Complementos de Estatística

LMAC, 2ºSemestre, 2021/2022

Projeto Computacional em R

- **Data de distribuição: 4 de abril 2022.**
- **Data de entrega: 20 de abril 2022 (até às 14:30h).**

Em todas as questões do projeto, a geração de números pseudo-aleatórios da $U(0,1)$ deve ser fixada com a instrução **set.seed(1)**. A interpretação e análise dos resultados obtidos devem ser baseadas em estatísticas amostrais sumárias (médias, quantis, variâncias, desvios padrão, amplitudes inter-quartis, erros quadrático médios, ...) e em representações gráficas (histogramas, distribuição empíricas, qq-plots, ...) preferencialmente obtidas com o pacote **ggplot2**.

1. Admitam um vetor aleatório $\mathbf{X} = (X_1, X_2, X_3) \sim Multinomial(n, p_1, p_2, p_3)$. A geração de $\mathbf{x}$ pode ser implementada recorrendo a números binomiais, a partir da transformação de números pseudo-aleatórios, que utilizam:

   (i) Somas de v.a's i.i.d. com $W \sim Ber(p)$ (**algoritmo (i)**);

   (ii) Duas variáveis aleatórias $U|(U \leq p)$ e $U|(U > p)$, onde $U \sim U(0,1)$ (**algoritmo (ii)**);

   (iii) Se $\{Y_i\}$ é uma sucessão de v.a's i.i.d. com $Y \sim Geo(p)$ e X é o menor inteiro tal que $\sum_{i=1}^{X+1} Y_i > n$, então $X \sim Bin(n,p)$ (**algoritmo (iii)**).

   Admitam que $n = 10, p_1 = 0.2, p_2 = 0.3$ e $p_3 = 0.5$.

   (a) Para gerar $\mathbf{x}$, implementem os três algoritmos anteriores. Utilizem também o gerador da multinomial disponível no R (**algoritmo (iv)**). Usando estes quatro algoritmos, gerem $\mathbf{x}_i$, para $i = 1, \ldots, m = 10000$, comparando os tempos das gerações. Estimem, para as quatro alternativas de geração, a média e a matriz de covariâncias de $\mathbf{X}$, assim como a $P(X_1 = 0, X_2 = 4, X_3 = 6)$. Comparem as estimativas obtidas com os resultados teóricos desta multinomial. Através de representações gráficas, avaliem qual dos quatro mecanismos de geração melhor se ajusta a esta distribuição multinomial. Qual foi o algoritmo que mostrou ser o mais eficiente para gerar $\mathbf{x}$?

   (b) Desenvolvam um estudo de simulação de Monte Carlo para estimar $\theta = P(\bar{X}_1 > 2)$. Para isso, selecionem um dos três primeiros algoritmos: **algoritmo (i)**, **algoritmo (ii)** ou **algoritmo (iii)**, e gerem $N = 10000$ amostras de dimensão $m = 100$ da variável aleatória $X_1 \sim Bin(10, 0.2)$. Obtenham a estimativa pontual de θ e a intervalar, a aproximadamente 95%:

      (i) sem redução de variância (Monte Carlo usual);

      (ii) com a técnica de redução de variância que utiliza a variável de controlo $U \sim U(0,1)$.

      Comparem, analisem e comentem os resultados obtidos.

2. Considerem o resultado seguinte que relaciona a distribuição Poisson-Gamma com a Binomial Negativa (no conjunto $I\!N_0$) :

   - Se $(X|Y = y) \sim Poi(y)$ e $Y \sim Gama(r, \frac{1-p}{p})$ então $X \sim BinNeg(r, 1-p)$, com $r \in I\!N$.

   Utilizando o resultado anterior, implementem um algoritmo (transformando números pseudo-aleatórios) para gerar $m = 10000$ valores de $X \sim BinNeg(6, 0.3)$ e repitam a mesma geração com o simulador disponível no R. Confrontem os tempos das duas geração, utilizem medidas sumárias e representações gráficas para analisar e comparar as duas gerações.

3. Considerem o par aleatório (X, Y) com função densidade de probabilidade conjunta:

$$f_{X,Y}(x, y) = e^{-x}e^{-y}e^{-e^{-y}}, \;\; y < x, x \in I\!R.$$

   Com base no método da transformação inversa, implementem um algoritmo para simular (x_i, y_i), $i = 1, \ldots, 1000$, e apresentem graficamente o resultado da vossa simulação.

4. Implementem um algoritmo com o método da aceitação-rejeição para gerar 10000 valores da variável aleatória $X \sim N(3, 2)$, tomando como candidatos valores de $Y \sim Exp(1)$. Em média, quantos números exponenciais serão necessários para gerar a amostra $(x_1, \ldots, x_{10000})$? Insiram um contador para contabilizar o número de iterações do algoritmo que efetivamente foram necessárias para obter a amostra $(x_1, \ldots, x_{10000})$ e comparem-no com o número esperado (teórico) de iterações. Utilizem representações gráficas para ilustrar a geração de $(x_1, \ldots, x_{10000})$, que permitam visualizar quais de os y's candidatos foram aceites/rejeitados.

5. O objetivo desta questão é implementarem um estudo de simulação de Monte Carlo para estimar $\mu = E(X)$, tentando reduzir a variância das estimativas com variáveis antitéticas. Considerem que $X \sim Exp(1)$, e simulem $N = 10000$ amostras de dimensão $n \in \{1, 2, 3, \ldots, 200\}$:

   (a) sem redução de variância (Monte Carlo usual);

   (b) com redução de variância com variáveis antitéticas.

   Usando o **ggplot2**, construam um gráfico de faixa (geom_ribbon) colocando no eixo dos xx a dimensão da amostra, n, e no eixo dos yy os valores gerados (sem e com redução de variância, usando diferentes cores) entre os quantis 0.025 e 0.975, destacando a média. Comparem, analisem e comentem os resultados obtidos.

Sobre o relatório:

- Não deve exceder 20 páginas, deve incluir índice e bibliografia.
- Deve conter a explicação da metodologia utilizada em cada uma das questões.
- O código do *software* R utilizado e o relatório devem ser enviados por e-mail para: `irodrig@math.tecnico.ulisboa.pt`.