Gibbs sampler for a Beta-Binomial model

1 Bayesian modeling: a general introduction

Unlike in the traditional frequentist framework, the Bayesian approach views parameters as random variables rather than fixed, unknown quantities. Therefore, parameters also have a probability distribution. Given a generic random variable y and a parameter or a set of parameters θ, from the Bayes theorem, we can write

$$\pi(\theta \mid y) = \frac{\pi(y \mid \theta)\ \pi(\theta)}{\pi(y)}$$

where $\pi(\theta \mid y)$ denote the posterior distribution of the parameter, $\pi(y \mid \theta)$ denote the likelihood distribution of the data, $\pi(\theta)$ denote the prior distribution on the parameter. Further, $\pi(y)$ is called the marginal distribution of y and is equal to $\int \pi(y \mid \theta)\ \pi(\theta) d\theta$. In such a context, we also say that $\pi(y)$ is a normalizing constant, since it 'scales' the posterior, making it a density function. Often in Bayesian statistics, we work up to proportionality so that the above expression is rewritten as

$$\pi(\theta \mid y) \propto \pi(y \mid \theta)\ \pi(\theta)$$

In trivial cases, the posterior distribution $\pi(\theta \mid y)$ can be identified and closed-form quantities of interest like a mean and a variance or quantiles can be computed. However, most of the time in practice, the posterior distribution is intractable or the analytical expressions are cumbersome to compute so that it is necessary to resort to *Markov chain Monte Carlo* (MCMC) techniques. Those rely on simulations from a Markov chain that will hopefully converge to the distribution of interest, often the posterior distribution. Some common MCMC algorithms include the Gibbs sampler, based on the work of Gelfand and Smith (1990) and the Metropolis-Hastings algorithm, after the work of Metropolis et al. (1953) and then of Hastings (1970).

2 The Beta-Binomial conjugacy

Suppose that a r.v. Y has a Binomial distribution (a Binomial likelihood for our data $\mathbf{y}$), characterized by the probability mass function:

$$f_p(y) = \pi(y \mid p) = \binom{n}{y} p^y (1-p)^{n-y}$$

where $\binom{n}{y} = \frac{n!}{y!(n-y)!}$ is the binomial coefficient representinng all the possible conbinations of y successes in n Bernoulli trials and p is the parameter representing the probability of success for each trial.

In Bayesian statistics, to model the uncertainty about p, a convenient prior would be the Beta distribution $Beta(\alpha, \beta)$, since its support is the interval $[0, 1]$. The Beta distribution has the following probability density function:

$$f_{\alpha,\beta}(p) = \pi(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

where α and β are shape parameters, $B(\alpha, \beta)$ is the Beta function and $\Gamma(\alpha)$ is the Gamma function, defined as $\int_0^\infty y^{\alpha-1} e^{-y} dy = (\alpha - 1)!$. The Beta density is a convenient choice for the distribution of p since it can have a wide variety of shapes (figure 1).

Let $p \sim Beta(\alpha, \beta)$. We have then

$$E[p] = \frac{\alpha}{\alpha+\beta} \qquad var(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

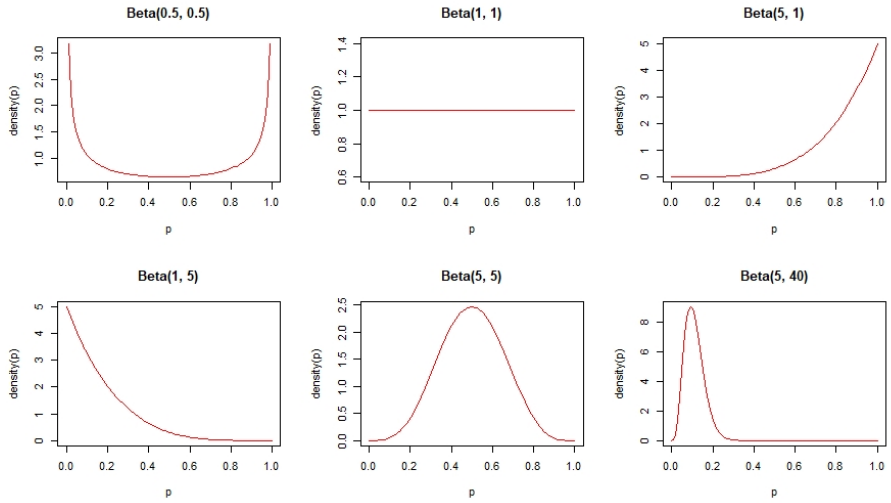

Figure 1: Some exampless of Beta densities for various shape parameters α and β.

Indeed,

$$
\begin{aligned}
E[p] &= \int_{-\infty}^{+\infty} p\ f_{\alpha,\beta}(p)\ dp = \int_0^1 p\ \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}\ dp \\
&= \frac{1}{B(\alpha,\beta)} \int_0^1 p^{(\alpha-1)+1}(1-p)^{\beta-1}\ dp \qquad\qquad B(\alpha,\beta) = \frac{\Gamma(\alpha)\ \Gamma(\beta)}{\Gamma(\alpha+\beta)} \\
&= \frac{1}{B(\alpha,\beta)} \int_0^1 p^{\alpha}(1-p)^{\beta-1}\ dp \qquad\qquad \text{note that } \int_0^1 p^{\alpha}(1-p)^{\beta-1}\ dp = B(\alpha+1,\beta) = \frac{\Gamma(\alpha+1)\ \Gamma(\beta)}{\Gamma(\alpha+1+\beta)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\ \Gamma(\beta)} \frac{\Gamma(\alpha+1)\ \Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\ \Gamma(\beta)} \frac{\alpha\Gamma(\alpha)\ \Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)} \\
&= \frac{\alpha}{(\alpha+\beta)}
\end{aligned}
$$

$$
\begin{aligned}
E[p^2] &= \int_{-\infty}^{+\infty} p^2\ f_{\alpha,\beta}(p)\ dp = \int_0^1 p^2\ \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}\ dp \\
&= \frac{1}{B(\alpha,\beta)} \int_0^1 p^{(\alpha-1)+2}(1-p)^{\beta-1}\ dp \\
&= \frac{1}{B(\alpha,\beta)} B(\alpha+2,\beta) \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\ \Gamma(\beta)} \frac{\Gamma(\alpha+2)\ \Gamma(\beta)}{\Gamma(\alpha+\beta+2)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\ \Gamma(\beta)} \frac{(\alpha+1)\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta+1)(\alpha+\beta)\Gamma(\alpha+\beta)} \\
&= \frac{\alpha(\alpha+\beta)}{(\alpha+\beta+1)(\alpha+\beta)}
\end{aligned}
$$

$$var(p) = E[p^2] - \big(E[p]\big)^2 = \frac{\alpha(\alpha+\beta)}{(\alpha+\beta+1)(\alpha+\beta)} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 = ... = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

The posterior distribution for p if Y is Binomial and a Beta prior is chosen for p will also have the functional form of a Beta r.v. That is why we say that Beta is the conjugate prior for a Binomial likelihood. Using the Bayes theorem, we have that

$$\begin{aligned}
\pi(p \mid y) &= \frac{\pi(y \mid p)\pi(p)}{\pi(y)} \\
&= \binom{n}{y} p^y (1-p)^{n-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\
&= \underbrace{\binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{\text{do NOT depend on } p} \quad p^y (1-p)^{n-y} p^{\alpha-1}(1-p)^{\beta-1} \\
&\propto p^{y+\alpha-1}(1-p)^{n-y+\beta-1}
\end{aligned}$$

and we find that $p \mid y \propto Beta(y+\alpha, n-y+\beta)$. The posterior mean and variance are given by

$$E[p \mid y] = \frac{y+\alpha}{n+\alpha+\beta} \qquad var(p \mid y) = \frac{(y+\alpha)(n-y+\beta)}{(n+\alpha+\beta)^2(n+\alpha+\beta+1)}$$

where y is the observed number of successes and n is the total number of observations. This new distribution is said to be Beta-Binomial.

3 Gibbs sampler for univariate Beta-Binomial model

Let us briefly introduce the general Gibbs sampling scheme. Suppose that we have a set of parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$. Then, each sampled parameter is obtained from its full conditional distribution (or less frequently from marginal distributions, as it is the case here) where the latest updated parameters are used and basic Gibbs sampling proceeds as follows:

1. Give initial values for the parameters, that is $\theta_1^{(0)}, ..., \theta_p^{(0)}$

2. At iteration i, sample each parameter from its marginal or full conditional distribution, that is

$$\begin{aligned}
\theta_1^{(i)} &\sim p(\theta_1 \mid \theta_2^{(i-1)}, \theta_3^{(i-1)}, ..., \theta_p^{(i-1)}) \\
\theta_2^{(i)} &\sim p(\theta_2 \mid \theta_1^{(i)}, \theta_3^{(i-1)}, ..., \theta_p^{(i-1)}) \\
&\vdots \\
\theta_p^{(i)} &\sim p(\theta_p \mid \theta_1^{(i)}, \theta_3^{(i)}, ..., \theta_{p-1}^{(i)})
\end{aligned}$$

A possible implementation of a Gibbs sampler for a Beta-Binomial model could be as follow:

1. Specify a number of simulations N

2. At iteration $i = 0$, initialize the algorithm

3. At iteration $i \geq 1$, repeat the following until N is achieved

(a) sample $\hat{y}$ from the marginal Binomial distribution

$$\hat{y}^{(i)} \mid p \sim Bin\big(n, \hat{p}^{(i-1)}\big)$$

(b) Sample $\hat{p}$ from the marginal Beta-Binomial distribution

$$\hat{p}^{(i)} \mid y \sim BetaBin\Big(\hat{y}^{(i)} + \alpha, n - \hat{y}^{(i)} + \beta \Big)$$

Provided that the algorithm converged to the target distribution, i.e. the joint posterior $\pi(y,p) \propto \binom{n}{y} p^{y+\alpha+1}(1-p)^{n-y+\beta-1}$, we get N couples $(\hat{y}^{(i)}, \hat{p}^{(i)})$ from which we can compute quantities of interest directly on the joint posterior or on the marginals.

4 Application: outcome of a political voting in Switzerland (implementation in R)

Switzerland has a semi-direct democratic system which allows its citizens to give their opinion on new laws in some circumstances. On June 13th 2021, Swiss electorate will vote on the popular initiative "For a Switzerland without artificial pesticides". Roughly, each Swiss citizen will be able to express his/her opinion (yes or no, accept or reject) on that new law and if a majority is reached ($> 50\%$ of yes voters), the new law will be in force. More on this votaing can be found here; https://www.admin.ch/gov/en/start/documentation/votes/20210613/popula-initiative-for-a-switzerland-without-artificial-pesticides.html

The outcome of the votatng is typically a Binomial experiment in the sense that we have a series of independent Bernoulli trials, with uncertainty on p, the parameter representing the probability of a yes vote. On April 29th, suppose we have a initial study conducted on $1,000$ Swiss citizens revealing that 44% of the people surveyed would accept the popular initiative and therefore vote yes. Let us further suppose that the error margin is about 6%. We could use those information to build a simple hierarchical Bayesian model. The model is as follows:

$$y_i \mid p \sim Bin\big(n, p_i\big)$$

$$p_i \sim Beta\big(\alpha, \beta\big)$$

For $i = 1, ..., N$. Since we have previous information from the initial surey, we define $\alpha = 29.92$ and $\beta = 38.08$ so that the prior distribution on p is centered about 0.44 and have standard deviation of about 6%. In R, we get:

```
> # The initial estimated percentage that a popular initiative is accepted is
> # 0.44% (440 yes on 1000 people surveyed), about 6% of error
>
> alpha = 29.92; beta = 38.08
> alpha/(alpha+beta) # mean centered on 0.44
[1] 0.44
>
> (alpha*beta) / ((alpha+beta)^2 * (alpha+beta+1)) # variance
[1] 0.003571014
> ((alpha*beta) / ((alpha+beta)^2 * (alpha+beta+1)))^0.5 # standard deviation
[1] 0.05975797
> # more of less 6%
>
> # our prior distribution on the proportion of yes voters

library(ggplot2)

> seq=seq(from = 0, to = 1, length = 100)
> q=dbeta(seq, 29.92, 38.08)
>
```

```
> df=data.frame(seq,q)
> ggplot(df, aes(seq)) +
+   geom_line(aes(y=q), colour="red3", size = 1.5) +
+   geom_text(x=0.7, y=5, label="mean = 0.44", size = 6)+
+   geom_text(x=0.7, y=4, label="sd = 0.06", size = 6)+
+   xlab("p")+ylab("density")+
+   xlim(0, 1)+ ylim(0, 7) +
+   ggtitle("Beta(29.92, 38.08) prior")
```

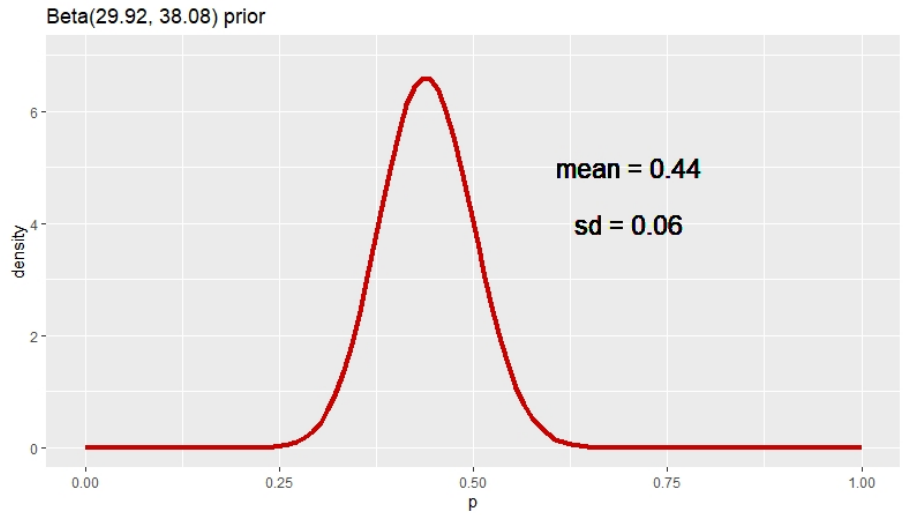

Figure 2: Informative Beta prior with α=29.92 and $\beta = 38.08$.

We are now ready to construct a Gibbs sampler as described in the previous section and run it. We use here $N = 200,000$ simulations, $50,000$ of which will be randomly discarded as a burn in. We use $n = 1,000$, which means that we are simulating outcomes on $1,000$ randomly Swiss citizens surveyed. We end up with a sample of $150,000$ couples of observations $(\hat{y}_i, \hat{p}_i)$ with $i = 1, ..., 150,000$ representing the simulated number of yes voters for corresponding simulated probability of yes votes.

```
> Gibbs_sampler <- function(N, burn, alpha, beta, n) {
+
+   # initialization
+   Y <- matrix(0, N, 2) # empty matrix to record the simulated values
+   Y[1,1] <- round(n*alpha/(alpha+beta)) # initial numerber of yes voters
+   Y[1,2] <- alpha/(alpha+beta) # initial observation the probablity of yes vote
+
+   # run MCMC
+   for(i in 2:N) {
+     Y[i,1] = rbinom(1, size = n, prob = Y[i-1,2]) # sample from the marginal Binomial dist.
+     Y[i,2] = rbeta(1, Y[i, 1] + alpha, n - Y[i, 1] + beta) # sample from marginal BetaBin dist.
+   }
+
+   Y <- Y[-sample((1:N), size = burn, replace = FALSE), ] # discard random observations
+
+   return(data.frame(y=Y[,1], p=Y[,2]))
+ }

> # run the simulations with N=200000 and 50000 randomly discarded observations
> set.seed(1986)
> random_sample <- Gibbs_sampler(N = 200000, burn = 50000,
+                                alpha = alpha,
+                                beta = beta, n = 1000)
```

```
> head(random_sample)
    y         p
1 440 0.4400000
2 455 0.4588798
3 432 0.4201363
4 395 0.3778682
5 375 0.3758049
6 369 0.3856567
> dim(random_sample)
[1] 150000      2
```

Next, we compute the posterior means to check if the chain has converged. We then compute a 95% credible interval for the estimator of the number of yes voters on simulated samples of size $1,000$ as we as for the parameter p. The results are displayed in the chunk below. In addition, we compute the probability that the popular initiative is accepted, which is about 0.16 or 16%.

```
> # check posterior means
> mean(random_sample$y);mean(random_sample$p)
[1] 438.9435
[1] 0.4389981
>
> # posterior 95% credible interval for y and p
> quantile(random_sample$y, probs = c(0.025,0.5,0.975))
 2.5%   50% 97.5%
  320   438   561
> quantile(random_sample$p, probs = c(0.025,0.5,0.975))
     2.5%       50%     97.5%
0.3242231 0.4385382 0.5563216
>
> # probability that the popular initiative is accepted (majority of the voters)
> length(which(random_sample$y > 500 )) / (length(random_sample$y))
[1] 0.1606733
```

We can now plot the joint posterior with the code below (in red is the proportion of samples of size $1,000$ for which we recorded more than 500 yes votes.

```
library(gridExtra)

htop <- ggplot(data=random_sample, aes(x=y)) +
  geom_histogram(aes(y=..density..), fill = "grey90", color = "black", binwidth = 0.3) +
  stat_density(colour = "red3", geom="line", size = 1.2, position="identity", show.legend=FALSE) +
  theme(axis.title.x = element_blank(),
        panel.background = element_blank())

blank <- ggplot() + geom_point(aes(1,1), colour="white") +
  theme(axis.ticks=element_blank(), panel.background=element_blank(), panel.grid=element_blank(),
        axis.text.x=element_blank(), axis.text.y=element_blank(), axis.title.x=element_blank(),
        axis.title.y=element_blank())

scatter <- ggplot(data=random_sample, aes(x=y, y=p)) +
  geom_point(size = 0.6, pch=3) +
  geom_smooth(method = "lm",se = FALSE, color="red4") +
  geom_point(data=random_sample2,
             aes(x=y,y=p),
             color='red',
             size=1)
theme_light()
```

```
hright <- ggplot(data=random_sample, aes(x=p)) +
  geom_histogram(aes(y=..density..), fill = "grey90", color = "black", binwidth = 0.05) +
  stat_density(colour = "red3", geom="line", size = 1.2, position="identity", show.legend=FALSE) +
  coord_flip() + theme(axis.title.y = element_blank(),
                       panel.background = element_blank())

grid.arrange(htop, blank, scatter, hright, ncol=2, nrow=2, widths=c(4, 1), heights=c(1, 4))
```

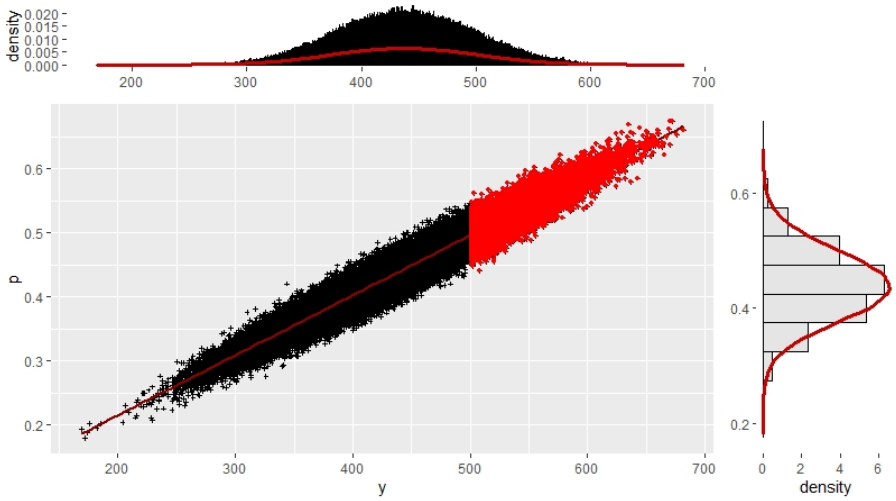

Figure 3: Joint posterior distribution.

Let us briefly display traceplots of the conditional chains and the convergence to the joint posterior to ensure that our results are trustworthy. We can use the following code, which produces the figures 4 and 5.

```
> # check for convergence of the chain using traceplots
> par(mfrow=c(2,1))
> plot(x=1:length(random_sample$y), y=random_sample$y, type="l")
> plot(x=1:length(random_sample$p), y=random_sample$p, type="l")
>
> # Bayesian convergence to the joint posterior
> y1 <- random_sample$y[1:500]
> p1 <- random_sample$p[1:500]
>
> par(mfrow=c(1,1))
> plot(y1, p1, type='l', main='Bayesian convergence to the joint posterior distribution',
+      xlab=expression(y[1]), ylab=expression(p[1]))
```

Eventually, we can plot the marginal distributions of the simulated number of yes voters in samples of size $1,000$ and the proportion of yes (figures 6 and 7).

```
> # Histogram y
> qplot(random_sample$y, geom="histogram",
+       fill=..count.., bins = length(seq(0.3,1,0.01))) +
+   scale_fill_gradient(low="firebrick1", high="firebrick4") +
+   ggtitle("Histogram of simulated number of yes voters on random samples of size n=1000") +
+   annotate(geom="text", x=600, y=5000, label="0.1606",
+            color="black") +
+   theme(plot.title = element_text(size=10),
+         axis.text.x = element_text(size = 10),
+         axis.title=element_text(size=8)) +
+   scale_x_continuous(name = "Number of yes voters") +
```

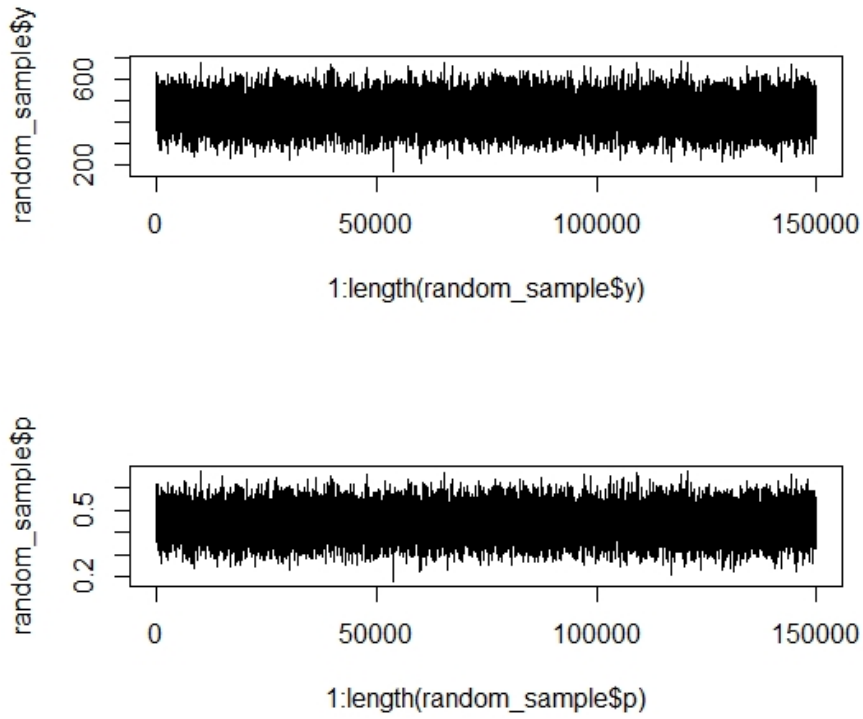

Figure 4: Traceplots to check the convergence of the chain.

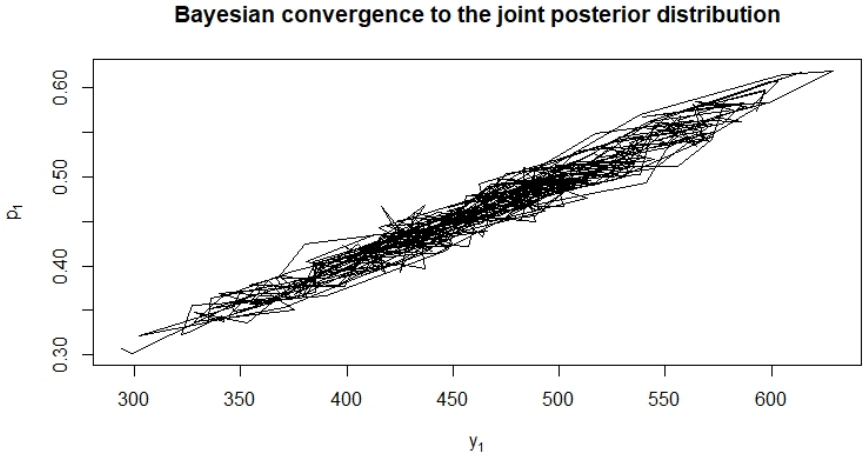

Figure 5: Convergence of $(\hat{y}, \hat{p})$

```
+   geom_vline(xintercept=500, color="black", size = 1.2,
+              linetype="dashed")
>
> # Histogram p
> qplot(random_sample$p, geom="histogram",
+       fill=..count.., bins = length(seq(0.3,1,0.01))) +
+   scale_fill_gradient(low="firebrick1", high="firebrick4") +
+   ggtitle("Histogram of simulated number of outcome p on random samples of size n=1000") +
+   annotate(geom="text", x=0.7, y=4000, label="0.1551",
```

```
+            color="black") +
+   theme(plot.title = element_text(size=10),
+         axis.text.x = element_text(size = 10),
+         axis.title=element_text(size=8)) +
+   scale_x_continuous(name = "percentage of yes") +
+   geom_vline(xintercept=0.5, color="black", size = 1.2,
+              linetype="dashed")
```

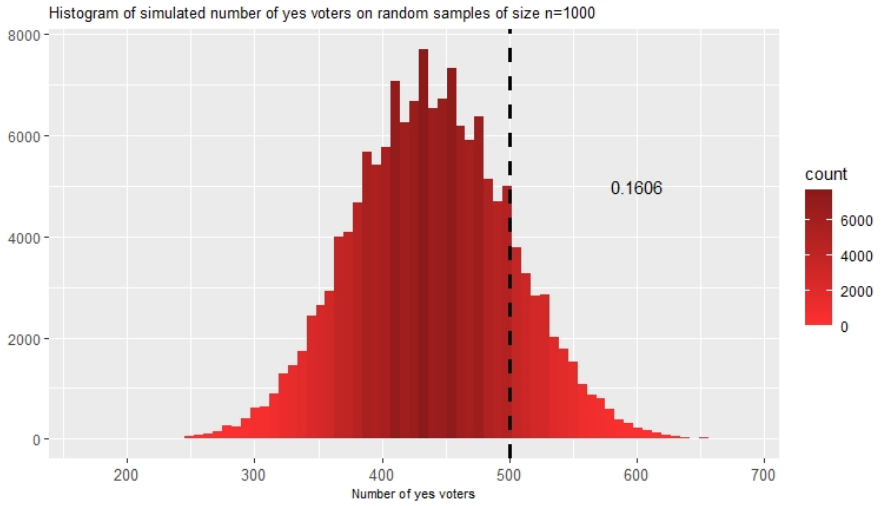

Figure 6: Conditional density of y

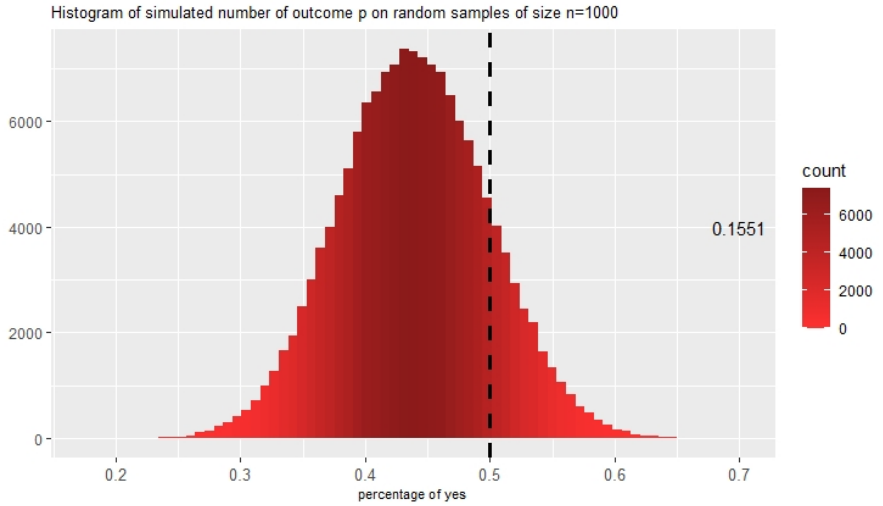

Figure 7: Conditional density of p.

The analysis should best be repeated as we gain more knowledge about the popular initiative (new surveys, expert opinions,...).

Julian Righ Sampedro, Sorens, 29.04.2021