# Trabalho - GA030 (Estatística)

Lorran de Araújo Durães Soares*

2024

## Introdução

Este texto refere-se à realização do trabalho da disciplina GA030 (Estatística) do Laboratório da pós graduação do Laboratório Nacional de Computação Científica, ministrada pelo prof. Marcio Rentes Borges. Serão apresentadas as questões propostas, seguidas da sua 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** |
|---|---|---|
| $Q \sim N(0, 2)$ | `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 o 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 o variável $X$, que tem origem em uma distribuição uniforme, teremos que média e variância serão dados 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 o variável $Y$, originado através de uma distribuição exponencial com $lambda = 0.05$, teremos que média e variância serão dados 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 o variável $T$, dado por uma distribuição binomial com $p = 0.40$ e $N = 15$, teremos que média e variância serão dados 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, a seguinte tabela 2 apresenta os resultado obtidos para 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:

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

e

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

pois eles são estimativas não tendencionas, de variância mínima e coerente. Logo, foram obtidos as seguintes estimativas presente 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

Pode ser observado após a realização deste exercício que, como cada conjunto de dados possui um número consideravelmente grande de amostras (5 milhões), então as estimativas para a média e a variância se aproximam dos valores reais teóricos.

## (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 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.

A visualização da figura acima evidencia que os histogramas estão condizentes com os respectivos modelos teóricos, pois os dados ficaram bem ajustados às pdf ou à 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:*

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

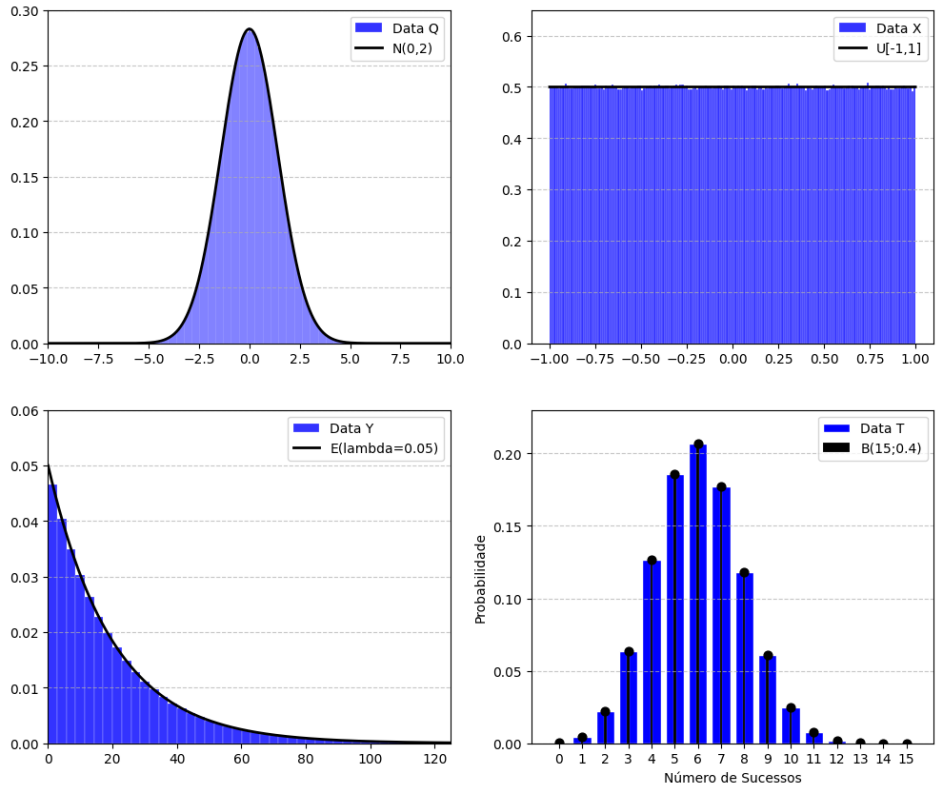


Figura 1 – Histograma de frequências relativas dos dados com distribuições sobrepostas

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 de 5, 10 e depois 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.

## (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 a variavel vinda da normal, é esperado que tal tal, presente na figura 3.

Para a variavel vinda da normal, é esperado que tal tal, presente na figura 4.

Para a variavel vinda da normal, é esperado que tal tal, presente na figura 5.

Para a variavel vinda da normal, é esperado que tal tal, presente na figura 6.

```python
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)
        # conferir se essa forma de calcular variancia ta certa (nao seria usando a media real)
        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)
```

✓ 1.8s Python

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

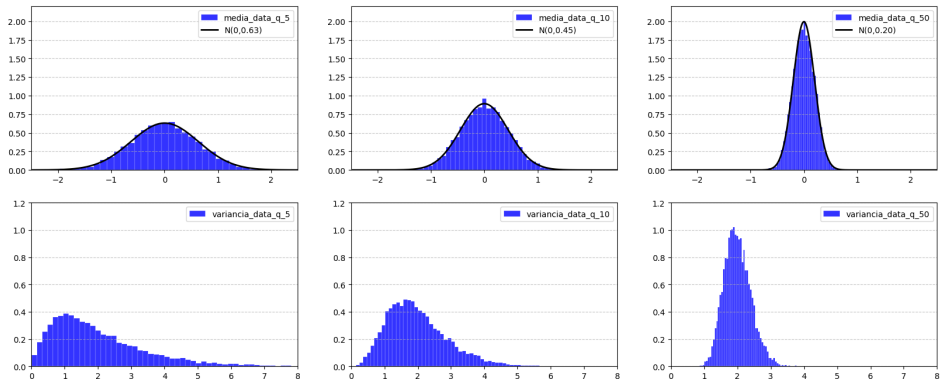


Figura 3 – Histograma de frequências das variáveis aleatórias média amostral e variância amostral com curva respectiva sobreposta

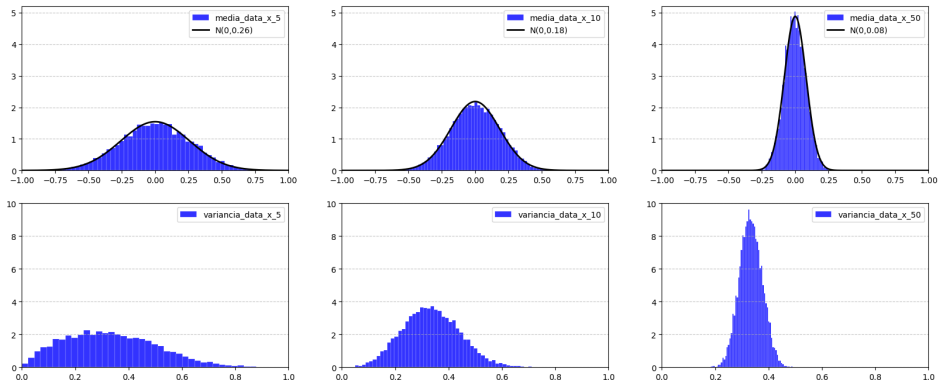


Figura 4 – Histograma de frequências das variáveis aleatórias média amostral e variância amostral com curva respectiva sobreposta

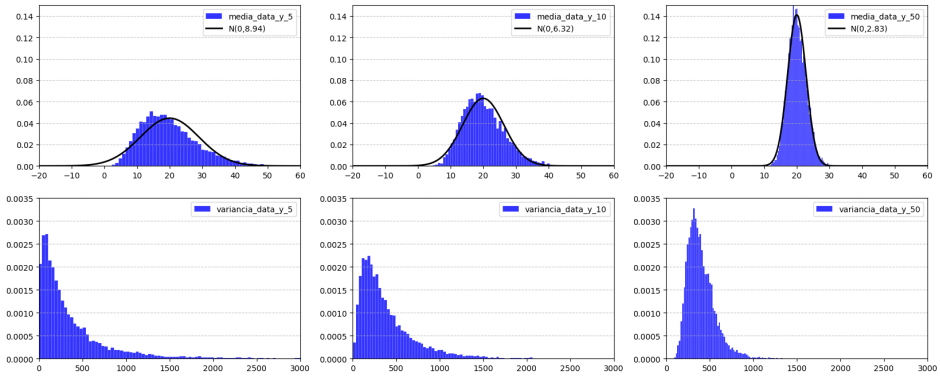


Figura 5 – Histograma de frequências das variáveis aleatórias média amostral e variância amostral com curva respectiva sobreposta

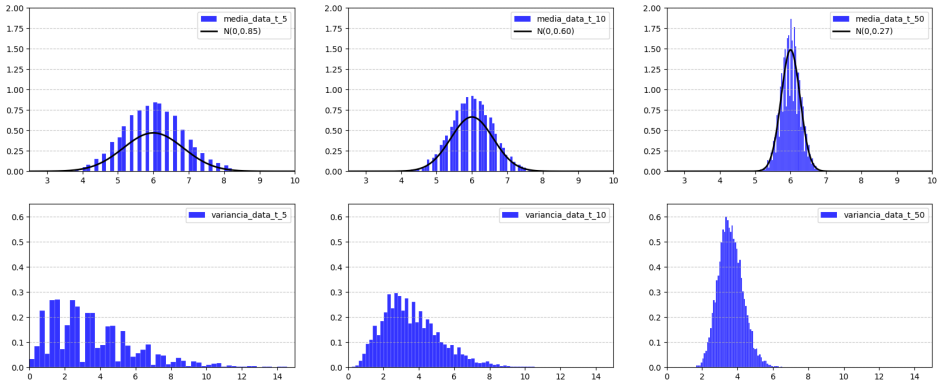


Figura 6 – Histograma de frequências das variáveis aleatórias média amostral e variância amostral com curva respectiva sobreposta

## (e)

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

**Resolução:**