Estatística Computacional - Lista 3

Questão 1. Indique o algoritmo para construir geradores de números aleatórios (partindo de números com distribuição uniforme [0,1]) para

a) Distribuição de Pareto (densidade $f(x|\alpha,\beta) = \frac{\beta\alpha^\beta}{x^{(\beta+1)}}\mathbf{I}_{(\alpha,\infty)}(\mathbf{x})$)

b) Distribuição Gumbel padrão(densidade $f(x) = e^{-(x+e^{-x})}$)

c) Distribuição F

d) Distribuição binomial negativa

e) Distribuição bivariada com parâmetros $p \in (0,1)$ e $n \in \{1,2,\cdots\}$ e com função massa conjunta

$$p(x,y|p) = \left(\begin{array}{c} n \\ x \end{array}\right)\left(\begin{array}{c} n \\ y \end{array}\right)\frac{x^y(n-x)^{n-y}p^x(1-p)^{n-x}}{n^n}, \qquad x,y \in \{0,1,\cdots,n\}.$$

obs: Não é necessário entregar a implementação, mas é interessante realizá-la para fins de estudo e conferência de resultados.

Questão 2. Se a variável aleatória X é um modelo de escala e locação, então sua função densidade tem a forma $f(x|a,b) = (1/b)f_0((x-a)/b)$, em que $f_0(x)$ é uma densidade que depende apenas de x, e não dos parâmetros de escala b e locação a.

a) Mostre que a geração de X , de uma variável X_0 com distribuição de f_0 é dada por $X = a + bX_0$.

b) Aplique o resultado do item a para gerar valores com distribuição $U[a-1/2; a+1/2]$, $U[0;b]$, e $U[a;b]$ a partir da $U[0;1]$.

Questão 3. Sejam $U_1, U_2, \cdots$ variáveis independentes com distribuição uniforme padrão. Seja

$$X = \{\text{menor n tal que } \sum_{i=1}^{n} U_i > 1\}$$

Utilize uma simulação de Monte Carlo para

a) Estimar $E(X)$. Qual o erro padrão desta estimativa?

b) Estimar $P(X \geq 10)$. Qual o erro padrão desta estimativa?

c) Estimar $P(X = 10)$. Qual o erro padrão desta estimativa?

(dica: utilize alguma distribuição como a beta como distribuição de importância)

Questão 4. Sejam

$$X = (6.2, 5.1, 7.6, 2.5, 3.5, 9.4, 4.1, 6.3, 3.0, 0.8)$$
$$Y = (6.9, 5.1, 7.5, 11.1, 10.9, 4.2, 10.5, 6.8, 12.3, 14.3)$$

e considere o modelo de regressão linear $Y = \beta_0 + \beta_1 X + \epsilon$. Sejam $\hat{\beta}_1$ e $\hat{\beta}_2$ os estimadores de mínimos quadrados de β_0 e β_1.

a) Utilize o bootstrap para construir um intervalo de confiança 95% para $\hat{\beta}_1$ e $\hat{\beta}_2$.

b) Assumindo que X é dado e $\epsilon \sim N(0,\sigma^2)$, utilize o bootstrap paramétrico para obter as mesmas estimativas.

c) Utilize um teste de permutação para verificar se $\beta_1 = 0$. É possível usar um teste de permutação para $\beta_0 = 0$?

Questão 5. Considere o exemplo de aula (e script do R) para a utilização de cadeias de Markov ocultas para modelar retornos da bolsa de valores. Nesse exemplo temos uma cadeia de Markov a tempo discreto modelando as transições entre 4 estados ocultos. Nesse caso, X_t representa o estado oculto da cadeia no dia $t \in \{1, \cdots, n\}$, $\pi = (\pi_1, \cdots, \pi_4)$ o vetor de probabilidades iniciais, tal que $P(X_1 = i) = \pi_i$, e $P = [p_{ij}]$ a matriz de transição da cadeia. Além disso, sejam Y_t as variáveis observáveis do processo, representadas pelos log retornos. As probabilidades de emissão $\phi(y_t|i) = f(Y_t = y_t|X_t = i)$ são tais que $Y_t|X_t = i \sim N(\mu_i, \sigma_i^2)$, para $i = 1, 2, 3, 4$.

1. Escolha valores inicias para os parâmetros do modelo P, σ_i, com $i = 1, 2, 3, 4$. Tome todos os valores de $\pi_i = 1/4$ e $\mu_i = 0$. Utilizando como base apenas o gerador de números (pseudo)aleatórios do R (runif), simule valores de Y_t, $t = 1, \cdots, 100$ para esta cadeia de Markov oculta. Descreva o algoritmo utilizado. (Dica: escolha valores distintos para os $\sigma's$ para facilitar a identificabilidade do seu modelo)

2. Utilize a função indicada no script de aula para estimar os parâmetros do modelo. Como as estimativas se comparam com os valores utilizados para a simulação?

3. Realize uma simulação de Monte Carlo para estimar a esperança dos estimadores por EM implementados nessa função. O que você pode afirmar sobre seu viés? (Dica: como os componentes de mistura não tem um ordenamento inerente, é importante reordenar os componentes dos vetores e matriz de transição para permitir a comparação. Uma sugestão é nomear os estados em ordem crescente de σ.)

   obs: Nessa questão você não deve utilizar outros geradores de números aleatórios que não os da função runif.

Questão 6. (Questão bônus) Na biologia molecular, o modelo Jukes Cantor é utilizado para descrever a probabilidade de mutações no DNA. Considere que temos uma amostra de N sequencias de DNA $X_1, \cdots, X_N$, em que cada sequência $X_i = (X_{i1}, \cdots, X_{iL})$ tem comprimento L. Os elementos dessas sequencias $X_{ij} \in \{A, G, C, T\}$. O número de mutações (alterações) que uma sequencia sofre em um tempo t tem distribuição Poisson(μtL), em que μ é chamada de taxa de mutação do modelo. O modelo também assume que todas as posições da sequência tem a mesma probabilidade de sofrer uma destas mutações, e que as mutações para todas as bases tem igual probabilidade.

Note, entretanto, que a evolução dos organismos (e portanto de seu DNA) não é completamente independente, pois os organismos descendem todos de um ancestral comum. A árvore da figura 1 representa essa relação. Nela os nós, numerados de 1 a 7, representam os organismos observados/atuais (1 a 4) e não observados/do passado (5 a 7). Os segmentos que ligam esses nós são chamados de ramos. Podemos interpretar a árvore da seguinte forma: O nó 7 representa o ancestral comum a toda a toda a árvore. Esse deu origem a duas linhagens que evoluíram de modo independente dando origem aos nós 4 e 6. Já o nó 6 deu origem a duas linhagens que evoluíram de modo independente para dar origem aos nós 3 e 5, e assim por diante. Os $\tau_1, \cdots \tau_6$ representam o tempo passado ao longo do seu respcetivo ramo da árvore.

Considere uma sequência de DNA aleatória de comprimento $L = 1000$ evoluindo ao longo da árvore da figura 1, para dar origem às sequências dos nós 1 a 7. Em cada ramo dá árvore, ela sofre mutações de acordo com o modelo Jukes Cantor, de modo condicionalmente independente dos demais ramos. Tome $\tau_i = 1$, para todo i e $\mu = 0.01$. Monte um estudo de Monte Carlo para estimar

a) Valor esperado e variância da proporção de posições na sequência em há variabilidade nos valores de X (bases) para os nós 1 a 4 (sequencias observadas). [chamamos isso de proporção de sítios polimórficos]

b) Valor esperado e variância da proporção de posições na sequência em que observamos pelo menos 3 valores distintos para X nos nós 1 a 4 (sequencias observadas).

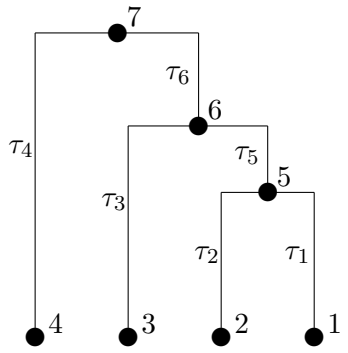

Figure 1: Exemplo de árvore filogenética com $N = 4$.