ISCTE-IUL
Licenciatura em Ciência de Dados

Método de Aceitação-Rejeição & Análise de Estimadores

Exercício realizado no âmbito da Unidade Curricular de Modelação Estocástica do 3º ano da Licenciatura em Ciência de Dados)

André Plancha, 105289
Andre_Plancha@iscte-iul.pt

João Francisco Botas, 104782
Joao_Botas@iscte-iul.pt

10 de outubro 2023
Versão 1.0.0

1)

A distribuição triangular é uma distribuição de probabilidades contínua com uma forma de triângulo. É definida por 3 propriedades: o vértice inferior esquerdo a, o vértice inferior direito b, e o vértice superior c. A sua função de distribuição de probabilidade (FDP) pode ser definida por (1).

$$\Delta(x \mid a, b, c) := \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & \text{se } a \leq x \leq c \\ \frac{2(b-x)}{(b-a)(b-c)} & \text{se } c \leq x \leq b \\ 0 & \text{em qualquer outro caso} \end{cases} \tag{1}$$

O método da aceitação-rejeição é um método de Monte Carlo com o objetivo de gerar números pseudo-aleatórios (NPAs) de uma distribuição não-uniforme. Consiste num processo iterativo que pretende aceitar ou rejeitar NPAs $\boldsymbol{X}$ geradas a partir do domínio $\{x : f(x) \neq 0\}$ de uma distribuição objetivo $f(x)$ de acordo com a probabilidade $P[X = x]$. Para isto, o processo irá gerar pontos da área de uma distribuição $C \cdot g(x)$ e verificar se esses também pertencem à área de $f(x)$. Ou seja, se (2) se observar então x é "aceite" como um NPA gerado da distribuição de $f(x)$.

$$C \cdot g(x) \cdot u(0,1) < f(x), C \in \{\lambda : \lambda g(x) \geq f(x)\} \tag{2}$$

Para o máximo da eficiência do algoritmo, o C escolhido deve ser o mais perto possível do máximo de $\frac{f(x)}{g(x)}$, pois o algoritmo terá maior probabilidade de gerar um número aceitável. Da mesma forma, é preferível que a forma de $g(x)$ seja uma parecida à forma de $f(x)$, sendo que a área de rejeição será menor. A distribuição $g(x)$ mais comum de escolher é a distribuição uniforme, devido à sua simplicidade, e à facilidade do cálculo do máximo de $\frac{f(x)}{g(x)}$(desde que seja fácil determinar o máximo de $f(x)$).

GERADOR ACEITAÇÃO-REJEIÇÃO($f(x)$, $g(x)$, C, $u([a,\ b])$, $D\{f(x)\}$):

ciclo
- $u \leftarrow u([0, 1])$
- $x \leftarrow u(D\{f(x)\})$
- **se** $u \leq \frac{f(x)}{Cg(x)}$ **produza** u
- **se não** continuar o ciclo

Algoritmo 1: Gerador Aceitação-Rejeição

O nosso objetivo foi gerar 10000 NPAs provenientes de $\Delta(x)$ usando o método da aceitação-rejeição. Decidimos usar como $g(x)$ a distribuição uniforme (com domínio de $[a, b]$), pois o máximo de $\Delta(x)$ será o parâmetro c, levando a que $C = \frac{f(c)}{g(c)}$. Na Figura 1 encontram-se os resultados das várias gerações executadas.

Para demonstrar a precisão dos resultados, decidimos usar vários parâmetros diferentes e comparar com estatísticas teóricas, elas sendo a média, mediana, variância, e forma da distribuição. Decidimos também mostrar uma simulação da geração e dos pontos aceites e rejeitados (canto superior direito da Figura 1).

Podemos obervar que as formas estão bastante semelhantes ao teórico, e as estatísticas amostrais analisadas são extremamente semelhantes às teóricas, provando assim a eficácia do gerador e a confiância que podemos ter nele. Os resultados completos encontram-se em anexo.

2)

O nosso objetivo neste problema foi avaliar e comparar o comportamento de dois estimadores para a assimetria: um usando quantis para o cálculo (s_1), e outro sendo o Coeficiente de Groeneveld e Meeden (s_2). Para isso,

```
Estatísticas teóricas e observadas


|desc                              |   mean| median| variance|
|:---------------------------------|------:|------:|--------:|
|Teórico (a = 0, b = 100, c = 50)  | 50.000| 50.000|  416.667|
|Observado (a = 0, b = 100, c = 50) | 49.798| 49.897|  417.380|
|Teórico (a = 0, b = 10, c = 5)    |  5.000|  5.000|    4.167|
|Observado (a = 0, b = 10, c = 5)  |  4.980|  4.990|    4.174|
|Teórico (a = 1, b = 2, c = 1.2)   |  1.400|  0.850|    0.047|
|Observado (a = 1, b = 2, c = 1.2)  |  1.398|  1.364|    0.047|*
|Teórico (a = 0, b = 20, c = 20)   | 13.333| 15.000|   22.222|
|Observado (a = 0, b = 20, c = 20)  | 13.320| 14.148|   22.419|
```

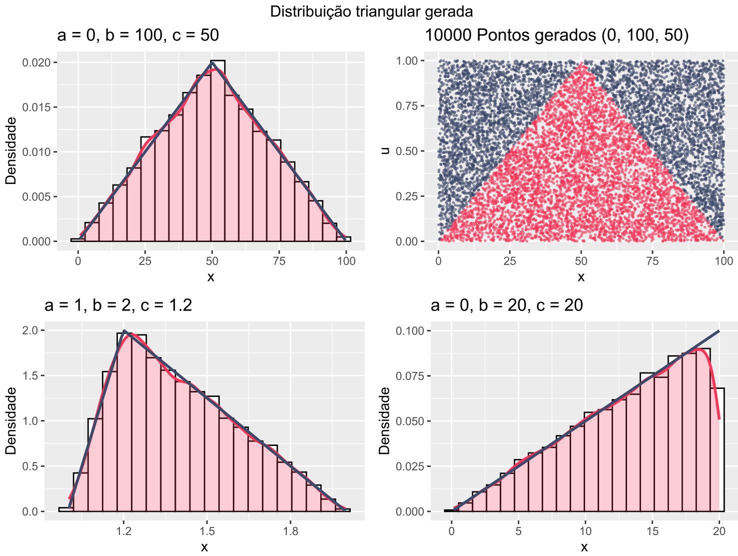

Figura 1: Distribuições geradas

usámos um gerador da distribuição *t-student*, para gerar 3 conjuntos de 100 amostras, cada conjunto com diferentes dimensões de amostras; e calculamos métricas de erros entre o valor teórico (para $\nu > 3$, a assimetira da distribuição será 0)[1], e os valores estimados. As métricas usadas foram o erro-padrão (EP) (desvio padrão de uma estimativa) e o erro quadrático médio (EQM).

O processo pode-se encontrar em anexo e na Figura 2 encontram-se os resultados obtidos.

Pudemos oberver que a dispersão das assimetrias estimadas, e consequentemente a distância inter-quartil é bastante inferior usando s_2 em vez de s_1. Pudemos notar também que os valores das métricas de erro são menores usando s_2. Para além disso, verificamos que quanto maior a dimensão das amostras, mais precisa fica a nossa mediana da assimetria teórica da distribuição. Com isto, podemos concluir que:

- Quanto maior a dimensão das amostras, maior a precisão e exatidão dos estimadores;
- O estimador s_2 tem maior precisão e exatidão na estimação da assimetria na distribuição *t-student*.[2]

	EP ↓		EQM ↓	
	s_1	s_2	s_1	s_2
100×20	0.287	**0.202**	0.082	**0.040**
100×100	0.134	**0.088**	0.018	**0.008**
100×1000	0.041	**0.031**	0.002	**0.001**

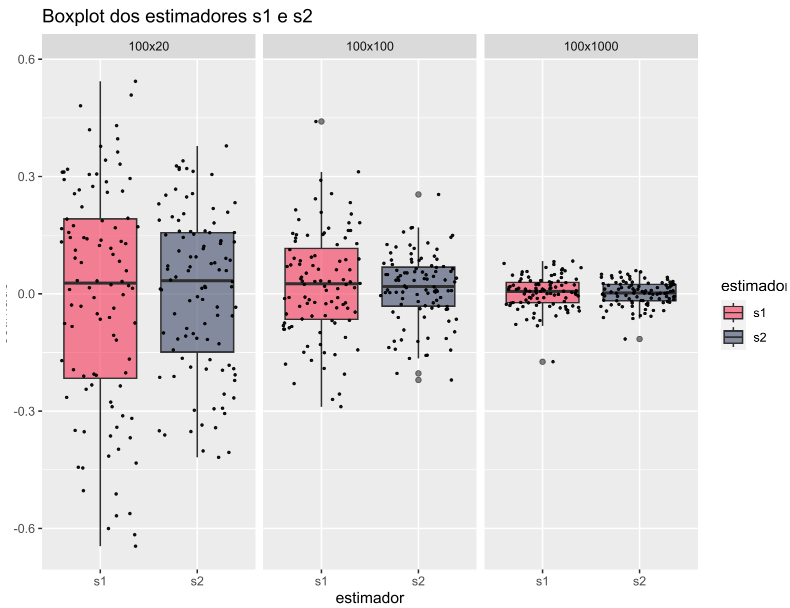

Figura 2: Resultados dos estimadores

[1]Nós escolhemos como propriedade da distribuição *t-student* usar graus de liberdade $\nu = 10$, para facilitar a comparação dos estimadores e para a distribuição não se aproximar tanto a uma normal.

[2]Pelo menos quando $\nu = 10$.