Trabalho – GA-030 Estatística

Roger de Souza Passos

[1]Laboratório Nacional de Computação Científica (LNCC)
Petrópolis – RJ – Brasil

rspassos@posgrad.lncc.br

1. Introdução

O presente trabalho tem como objetivo a fixação das ideias introduzidas na disciplina GA-030. Para isso, utilizaremos dados armazenados em quatro arquivos, que contêm amostras de diferentes variáveis aleatórias, conforme a tabela 1.

Variável	Arquivo	Distribuição
$Q \sim \mathbb{N}(1,2)$	`data1q.dat`	Normal
$X \sim \mathbb{U}[-3,3]$	`data1x.dat`	Uniforme
$Y \sim \mathbb{E}(\lambda = 0.1)$	`data1y.dat`	Exponencial
$T \sim \mathbb{B}(20, 0.20)$	`data1t.dat`	Binomial

Tabela 1. Descrição das variáveis aleatórias. Cada arquivo possui 10^6 pontos amostrais.

2. Exercícios

(a) Dado que conhecemos a distribuição de probabilidades de cada variável aleatória e os parâmetros que as caracterizam (tabela 1), calcule a expectância e a variância (teóricas) de cada uma delas.

Utilizando as definições vistas na disciplina:

Variável	Expectância	Variância
Q	$E[Q] = \mu_Q = 1$	$Var[Q] = \sigma_Q^2 = 2$
X	$E[X] = \dfrac{(a+b)}{2} = \dfrac{(-3+3)}{2} = 0$	$Var[X] = \dfrac{(b-a)^2}{12} = \dfrac{36}{12} = 3$
Y	$E[Y] = \dfrac{1}{\lambda} = \dfrac{1}{0.1} = 10$	$Var[Y] = \dfrac{1}{\lambda^2} = \dfrac{1}{0.1^2} = 100$
T	$E[T] = Np = 20 \cdot 0.20 = 4$	$Var[T] = Np(1-p) = 20 \cdot 0.20 \cdot 0.80 = 3.2$

Tabela 2. Expectância e variância teóricas das variáveis aleatórias.

(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.

Este e os demais experimentos computacionais foram realizados utilizando Python 3.11.2 com auxílio das bibliotecas NumPy e Matplotlib em suas versões 2.3.1 e 3.10.3 respectivamente. As estimativas para a média ($\hat{\mu}$) e variância ($\hat{\sigma}^2$) foram calculadas utilizando as expressões para uma realização da média amostral e da variância amostral:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} a_i \tag{1}$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \hat{\mu})^2 \tag{2}$$

onde $n = 10^6$ é o tamanho da amostra para cada variável e a_i corresponde a um ponto amostral da variável correspondente. Os resultados obtidos são apresentados na tabela 3.

Variável	$\hat{\mu}$	$\hat{\sigma}^2$
Q	0.99953775	2.00302271
X	−0.00217247	3.00119156
Y	9.98905957	99.56633665
T	4.00196200	3.20480736

Tabela 3. Resultados para as estimativas de média e variância.

Ao comparar estes resultados com os valores teóricos apresentados na tabela 2, observa-se concordância entre a teoria e o experimento numérico. A pequena discrepância observada nos valores estimados em relação aos teóricos é esperada devido à natureza das variáveis aleatórias. Isso porque mesmo com 10^6 pontos amostrais, as flutuações estatísticas permanecem, ainda que em magnitude reduzida.

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

A figura 1 apresenta os histogramas dos dados de cada variável, com as alturas normalizadas para que a área do histograma some 1 (`density = True`[1]) sobrepostos com suas respectivas distribuições teóricas (tabela 1). Em todos os casos, as distribuições empíricas mostram excelente concordância com os modelos teóricos, validando a qualidade dos dados amostrais.

[1] Matplotlib – Normalizing histograms: density and weight

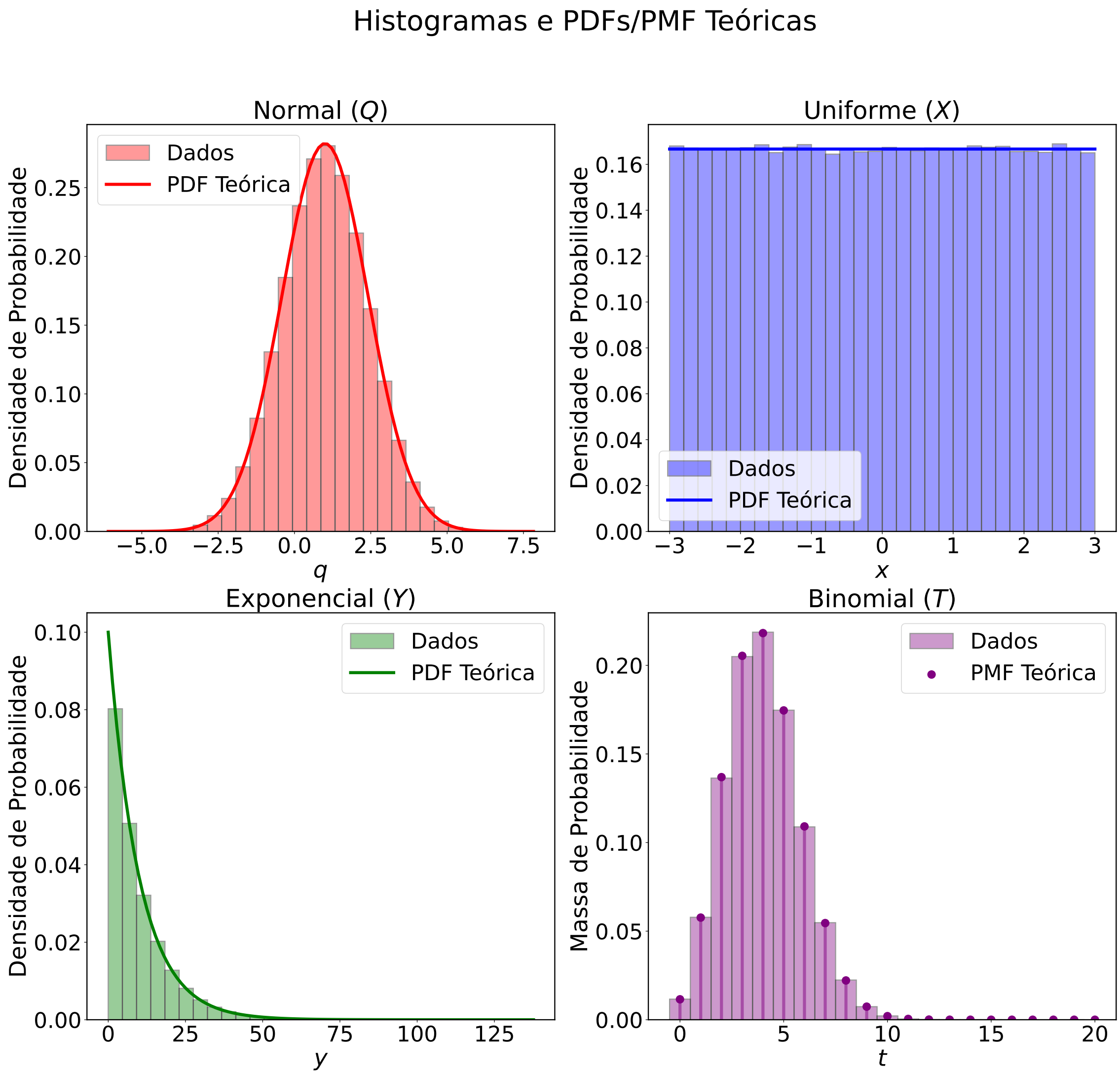

Figura 1. Histogramas normalizados das variáveis aleatórias com suas respectivas distribuições teóricas sobrepostas.

(d) 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.

O código utilizado para geração das amostras e cálculo das estatísticas é apresentado a seguir:

Código 1. Código de amostragem e cálculo das estatísticas do item (d)

```
data = {
    'normal': data_q,
    'uniform': data_x,
    'exponential': data_y,
    'binomial': data_t
}

ns = [5, 10, 50]
N = 10000

sample_mean = {dist: {n:[] for n in ns} for dist in data}
sample_variance = {dist: {n:[] for n in ns} for dist in data}

for dist in data:
    for n in ns:
        for _ in range(N):

            sample = np.random.choice(data[dist], size=n)

            sample_mean[dist][n].append(np.mean(sample))

            sample_variance[dist][n].append(np.var(sample, ddof=1))
```

A tabela 4 apresenta as estimativas dos valores esperados para as estatísticas $\bar{W}^{(n)}$ e $S_W^{2(n)}$, calculadas a partir de $N = 10^4$ amostras de tamanho n. Ao tomar a diferença absoluta entre as estimativas e os valores teóricos, observa-se que as médias amostrais aproximam-se dos valores teóricos (μ_W), com diferenças absolutas menores que 0.081 em todos os casos. As variâncias amostrais também convergem estatisticamente para os valores teóricos (σ_W^2). Isso demonstra que, conforme esperado pela teoria, ambos estimadores são não-tendenciosos para seus respectivos parâmetros populacionais - isto é, $E[\bar{W}^{(n)}] \to \mu_W$ e $E[S_W^{2(n)}] \to \sigma_W^2$, independentemente do tamanho amostral n. Em ambos os casos, a variável aleatória exponencial (Y) apresenta uma maior diferença absoluta em relação às demais, o que é consistente com sua maior variabilidade intrínseca ($\sigma_Y^2 = 100$).

Variável (W)	n	$E[\bar{W}^{(n)}]$	μ_W	$\lvert E[\bar{W}^{(n)}] - \mu_W \rvert$	$E[S_W^{2(n)}]$	σ_W^2	$\lvert E[S_W^{2(n)}] - \sigma_W^2 \rvert$
	5	1.005343	1	0.005343	1.983100	2	0.016900
Q	10	0.999830	1	0.000170	1.987086	2	0.012914
	50	0.998441	1	0.001559	2.009807	2	0.009807
	5	-0.000911	0	0.000911	2.984214	3	0.015786
X	10	0.000660	0	0.000660	2.998373	3	0.001627
	50	0.004270	0	0.004270	3.005434	3	0.005434
	5	9.959956	10	0.040044	98.792593	100	1.207407
Y	10	9.919729	10	0.080271	98.718798	100	1.281202
	50	9.986716	10	0.013284	99.533197	100	0.466803
	5	4.007920	4	0.007920	3.205340	3.2	0.005340
T	10	4.001510	4	0.001510	3.192012	3.2	0.007988
	50	3.999778	4	0.000222	3.206950	3.2	0.006950

Tabela 4. Resultados das estimativas de média e variância utilizando as variáveis aleatórias média ($\bar{W}^{(n)}$) e variância ($S_W^{2(n)}$) amostrais para diferentes tamanhos de amostra n.

(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).**

As figuras 2 a 5 apresentam os histogramas das médias amostrais considerando diferentes tamanhos de amostra, conforme o código do exercício anterior. Além disso, as figuras também mostram a função densidade de probabilidade teórica esperada, que corresponde à $\mathbb{N}(\mu_W, \frac{\sigma_W^2}{n})$, onde W representa a variável aleatória de origem. As figuras 7 a 10 exibem os histogramas das variâncias amostrais para diferentes tamanho de amostra. Já a figura 6 mostra a variância amostral da variável de origem normal Q, que quando escalada por um fator $\frac{(n-1)}{\sigma_Q^2}$, apresenta uma distribuição qui-quadrado com $(n-1)$ graus de liberdade.

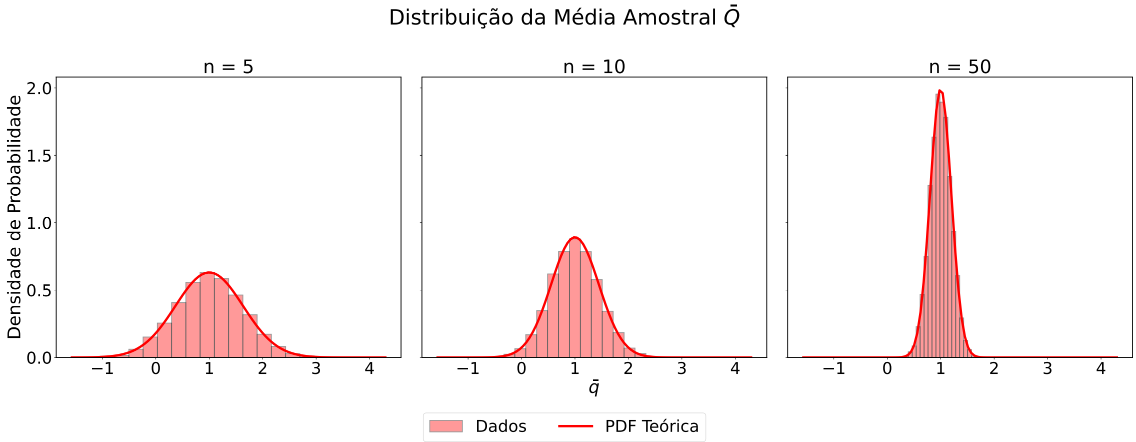

Figura 2. Histograma da média amostral $\bar{Q}$ e PDF teórica $\mathbb{N}(\mu_Q, \frac{\sigma_Q^2}{n})$.

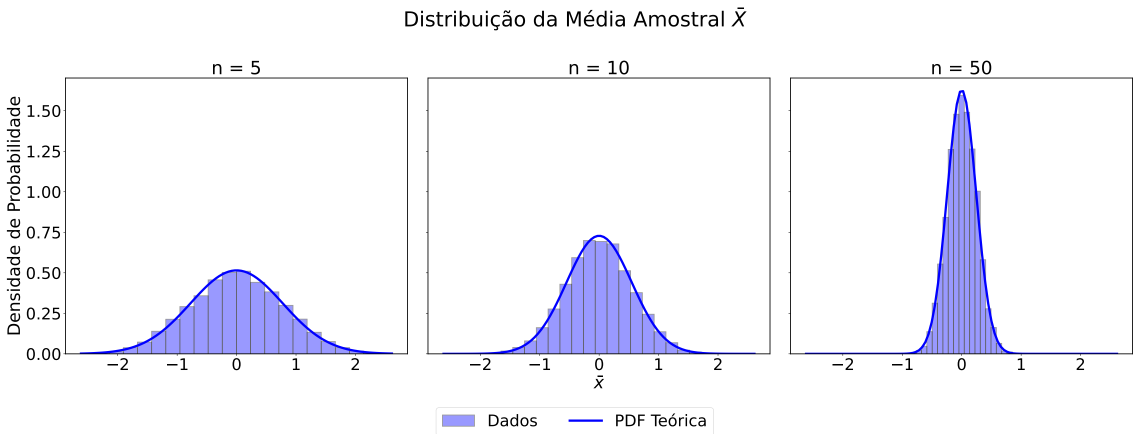

Figura 3. Histograma da média amostral $\bar{X}$ e PDF teórica $\mathbb{N}(\mu_X, \frac{\sigma_X^2}{n})$.

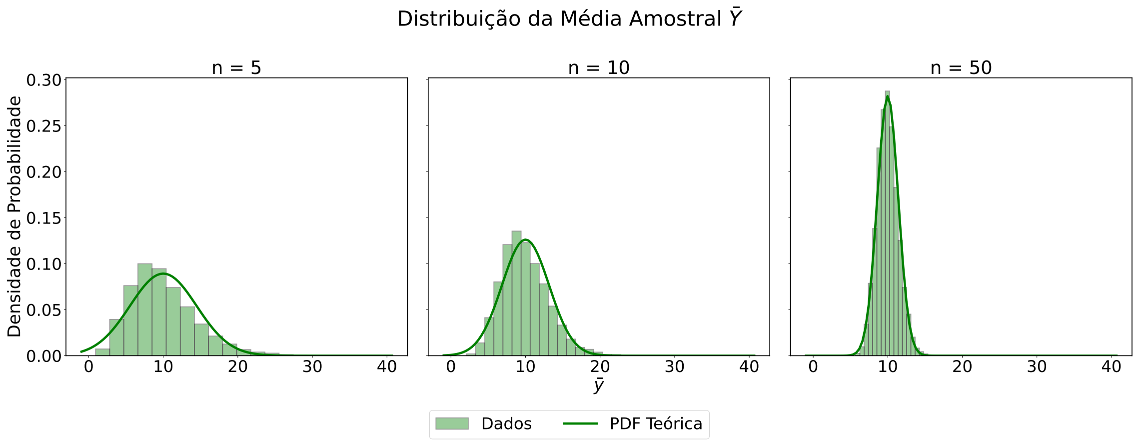

Figura 4. Histograma da média amostral $\bar{Y}$ e PDF teórica $\mathbb{N}(\mu_Y, \frac{\sigma_Y^2}{n})$.

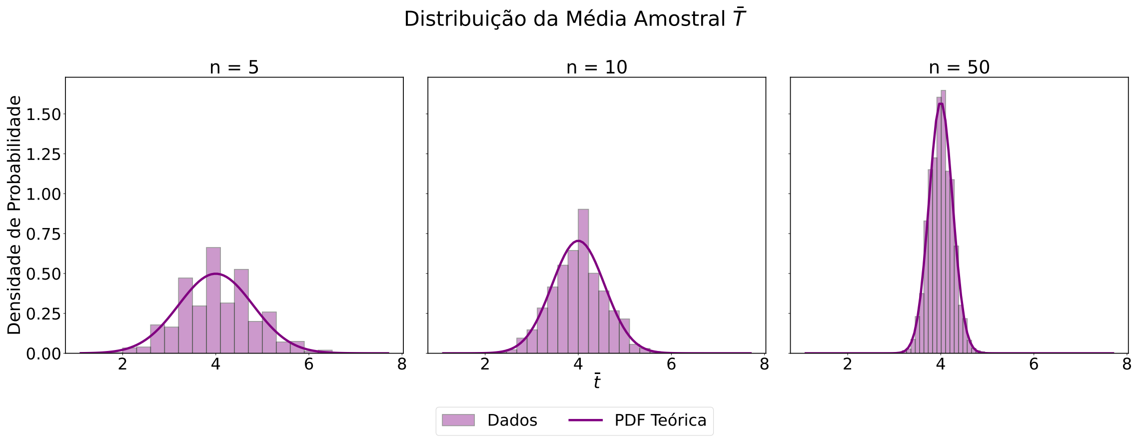

Figura 5. Histograma da média amostral $\bar{T}$ e PDF teórica $\mathbb{N}(\mu_T, \frac{\sigma_T^2}{n})$.

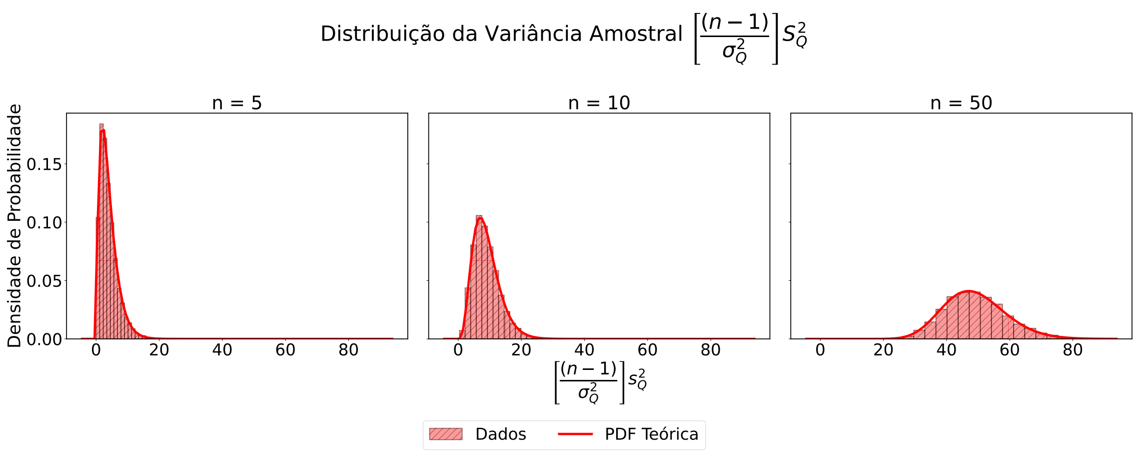

Figura 6. Histograma da variância amostral S_Q^2 escalada e PDF teórica $\chi^2_{(n-1)}$.

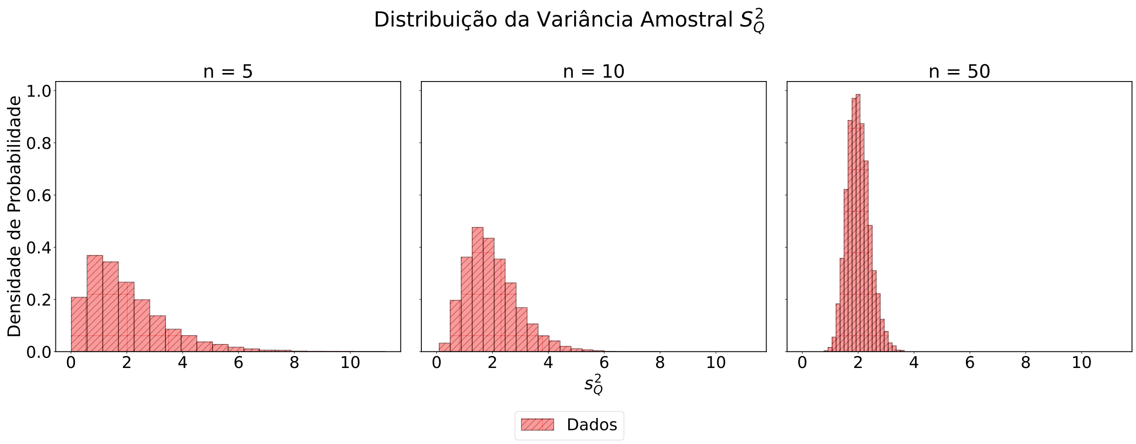

Figura 7. Histograma da variância amostral S_Q^2.

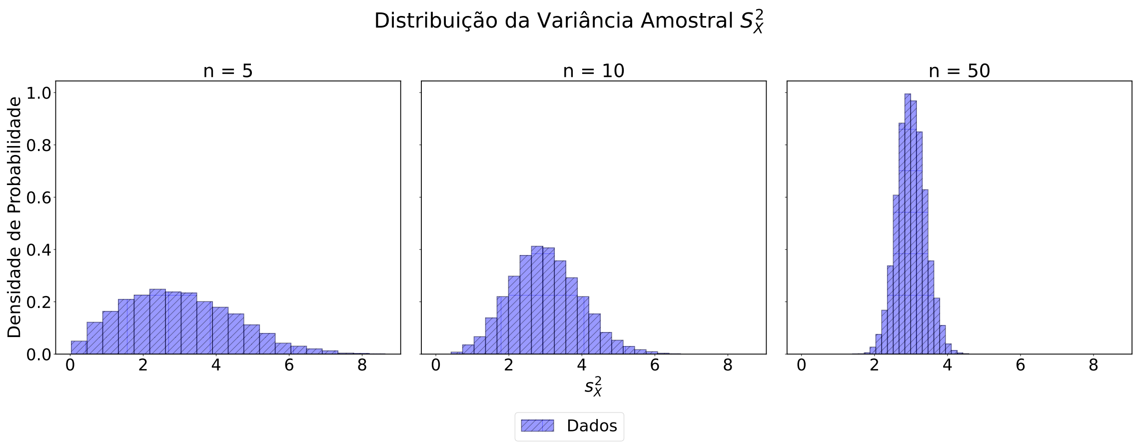

Figura 8. Histograma da variância amostral S_X^2.

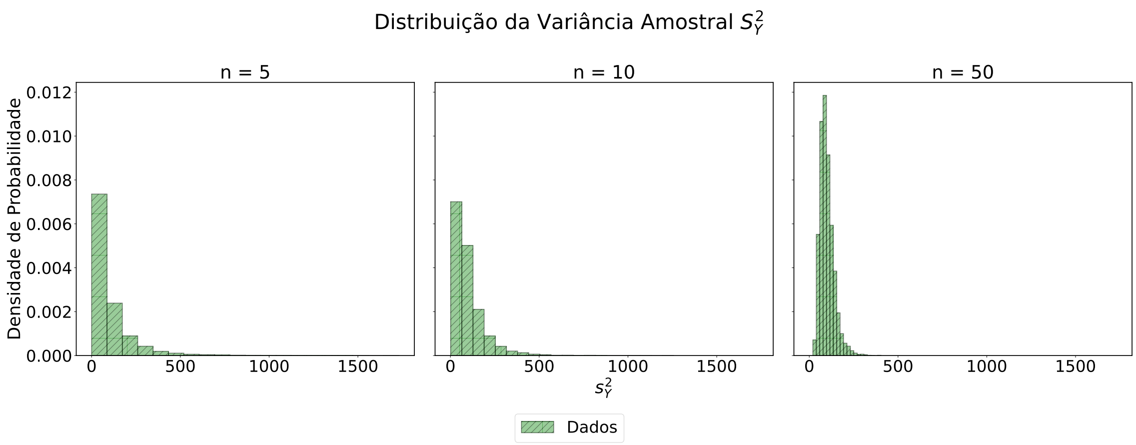

Figura 9. Histograma da variância amostral S_Y^2.

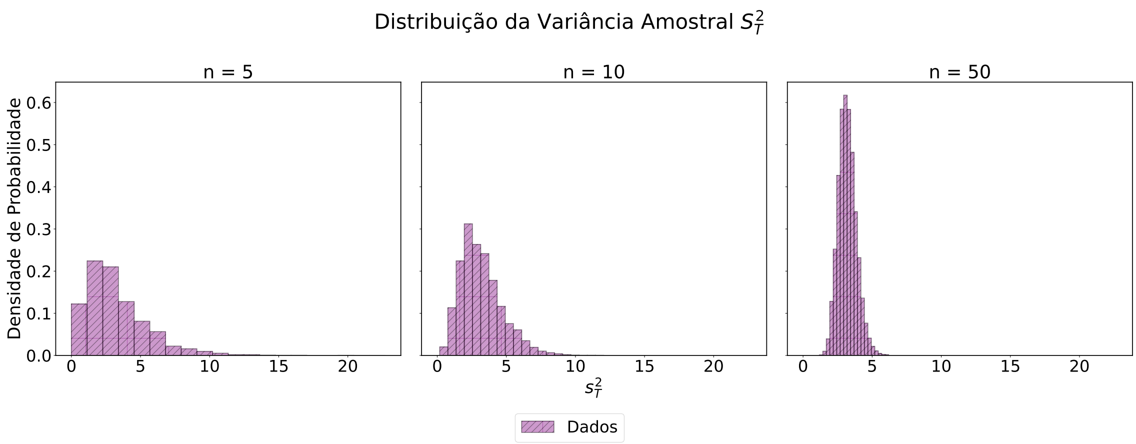

Figura 10. Histograma da variância amostral S_T^2.

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

Com relação aos histogramas das médias amostrais (figuras 2 a 5), é possível observar que:

i. $\bar{Q}$ aproxima-se de uma normal independentemente de n e $\bar{X}$ segue uma tendencia parecida apesar de não previsto teoricamente, devido à distribuição de origem uniforme ser simétrica;

ii. $\bar{Y}$ e $\bar{T}$ só aproximam-se melhor da normal para $n = 50$, ilustrando o Teorema do Limite Central;

iii. A variância das médias amostrais diminui proporcionalmente a $1/n$, como esperado pela fórmula $\sigma^2_{\bar{W}} = \sigma^2_W/n$.

Já em relação às distribuições das variâncias amostrais:

i. Para a variável normal Q, a distribuição da variância amostral segue exatamente a teoria, com a versão escalada seguindo uma $\chi^2_{(n-1)}$ (figura 6).

ii. Para as outras variáveis (figuras 7 a 10), a distribuição da variância amostral se concentra em torno do valor esperado teórico, com redução da variância à medida que n aumenta.

Referências

[Borges 2025] Borges, M. R. (2025). GA-030 Estatística. https://lncc.br/~mrborges/. Laboratório Nacional de Computação Científica (LNCC). Acesso em: Agosto de 2025.