

Consideremos que foi fixada uma semente igual a 1166. O objetivo deste exercício é o de gerar 2500 amostras de tamanho n , para cada $n \in \{30, 50, 100, 200, 300, 500, 1000\}$, de uma distribuição de Bernoulli com parâmetro igual a 0.5. De seguida, usar dois métodos distintos que calculam intervalos de confiança de aproximação para o parâmetro mencionado, o que permite obter a diferença entre as amplitudes desses intervalos. Por fim, calcular a média das 2500 diferenças para cada n . Para tal, recorreu-se ao seguinte trecho de código em R:

```
1 library("ggplot2")
2 library("Rlab")
3
4 SEED <- 1166
5 SAMPLE_COUNT <- 2500
6 BERNOULLI_P <- 0.5
7 CONF_LEVEL <- 0.97
8 N <- c(30, 50, 100, 200, 300, 500, 1000)
9 set.seed(SEED)
10
11 method_1 <- function(samples, conf_level) {
12   len <- length(samples)
13   mean <- mean(samples)
14   z <- qnorm((1 + conf_level) / 2)
15   denom <- 2 * (1 + z**2 / len)
16   upper <- ((2 * mean + z**2 / len) + sqrt(4 * mean * z**2 * (1 - mean) / len + z**4 / len**2)) / denom
17   lower <- ((2 * mean + z**2 / len) - sqrt(4 * mean * z**2 * (1 - mean) / len + z**4 / len**2)) / denom
18   return(abs(upper - lower))
19 }
20
21 method_2 <- function(samples, conf_level) {
22   len <- length(samples)
23   mean <- mean(samples)
24   upper <- mean + qnorm(1 - (1 - conf_level) / 2) * sqrt(mean * (1 - mean) / len)
25   lower <- mean - qnorm(1 - (1 - conf_level) / 2) * sqrt(mean * (1 - mean) / len)
26   return(abs(upper - lower))
27 }
28
29 df <- data.frame()
30 for (n in N) {
31   method_diffs <- c()
32   for (i in 1:SAMPLE_COUNT) {
33     samples <- rbern(n, BERNOULLI_P)
34     diff <- method_2(samples, CONF_LEVEL) - method_1(samples, CONF_LEVEL)
35     method_diffs <- append(method_diffs, diff)
36   }
37   mean_diffs <- mean(method_diffs)
38   df <- rbind(df, data.frame(n = n, difference = mean_diffs))
39 }
40
41 ggplot(df, aes(x = n, y = difference)) +
42   geom_line(color = "#e76f51") +
43   geom_point(color = "#e76f51") +
44   xlab("Dimensão da Amostra") +
45   ylab("Variação das Diferenças Médias") +
46   labs(title = "Relação entre Variação das Diferenças Médias e Dimensão da Amostra",
47        subtitle = sprintf("semente = %d | k = %d | p = %.2f | γ = %.2f",
48                            SEED, SAMPLE_COUNT, BERNOULLI_P, CONF_LEVEL))
```

O gráfico obtido permite concluir que a relação entre as duas variáveis em causa é inversamente proporcional. Deste modo, à medida que o tamanho das 2500 amostras aumenta, os métodos 1 e 2 apresentam, em média, intervalos de confiança com uma amplitude cada vez mais próxima.

As diferenças médias foram obtidas subtraindo a amplitude do método 2 pela a do método 1 para cada amostra e posteriormente calculando a média dessas diferenças. Assim, já que todos os valores no eixo dos yy são positivos, conclui-se que, em geral, para amostras de dimensões mais pequenas o método 1 é mais favorável para aproximar o parâmetro p , pois o seu intervalo de confiança tem menor amplitude. Porém, para amostras de dimensões maiores, ambos os métodos garantem um intervalo de confiança com amplitude semelhante.

