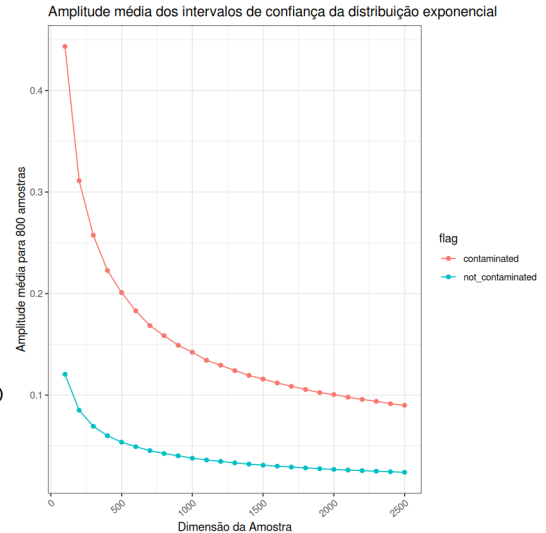


Exercício 10 - Projeto Computacional PE 2022

Diogo Gaspar, 99207

Consideremos como premissas que foram fixadas uma semente em 301 e um conjunto de tamanhos de amostras $\{100, 200, \dots, 2500\}$. O objetivo deste exercício passa por gerar 800 amostras com distribuição exponencial de valor esperado $\frac{1}{\lambda} = \frac{1}{3.13}$ para cada tamanho supra-mencionado. De seguida, substituir 25% das observações de cada amostra por outras geradas de uma população que modela a distribuição dos *outliers*, tal que $\lambda_c = 0.53$. Para cada amostra, construir um intervalo de confiança para o inverso do valor esperado, com nível de confiança $1 - \alpha = 0.94$. Por fim, para cada tamanho de amostra (contaminada e não contaminada), calcular a média da amplitude de todos os intervalos de confiança obtidos. Para tal, recorreu-se ao seguinte trecho de código R (utilizando as bibliotecas *ggplot2*, *dplyr* e *tidyr*):

```
1  set.seed(301)
2  m <- 800
3  lambda_not_contaminated <- 3.13
4  lambda_contaminated <- 0.53
5  alpha <- 1 - 0.94
6  dimensions <- seq(100, 2500, 100)
7  epsilon <- 0.25
8
9  calculate_mean_widths <- function(n) {
10   not_contaminated <- c()
11   contaminated <- c()
12   for (i in 1:m) {
13     contaminated_amount <- floor(n * epsilon)
14     nc_exp <- rexp(n, rate=lambda_not_contaminated)
15     c_exp <- rexp(contaminated_amount, rate=lambda_contaminated)
16     c_exp <- c(c_exp[0:contaminated_amount], nc_exp[contaminated_amount:n])
17
18     nc_upper_bound <- ((1 + qnorm(1-alpha/2)/sqrt(n))/mean(nc_exp))
19     nc_lower_bound <- ((1 - qnorm(1-alpha/2)/sqrt(n))/mean(nc_exp))
20     c_upper_bound <- ((1 + qnorm(1-alpha/2)/sqrt(n))/mean(c_exp))
21     c_lower_bound <- ((1 - qnorm(1-alpha/2)/sqrt(n))/mean(c_exp))
22     not_contaminated <- c(not_contaminated, abs(nc_upper_bound - nc_lower_bound))
23     contaminated <- c(contaminated, abs(c_upper_bound - c_lower_bound))
24   }
25   return(c(mean(not_contaminated), mean(contaminated)))
26 }
27
28 not_contaminated <- c()
29 contaminated <- c()
30 for (n in dimensions) {
31   mean_widths <- calculate_mean_widths(n)
32   not_contaminated <- c(not_contaminated, mean_widths[1])
33   contaminated <- c(contaminated, mean_widths[2])
34 }
35
36 df = data.frame(dimensions, not_contaminated, contaminated)
37 df <- rename(df, "Não Contaminado" = "not_contaminated", "Contaminado" = "contaminated")
38 df <- pivot_longer(df, "Não Contaminado":"Contaminado")
39 df <- rename(df, "contaminated" = name, mean_widths = value)
40
41 ggplot(df, aes(x = dimensions, y = mean_widths, colour = contaminated)) +
42   geom_line() +
43   geom_point() +
44   labs(x = "Dimensão da Amostra", y = "Amplitude média para 800 amostras") +
45   ggtitle("Amplitude média dos intervalos de confiança da distribuição exponencial") +
46   theme_bw() +
47   scale_colour_brewer(palette = "Set1") +
48   theme(axis.text.x = element_text(angle = 40, hjust=1))
```



Note-se que ambas as curvas, para amostras não contaminadas e contaminadas, seguem destinos semelhantes: começam relativamente elevadas, eventualmente acabando por começar a estabilizar próximo de 2500. Mais, note-se que amostras com indivíduos não contaminados apresentam amplitude média razoavelmente maior, levando portanto à conclusão de que amostras com indivíduos contaminados têm maior grau de confiança (tal deve-se, também, ao facto de $\lambda > \lambda_c$).