# Bayesian Analysis

*Note: This report contains some research questions. The answers are provided with respect to the research questions.*

## Introduction

The dataset consists of birth recordings of the 100 2-children families. If a child is a male, it is recorded as 1 and if it is a female, it is recorded as 0. The dataset and the library are loaded. The seed is used to reproduce the same results. The other lines of code are used to speed up the performance

```
library(rstan)
library(bayesplot)
family = readRDS("./family.rds")
rstan_options(auto_write=T)
options(mc.cores=parallel::detectCores())
SEED <- 48927
```

**Question 1: What is the posterior distribution for the probability of a birth being a boy?**

As we see, the data is in the form of binary. In other words, the outcome is binary. Hence we can assume a Bernoulli distribution with parameter theta. Here theta represents the probability of having a boy child. Since this is a probability, we can assume a beta distribution with parameters 1 and 1. In other words, a uniform distribution over the interval 0 and 1(given in the question).

Since the firstborn and second are two independent events, We are going to construct 2 models for having firstborn as a boy, second born child as a boy. We will construct the third model for having a boy child irrespective of the order(i.e first or second born).

```
data = list(N = 100,y=family$birth1 | family$birth2 , y1=family$birth1,y2=fam
ily$birth2)
fit = stan(file="model.stan",data=data,seed = SEED)
print(fit)
```

```
## Inference for Stan model: model.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##           mean se_mean   sd    2.5%     25%     50%     75%   97.5% n_eff
Rhat
## theta     0.89    0.00 0.03    0.82    0.87    0.90    0.91    0.94  3810
1
## theta1    0.51    0.00 0.05    0.41    0.48    0.51    0.54    0.61  3797
1
```

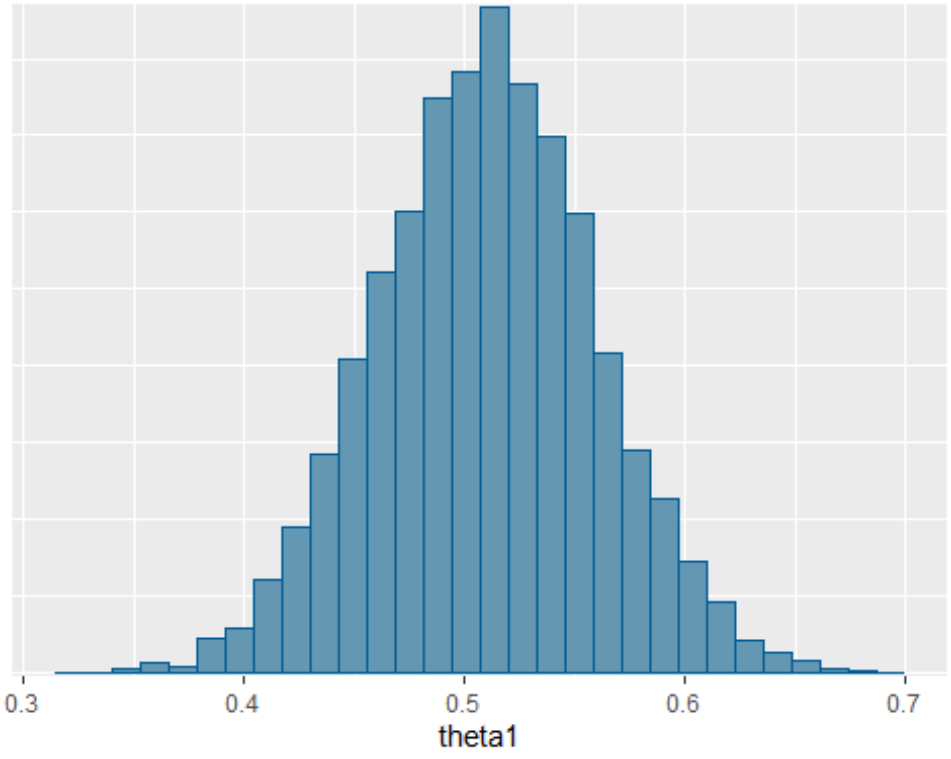
```
## theta2    0.60    0.00 0.05    0.50    0.57    0.60    0.63    0.69  3901
1
## lp__   -175.80    0.03 1.24 -178.99 -176.38 -175.48 -174.87 -174.40  2034
1
##
## Samples were drawn using NUTS(diag_e) at Tue Jun 23 15:13:14 2020.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

We can see Rhat value as 1 for all the parameters. This means that the chains are converged

Let us see the distribution of first born child as a male.

```
draws = as.data.frame(fit)
mcmc_hist(draws, pars='theta1')
```

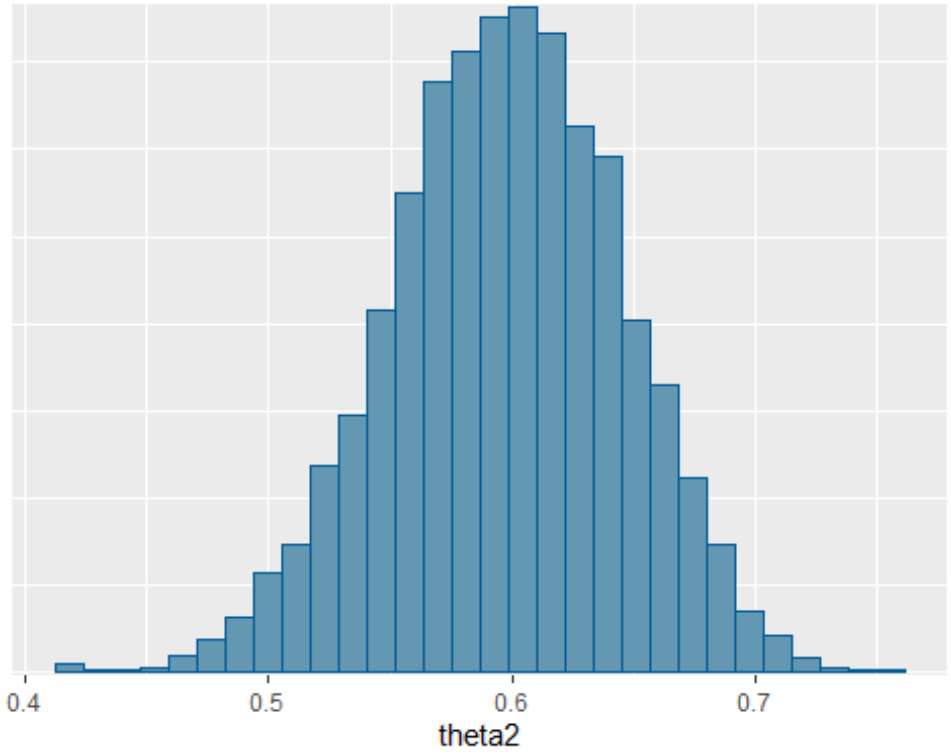
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




From the plot, we can say the average probability for having firstborn child as a male is approximately 50%. The 95% CI falls around 0.45 to 0.55.

```
mcmc_hist(draws, pars='theta2')
```

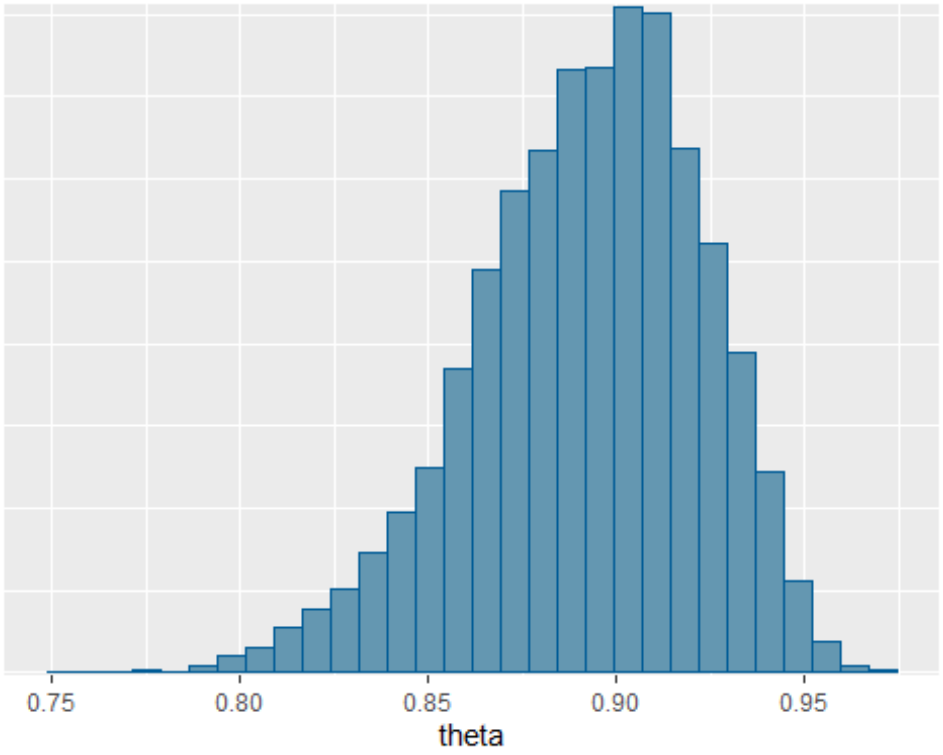
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




The average probabilty probability of having secondborn child as male is approximately 60%. The 95% CI falls around 0.55 and 0.65 approximately.

```
mcmc_hist(draws, pars='theta')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




The above plot is skewed, according the above plot the mean probability of having a boy child irrespective of the order (i.e firstborn or second born) is approximately 90%.

We can conclude that the probability of having a boy child in a family is 90%. To have first born as a boy child, the probabilty is 50% and for the second born as a boy child, the probability is 60%.

We are going to use the extract() function to get the samples from the posterior distribution.

```
CI = matrix(NA,ncol =6,nrow=3 )
rownames(CI) = c("Theta1","Theta2","Theta")
colnames(CI) = c("60% lower","60% upper","80% lower",
                 "80% upper","90% lower","90% upper")

CI = as.data.frame(CI)


samples = extract(fit)
theta = samples$theta[1:1000]
theta1 = samples$theta1[1:1000]
theta2 = samples$theta2[1:1000]

CI[1,] =c(HDInterval::hdi(theta1, credMass = 0.6)[1:2],
HDInterval::hdi(theta1, credMass = 0.8)[1:2],
HDInterval::hdi(theta1, credMass = 0.9)[1:2])
```

```
CI[2,] = c(HDInterval::hdi(theta2, credMass = 0.6)[1:2],
HDInterval::hdi(theta2, credMass = 0.8)[1:2],
HDInterval::hdi(theta2, credMass = 0.9)[1:2])


CI[3,] = c(HDInterval::hdi(theta, credMass = 0.6)[1:2],
HDInterval::hdi(theta, credMass = 0.8)[1:2],
HDInterval::hdi(theta, credMass = 0.9)[1:2])


round(CI,2)
```

```
##        60% lower 60% upper 80% lower 80% upper 90% lower 90% upper
## Theta1      0.46      0.55      0.44      0.57      0.42      0.59
## Theta2      0.57      0.64      0.54      0.66      0.52      0.68
## Theta       0.88      0.93      0.86      0.93      0.85      0.94
```

For theta1: 60% of the posterior samples fall under the range 0.46 to 0.55. 80% of the posterior samples fall under the range 0.44-0.57 and 90% of the posterior samples fall under the interval of 0.42 - 0.59.

For theta2: 60% of the posterior samples fall under the range 0.57 to 0.64. 80% of the posterior samples fall under the range 0.54-0.66 and 90% of the posterior samples fall under the interval of 0.52 - 0.68.

For theta : 60% of the posterior samples fall under the range 0.88 to 0.93. 80% of the posterior samples fall under the range 0.86-0.93 and 90% of the posterior samples fall under the interval of 0.85 - 0.94.

**Question 2:Draw a sample of size 1,000 from the posterior predictive distribution. How does this compare with the observed number of boys in the dataset? What should you conclude from this analysis?**

To draw a sample of size 1000 from posterior prediction distribution, we can use base R. Let us extract the posterior distribution of the parameters from the stan object and perform posterior predictive distribution.

Let us also check the proportion of boys in posterior predictive distribution and proportion of boys in the observed dataset.

```
proportion = function(draws){

  tab = table(draws)
  return(prop.table(tab))
}



first = rep(NA,1000)
```

```r
second = rep(NA,1000)
boys = rep(NA,1000)

## size = 1 in a binomial distribution represents the bernoulli distribution

for(i in 1:1000){

  first[i] = rbinom(1, 1,theta1[i])
  second[i] = rbinom(1, 1,theta2[i])
  boys[i] = rbinom(1, 1,theta[i])

}



proportion(first)

## draws
##     0     1
## 0.479 0.521

proportion(family$birth1)

## draws
##    0    1
## 0.49 0.51

proportion(second)

## draws
##     0     1
## 0.389 0.611

proportion(family$birth2)

## draws
##   0   1
## 0.4 0.6

proportion(boys)

## draws
##     0     1
## 0.108 0.892

proportion(family$birth1|family$birth2)

## draws
## FALSE  TRUE
##   0.1   0.9
```

The observed proportion and the sampled proportion are very close to each other in all the 3 cases. Hence the posterior distribution has captured the dataset well.

**Question 3 :We will now consider how useful our posterior distribution from question 1 is to predict only the number of first born boys. To do this, simulate a sample of 1,000 replicates of 100 births. Compare this to the actual number of first born boys. What do you conclude?**

The distribution is replicated for 1000 using sampled thetas of size 1000. A matrix is constructed to store the results of 1000 replications. Each row consists of 100 samples. Each sample represents the firstborn child.

```
first_born = matrix(NA,nrow = 1000,ncol=100)

for(i in 1:1000){

  first_born[i,] = rbinom(100,1,theta1[i])
}
total_boys1 = rowSums(first_born)


mean(total_boys1)

## [1] 50.947

length(family$birth1[family$birth1==1])

## [1] 51
```

We took the summation of each row to have the count of the firstborn boys in each replication. The mean of the total number of boys for 1000 replication is 51. The observed total number of boys in the firstborn category is also 51. We can easily conclude that the posterior distribution of theta1 has captured the distribution of the first born children well.

**Question 4: The posterior distribution from question 1 (the posterior distribution of the probability of a boy, given 200 births) assumes that the sex of first and second born are independent. Consider how you can assess this assumption of independence using a posterior predicitive approach. [Hint: explore the posterior predictive distribution of second boys that followed a first born female.]**

We can use the idea of conditional Probability, we can extract the families where firstborn is a girl and the second born is the boy. If we can prove that Pr(second born=boy/ firstborn=girl) = Pr(Secondborn = boy), then the sex of first and second born are independent.

To prove the above statement, we can use the Chi-sq test. If the p-value is greater than 0.05 then the above line is true. In other words, the assumption will be true.

```
girls_theta = theta1[first==0 & second==1]
girls_theta = 1- girls_theta
```

```r
boys_theta = theta2[first==0 & second==1]

mat = matrix(NA,nrow=length(girls_theta),ncol = 2)
mat[,1] = girls_theta
mat[,2] = boys_theta
data = as.data.frame(mat)

chisq.test(data) ## Expected data as Girls samples and Observed as Boys Sampl
es.

## Warning in chisq.test(data): Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 1.531, df = 285, p-value = 1
```

We can see that the p value >0.05. We can now conclude that the sex of first and second born are independent.