Atividade 1 - ME524

Vitor Ribas Perrone

RA: 245040

Campinas, 2024

1 Introdução

Em Silveira et al. (2009), foi realizado um estudo para determinar o número de onças no Parque Nacional da Serra da Capivara, que possui área de $524km^2$. A partir dos dados coletados, o estudo será replicado utilizando uma abordagem com inferência Bayesiana e MCMC.

2 Primeira Abordagem

Após a coleta dos dados ao longo dos 6 dias em cada uma das 14 armadilhas, sendo n_i o número de indivíduos capturados pela armadilha i e m_j o número de indivíduos capturados pela armadilha j que já haviam sido catalogados previamente, os resultados obtidos foram:

Tabela 1: Dados da Coleta em cada uma das Armadilhas

i/j	1	2	3	4	5	6	7	8	9	10	11	12	13	14
n_i	2	4	2	4	4	2	3	4	2	3	5	10	2	5
m_j		0	2	4	4	2	3	2	2	3	5	6	2	5

A partir da Tabela 1, é possível computar o número de indivíduos registrados, dado por

$$r = \sum_{i=1}^{14} n_i - \sum_{j=2}^{14} m_j = 12$$

Com a coleta de dados feita, inicia-se a etapa inferencial, pois a função de verossimilhança é dada por

$$\mathcal{L}(N, p_1, \ldots, p_{14}|n_1, \ldots, n_{14}, m_2, \ldots, m_{14}) \propto \frac{N!}{(N-r)!} \prod_{i=1}^{14} p_i^{n_i}(1-p_i)^{N-n_i} \mathbf{1}_{\{0,1,2,3,\ldots\}}(\mathrm{N}) \mathbf{1}_{[0,1]}(p_i)$$

e as distribuições a priori $N \sim Poisson(\lambda)$ e $p_i \sim U(0,1)$, dadas por

$$\pi(N) = \frac{e^{-\lambda}\lambda^N}{N!} \mathbf{1}_{\{0,1,2,3,\ldots\}}(\mathrm{N}) \; ; \pi(p_i) = \mathbf{1}_{[0,1]}(p_i)$$

Sendo assim, é possível obter a distribuição a posteriori $f(N, p_1, ..., p_{14}|n_1, ..., n_{14}, m_2, ..., m_{14})$,

$$f(N, p_1, ..., p_{14}|n_1, ..., n_{14}, m_2, ..., m_{14}) \propto \mathcal{L}(N, p_1, \ldots, p_{14}|n_1, \ldots, n_{14}, m_2, \ldots, m_{14})\pi(N) \prod_{i=1}^{14} \pi(p_i)$$

e, a partir dela, é possível obter as condicionais completas de $N, p_1, ..., p_{14}$ ao selecionar apenas os termos que dependem de cada variável em questão.

Antes de propriamente calcular a distribuição a posteriori, é necessário realizar um adendo sobre a indicadora de N, que na função de verossimilhança é dada por $\mathbf{1}_{\{0,1,2,3,\ldots\}}(\mathrm{N})$. Entretanto após realizar uma coleta de dados, não faz mais sentido considerar que os valores que N pode assumir começam do 0, pois algumas onças já foram observadas, então a indicadora da distribuição a posteriori de N é dada por $\mathbf{1}_{\{r,r+1,r+2,\ldots\}}(\mathrm{N})$.

Logo, a distribuição a posteriori é:

$$f(N, p_1, ..., p_{14}|n_1, ..., n_{14}, m_2, ..., m_{14}) \propto \frac{N!}{(N-r)!} \prod_{i=1}^{14} p_i^{n_i}(1-p_i)^{N-n_i} \frac{e^{-\lambda}\lambda^N}{N!} \mathbf{1}_{\{r,r+1,r+2,\ldots\}}(\mathrm{N}) \mathbf{1}_{[0,1]}(p_i)$$

Obtida a distribuição a posteriori, são selecionados os núcleos de cada condicional completa:

Como $(1-p_i)^{N-n_i} = \dfrac{(1-p_i)^N}{(1-p_i)^{n_i}}$,

$$\pi(N|p_1, ..., p_{14}, n_1, ..., n_{14}, m_2, ..., m_{14}) \propto \frac{N!}{(N-r)!} \prod_{i=1}^{14} (1-p_i)^N \frac{\lambda^N}{N!} \mathbf{1}_{\{r,r+1,r+2,\ldots\}}(\mathrm{N})$$

$$\iff \pi(N|p_1, ..., p_{14}, n_1, ..., n_{14}, m_2, ..., m_{14}) \propto \frac{\left[\lambda \prod_{i=1}^{14} (1-p_i)\right]^N}{(N-r)!} \mathbf{1}_{\{r,r+1,r+2,\ldots\}}(\mathrm{N})$$

Portanto, a condicional completa de N segue uma distribuição de Poisson Truncada iniciando no r e com taxa $\lambda \prod_{i=1}^{14}(1-p_i)$. Entretanto, para facilitar a implementação, é análogo considerar que $N - r \sim Poisson\left(\lambda \prod_{i=1}^{14}(1-p_i)\right)$. Logo, para amostrar N, basta obter uma amostra de $Poisson\left(\lambda \prod_{i=1}^{14}(1-p_i)\right)$ e somar $r = 12$.

Já para os p_i , a condicional completa é dada por

$$\pi(p_i|N, p_{\{k:k\neq i\}}, n_1, ..., n_{14}, m_2, ..., m_{14}) \propto p_i^{n_i}(1-p_i)^{N-n_i}\mathbf{1}_{[0,1]}(p_i)$$

Portanto, a condicional completa de p_i segue uma distribuição Beta$(n_i + 1, N - n_i + 1)$ $\forall\ i \in \{1, 2, ..., 14\}$.

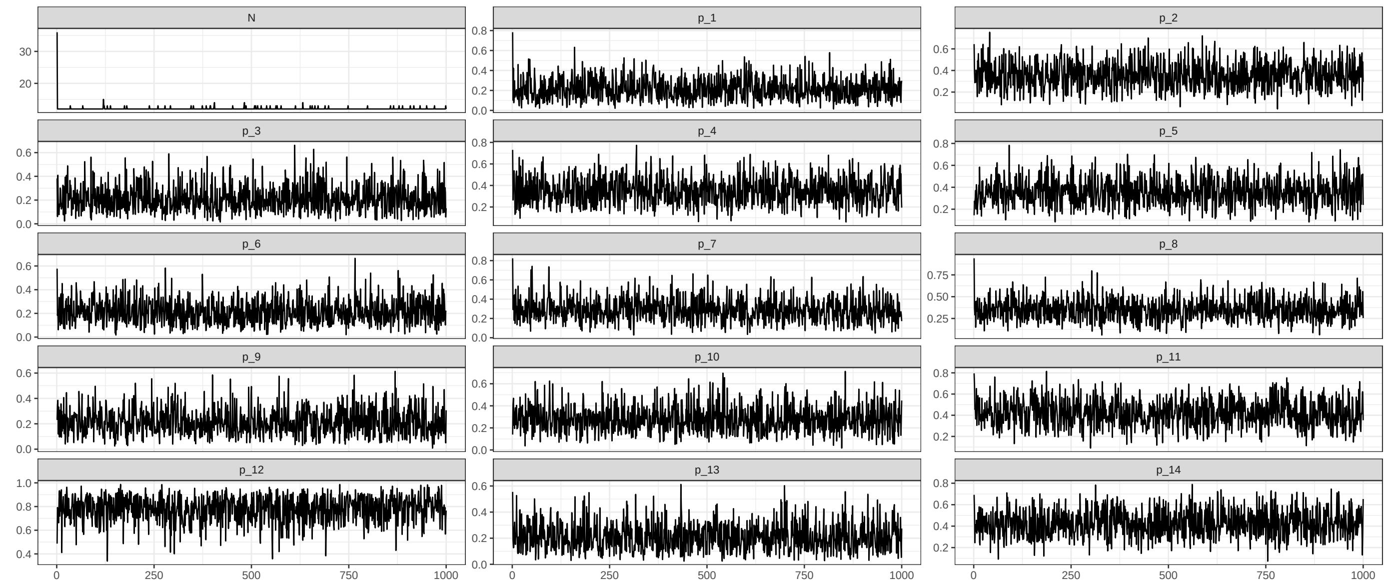
Com todas as contas devidamente realizadas, é possível obter amostragens da distribuição a posteriori por meio do Amostrador de Gibbs, isto é, amostrando sequencialmente as variáveis aleatórias de interesse até atingir convergência da seguinte maneira:

Partindo de $N^{(0)}, p_1^{(0)}, ..., p_{14}^{(0)}$ gerados pelas prioris, o k-ésimo passo do amostrador é:

$$\begin{gathered}
N^{(k)} \sim Poisson\left(\lambda \prod_{i=1}^{14}(1-p_i^{(k-1)})\right) + r \\
p_1^{(k)} \sim Beta(n_1 + 1, N^{(k)} - n_1 + 1) \\
... \\
p_{14}^{(k)} \sim Beta(n_{14} + 1, N^{(k)} - n_{14} + 1)
\end{gathered}$$

Considerando $\lambda = 30$, é possível gerar amostras da Posteriori. Gerando $N^{(0)}, p_1^{(0)}, ..., p_{14}^{(0)}$ a partir das prioris, é possível analisar a convergência do algoritmo por meio da Figura 1, em que foi amostrada uma sequência de 1000 termos e foram realizados gráficos de linhas. Sobre o N, fica bem visível a convergência muito rápida para valores muito próximos de 12, mas na maioria das vezes ainda 12, o que se dá ao fato da taxa ser muito baixa devido ao produtório. Para os p_i's, não é tão imediato visualizar a convergência, especialmente pelo fato da amplitude de valores possíveis ser baixa, mas em cada gráfico existe uma tendência central que está sendo mantida, tornando razoável assumir que está convergindo para a distribuição de interesse.

Figura 1: Gráficos de Linhas de 1 Amostra Algoritmo de Gibbs com Prioris Uniformes



3 Segunda Abordagem

Com a primeira implementação realizada, é possível refazer o estudo com outra abordagem, agora considerando o fato de que onças são animais mais espertos. Ao escolher distribuições uniformes como prioris para as probabilidades de captura de cada armadilha, está se assumindo que a probabilidade de captura tem peso igual no intervalo [0,1] e que o valor esperado é $1/2$. Assim sendo, uma abordagem possível é considerar as prioris como Betas, em particular, foi feita a escolha da Beta(2,5), pois como b > a, a distribuição é concentrada na parte esquerda e sua esperança é $\dfrac{a}{a+b} = \dfrac{2}{7}$, levando em consideração o comportamento dos animais.

Na nova abordagem, a densidade de cada p_i é dada por

$$\pi(p_i) = p_i^{2-1}(1-p_i)^{5-1}\mathbf{1}_{[0,1]}(p_i)$$

e, com a alteração das prioris, a distribuição a posteriori agora é dada por

$$f(N, p_1, ..., p_{14}|n_1, ..., n_{14}, m_2, ..., m_{14}) \propto \frac{N!}{(N-r)!}\prod_{i=1}^{14} p_i^{n_i}(1-p_i)^{N-n_i}\frac{e^{-\lambda}\lambda^N}{N!}\prod_{i=1}^{14} p_i^{2-1}(1-p_i)^{5-1}\mathbf{1}_{\{r,r+1,r+2,...\}}(\mathrm{N})\mathbf{1}_{[0,1]}(p_i)$$

Com isso, a condicional completa de N não passa por nenhuma alteração, mas as condicionais completas de p_i agora são da forma

$$\pi(p_i|N, p_{\{k:k\neq i\}}, n_1, ..., n_{14}, m_2, ..., m_{14}) \propto p_i^{n_i}(1-p_i)^{N-n_i}p_i^{2-1}(1-p_i)^{5-1}\mathbf{1}_{[0,1]}(p_i)$$
$$\iff \pi(p_i|N, p_{\{k:k\neq i\}}, n_1, ..., n_{14}, m_2, ..., m_{14}) \propto p_i^{n_i+2-1}(1-p_i)^{N-n_i+5-1}$$

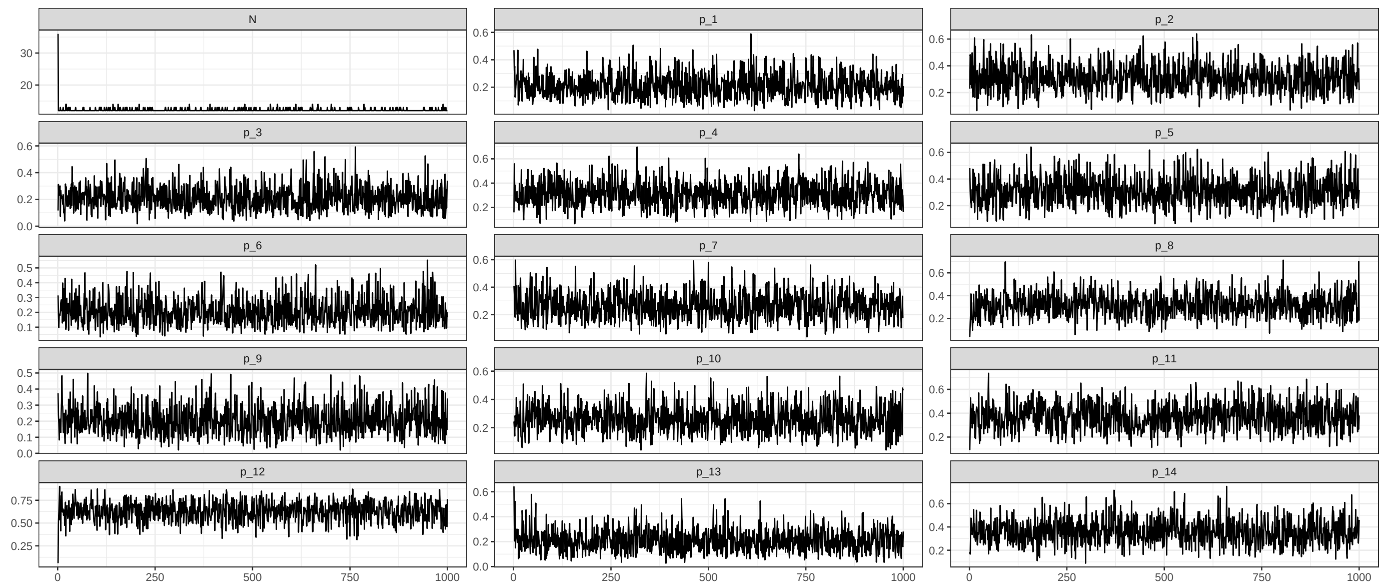
Sendo assim, as posteriori de p_i agora seguem uma distribuição Beta$(n_i + 2, N - n_i + 5)$ $\forall\ i \in \{1, 2, ..., 14\}$.

Com todas as contas devidamente realizadas, é possível novamente obter amostras por meio do Amostrador de Gibbs. Partindo de $N^{(0)}, p_1^{(0)}, ..., p_{14}^{(0)}$ gerados pelas prioris, o k-ésimo passo do amostrador é:

$$N^{(k)} \sim Poisson\left(\lambda\prod_{i=1}^{14}(1-p_i^{(k-1)})\right) + r$$
$$p_1^{(k)} \sim Beta(n_1 + 2, N^{(k)} - n_1 + 5)$$
$$...$$
$$p_{14}^{(k)} \sim Beta(n_{14} + 2, N^{(k)} - n_{14} + 5)$$

Considerando $\lambda = 30$ como no primeiro caso, é possível fazer uma análise análoga para determinar a convergência por meio dos gráficos presentes na Figura 2. Sobre N, também converge rapidamente para valores em torno de 12, mas existe uma probabilidade maior de existirem mais de 12 onças em comparação ao caso anterior. Já sobre os p_i's, a convergência se dá de maneira similar ao caso anterior, o que deixa bem razoável assumir que está convergindo.

Figura 2: Gráficos de Linhas 1 Amostra Algoritmo de Gibbs com Prioris Betas



4 Comparação dos Métodos

Como ambas as implementações estão convergindo, a fim de realizar uma comparação mais específica entre os diferentes resultados obtidos, foram geradas 500 amostras de cada abordagem pelo Algoritmo de Gibbs e realizados histogramas, contidos nas figuras 3 e 4.

A diferença de abordagem nos p_i's implica de maneira significante o resultado ao comparar as posterioris de cada um deles. Afinal, ao supor inicialmente que as probabilidades de captura são baixas, isso se reflete na posteriori tornando a distribuição mais concentrada em valores baixos, especialmente quando os dados também se comportam de tal maneira.

Essa mudança implica diretamente no valor principal que desejamos descobrir com o estudo, o total da população N, visto que quanto mais difícil capturar os animais, maior a probabilidade de existirem mais que não foram catalogados. Isso fica bem evidente ao comparar os histogramas de N, em que um deles mal apresenta amostras valendo 13 e 14 enquanto no outro a representação está maior, mesmo que não tanto.

Figura 3: Histogramas 500 Amostras Algoritmo de Gibbs com Prioris Uniformes

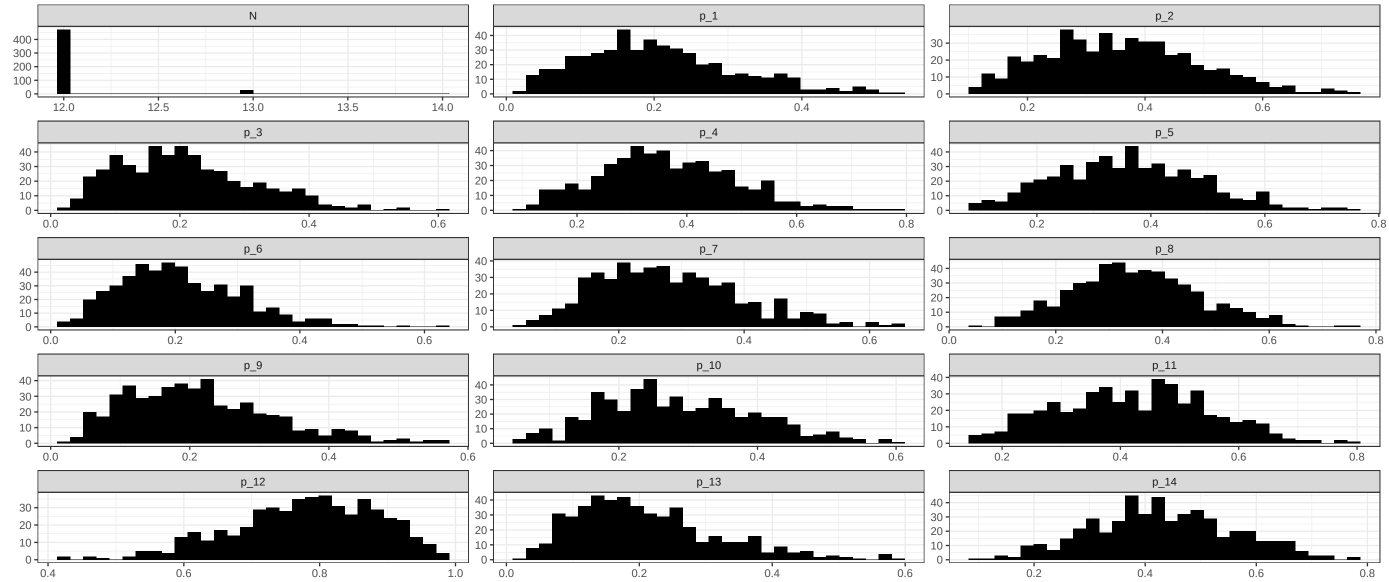
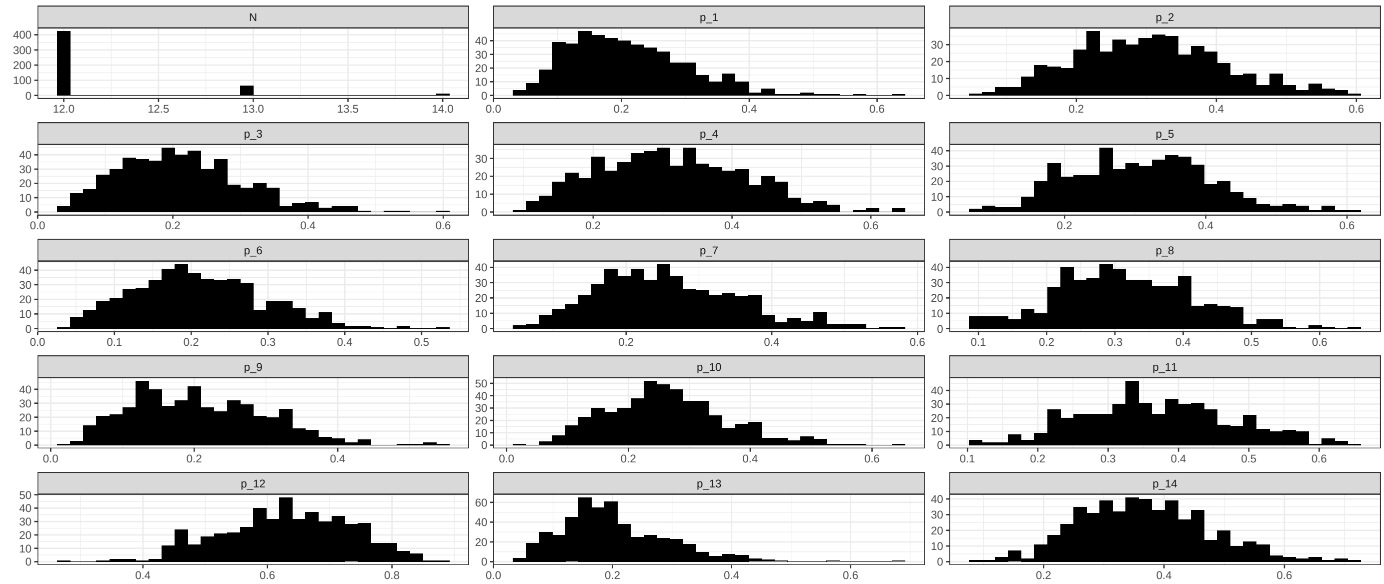


Figura 4: Histogramas 500 Amostras Algoritmo de Gibbs com Prioris Betas



Por fim, a conclusão do estudo é que em ambas as abordagens, a maior probabilidade é de que todas as onças foram capturadas e de fato só existem 12 onças, isto é, $N = 12$. Entretanto, existe uma probabilidade de que existam mais, mas muito dificilmente passa das 14, mesmo na abordagem mais otimista.

Em Silveira et al. (2009), o resultado obtido foi "14 ± SE 3.6 jaguars in an area of 524 km^2". Ao analisar os resultados obtidos com as duas abordagens implementadas com MCMC, eles estão presentes nesse intervalo de confiança, então é possível dizer que os estudos foram complementares.

Entretanto, na abordagem implementada com MCMC, não era possível obter valores de N abaixo de 12, então pode-se dizer que para obter uma generalização para demais populações, a abordagem de Silveira et al. (2009) é mais eficaz, enquanto o estudo por MCMC foi mais direcionado em estimar o total de onças dos dados coletados.

5 Códigos

Todas as implementações foram realizadas em linguagem R por meio do Software RStudio e todo o código produzido segue abaixo com comentários para facilitar a interpretação. Além disso, por motivos de reprodutibilidade, foi fixada a semente 524 antes de cada etapa que envolvia geração de números pseudo aleatórios.

```r
library(tidyverse)

#Dados Coletados
n <- c(2,4,2,4,4,2,3,4,2,3,5,10,2,5)
m <- c(0,2,4,4,2,3,2,2,3,5,6,2,5)
r <- sum(n) - sum(m)

#Funcao que Executa o Amostrador de Gibbs no Primeiro caso em Questao
amostradorGibbs <- function(k) {

  #Matriz em que serao Guardados Resultados das Iteracoes
  matriz <- matrix(NA, nrow = k, ncol = 15)

  #Amostras Iniciais Baseadas na Priori
  matriz[1, 1] <- rpois(1, 30)
  matriz[1,2:15] <- runif(14, 0, 1)

  for (i in 2:k) {
    matriz[i, 1] <- rpois(1, 30*prod(1 - matriz[i-1, 2:15])) + 12
    matriz[i, 2:15] <- sapply(n, function(ni) rbeta(1, ni + 1, matriz[i,1] - ni + 1)) }
  return(matriz) }

#Implementando uma Amostra
set.seed(524)
amostra <- amostradorGibbs(1000)
colnames(amostra) <- c("N", paste0("p_", 1:14))
as.data.frame(amostra) %>% mutate("indice" = 1:1000) %>%
  pivot_longer(1:15, names_to = "Var") %>%
  mutate(Var = factor(Var, levels = c("N", paste0("p_", 1:14)))) %>%
  ggplot() + geom_line(aes(x = indice, y = value)) +
  facet_wrap(~Var, scales = "free_y", ncol = 3) +
  theme_bw() + labs(x = "", y = "")

#Realizando 500 amostras para analisar
amostras <- matrix(NA, nrow = 500, ncol = 15)
set.seed(524)
for (i in 1:500) amostras[i,] <- amostradorGibbs(1000)[1000,]

colnames(amostras) <- c("N", paste0("p_", 1:14))
as.data.frame(amostras) %>% pivot_longer(1:15, names_to = "Var") %>%
  mutate(Var = factor(Var, levels = c("N", paste0("p_", 1:14)))) %>%
  ggplot() + geom_histogram(aes(x = value), fill = "black") +
  facet_wrap(~Var, scales = "free", ncol = 3) +
  theme_bw() + labs(x = "", y = "")

#Amostrador de Gibbs no segundo caso em Questao
amostradorGibbs2 <- function(k) {

  #Matriz em que serao Guardados Resultados das Iteracoes
  matriz <- matrix(NA, nrow = k, ncol = 15)

  #Amostras Iniciais Baseadas na Priori
  matriz[1, 1] <- rpois(1, 30)
  matriz[1,2:15] <- rbeta(14, 2, 5)

  for (i in 2:k) {
    matriz[i, 1] <- rpois(1, 30*prod(1 - matriz[i-1, 2:15])) + 12
    matriz[i, 2:15] <- sapply(n, function(ni) rbeta(1, ni + 2, matriz[i,1] - ni + 5)) }
```

```r
  return(matriz) }

#Implementando uma amostra
set.seed(524)
amostra2 <- amostradorGibbs2(1000)
colnames(amostra2) <- c("N", paste0("p_", 1:14))

as.data.frame(amostra2) %>% mutate("indice" = 1:1000) %>%
  pivot_longer(1:15, names_to = "Var") %>%
  mutate(Var = factor(Var, levels = c("N", paste0("p_", 1:14)))) %>%
  ggplot() + geom_line(aes(x = indice, y = value)) +
  facet_wrap(~Var, scales = "free_y", ncol = 3) +
  theme_bw() + labs(x = "", y = "") +


#Realizando 500 amostras para analisar
set.seed(524)
amostras2 <- matrix(NA, nrow = 500, ncol = 15)
for (i in 1:500) amostras2[i,] <- amostradorGibbs2(1000)[1000,]

colnames(amostras2) <- c("N", paste0("p_", 1:14))
as.data.frame(amostras2) %>% pivot_longer(1:15, names_to = "Var") %>%
  mutate(Var = factor(Var, levels = c("N", paste0("p_", 1:14)))) %>%
  ggplot() +
  geom_histogram(aes(x = value), fill = "black") +
  facet_wrap(~Var, scales = "free", ncol = 3) +
  theme_bw() + labs(x = "", y = "")
```

6 Referências

L. Silveira, A. T. Jácomo, S. Astete, R. Sollmann, N. M. Tôrres, M. M. Furtado, and J. Marinho-Filho. Density of the near threatened jaguar Panthera onca in the caatinga of north-eastern Brazil. Oryx, 44(1): 104–109, 2009.