Notas de aula sobre os métodos da Exclusão e da Inversão

Tratamento Estatístico de Dados em Física Experimental (TEFE)

Material elaborado por Zwinglio Guimarães, Leonardo Werneck, Leandro Mariano e Daniel Lessa

Quando queremos estudar numericamente o que ocorre quando os dados seguem uma determinada função densidade de probabilidade devemos ser capazes de gerar dados no computador que satisfaçam esta função densidade de probabilidade. Em geral, a forma mais simples de se gerar estes dados pseudoaleatórios é através do "método da exclusão" também conhecido como "método da rejeição". Contudo, além deste método ser computacionalmente pouco eficiente, ele só pode ser usado se o domínio da função densidade de probabilidade for finito. Uma solução mais completa para este problema é a utilização do chamado "método da inversão", que além de ser computacionalmente mais eficiente, ainda permite gerar dados cujo domínio é ilimitado. Aqui comentaremos um pouco mais sobre estes dois métodos, dando exemplos concretos de como utilizá-los e alguns de seus resultados.

Suponha que desejamos criar dados que sigam a seguinte função densidade de probabilidade:

$$f(x) = \begin{cases} Ax \ , & se\ 0 \leq x \leq 1 \\ 0, & caso\ contrário \end{cases} \tag{1}$$

Para que esta seja uma função densidade de probabilidade, devemos normalizá-la:

$$\int_{-\infty}^{\infty} dx\, f(x) = 1 = A \int_{0}^{1} dx\, x = \frac{A}{2} \Rightarrow \boxed{A = 2} \tag{2}$$

Utilizando (2), podemos escrever a função a ser simulada no computador

$$f(x) = \begin{cases} 2x \ , & se\ 0 \leq x \leq 1 \\ 0, & caso\ contrário \end{cases} \tag{3}$$

O esboço desta função encontra-se na Figura 1. Gostaríamos de simular dados de tal forma que estes estejam sujeitos à esta distribuição de probabilidade, ou seja, após simularmos diversos dados, devemos obter um histograma muito próximo dela quando fizermos um histograma dos dados simulados.

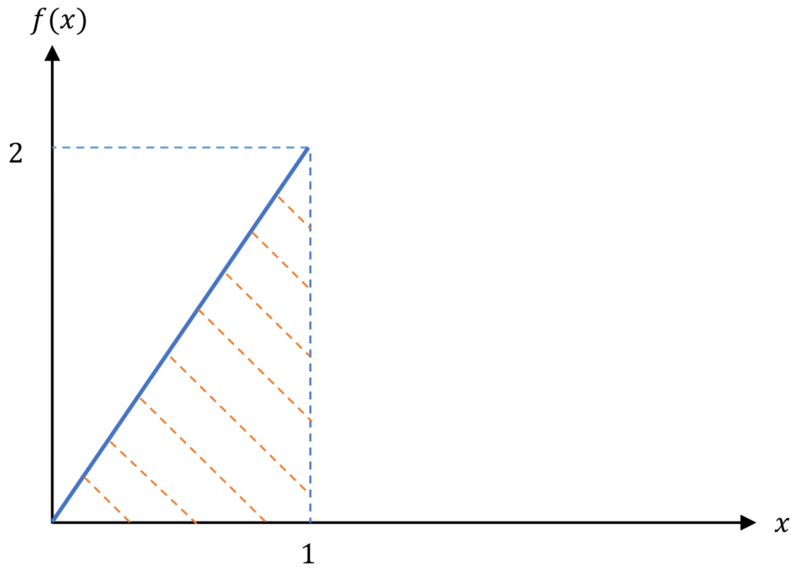

Figura 1 *– Esboço da função densidade de probabilidade definida pela equação (3).*

<u>1 - O Método da Exclusão</u>

O método da exclusão consiste em gerar pontos candidatos uniformemente distribuídos ao longo do domínio da função desejada e, em seguida, submetê-los a um teste aleatório cuja probabilidade de sucesso depende do valor da função densidade de probabilidade naquele ponto. Os pontos que não passam no teste são descartados (daí o nome “exclusão” ou “rejeição”).

Para o caso de nosso problema, devemos gerar pontos candidatos $x \in [0,1]$ e uma variável auxiliar para teste $y \in [0,2]$, e verificar se $y \leq f(x) = 2x$. Caso essa condição seja satisfeita, armazenamos a coordenada x do ponto gerado, caso contrário o dado é excluído e um novo é gerado. A Figura 2, mostra o histograma de uma geração de 1×10^6 pontos usando este método.

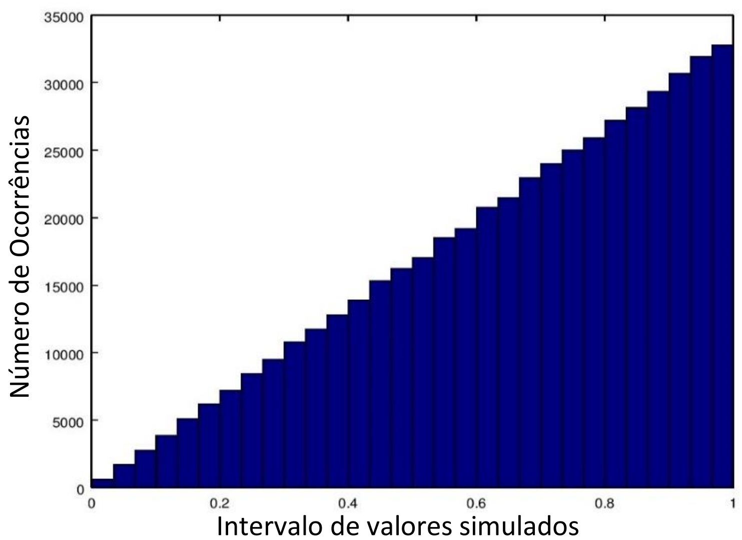

Figura 2: *Histograma do número de ocorrência pelo intervalo de valores simulados, através do método da exclusão, para* 1×10^6 *pontos simulados.*

2- O Método da Inversão

O método da inversão utiliza a função probabilidade acumulada, $g(x)$, definida por:

$$g(x) = \int_{-\infty}^{x} dx' \, f(x') \tag{4}$$

A função de probabilidade acumulada $g(x)$ é definida pela integral da função densidade de probabilidade entre $-\infty$ e x, ou seja, ela representa a probabilidade de obtermos um valor no intervalo $x' \in [-\infty, x]$. Note que, quando x tende infinito, a função de probabilidade acumulada tende a 1. Dessa forma, os valores possíveis para a função probabilidade acumulada estão obrigatoriamente entre 0 e 1 (compare com a condição de normalização que utilizamos em (2)).

O método da inversão consiste em se gerar um valor para $g(x)$ com distribuição uniforme entre 0 e 1 e usar a função inversa g^{-1} para encontrar o valor de x correspondente. Os valores de x gerados desta forma tem densidade de probabilidade acumulada igual à função $g(x)$ e, portanto, seguem a função densidade de probabilidade $f(x)$.

Note que neste método o domínio não precisa ser limitado, porém é preciso determinar a inversa da função de probabilidade acumulada. Se a função inversa não for determinada de forma analítica, um método numérico para achar a função inversa pode ser usado, porém, a eficiência computacional será muito prejudicada.

No caso do nosso exemplo, a função de probabilidade acumulada pode ser escrita como:

$$g(x) = \int_{0}^{x} dx' 2x' = x^2 \tag{5}$$

Assim, a aplicação do método da inversão consiste em gerar números aleatórios entre 0 e 1 para $g(x)$ e, a partir destes números, encontrar o x correspondente através da função inversa:

$$\boxed{x = g^{-1} = \sqrt{g(x)}} \tag{6}$$

A Figura 3 ilustra graficamente o processo de inversão. O método ilustrado na Figura 3 pode ser repetido diversas vezes para gerar um conjunto de pontos que satisfaçam a função densidade de probabilidade $f(x)$. O exemplo mostrado na Figura 4 corresponde a uma simulação de 1×10^6 pontos gerados.

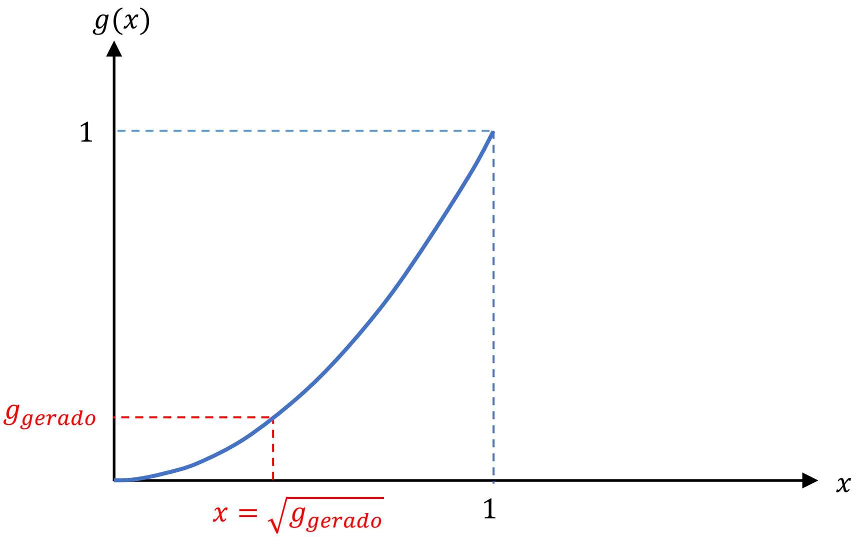

Figura 3: *Ilustração gráfica do método de inversão*

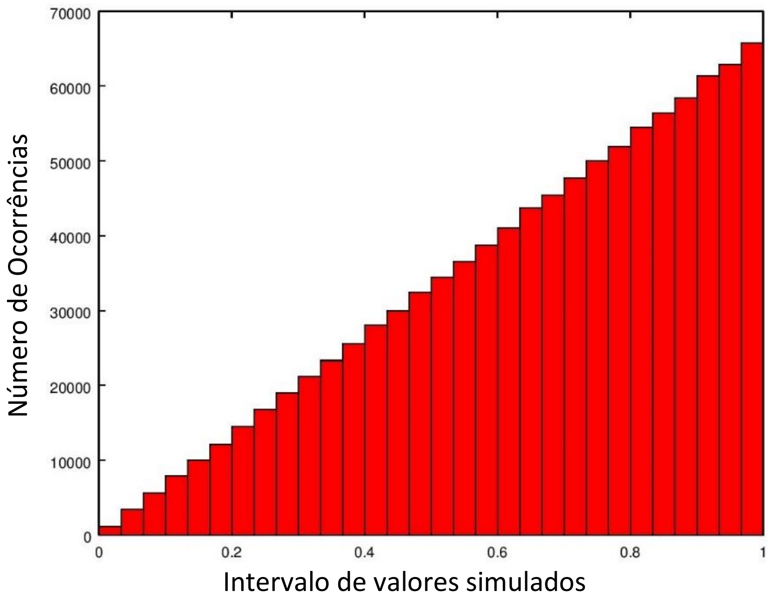

Figura 4: *Histograma do número de ocorrência pelo intervalo de valores simulados, através do método da inversão, para* 1×10^6 *pontos simulados.*