Trabalho - GA030 (Estatística)

Lorran de Araújo Durães Soares*

2024

Introdução

Este texto refere-se à realização do trabalho da disciplina GA030 (Estatística) do curso de pós graduação do Laboratório Nacional de Computação Científica (LNCC), ministrada pelo prof. Marcio Rentes Borges. Serão apresentadas as questões propostas, seguidas da sua respectiva resolução. Clique aqui para acessar o código referente à realização de todos os exercícios.

Questão 1

Após abordarmos a *Lei dos Grandes Números* e o *Teorema do Limite Central*, chegamos a um ponto crucial do curso: a estimação de parâmetros (desconhecidos) associados à distribuição de probabilidade de uma variável aleatória.

O presente trabalho tem como objetivo a fixação das ideias introduzidas até aqui. Para isso, utilizaremos dados armazenados em quatro arquivos, que contêm amostras de diferentes variáveis aleatórias, conforme a Tabela 1.

Variável	**Arquivo**	**Distribuição**
$Q \sim \mathcal{N}(0, 2)$	`data1q.dat`	Normal
$X \sim U[-1, 1]$	`data1x.dat`	Uniforme
$Y \sim E(\lambda = 0.05)$	`data1y.dat`	Exponencial
$T \sim B(15, 0.40)$	`data1t.dat`	Binomial

Tabela 1 – Tabela de dados

(a)

Dado que conhecemos a distribuição de probabilidades de cada variável aleatória e os parâmetros que as caracterizam (Tabela 1), calcule a expectativa e a variância (teóricas) de cada uma delas, usando as definições que vimos em aula.

Resolução:

Usando as definições dadas em aula e utilizando os parâmetros presentes na tabela 1, iremos calcular a média e a variância teórica de cada variável de dados.

*lorranspbr@gmail.com

- Para a variável Q, não será necessário cálculos, pois os próprios parâmetros da curva normal fornecem à sua média e variância. Logo, $\mu_Q = 0$ e ${\sigma_Q}^2 = 2$.
- Para a variável X, que tem origem em uma distribuição uniforme, teremos que a média e a variância serão dadas por:

$$\mu_X = \frac{-1+1}{2} \Rightarrow \mu_X = 0$$

$${\sigma_X}^2 = \frac{(1-(-1))^2}{12} = \frac{4}{12} \Rightarrow {\sigma_X}^2 = \frac{1}{3}$$

- Para a variável Y, originado através de uma distribuição exponencial com $\lambda = 0.05$, teremos que a média e a variância serão dadas por:

$$\mu_Y = \frac{1}{\lambda} = \frac{1}{0.05} \Rightarrow \mu_y = 20$$

$${\sigma_Y}^2 = \frac{1}{\lambda^2} = \frac{1}{0.05^2} = \frac{1}{0.0025} \Rightarrow {\sigma_Y}^2 = 400$$

- Para a variável T, dada por uma distribuição binomial com $p = 0.40$ e $N = 15$, teremos que a média e a variância serão dadas então por:

$$\mu_T = N \cdot p = 15 \cdot 0.4 \Rightarrow \mu_T = 6$$

$${\sigma_T}^2 = N \cdot p(1-p) = 15 \cdot 0.4(1-0.4) \Rightarrow {\sigma_T}^2 = 3.6$$

Logo, com os resultados obtidos, a tabela 2 apresenta a média e variância teóricas de cada uma das variáveis.

Variável	**Média**	**Variância**
$Q \sim N(0,2)$	0	2
$X \sim U[-1,1]$	0	$\frac{1}{3}$
$Y \sim E(\lambda = 0.05)$	20	400
$T \sim B(15, 0.40)$	6	3.6

Tabela 2 – Média e Variância dos dados

(b)

Utilize o R (ou outro programa) para ler cada arquivo e calcule estimativas para a média e a variância do conjunto de dados (usando todos os dados disponíveis nos arquivos). Em seguida, compare com os resultados obtidos no exercício anterior. Faça comentários.

Resolução:

Para fazer a estimativa da média e da variância de cada variável, iremos usar os estimadores clássicos, pois estes se tratam de estimativas não tendencionas, de variância mínima e coerentes:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Foi então construído um programa na linguagem Python, que foi então utilizado para obter as estimativas para a média e para a variância de cada variável aleatória. Os resultados estão presentes na tabela 3, apresentadas aqui com três casas decimais de precisão.

Variável	**Média**	**Variância**
$Q \sim N(0, 2)$	0	2
$X \sim U[-1, 1]$	0	0.333
$Y \sim E(\lambda = 0.05)$	20.007	400.37
$T \sim B(15, 0.40)$	6	3.602

Tabela 3 – Estimativas para a Média e Variância dos dados

Após a realização deste exercício, pode ser observado que as estimativas para a média e para a variância se aproximam consideravelmente dos valores reais. Isso acontece pois o número de amostras de cada conjunto de dados é suficientemente grande. Neste caso, cada variável aleatória possui 5 milhões de amostras, tornando possível estimar a média e a variância com a precisão apresentada.

(c)

Construa os histogramas com as frequências relativas de cada uma das variáveis, verificando se estes são condizentes com os modelos teóricos (Tabela 1).

Resolução:

Para cada variável, foi construído o histograma das frequências relativas, normalizando os dados para que o histograma, dada a grande quantidade de dados, se aproxime da função de densidade de probabilidade (pdf) ou da função de probabilidade (pf) correspondente à distribuição de origem dos dados. Para realizar essa comparação, a pdf ou pf teórica foi sobreposta ao histograma, com o intuito de verificar se este está condizente com o modelo teórico. A figura 1 apresenta o resultado obtido. Para as variáveis contínuas, as partições (bins) foram ajustados automaticamente.

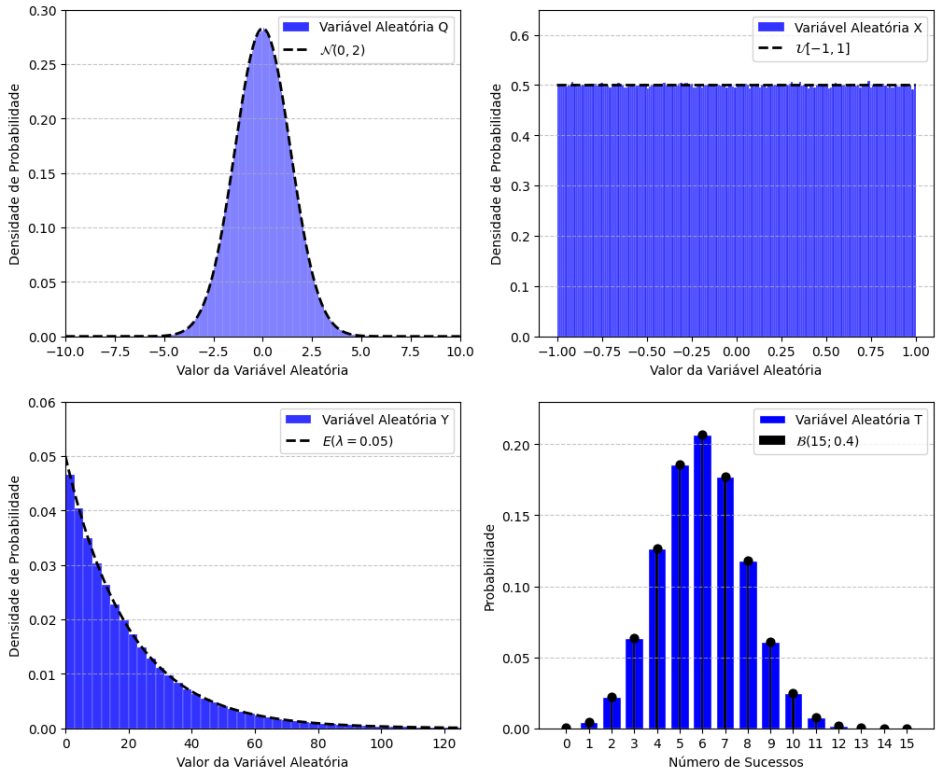

Figura 1 – Histograma de frequências relativas dos dados com distribuições sobrepostas

A visualização da figura acima evidencia que os histogramas estão condizentes com os respectivos modelos teóricos, pois os dados ficaram bem ajustados às respectivas pdf ou pf em todos os casos.

(d)

Considere cada uma das amostras das variáveis aleatórias, contidas nos arquivos, e suas diferentes distribuições de probabilidades. Tome amostras aleatórias de tamanho n (n = 5, 10 e 50) de cada uma das variáveis aleatórias e construa as variáveis aleatórias (estatísticas):

- *média amostral:*

$$\bar{W}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} W_i$$

- *variância amostral:*

$$S_W{}^{2(n)} = \frac{1}{n-1}\sum_{i=1}^{n}(W_i - \bar{W}_n)^2$$

onde $W = Q, X, Y$ ou T. Use 10000 amostras simples (pontos amostrais) para gerar as variáveis aleatórias média amostral e variância amostral. Obs.: Lembre-se das características que as amostras aleatórias devem ter. Apresente o código.

Resolução:

Para a realização desta questão, foram realizados 1000 sorteios aleatórios de 5, 10 e posteriormente 50 dados de cada variável, construindo então as variáveis aleatórias média e variância amostrais. A figura 2 mostra a codificação realizada para a realização desta questão.

```
        n_samples = [5,10,50]

        for key, dat in data.items():
            for n in n_samples:
                medias = []
                variancias = []

                amostras = np.random.choice(dat, size=(10000, n), replace=
                    False)
                medias = amostras.mean(axis=1)
                variancias = amostras.var(axis=1, ddof=1)

                if key == 'data_q':
                    medias_amostrais_q[f'{key}_{n}'] = medias
                    variancias_amostrais_q[f'{key}_{n}'] = variancias
                elif key == 'data_x':
                    medias_amostrais_x[f'{key}_{n}'] = medias
                    variancias_amostrais_x[f'{key}_{n}'] = variancias
                elif key == 'data_y':
                    medias_amostrais_y[f'{key}_{n}'] = medias
                    variancias_amostrais_y[f'{key}_{n}'] = variancias
                elif key == 'data_t':
                    medias_amostrais_t[f'{key}_{n}'] = medias
                    variancias_amostrais_t[f'{key}_{n}'] = variancias
                else:
                    raise(ValueError)
```

Figura 2 – Código para cálculo das variáveis média amostral e variância amostral

(e)

Usando o código da questão anterior, construa os histogramas de frequências das variáveis aleatórias média amostral e variância amostral, para os diferentes valores de n e compare com as distribuições teóricas esperadas para estas variáveis. Faça isso para as variáveis (Q, X, Y e T).

Resolução:

Para esta questão, é esperado que todas as médias amostrais se aproximem de uma curva normal, com mesma média que a variável de origem e com variância igual a da origem dividida por n, onde n se trata do número de amostras que caracterizam a média amostral, neste caso, 5, 10 ou 50. Já para a variância, no caso da variável originada da distribuição normal, é esperado que, multiplicada por um fator, ela se caracterize por uma distribuição qui-quadrada, com $n-1$ graus de liberdade. Para as outras variáveis, não é esperado que a variância amostral se comporte de uma maneira específica.

Logo, foram construídos os histogramas e as curvas sobrepostas esperadas, presentes nas figuras 3, 4, 5 e 6, onde cada coluna apresenta o histograma da média amostral (azul) e da variância amostral (vermelha) com n igual a 5, 10 e 50, respectivamente.

Observando a sobreposição das curvas, pode se notar que os histogramas se aproximam satisfatóriamente das distribuições teóricas esperadas para cada variável.

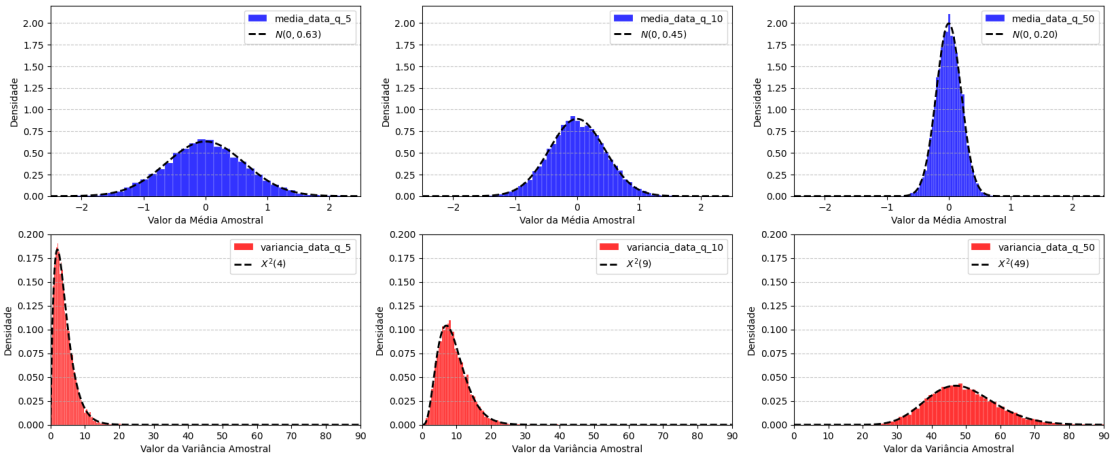

Figura 3 – Histograma das variáveis média e variância amostrais com origem Q

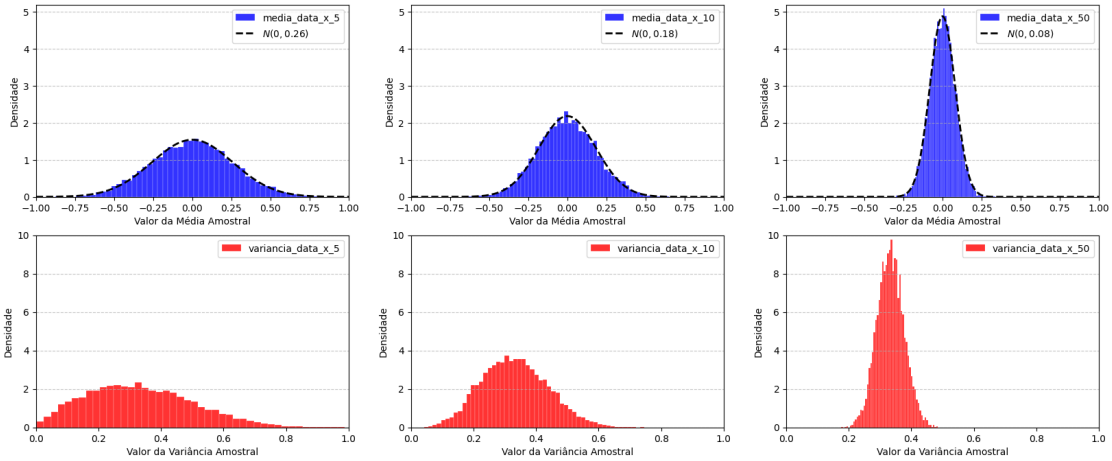

Figura 4 – Histograma das variáveis média e variância amostrais com origem X

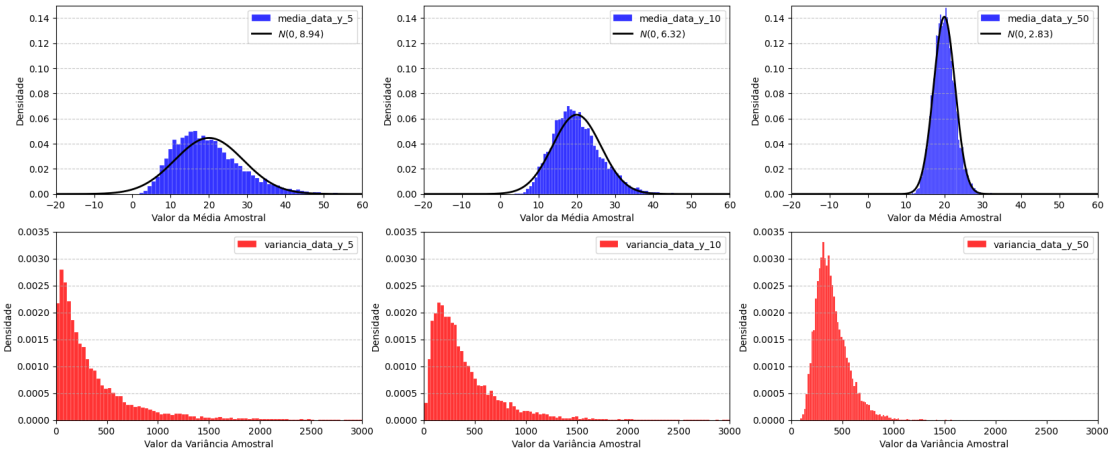

Figura 5 – Histograma das variáveis média e variância amostrais com origem Y

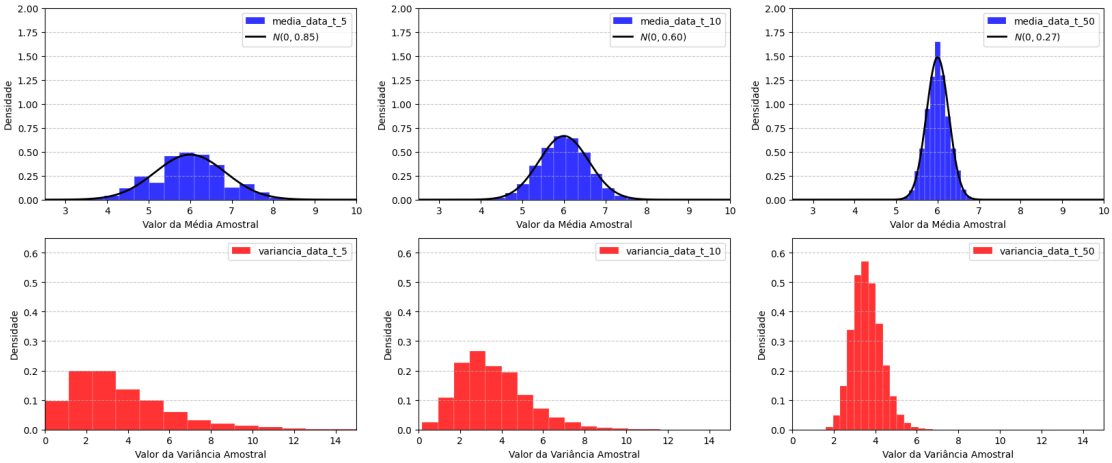

Figura 6 – Histograma das variáveis média e variância amostrais com origem T

(e)

Compare os histogramas, para os diferentes valores de n, e discuta os resultados.

Resolução:

Observando as figuras, pode se concluir que quanto maior era a amostra, mais facilmente o histograma se ajustou à sobreposição da curva esperada. Embora não se esperasse nada das outras variâncias, elas também obtiveram uma distribuição blabla parecendo normal e tals.