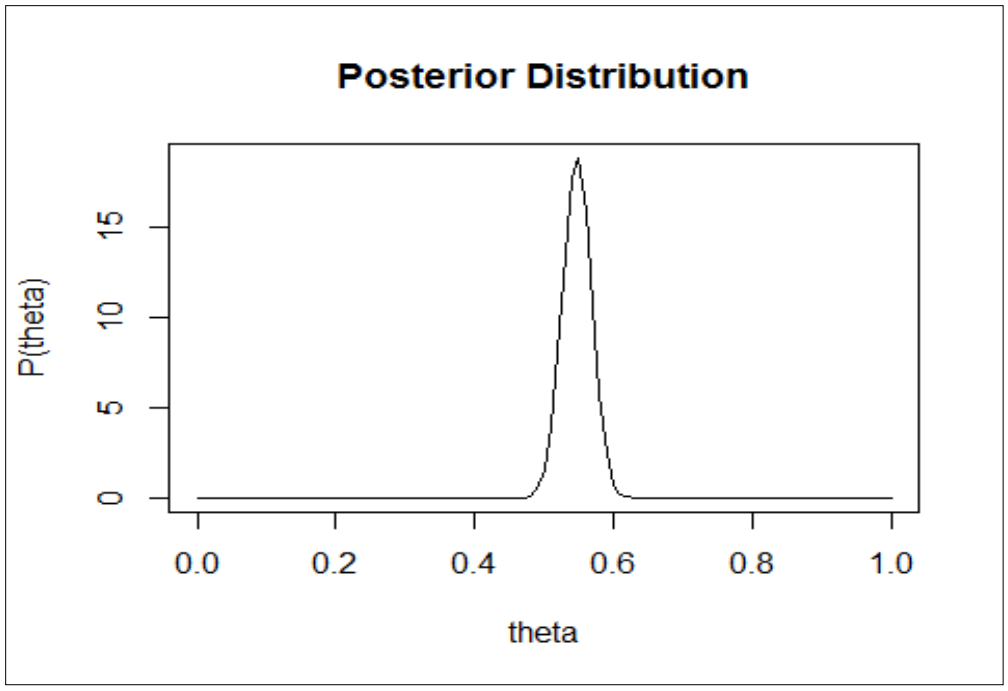
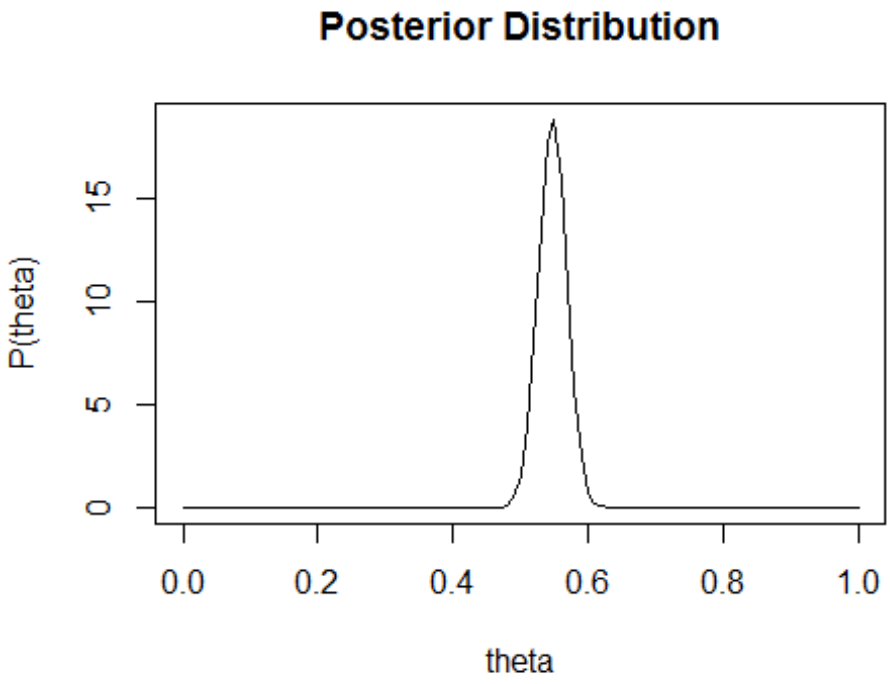



Figure 1: The probability density function obtained for $\theta$.

The expected value of $\theta$ for the posterior distribution was 0.547. Also, as can be seen in the plot of the posterior distribution for $\theta$, it makes sense that a little more than half the number of incoming freshmen are expected to want to take calculus 1 at next fall. Next time, I would like to use a bernoulli distribution for the likelihood and see what answers I get and compare the answers since the bernoulli distribution would model each students' choices, instead of assessing all the students as the same.

## Appendix

```
library(rmutil)
posterior <- function(theta) {
  x <- 305
  N <- 557
  logpost <- (log(theta^x)+log((1-theta)^(N-x)))-((lgamma(x+1)+lgamma(N-x+1))
-(lgamma(N+2)))
  post <- exp(logpost)
  return(post)
}
plot(Vectorize(posterior),main="Posterior Distribution",
  xlab="theta", ylab="P(theta)")
```




```
expValue <- function(theta){
  return(theta*posterior(theta))
}
meantheta <- integrate(expValue, 0, 0.999)
meantheta

0.5474061 with absolute error < 2.5e-05
```