



LABORATÓRIO NACIONAL DE COMPUTAÇÃO CIENTÍFICA
MESTRADO EM MODELAGEM COMPUTACIONAL

Trabalho (GA030 - Estatística)

Lorran de Araújo Durães Soares

Petrópolis - RJ

2024

Lorran de Araújo Durães Soares

Trabalho (GA030 - Estatística)

Trabalho apresentado como parte dos critérios de avaliação da
disciplina GA030 - Estatística.

Professor(a): Marcio Rentes Borges

Petrópolis - RJ

2024

Sumário

1	Introdução	2
2	Questão 1	2
2.1	(a)	2
2.2	(b)	3
2.3	(c)	4
2.4	(d)	5
2.5	(e)	7
2.6	(f)	9

1 Introdução

Este texto refere-se ao trabalho realizado na disciplina GA030 (Estatística), do curso de pós-graduação oferecido pelo Laboratório Nacional de Computação Científica (LNCC), sob a orientação do professor Marcio Rentes Borges. Neste documento, serão apresentadas as questões propostas pelo trabalho, seguidas de suas respectivas resoluções. Clique aqui para acessar o código referente à realização de todos os exercícios.

2 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

2.1 (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 [1] e utilizando os parâmetros presentes na tabela 1, iremos calcular a média e a variância teórica de cada variável aleatória:

- 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 aleatórias.

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

2.2 (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 estimar a média e a variância de cada variável, utilizaremos os estimadores clássicos,

pois eles são estimativas não tendenciosas, de variância mínima e consistentes:

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

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

Foi desenvolvido um programa em Python para calcular as estimativas da média e da variância de cada variável aleatória. Os resultados obtidos estão apresentados na Tabela 3, com precisão de três casas decimais.

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

A realização deste exercício revelou que as estimativas da média e da variância encontradas foram altamente precisas, aproximando-se dos valores reais teóricos. Essa precisão é garantida pelo grande número de amostras em cada conjunto de dados. Com 5 milhões de amostras por variável aleatória, é possível calcular a média e a variância com elevado grau de exatidão.

2.3 (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 aleatória, foi construído o histograma das frequências relativas, normalizado de forma que, devido à grande quantidade de dados, o histograma se aproximasse da função de densidade de probabilidade (PDF) ou da função de probabilidade (PMF) correspondente à distribuição teórica dos dados. Para verificar a consistência com o modelo teórico, a PDF ou PMF teórica foi sobreposta ao histograma. A Figura 1 apresenta os resultados obtidos. No caso de variáveis contínuas, as partições (bins) foram ajustadas automaticamente.

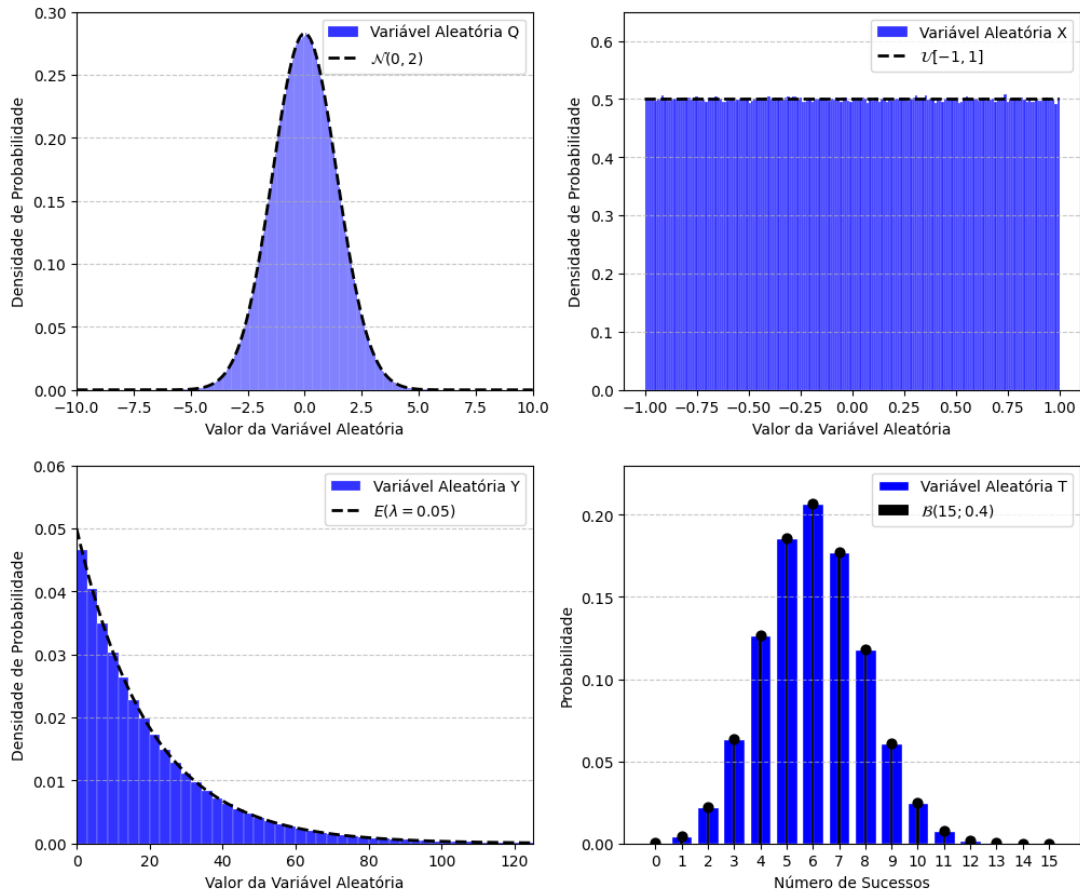


Figura 1: Histograma de frequências relativas dos dados com distribuições sobrepostas

A análise da Figura 1 demonstra que os histogramas estão alinhados com seus respectivos modelos teóricos, uma vez que os dados apresentaram um ajuste adequado às PDFs ou PMFs em todos os casos.

2.4 (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 resolver esta questão, foram realizados 1000 sorteios aleatórios contendo 5, 10 e, posteriormente, 50 valores de cada variável aleatória. Com esses sorteios, foram construídas as variáveis aleatórias referentes à média amostral e à variância amostral utilizando as fórmulas descritas no enunciado da questão. A Figura 2 apresenta o código utilizado para implementar esse procedimento.

```
1 n_samples = [5,10,50]
2
3 for key, dat in data.items():
4     for n in n_samples:
5         medias = []
6         variancias = []
7
8         amostras = np.random.choice(dat, size=(10000, n), replace=False)
9         medias = amostras.mean(axis=1)
10        variancias = amostras.var(axis=1, ddof=1)
11
12        if key == 'data_q':
13            medias_amostrais_q[f'{key}_{n}'] = medias
14            variancias_amostrais_q[f'{key}_{n}'] = variancias
15        elif key == 'data_x':
16            medias_amostrais_x[f'{key}_{n}'] = medias
17            variancias_amostrais_x[f'{key}_{n}'] = variancias
18        elif key == 'data_y':
19            medias_amostrais_y[f'{key}_{n}'] = medias
20            variancias_amostrais_y[f'{key}_{n}'] = variancias
21        elif key == 'data_t':
22            medias_amostrais_t[f'{key}_{n}'] = medias
23            variancias_amostrais_t[f'{key}_{n}'] = variancias
24        else:
25            raise(ValueError)
```

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

2.5 (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, espera-se que as médias amostrais sigam uma distribuição normal, com a mesma média da variável de origem e com variância igual à variância da variável aleatória de origem dividida por n , onde n representa o número de amostras que compõem a média amostral (neste caso, $n = 5$, $n = 10$ ou $n = 50$).

No caso da variância, para variáveis originadas de distribuições normais, ao ser multiplicada pelo fator $\frac{n-1}{\sigma^2}$, onde σ^2 é a variância da variável de origem, espera-se que ela siga uma distribuição qui-quadrado com $n - 1$ graus de liberdade. Para as demais variáveis, não há uma expectativa específica sobre o comportamento da variância amostral.

Dessa forma, foram construídos histogramas e sobrepostas as curvas teóricas esperadas. Os resultados estão apresentados nas Figuras 3, 4, 5 e 6. Cada coluna dessas figuras exibe o histograma da média amostral (em azul) e o da variância amostral (em vermelho) para $n = 5$, $n = 10$ e $n = 50$, respectivamente.

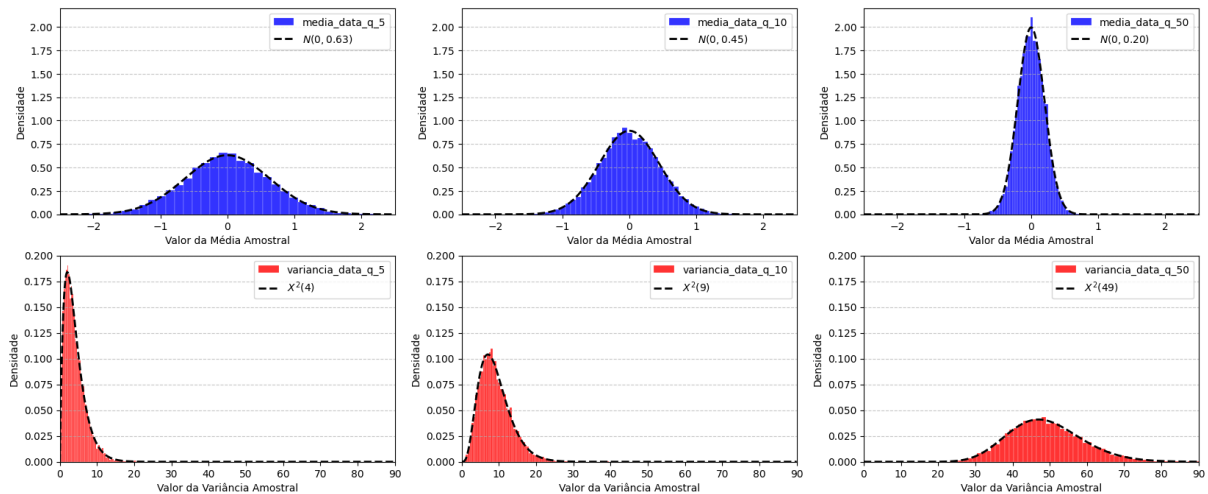


Figura 3: Histograma das variáveis média e variância amostrais com origem Q

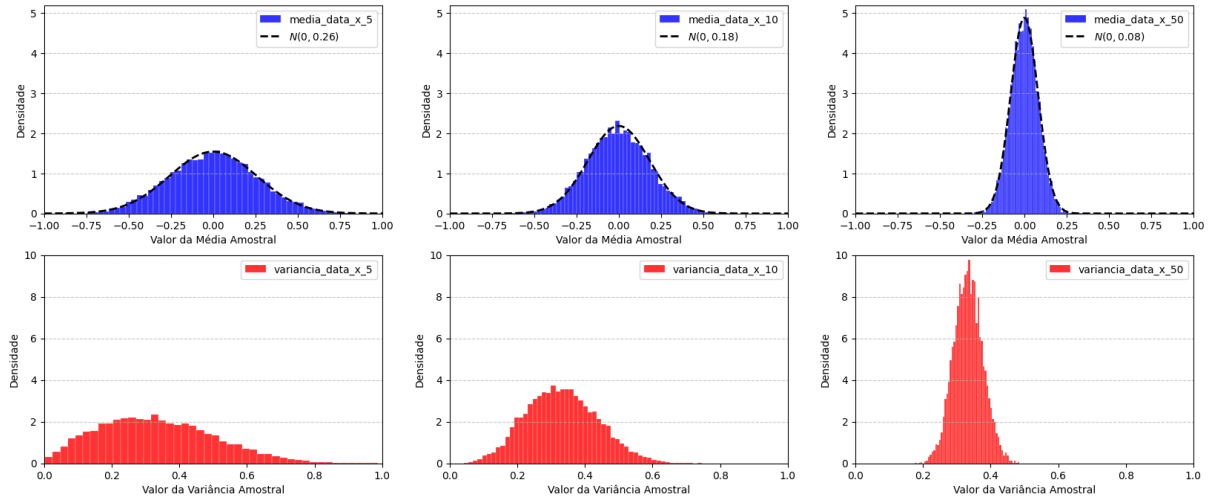


Figura 4: Histograma das variáveis média e variância amostrais com origem X

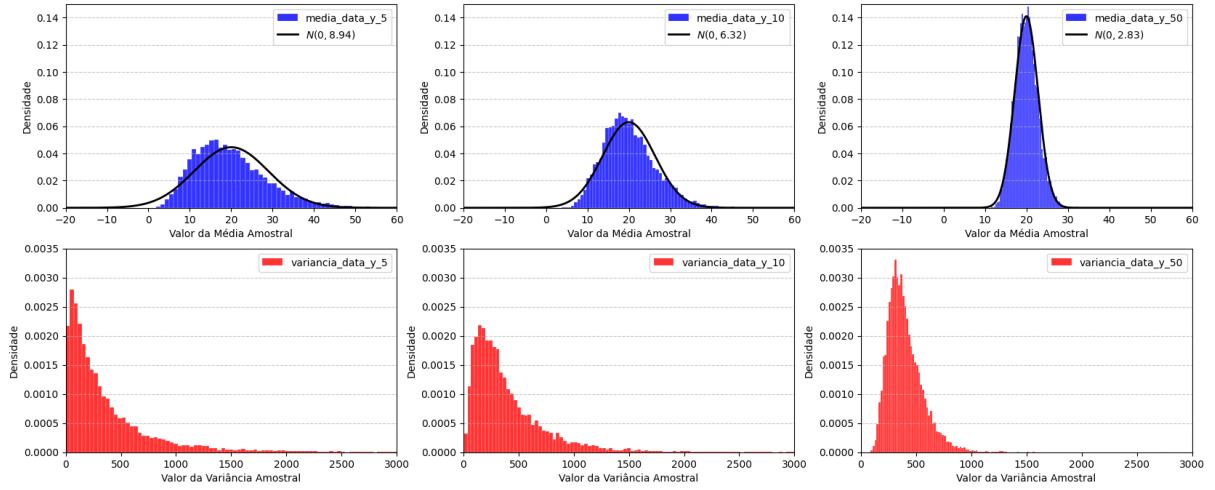


Figura 5: Histograma das variáveis média e variância amostrais com origem Y

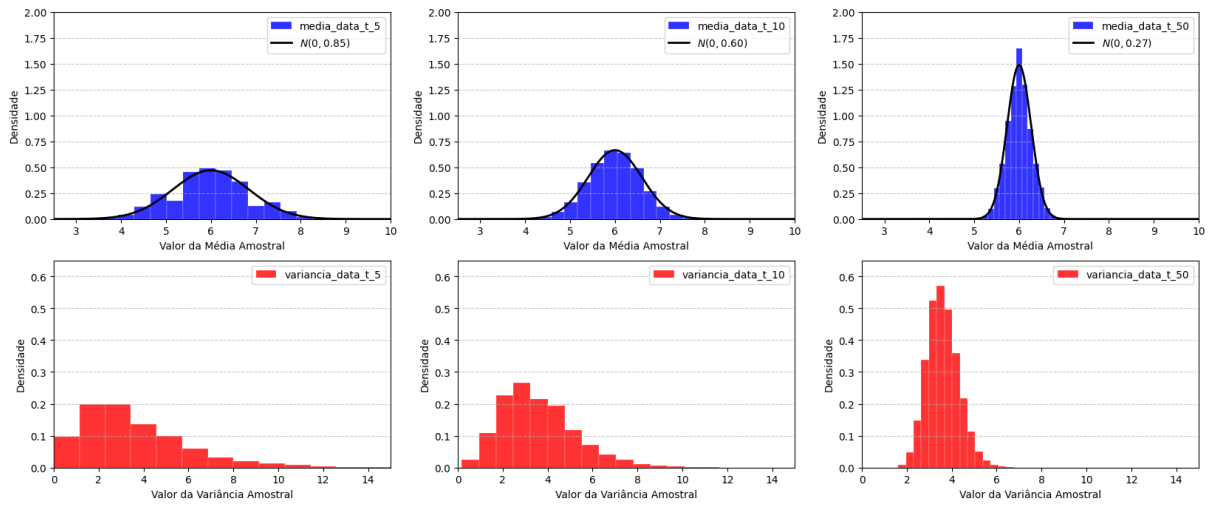


Figura 6: Histograma das variáveis média e variância amostrais com origem T

Ao observar a sobreposição das curvas, nota-se que os histogramas se ajustam de forma satisfatória às distribuições teóricas esperadas para cada variável. Esse comportamento indica que as médias amostrais seguem uma distribuição aproximadamente normal, com a mesma média da variável original e variância reduzida por um fator de n .

Além disso, a variância amostral, no caso da variável originada de uma distribuição normal, apresenta o comportamento esperado de uma distribuição qui-quadrada, com $n - 1$ graus de liberdade.

Esses resultados corroboram a consistência das estimativas para as médias e variâncias amostrais, conforme esperado teoricamente.

2.6 (f)

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

Resolução:

Ao observar os histogramas, podemos concluir então sobre a média amostral no geral que:

- Para $n = 5$ o histograma da média amostral apresenta uma distribuição mais dispersa, o que é esperado devido ao tamanho pequeno da amostra.
- Para $n = 10$, a média amostral já mostra uma tendência mais aproximada de uma distribuição normal, com menor dispersão.
- Para $n = 50$, o histograma se aproxima ainda mais de uma distribuição normal, o que é uma confirmação do teorema central do limite, já que a média amostral deve convergir para uma distribuição normal à medida que o tamanho da amostra aumenta.

Além disso, observando os resultados sobre a variância:

- Para $n = 5$, a variância amostral apresenta uma grande dispersão, e isso é esperado, pois amostras pequenas tendem a apresentar maior variabilidade.
- Para $n = 10$, a variância amostral começa a se estabilizar, com uma distribuição mais concentrada.
- Para $n = 50$, no caso da variável original normal, a variância amostral já está bem ajustada ao modelo teórico de uma distribuição qui-quadrada com $n - 1$ graus de liberdade.

No entanto, é interessante observar que, embora não seja esperado que as variâncias amostrais de dados que não seguem uma distribuição normal apresentem um comportamento específico, à medida que o valor de n aumenta, essas variâncias acabam convergindo visualmente para uma

distribuição que se assemelha a uma distribuição normal ou qui-quadrada. Esse comportamento pode ser atribuído ao aumento do tamanho da amostra, o que pode gerar uma estabilização dos dados e uma aproximação com distribuições conhecidas, mesmo quando os dados de origem não seguem essas distribuições. Isso reforça a ideia de que, em amostras grandes, as distribuições amostrais podem exibir padrões de comportamento mais previsíveis.

Em resumo, como esperado, conforme o valor de n aumenta, os histogramas das médias amostrais se aproximam de uma distribuição normal e a variância amostral, no caso da variável de origem normal, se ajusta mais precisamente à distribuição qui-quadrada teórica. Este comportamento é consistente com os resultados teóricos e confirma a adequação dos métodos de estimação para médias e variâncias amostrais.

Referências

- [1] BORGES, M. R. Estatística, 2024. <https://lncc.br/~mrborges/>. Acessado em: 22 nov. 2024.