Teorema central do limite

Fábio A. Bocampagni

setembro de 2023

Introdução

0.1 Somar dois números aleatórios aumenta a aleatoriedade?

Para responder a essa pergunta, consideremos as condições impostas pelo enunciado:

Enunciado 0.1. *Para simplificar, ao invés de somarmos números de 0 a 100, vamos somar números que sejam 0 ou 1. Ou seja, a cada lance de uma moeda, se der CARA somamos 0 e se der COROA somamos 1.*

Seja M o número de moedas jogadas.

Seja S o valor da soma das M moedas.

Seja $\frac{S}{M}$ a soma normalizada do número de moedas jogadas.

Questão 01

Quando M=2, quais os possíveis valores de S? Quais os possíveis valores de S/M?

Resolveremos primeiro para os casos de S. Para $M = 2$, teremos 4 casos. É natural pensar que sempre teremos uma relação de 2^M valores possíveis. Sendo assim, é natural encontrar os valores para os outros casos de M.

Agora, para o caso onde temos a normalização por M, ou seja, $\frac{S}{M}$ o número de valores possíveis é delimitado pelo tamanho de M. Sendo assim, os valores possíveis sempre serão iguais aos valores de M.

Um outro ponto que vale ressaltar é a medida que M cresce, pode-se aproximar a variável aleatória S de uma distribuição de probabilidades normal. Analogamente, $\frac{S}{M}$ pode ser vista como uma normal reduzida. Como a normal reduzida tem uma variância que decai com o aumento de M, pode-se entender que sua incerteza diminuiu. Ao passo que S, nesse caso sendo aproximado por uma normal padrão, tem uma incerteza maior pela ausência de normalização.

Questão 02

Quando M=2, qual a variância de S? Qual a variância de S/M?

Para calcular a variância de S quando $M = 2$, você pode usar a fórmula da variância, que é dada por:

$$\text{Var}(S) = E(S^2) - [E(S)]^2$$

Primeiro, calculemos $E(S)$, que é a média da soma das moedas jogadas quando $M = 2$.

Cada moeda é lançada e pode ter 2 resultados: CARA (0) ou COROA (1). Portanto, há $2^2 = 4$ possíveis resultados para o par de moedas (0,0), (0,1), (1,0) e (1,1).

A média de S é calculada somando todos esses resultados possíveis e dividindo pelo número de resultados:

$$E(S) = \frac{0 + 1 + 1 + 2}{4} = \frac{4}{4} = 1$$

Agora, calculemos $E(S^2)$, que é a média dos quadrados da soma das moedas jogadas:

$$E(S^2) = \frac{0^2 + 1^2 + 1^2 + 2^2}{4} = \frac{0 + 1 + 1 + 4}{4} = \frac{6}{4} = 1.5$$

Agora, podemos usar a fórmula da variância:

$$\text{Var}(S) = E(S^2) - [E(S)]^2 = 1.5 - 1^2 = 1.5 - 1 = 0.5$$

Portanto, a variância de S quando $M = 2$ é 0.5.

Agora, vamos calcular a variância de $\frac{S}{M}$ para $M = 2$.

A média de $\frac{S}{M}$ é igual à média de S dividida por M, uma vez que S é a soma das moedas e M é o número de moedas jogadas:

$$E\left(\frac{S}{M}\right) = \frac{E(S)}{M} = \frac{1}{2} = 0.5$$

Agora, calculemos $E\left(\left(\frac{S}{M}\right)^2\right)$, que é a média dos quadrados de $\frac{S}{M}$:

$$E\left(\left(\frac{S}{M}\right)^2\right) = \left(\frac{1}{2}\right)^2 = \frac{1}{4} = 0.25$$

Finalmente, podemos calcular a variância de $\frac{S}{M}$ usando a fórmula da variância:

$$\text{Var}\left(\frac{S}{M}\right) = E\left(\left(\frac{S}{M}\right)^2\right) - \left[E\left(\frac{S}{M}\right)\right]^2 = 0.25 - 0.5^2 = 0.25 - 0.25 = 0$$

Portanto, a variância de $\frac{S}{M}$ quando $M = 2$ é 0.

A variância de S é sempre maior que a variância de $\frac{S}{M}$ e cresce a medida que M aumenta.

Questão 03

Pesquise sobre o conceito de entropia. Quando M=2, qual a entropia de S? Qual a entropia de S/M?

Presente em várias áreas da ciência, a entropia é um conceito de medição de incerteza. Na física, pode-se fazer paralelos com o grau de desordem de um sistema. Na teoria da informação, entropia é uma medida da incerteza ou imprevisibilidade em uma fonte de informação. Quanto maior a entropia, mais imprevisível é a fonte. A entropia também está relacionada com a quantidade média de informação necessária para codificar eventos de uma fonte de dados. Já na em probabilidade e estatística a entropia é usada para medir a dispersão ou a impureza de uma distribuição de probabilidade. Em árvores de decisão e aprendizado de máquina, a entropia é frequentemente usada para avaliar a impureza de um conjunto de dados e orientar a escolha das melhores divisões em árvores de decisão.

Para calcular a entropia de S e S/M quando $M = 2$, primeiro precisamos determinar as probabilidades de ocorrência de cada valor possível de S e S/M e, em seguida, usar essas probabilidades para calcular a entropia.

Entropia de S (Soma das Moedas):

Quando $M = 2$, temos quatro possíveis resultados para S: 0, 1, 1 e 2. Para calcular a entropia de S, precisamos determinar as probabilidades de cada um desses resultados:

$$P(S = 0) = \frac{1}{4} \quad \text{(pois há uma maneira de obter } S = 0\text{)}$$

$$P(S = 1) = \frac{2}{4} \quad \text{(há duas maneiras de obter } S = 1\text{)}$$

$$P(S=2) = \frac{1}{4} \quad \text{(há uma maneira de obter } S=2\text{)}$$

A entropia de S é calculada usando a fórmula da entropia de Shannon:

$$H(S) = -\sum_i P(s_i) \cdot \log_2(P(s_i))$$

Calculando a entropia:

$$\begin{aligned}
H(S) &= -\left(\frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right) + \frac{2}{4} \cdot \log_2\left(\frac{2}{4}\right) + \frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right)\right) \\
H(S) &= -\left(\frac{1}{4} \cdot (-2) + \frac{2}{4} \cdot (-1) + \frac{1}{4} \cdot (-2)\right) \\
H(S) &= -\left(-\frac{1}{2} - \frac{1}{2} - \frac{1}{2}\right) \\
H(S) &= -\left(-\frac{3}{2}\right) \\
H(S) &= \frac{3}{2}
\end{aligned}$$

Portanto, a entropia de S quando $M = 2$ é $\frac{3}{2}$ bits.

Entropia de S/M (Soma Normalizada):

Para S/M, temos três possíveis resultados: 0, 0.5 e 1. Vamos calcular as probabilidades de cada um desses resultados:

$$P(S/M = 0) = \frac{1}{4} \quad \text{(uma maneira de obter } S/M = 0\text{)}$$

$$P(S/M = 0.5) = \frac{2}{4} \quad \text{(duas maneiras de obter } S/M = 0.5\text{)}$$

$$P(S/M = 1) = \frac{1}{4} \quad \text{(uma maneira de obter } S/M = 1\text{)}$$

A entropia de S/M é calculada usando a fórmula da entropia de Shannon:

$$H(S/M) = -\sum_i P(s_{m_i}) \cdot \log_2(P(s_{m_i}))$$

Calculando a entropia:

$$\begin{aligned}
H(S/M) &= -\left(\frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right) + \frac{2}{4} \cdot \log_2\left(\frac{2}{4}\right) + \frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right)\right) \\
H(S/M) &= -\left(\frac{1}{4} \cdot (-2) + \frac{2}{4} \cdot (-1) + \frac{1}{4} \cdot (-2)\right) \\
H(S/M) &= -\left(-\frac{1}{2} - \frac{1}{2} - \frac{1}{2}\right) \\
H(S/M) &= -\left(-\frac{3}{2}\right) \\
H(S/M) &= \frac{3}{2}
\end{aligned}$$

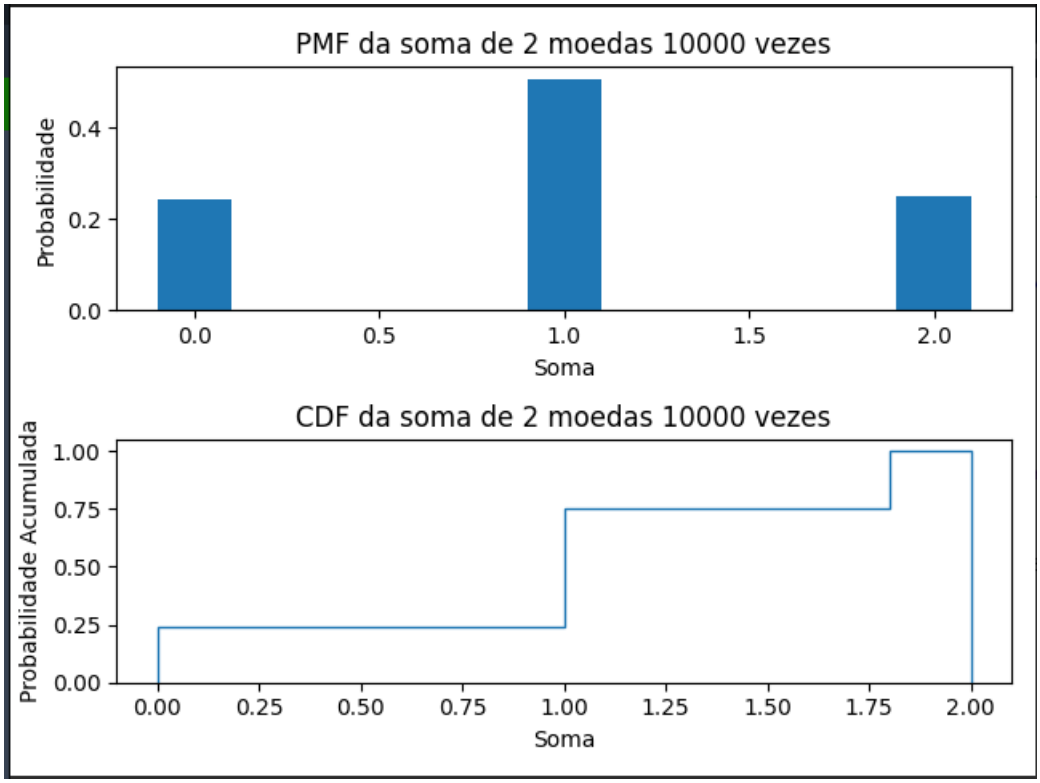
Portanto, a entropia de S/M quando $M = 2$ também é $\frac{3}{2}$ bits.

Questão 04

Seja M=2. Faça um simulador que colha 10,000 amostras de S. Plote o histograma e a CDF das amostras colhidas.

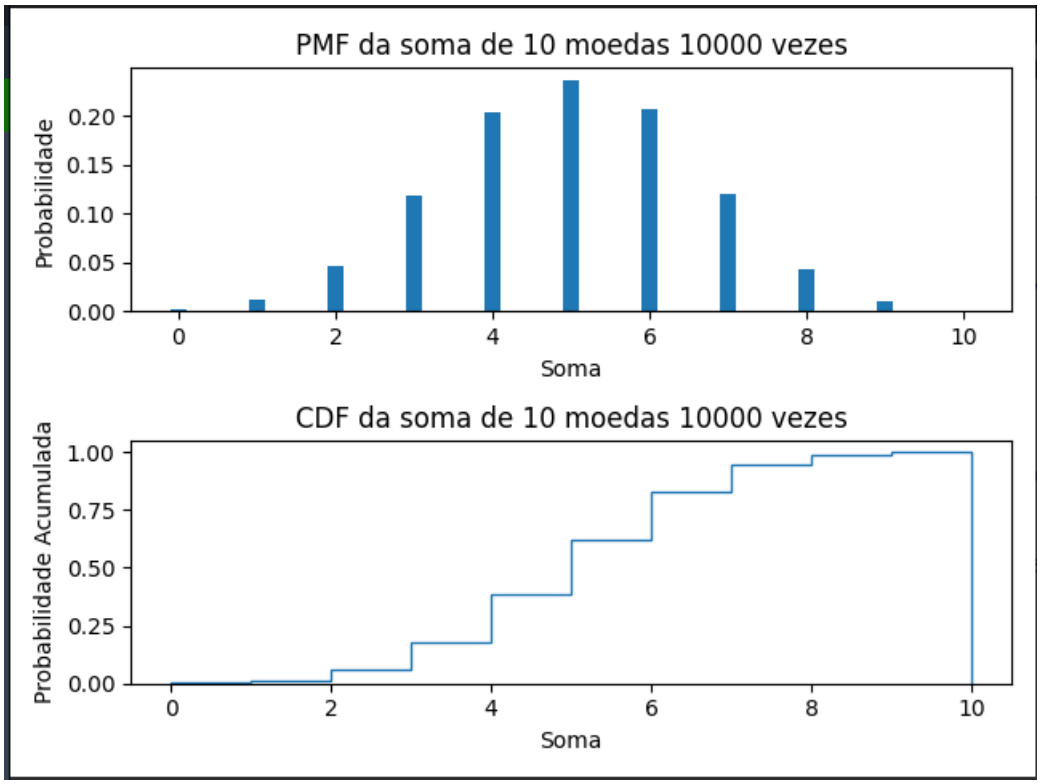
```python
from random import randint
import matplotlib.pyplot as plt
import numpy as np

AGES = 10000

def simulador(M):
    generations = []
    for _ in range(AGES):
        coins = [randint(0, 1) for _ in range(M)]
        generations.append(sum(coins))

    plt.subplot(2, 1, 1)
    bins = np.arange(-0.5, M + 1.5, 1)  # Centralize as barras
    plt.hist(generations, bins=bins, rwidth=0.2, align='mid', density=True)
    plt.xlabel('Soma')
    plt.ylabel('Probabilidade')
    plt.title(f'PMF da soma de {M} moedas {AGES} vezes')

    plt.subplot(2, 1, 2)
    plt.hist(generations, cumulative=True, align='mid', density=True, histtype='step')
    plt.xlabel('Soma')
    plt.ylabel('Probabilidade Acumulada')
    plt.title(f'CDF da soma de {M} moedas {AGES} vezes')

    plt.tight_layout()
    plt.show()
```

Executando o simulador para $M = 2$, podemos obter o seguinte resultado:

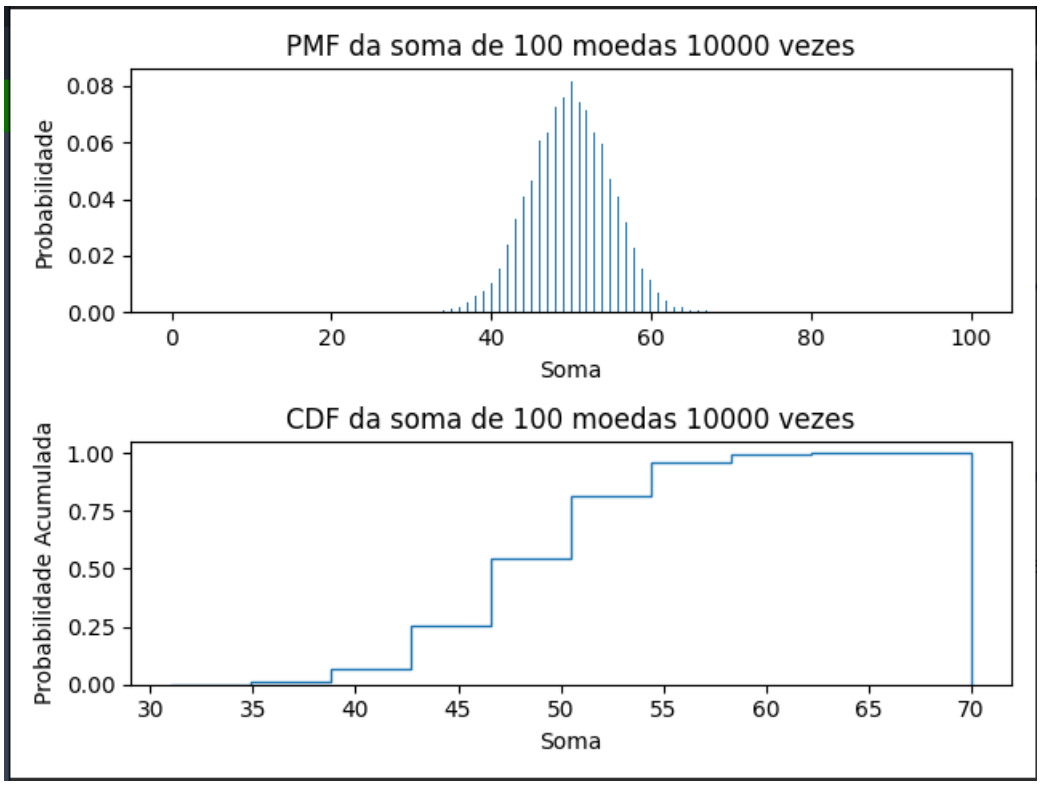


Analogamente, se executarmos para $M = 10$, podemos notar que a curva começa a se aproximar

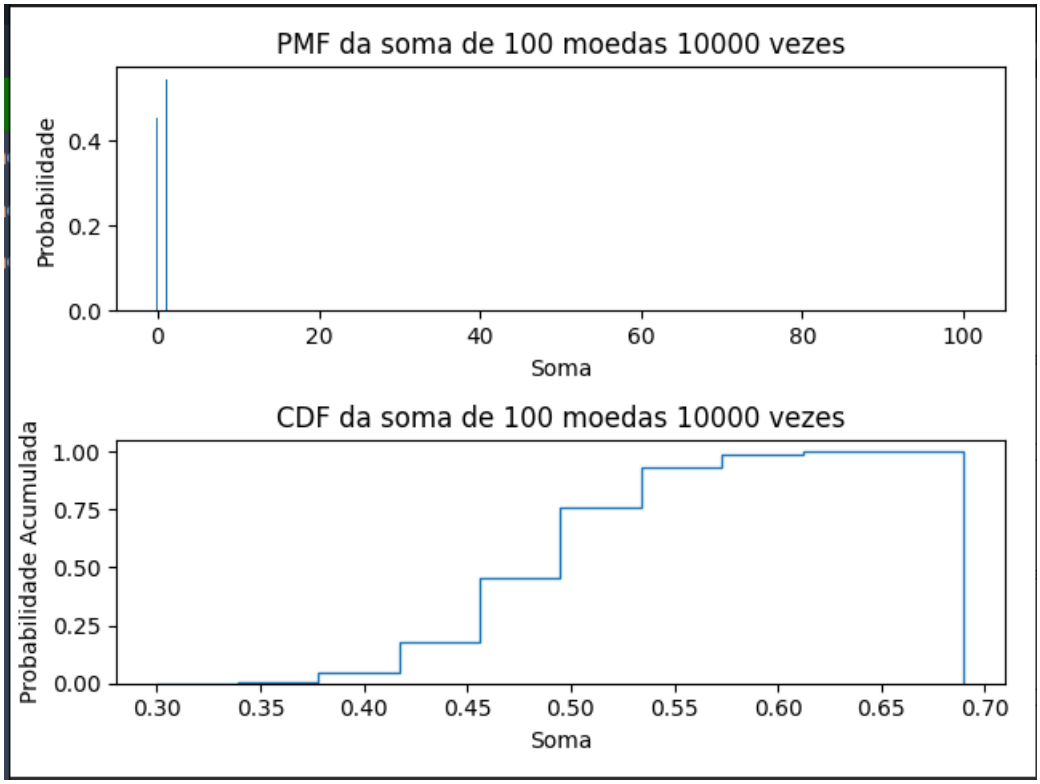
do que conhecemos de uma gaussiana.



Com $M = 100$ claramente nota-se traços, se não totalmente, o desenho da curva gaussiana. O que mais uma vez mostra que podemos aproximar S de uma distribuição normal.



Agora, no que tange $\frac{S}{M}$ à medida que M aumenta, a distribuição de $\frac{S}{M}$ tende a se concentrar mais em torno de 0.5 (a média de uma distribuição de Bernoulli), mas também fica mais estreita devido à redução da variância.



Observações finais

Essa sessão é destinada a responder o restante das questões levantadas, sem esgotar o assunto nelas inseridos, porém, ainda assim, os tratar de maneira precisa.

Enunciado do Teorema Central do Limite: O Teorema Central do Limite (TCL) afirma que a soma (S) de N variáveis aleatórias independentes (X), com qualquer distribuição e variâncias semelhantes, é uma variável com distribuição que se aproxima da distribuição de Gauss (distribuição normal) quando N aumenta.

Variância de S prevista pelo TCL:
De acordo com o TCL, a variância de S (soma das moedas) é dada por σ^2, onde σ^2 é a variância das moedas (0 ou 1). Podemos perceber nesse caso, mais uma vez, que a variância de S aumenta conforme M aumenta pois a variância depende da esperança, que depende de M

Variância de S/M prevista pelo TCL:
A variância de S/M (soma normalizada, ou média amostral) é obtida dividindo a variância de S por n. Portanto, a variância prevista de S/M pelo TCL é $\frac{\sigma^2}{n}$, onde podemos notar que quando M, ou nesse caso n aumentam, a variância diminui.

Comentários gerais sobre o vídeo mencionado
O vídeo mencionado é bem construído e apresenta o tópico de forma descontraída. O ponto de atenção fica no que tange a conclusão da pergunta feita. A incerteza irá aumentar se considerar um

S não normalizado, pois como vimos, a variância nesse caso aumenta conforme o número de moedas somadas aumenta. Agora, se considerarmos a média amostral, ou seja, a soma normalizada, a variância irá diminuir, diminuindo assim a incerteza, fazendo com que os valores tendam a um limiar.