

# Analizando a Complexidade Computacional de Problemas de Medidas de Tendência Central e Dispersão

**Aline B. Loreto\***,

Departamento de Matemática, UNISC,  
96815-900, Santa Cruz do Sul, RS  
E-mail: abl@unisc.br

**Marcília A. Campos,**

Centro de Informática, UFPE  
50732-970, Recife, PE  
E-mail: mac@cin.ufpe.br

**Dalcídio M. Claudio,**

Faculdade de Matemática, PUCRS  
90619-900, Porto Alegre, RS  
E-mail: dalcidio@inf.pucrs.br

**Laira V. Toscani**

Instituto de Informática, PPGC\*, UFRGS  
91501-970, Porto Alegre, RS  
E-mail: laira@inf.ufrgs.br.

O presente trabalho apresenta: (i) investigação da complexidade computacional em problemas das medidas de tendência central média intervalar e moda intervalar, e das medidas de dispersão variância intervalar, desvio padrão intervalar, coeficiente de variação intervalar, covariância intervalar e coeficiente de correlação intervalar; (ii) abordagem intervalar de medidas de tendência central e de medidas de dispersão, (iii) classificação quanto a classe de complexidade dos problemas dos indicadores estatísticos descritivos e (iv) forma de representação dos valores reais em valores intervalares, de tal modo que não ocorram superestimação nos intervalos solução.

A literatura ([2], [3], [4], [5]) mostra que foram realizadas pesquisas somente em problemas das medidas de dispersão variância intervalar, covariância intervalar e coeficiente de correlação intervalar, e que a utilização da computação intervalar na solução de problemas de medidas de dispersão intervalar sempre fornece solução com intervalos superestimados (intervalos com amplitude grande), e que ao procurar uma solução com intervalos de amplitude pequena (através da computação da imagem intervalar), o problema passa a pertencer a classe de problemas NP-Difícil.

Observa-se que a NP-dificuldade em problemas de medidas de dispersão está relacionada com o pro-

cessamento dos dados de entrada (pois tem-se que considerar todos os valores compreendidos entre  $\underline{x}$  e  $\bar{x}$  do intervalo  $\mathbf{x}$  para encontrar os extremos do intervalo solução). Outro fato importante, e que completa a caracterização dos problemas como NP-Difíceis, é que os problemas em questão são problemas de decisão, pois se deseja saber se existem intervalos  $y = [\underline{y}, \bar{y}]$  que contenham as soluções aproximadas destes problemas.

## **Abordagem Intervalar de Medidas de Tendência Central e Dispersão**

Sejam  $X$  e  $Y$  variáveis aleatórias populacionais seguindo a lei da probabilidade  $f_X(\theta_X)$  e  $f_Y(\theta_Y)$ , respectivamente. Sejam  $\mu_x$ ,  $\mu_y$ ,  $\sigma_x^2$  e  $\sigma_y^2$  as esperanças matemáticas e variâncias das variáveis aleatórias  $X$  e  $Y$ . Adicionalmente, sejam  $\{x_1, \dots, x_n\}$  e  $\{y_1, \dots, y_m\}$  amostras aleatórias de  $X$  e  $Y$ .

Uma forma de se considerar não somente os erros numéricos mas também os erros aleatórios inerentes aos valores  $x_1, \dots, x_n$  e  $y_1, \dots, y_m$  é considerar que, para  $i = 1, \dots, n$  e  $j = 1, \dots, m$ ,  $x_i \in [\underline{x}_i, \bar{x}_i]$  e  $y_j \in [\underline{y}_j, \bar{y}_j]$ . Neste caso, consideram-se os dados de entrada  $\{x_1, \dots, x_n\}$ ,  $x_1 \in \mathbf{x}_1, \dots, x_n \in \mathbf{x}_n$ , ou seja, domínios intervalares onde  $\mathbf{x}_1 = [\underline{x}_1, \bar{x}_1], \dots, \mathbf{x}_n = [\underline{x}_n, \bar{x}_n]$ . Nestas situações, para diferentes valores possíveis  $x_i \in \mathbf{x}_i$ , obtém-se diferentes valores da média  $me$ , moda  $mo$ , variância  $va$ , desvio padrão

$dp$ , coeficiente de variação  $cv$ , covariância  $co$  e coeficiente de correlação  $cc$ .

Considerando que tem-se dados de entrada intervalares, deseja-se conhecer quais são os intervalos da média intervalar  $\mathbf{ME}_v$ , moda intervalar  $\mathbf{MO}_v$ , variância intervalar  $\mathbf{VA}_v$ , desvio padrão intervalar  $\mathbf{DP}_v$ , coeficiente de variação intervalar  $\mathbf{CV}_v$ , covariância intervalar  $\mathbf{CO}_v$  e coeficiente de correlação intervalar  $\mathbf{CC}_v$  dos possíveis valores da média  $me$ , moda  $mo$ , variância  $va$ , desvio padrão  $dp$ , coeficiente de variação  $cv$ , covariância  $co$  e coeficiente de correlação  $cc$ .

A seguir apresenta-se uma abordagem intervalar para os indicadores estatísticos descritivos média, moda, variância, coeficiente de variação, covariância e coeficiente de correlação:

- **Média aritmética Intervalar:**

$$\mathbf{ME}_v = [\underline{me}, \overline{me}] = \frac{1}{n-1} \left[ \sum_{i=1}^n \underline{x}_i, \sum_{i=1}^n \overline{x}_i \right].$$

- **Moda Intervalar:** Se existe um valor modal para os dados reais, então

$$\mathbf{MO}_v = [\underline{mo}, \overline{mo}] = [mo\{\underline{x}_i\}, mo\{\overline{x}_i\}],$$

$$1 \leq i \leq n.$$

- **Variância Intervalar:**

$$\mathbf{VA}_v = [\underline{va}, \overline{va}] = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{ME})^2.$$

- **Desvio padrão Intervalar:**

$$\mathbf{DP}_v = [\underline{dp}, \overline{dp}] = +\sqrt{[\underline{va}, \overline{va}]} = [+ \sqrt{\underline{va}}, + \sqrt{\overline{va}}].$$

- **Coeficiente de variação Intervalar:** Se  $0 \notin [\underline{me}, \overline{me}]$ , então

$$\mathbf{CV}_v = [\underline{cv}, \overline{cv}] = \frac{\mathbf{DP}}{\mathbf{ME}} = \frac{[\underline{dp}, \overline{dp}]}{[\underline{me}, \overline{me}]}.$$

- **Covariância Intervalar:**

$$\mathbf{CO}_v = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{ME}_X)(\mathbf{y}_i - \mathbf{ME}_Y),$$

onde  $\mathbf{ME}_X$  é a média intervalar dos valores de  $X$  e  $\mathbf{ME}_Y$  a média intervalar dos valores de  $Y$ .

- **Coeficiente de correlação Intervalar:** Se  $0 \notin \mathbf{DP}_X \mathbf{DP}_Y$ , então

$$\mathbf{CC}_v = [\underline{cc}, \overline{cc}] = \frac{\mathbf{CO}}{\mathbf{DP}_X \mathbf{DP}_Y},$$

onde  $\mathbf{CO}$  é a covariância intervalar,  $\mathbf{DP}_X$  o desvio padrão intervalar dos valores de  $X$  e  $\mathbf{DP}_Y$  o desvio padrão intervalar dos valores de  $Y$ .

A partir das expressões intervalares definidas para os indicadores estatísticos, propõe-se algoritmos para a solução dos problemas de computar os intervalos de medidas de tendência central intervalar e dispersão intervalar. Tais algoritmos utilizam a aritmética intervalar definida por Moore [11] e a extensão intervalar [12].

### Técnicas de Computação Intervalar

Na computação intervalar, pode-se calcular o intervalo solução  $\mathbf{y} = [\underline{y}, \overline{y}] = f(\mathbf{x}_1, \dots, \mathbf{x}_n)$  através de métodos de aproximação, técnicas de otimização, extensão intervalar e, ainda, por métodos considerados mais sofisticados [8] como forma centrada.

Dentre os métodos de aproximação existentes, descreve-se o Método de Linearização de uma função. Quando é suficiente obter os valores *aproximados* dos extremos  $\underline{y}$  e  $\overline{y}$  do intervalo  $\mathbf{y}$ , pode-se *linearizar* a função  $f(x_1, \dots, x_n)$ , isto é, representar  $x_i$  como  $x_i = \tilde{x}_i - \Delta x_i$ , e expandir a expressão resultante  $f(x_1, \dots, x_n) = f(\tilde{x}_1 - \Delta x_1, \dots, \tilde{x}_n - \Delta x_n)$  em série de Taylor em relação a  $\Delta x_i$ , omitindo termos quadráticos e de ordem superior nesta expansão. Como resultado obtém-se uma fórmula aproximada  $f(x_1, \dots, x_n) \approx a_0 + a_1 \cdot \Delta x_1 + \dots + a_n \cdot \Delta x_n$ , onde  $a_0 = f(\tilde{x}_1, \dots, \tilde{x}_n) = \tilde{y}$  e  $a_1 = -\frac{\partial f}{\partial x_i}(\tilde{x}_1, \dots, \tilde{x}_n)$ . Para a função linear resultante (aproximada), calcula-se o intervalo  $\mathbf{y}$  da seguinte maneira: i)  $\underline{y} = \tilde{y} - |a_1| \cdot \Delta_1 - \dots - |a_n| \cdot \Delta_n$  e ii)  $\overline{y} = \tilde{y} + |a_1| \cdot \Delta_1 + \dots + |a_n| \cdot \Delta_n$ .

Segundo Kreinovich *et al* [8], para muitos problemas práticos métodos de aproximação não são suficientes, pois deve-se ter uma estimativa garantida para o intervalo  $\mathbf{y}$ , e devido a uma série de fatores como: i) medições de erros relativamente grandes, de modo que seus quadrados não podem ser omitidos seguramente; ii) a função  $f(x_1, \dots, x_n)$  que descreve a relação entre quantidades medidas diretamente  $x_i$  e a quantidade desejada  $y$ , pode não ser linear; iii) muitos algoritmos de processamento de dados processam valores em grande quantidade, ou seja, quando eles processam os valores medidos em diferentes momentos de tempo e iv) existem problemas nos quais necessita-se de uma estimativa garantida, pois uma superestimação poder ser desastrosa; métodos de aproximação não são muito empregados.

Uma das técnicas de otimização é o cálculo do ponto de máximo e de mínimo de uma função, também conhecido como Imagem Intervalar [12]. O problema de encontrar os extremos do intervalo solução é um problema de otimização: o extremo inferior  $\underline{y}$  é a solução do problema de *minimização*  $f(x_1, \dots, x_n) \rightarrow \min$  sobre as condições  $\underline{x}_i \leq x_i \leq \overline{x}_i$ ,  $1 \leq i \leq n$  (onde  $\underline{x}_i = \tilde{x}_i - \Delta_i$  e  $\overline{x}_i = \tilde{x}_i + \Delta_i$ ), e o extremo superior  $\overline{y}$  é a solução do problema de *maximização*  $f(x_1, \dots, x_n) \rightarrow \max$  sobre as condições  $\underline{x}_i \leq x_i \leq \overline{x}_i$ ,  $1 \leq i \leq n$ . Para encontrar o máximo, é suficiente encontrar o ponto cuja derivada é igual a 0. Calcular o valor  $f(x)$  de

todos os pontos “candidatos” e de todos os extremos significa encontrar todos os valores de  $f(x)$  de todos estes pontos, o maior valor de  $f(x)$  é o máximo desejado (correspondentemente, o menor dos valores de  $f(x)$  é o mínimo desejado). Para uma aplicação proveitosa desta técnica deve-se considerar o tipo de função, se a função  $f(x)$  é muito complicada, então a equação  $df/dx = 0$  também será muito complicada e, por essa razão, difícil de resolver, porém para funções razoavelmente simples este método é muito eficiente.

Ratschek [14] afirma que a maioria dos métodos de otimização suportam no mínimo dois defeitos (falhas). O primeiro defeito é que o método não garante que os pontos de mínimo possam ser encontrados para uma dada tolerância. Isto significa que os resultados são subjulgados para duvidar de sua validade. O segundo defeito é que o método, dependendo das condições da função, permite encontrar somente o mínimo local ao invés do global. Estes defeitos dificultam a solução de problemas de otimização global. A otimização global é considerada, por essa razão, um assunto intratável.

Segundo Ferson *et al* [3], historicamente o primeiro método para computar o intervalo solução é o método o qual chamamos de extensão intervalar [12]. Este método está baseado no fato que dentro do computador, todo algoritmo consiste de operações elementares (aritméticas e lógicas). Para cada operação elementar  $f(a, b)$ , se conhecemos os intervalos  $\mathbf{a}$  e  $\mathbf{b}$  para  $a$  e  $b$ , podemos computar a imagem exata  $f(\mathbf{a}, \mathbf{b})$  através da aritmética intervalar definida por Moore em [11]. Na extensão intervalar, repetimos a computação formando o programa  $f$  passo-a-passo, substituindo cada operação elementar de números reais pela correspondente operação da aritmética intervalar. Em alguns casos o intervalo solução é exato, porém outros casos mais complexos (como o problema de computar o intervalo da variância, por exemplo) obtém-se intervalos solução superestimados.

Outro método, considerado sofisticado, é a forma centrada [14]. Estima-se o intervalo solução de  $f(\mathbf{x}_1, \dots, \mathbf{x}_n)$  de uma função sobre um *box*  $\mathbf{x}_1 \times \dots \times \mathbf{x}_n$  como  $f(\mathbf{x}_1, \dots, \mathbf{x}_n) \subseteq f(\tilde{x}_1, \dots, \tilde{x}_n) + \sum_{i=1}^n \frac{\partial f}{\partial x_i}(\mathbf{x}_1, \dots, \mathbf{x}_n) \cdot [-\Delta_i, \Delta_i]$ , onde  $\tilde{x}_i = (x_i + \bar{x}_i)/2$  é o ponto médio e  $\Delta_i = (x_i - \bar{x}_i)/2$  é o raio do intervalo  $\mathbf{x}$ . Quando todos os intervalos são os mesmos, a forma centrada não fornece o intervalo desejado, retorna um intervalo centrado no ponto  $f(\tilde{x}_1, \dots, \tilde{x}_n)$ . Dessa forma, como resultado de aplicação da forma centrada, obtém-se um intervalo centrado em 0 (zero), isto é, o extremo inferior do intervalo é negativo. O extremo superior produzido pela forma centrada é diferente do extremo superior do intervalo desejado.

Dentre os métodos citados acima, escolhe-se a extensão intervalar por se adequar às expressões dos indicadores estatísticos com abordagem intervalar

(definidos anteriormente), e por tornar mais acessível a projeção de algoritmos imediatos (algoritmos de fácil construção) [15].

### Complexidade Computacional de Problemas

O termo complexidade, no contexto de algoritmos, refere-se aos requerimentos de recursos necessários para que um algoritmo possa resolver um problema sob o ponto de vista computacional, ou seja, à quantidade de trabalho despendido pelo algoritmo [15]. Quando o recurso é o tempo, são escolhidas uma ou mais operações fundamentais e então são contados os números de execuções desta operação fundamental na execução do algoritmo. Segundo Toscani [15] a escolha de uma operação como operação fundamental é aceitável se o número de operações executadas pelo algoritmo é proporcional ao número de execuções da operação fundamental.

A complexidade também pode ser vista como uma propriedade do problema, o que significa dar uma medida independente do tratamento dado ao problema, independente do caminho percorrido na busca da solução, portanto independente de algoritmos. Alguns problemas são bem comportados, isto é, permitem chegar a limites de complexidades bem definidos, outros estão em classes com contornos não bem claros [15].

Quanto a complexidade do problema, são definidas várias classes, como: P, NP, NP-Completo e NP-Difícil, entre outras. Um problema (da classe P) é considerado tratável se existe um algoritmo determinístico de tempo polinomial que resolve todas as instâncias deste. Um problema é dito intratável se a sua complexidade inferior (melhor algoritmo possível) não é polinomial [15]. Para um problema pertencer a classe NP significa que existe pelo menos um algoritmo não-determinístico que o resolve em tempo polinomial. Um problema (não necessariamente da classe NP) é chamado NP-Difícil se todo o problema da classe NP pode ser reduzido a este. Se um problema da classe NP é NP-Difícil, este pode ser chamado de NP-Completo. Para problemas das classes NP-Completo e NP-Difícil não se espera encontrar um algoritmo eficiente para verificá-lo. Qualquer problema NP-Difícil pertencente ou não à classe NP, tem a propriedade de não ser resolvido em tempo polinomial, a menos que P = NP [7][8].

Identificar a tratabilidade e a intratabilidade dos problemas, mesmo aqueles que possuem algoritmos imediatos [15], é de extrema importância para os projetistas de algoritmos, pois conhecendo as classes de complexidade a que os problemas pertencem, os projetistas poderiam ter uma medida real quanto às soluções disponíveis e a expectativa de melhorar esses resultados.

### Análise da Complexidade dos Problemas de Medidas de Tendência Central e Dis-

## persão

Para a investigação da complexidade de problemas de medidas de tendência central intervalar e dispersão intervalar definimos um novo problema, para cada indicador estatístico descritivo intervalar, através da definição do domínio intervalar de funções intervalares com variáveis intervalares.

Na literatura, os problemas dos indicadores estatísticos variância, covariância e coeficiente de variação são definidos considerando variáveis reais  $x$  e  $y$ ,  $n$  valores reais  $x_1, \dots, x_n$  (onde  $x_1 \in \mathbf{x}_1, \dots, x_n \in \mathbf{x}_n$ ),  $n$  valores reais  $y_1, \dots, y_m$  (onde  $y_1 \in \mathbf{y}_1, \dots, y_n \in \mathbf{y}_n$ ) e, como domínio, intervalos  $\mathbf{x}_i$  contidos no domínio da função  $f(x_1, \dots, x_n)$ ,  $x_1 \in \mathbf{x}_1, \dots, x_n \in \mathbf{x}_n$ , ou seja,  $\mathbf{x}_i \subset \text{Dom}(f)$ . A busca da solução destes problemas é através da computação da imagem intervalar.

No presente trabalho considera-se, para os problemas de medidas de tendência central e dispersão, variáveis intervalares  $\mathbf{x}$  e  $\mathbf{y}$ ,  $n$  valores intervalares  $\mathbf{x}_1, \dots, \mathbf{x}_n$ ,  $n$  valores intervalares  $\mathbf{y}_1, \dots, \mathbf{y}_n$  e domínio composto por valores intervalares, ou seja,  $f(\mathbf{x}_1, \dots, \mathbf{x}_n)$ ,  $\mathbf{x}_i = \text{Dom}(f) \subseteq \mathfrak{R}$ . Para a solução destes problemas, aplica-se a extensão intervalar [12].

Por meio da análise da complexidade computacional verificamos que os problemas de medidas de tendência central intervalar e dispersão intervalar pertencem à classe de problemas P.

A Tabela 1 apresenta a comparação entre os resultados de complexidade dos problemas dos indicadores estatísticos descritivos considerando a utilização da extensão intervalar (nossos resultados) e a imagem intervalar (estado da arte).

Problemas dos Indicadores Intervalares	Complexidade dos Problemas com Extensão	Complexidade dos Problemas com Imagem
	Intervalar	Intervalar
Média	P	-
Moda	P	-
Variância	P	NP-Difícil
Desvio Padrão	P	-
Coef. de Variação	P	-
Covariância	P	NP-Difícil
Coef. de Correlação	P	NP-Difícil

Tabela 1: Problemas dos Indicadores Intervalares, complexidade dos problemas com extensão intervalar e com imagem intervalar.

Utilizamos o símbolo “-” na Tabela 1 para indicar que não foram encontrados resultados de complexidade para os referidos problemas intervalares.

### Verificação da Superestimação

Segundo Ratschek *et al* [14], os computadores utilizam uma aritmética chamada aritmética de ponto flutuante. Nesta aritmética números reais são aproximados por um subconjunto de números reais chamados representação numérica da máquina. Devido esta representação são gerados dois tipos de

erros: i) ocorre quando uma entrada de valor real é aproximada por um número de máquina e ii) é causado por resultados intermediários aproximados pelos números de máquina. A aritmética intervalar fornece uma ferramenta para estimar e controlar esses erros *automaticamente*. No lugar de aproximar um valor real  $x$  por um número de máquina, o valor real  $x$ , usualmente desconhecido, é aproximado por um intervalo  $\mathbf{x}$  tendo número de máquina nos extremos inferior e superior. O intervalo  $\mathbf{x}$  contém o valor  $x$ . O comprimento (ou diâmetro) deste intervalo pode ser usado como medida para qualidade da aproximação. Os cálculos são executados usando intervalos no lugar de números reais e, conseqüentemente, a aritmética real é substituída pela aritmética intervalar. Os erros, na computação intervalar, podem ser estimados por *erro absoluto*:  $|x - m(\mathbf{x})| < w(\mathbf{x})/2$ , onde  $w(\mathbf{x}) = \bar{x} - \underline{x}$  é o diâmetro do intervalo  $\mathbf{x}$ , e por *erro relativo*:  $\left| \frac{x - m(\mathbf{x})}{x} \right| \leq \frac{w(\mathbf{x})}{2 \min|\mathbf{x}|}$  se  $0 \notin \mathbf{x}$ .

Para certificação da não ocorrência de superestimação nos intervalos solução, apresentam-se exemplos de cálculos intervalares usando sistema de ponto flutuante e arredondamento direcionado [9].

Os valores reais  $\{x_1, \dots, x_n\}$ , das amostras aleatórias de uma população, são representados por intervalos  $\mathbf{x}_i = [x_i - \delta, x_i + \delta]$ , onde  $\delta$  é a precisão escolhida para os valores reais.

Na verificação da superestimação no intervalo solução, deve-se considerar o valor  $x_k \in \mathfrak{R}$ , o intervalo  $\mathbf{x}_k = [\underline{x}_k, \bar{x}_k]$  e uma dada exatidão  $\varepsilon$ . A análise do erro é realizada através das seguintes medidas de erros absolutos:

- i)  $|x_k - m(\mathbf{x}_k)| < \varepsilon$ , onde  $m(\mathbf{x}_k) = (\underline{x}_k + \bar{x}_k)/2$ , ou seja, o ponto médio do intervalo  $\mathbf{x}_k$ ;
- ii)  $|x_k - m(\mathbf{x}_k)| < w(\mathbf{x}_k)/2$ , onde  $w(\mathbf{x}_k) = \bar{x}_k - \underline{x}_k$ , isto é, o diâmetro do intervalo  $\mathbf{x}_k$ .

Estamos considerando estas medidas de erros para realizar uma completa comparação entre os erros obtidos após o processamento de operações aritméticas intervalares.

Exemplificando, consideramos a altura de uma turma de 21 alunos da 5a. série da escola Imaculada Conceição, localizada na cidade de Cachoeira do Sul, RS:  $\{1.44, 1.31, 1.53, 1.45, 1.51, 1.43, 1.46, 1.42, 1.40, 1.41, 1.49, 1.47, 1.50, 1.60, 1.49, 1.53, 1.51, 1.60, 1.48, 1.32, 1.46\}$ . Representamos estes valores em intervalos com precisão  $\delta = 0.005$ :  $\{[1.435, 1.445], [1.305, 1.315], [1.525, 1.535], [1.445, 1.455], [1.505, 1.515], [1.425, 1.435], [1.455, 1.465], [1.415, 1.425], [1.395, 1.405], [1.405, 1.415], [1.485, 1.495], [1.465, 1.475], [1.495, 1.505], [1.595, 1.605], [1.485, 1.495], [1.525, 1.535], [1.505, 1.515], [1.595, 1.605], [1.475, 1.485], [1.315, 1.325], [1.455, 1.465]\}$ .

Os cálculos dos indicadores estatísticos intervalares foram realizados utilizando sistema de ponto flutuante F(10, 4, -10, 10) com arredondamento

direcionado [9]. As operações intervalares envolvidas podem ser encontradas em Moore [11]. Como exatidão, na medida do erro, considera-se  $\varepsilon = 10^{-2}$ .

Na Tabela 2 apresentam-se os resultados dos indicadores estatísticos com dados reais e dados intervalares.

Problemas dos Indicadores	Dados Reais	Dados Intervalares
Média	1.5405	[1.5352, 1.5458]
Moda	1.46, 1.49, 1.51, 1.53, 1.60	[1.455, 1.465], [1.485, 1.495], [1.505, 1.515], [1.525, 1.535], [1.595, 1.605]
Variância	0.01105	[0.009339, 0.01299]
Desvio Padrão	0.1051	[0.09663, 0.1139]
Coef. de Variação	0.06824	[0.06251, 0.07422]

Tabela 2: Problemas dos Indicadores, dados reais e dados intervalares.

Na Tabela 3 apresentamos os erros obtidos ao calcularmos os indicadores estatísticos com valores intervalares.

Problemas dos Indicadores	Erros (i)	Absolutos (ii)
Média	0.005250	$0 < 0.005250$
Moda	0.005	$0 < 0.005$
Variância	0.001933	$0.0001103 < 0.001823$
Desvio Padrão	0.008824	$0.0001681 < 0.008656$
Coef. de Variação	0.005981	$0.0001291 < 0.005852$

Tabela 3: Problemas dos Indicadores e erros absolutos.

Para calcular os indicadores estatísticos covariância e coeficiente de correlação precisa-se de dois conjuntos de dados. Dessa forma, considera-se a altura de 21 alunos da turma da 6a. série da mesma escola citada anteriormente: {1.48, 1.44, 1.50, 1.60, 1.41, 1.65, 1.55, 1.56, 1.54, 1.55, 1.53, 1.51, 1.68, 1.49, 1.38, 1.50, 1.48, 1.42, 1.56, 1.56, 1.55}. Representamos estes valores em intervalos com a mesma precisão  $\delta = 0.005$ : {[1.475, 1.485], [1.435, 1.445], [1.495, 1.505], [1.595, 1.605], [1.405, 1.415], [1.645, 1.655], [1.545, 1.555], [1.555, 1.565], [1.535, 1.545], [1.545, 1.555], [1.525, 1.535], [1.505, 1.515], [1.675, 1.685], [1.485, 1.495], [1.380, 1.390], [1.495, 1.505], [1.475, 1.485], [1.415, 1.425], [1.555, 1.565], [1.555, 1.565], [1.545, 1.555]}.

A Tabela 4 apresenta os resultados da covariância e coeficiente de correlação com dados reais e dados intervalares.

Problemas dos Indicadores	Dados Reais	Dados Intervalares
Covariância	0.004495	[0.002688, 0.006439]
Coef. de Correlação	0.3975	[0.2027, 0.6769]

Tabela 4: Problemas dos Indicadores, dados reais e dados intervalares (covariância e coeficiente de correlação).

Na Tabela 5 apresentamos os erros obtidos ao calcularmos a covariância e coeficiente de correlação com valores intervalares.

Problemas dos Indicadores	Erros (i)	Absolutos (ii)
Covariância	0.001944	$0.00006823 < 0.001876$
Coef. de Correlação	0.2793	$0.04229 < 0.2371$

Tabela 5: Problemas dos Indicadores e erros absolutos (covariância e coeficiente de correlação).

Os cálculos reais dos indicadores estatísticos foram realizados no software NetBook [13]. Os cálculos intervalares foram realizados no software Maple Intervalar [10].

### Considerações Finais

As pesquisas desenvolvidas estão concentradas em alguns indicadores estatísticos como variância e covariância, talvez por serem os indicadores mais utilizados ou mais comuns na área de estatística.

Conforme comentado anteriormente, e descrito na literatura, a utilização da computação intervalar para calcular o intervalo da variância, covariância e coeficiente de correlação sempre fornece intervalos superestimados (intervalos com amplitude grande), e que ao utilizar a imagem intervalar os problemas destes indicadores estatísticos passam a pertencer a classe de problemas NP-Difícil.

Por não existir na literatura uma abordagem intervalar da estatística descritiva (existe apenas para alguns indicadores descritivos), o presente trabalho aborda a nível intervalar, praticamente, todos os indicadores estatísticos descritivos.

Preocupados com os resultados de NP-Dificuldade obtidos para os problemas de computar os intervalos da variância, covariância e coeficiente de correlação, concentramos nossas pesquisas em métodos da computação intervalar que tornasse possível a computação dos intervalos destes indicadores, e dos demais como média, moda, desvio padrão e coeficiente de variação. Dos métodos de computação intervalar existentes na bibliografia, escolhemos a extensão intervalar por ser mais acessível na implementação dos algoritmos, e por se adaptar aos problemas de estatística descritiva com dados intervalares.

No trabalho desenvolvido, observa-se que:

- Para o indicador média intervalar, o valor intervalar é o mesmo se considerar a computação da imagem intervalar ou a extensão intervalar, ou seja, não existe intervalo com problema de superestimação. Isto devido a média apresentar uma simples operação aritmética (adição);
- Para o indicador moda intervalar não necessitamos utilizar métodos de computação intervalar (imagem intervalar ou a extensão intervalar), pois não apresenta operações aritméticas entre

intervalos. Os valores intervalares para este indicador são os intervalos dos dados de entrada, não ocorrendo problema de intervalos superestimados;

- Os indicadores variância, covariância e coeficiente de correlação (analisados na literatura) tornam-se possíveis de serem calculados utilizando a extensão intervalar. Para estes indicadores verificamos a não ocorrência de superestimação nos intervalos calculados, fato este verificado nos cálculos realizados e apresentados nas Tabelas 2, 3, 4 e 5;
- O desvio padrão e o coeficiente de variação, que antes não eram possíveis de serem calculados devido a variância ser um problema NP-Difícil, são calculados através de operações aritméticas intervalares. Os valores intervalares obtidos não são superestimados, conforme Tabelas 2 e 3.

Se estes indicadores fossem analisados como os indicadores estatísticos variância, covariância e coeficiente de correlação, ou seja, considerando a computação da imagem intervalar para computar os valores intervalares, a expectativa de solução seria a de problemas NP-Difíceis. Entretanto, a solução encontrada neste trabalho é que os problemas destes indicadores estatísticos pertencem à classe de problemas P.

As principais contribuições do presente trabalho referem-se a: (i) definição de medidas de tendência central e dispersão a nível intervalar; (ii) possibilidade de calcular os valores intervalares da variância, covariância e coeficiente de correlação; (iii) análise da complexidade dos problemas dos indicadores estatísticos descritivos; (iv) classificação quanto a classe de complexidade dos problemas de medidas de tendência central e dispersão intervalares e (v) possibilidade de obter intervalos solução sem superestimação com a extensão intervalar devido a abordagem intervalar desenvolvida para os indicadores estatísticos descritos e a forma de representação dos valores reais em valores intervalares.

## Referências

- [1] M. A. Campos, R. A. Faria, “Definição de alguns Indicadores Estatísticos usando Intervalos”, em XI Congresso Nacional de Matemática Aplicada e Computacional, SBMAC, Ouro Preto, MG, 1988.
- [2] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, M. Aviles, “Exact Bounds on Sample Variance of Interval Data”, Proc. Extended Abstracts of the 2002 SIAM Workshop on Validated Computing, pp. 67-69, Toronto, Canada, 2002.
- [3] S. Ferson, L. Ginzburg, V. Kreinovich, J. Lopez, “Absolute Bounds on the Mean of Sum, Product, etc.: A Probabilistic Extension of Interval Arithmetic”, Proc. Extended Abstracts of the 2002 SIAM Workshop on Validated Computing, pp. 70-72, Toronto, Canada, 2002.
- [4] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, M. Aviles, Computing Variance for Interval Data is NP-Hard, *ACM SIGACT News*, 33(2) (2002) 108-118.
- [5] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, M. Aviles, Exact Bounds on Finite Populations of Interval Data, *Reliable Computing*, 11(3) (2004) 207-233.
- [6] S. Lipschutz, “Probabilidade”, McGraw-Hill, São Paulo, 1972.
- [7] M. E. Garey, D. S. Johnson, “Computers and intractability: a guide to the theory of NP-completeness”, Freeman, San Francisco, 1979.
- [8] V. Kreinovich, “Computational complexity and feasibility of data processing and interval computations”, KLUWER, 1998.
- [9] U. W. Kulisch, W. L. Miranker, “Computer Arithmetic in Theory and Practice”, Academic Press, New York, 1981.
- [10] H. Grimmer, “Interval Arithmetic in Maple with intpakX”, To appear in: PAMM - Proceedings in Applied Mathematics and Mechanics, GAMM-Conference Augsburg 2002, Wiley-InterScience.
- [11] R. E. Moore, “Interval Analysis”, Prentice-Hall, Englewood Cliffs, New Jersey, 1966.
- [12] R. E. Moore, “Methods and Applications of Interval Analysis”, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1979.
- [13] M. A. Campos, E. L. Silva, D. C. Pedrosa, J. A. Loureiro, J. L. C. Silva, C. A. Ferraz, “Net-Book: uma ferramenta para avaliação de desempenho em redes de comunicação”, em Salão de Ferramentas, Simpósio Brasileiro de Redes de Computadores, pp. 967-974, 2004.
- [14] H. Ratschek, J. Rokne, “New Computer Methods for Global Optimization”, Ellis Horwood Limited, Great Britain, 1988.
- [15] L. V. Toscani, P. A. Veloso, “Complexidade de Algoritmos: análise, projetos e métodos”, Sagra-Luzzato, Porto Alegre, Instituto de Informática da UFRGS, 2001.