

INDICADORES ESTATÍSTICOS DESCRITIVOS INTERVALARES EM PYTHON

LUCAS MENDES TORTELLI¹; DIRCEU A. MARASCHIN JR. ²

ALICE FONSECA FINGER³;

ALINE BRUM LORETO⁴

¹Universidade Federal de Pelotas – lmortelli@inf.ufpel.edu.br

²Universidade Federal de Pelotas – dmaraschin@inf.ufpel.edu.br

³Universidade Federal do Pampa – alicefinger@unipampa.edu.br

⁴Universidade Federal de Santa Maria – aline.loreto@ufsm.br

1. INTRODUÇÃO

A computação científica é utilizada quando é necessário grande exatidão e confiabilidade (CAMPOS, 2002). Ao representar dados numéricos no sistema de ponto flutuante, estes são aproximados para um subconjunto finito dos números reais (RATSCHEK, 1988). Através desta aproximação de valores são ocasionados erros pela impossibilidade da máquina representá-los. O valor impreciso adquirido do término do processamento consiste no acúmulo de todos os erros que ocorreram no decorrer das operações.

A estatística descritiva tem como objetivo a representação e organização de dados de uma população ou amostra. As medidas descritivas servem para reduzir o conjunto de observações, para um pequeno grupo de valores que fornece toda a informação relevante. Devido a esta sensibilidade de correteza nos dados, todas as etapas que compreendem a análise estatística requerem precisão. Principalmente no momento do processamento erros originam-se a cada operação, não garantindo desta maneira confiabilidade nos resultados.

A matemática intervalar proposta por Moore (1966) surge para automatizar o erro computacional científico com limites confiáveis. Nesta representação numérica, os valores reais são representados através de um intervalo. Desta maneira todo erro originado para tratamento dos valores como o arredondamento e truncamentos, serão tratados pela aritmética intervalar e pelos arredondamentos direcionados. O tamanho deste intervalo pode ser utilizado como fator de avaliação da qualidade do intervalo obtido (RATSCHEK, 1988).

Este artigo visa apresentar a biblioteca estatística descritiva intervalar realizado para a linguagem Python. Todos os indicadores utilizados neste trabalho foram previamente apresentados e provados por Loreto (2006). Todas as operações serão certificadas quanto sua exatidão, através de dados fornecidos pelo IBGE sobre a incidência de AIDS no estado do Rio Grande do Sul, compreendido no período de 1990 até 2008.

2. ARITMÉTICA INTERVALAR

A aritmética intervalar surgiu da necessidade de obter resultados mais precisos e com menor erro possível. Desta maneira, a aritmética intervalar ganhou força até sua definição formal feita por Moore (1966). O objetivo de representar valores de ponto flutuante em intervalos numéricos está na análise do erro computacional gerado, e assim poder controlá-lo com limites confiáveis. Estes limites que consistem nos extremos dos intervalos são dois números de ponto flutuante que encapsulam o valor real x no intervalo X , ou seja, $X = [x_1 ; x_2]$. A aritmética intervalar utiliza-se de um arredondamento direcionado (RATSCHEK, 1988), para reduzir os possíveis ajustes incorretos que a máquina pode realizar para obtenção da solução. O arredondamento direcionado são ajustes realizados

nos limites do intervalo, arredondando-os para o menor e para o maior número de máquina que contém o resultado das operações, obtendo-se assim um intervalo de máquina, com diâmetro mínimo. As operações básicas aritméticas que compõem os conceitos de utilização de intervalos são apresentadas abaixo.

- $[x_1, x_2] + [y_1, y_2] = [x_1 + y_1, x_2 + y_2]$
- $[x_1, x_2] - [y_1, y_2] = [x_1 - y_2, x_2 - y_1]$
- $[x_1, x_2] \cdot [y_1, y_2] = [\min(x_1 y_1, x_1 y_2, x_2 y_1, x_2 y_2), \max(x_1 y_1, x_1 y_2, x_2 y_1, x_2 y_2)]$
- $[x_1, x_2] / [y_1, y_2] = [x_1, x_2] \cdot [1/y_2, 1/y_1]$, se $0 \notin [y_1, y_2]$

Em se tratando de esforço computacional necessário para realizar operações intervalares, adota-se neste trabalho o método que contém a menor complexidade para solucionar problemas intervalares, no qual é denominada de extensão intervalar. Esta realiza a troca de todo valor numérico real de uma expressão pelos seus respectivos valores intervalares. Esta forma de solucionar problemas computacionais pertence à classe P de problemas tratáveis, uma vez que somente os limites do intervalo são utilizados (LORETO, 2006).

3. A BIBLIOTECA ESTATÍSTICA DESCRITIVA INTERVALAR

A estatística descritiva tem como principal objetivo analisar o conjunto de dados fornecido, seja por meio de tabelas, gráficos ou medidas descritivas. O processo de análise deve ser criterioso desde a etapa de seleção de amostra, pois utilizar quaisquer dados disponíveis ou utilizar dados que foram obtidos com pouco cuidado, pode levar a soluções pouco confiáveis, até a etapa de sintetização de todo conjunto de dados em valores numéricos. Estes valores, chamados de medidas descritivas, visam facilitar a interpretação dos dados através de medidas de tendência central, variabilidade, simetria, valores discrepantes e entre outros.

A biblioteca utiliza as expressões intervalares apresentadas na seção anterior nos indicadores descritivos e contém funções básicas para o funcionamento do tipo intervalar no ambiente IntPy(VARJÃO, 2011). Algumas funções não serão apresentadas por não estarem ligadas diretamente com o tipo intervalar e também por serem funções internas da linguagem Python.

Todas as operações disponíveis pela biblioteca são apresentadas nas Tabelas 1 e 2. A Tabela 2 apresenta as operações da estatística descritiva que foram implementadas com intervalos.

Na aritmética intervalar tem-se ferramentas para estimar e controlar os erros gerados inerente as aproximações numéricas. Estas medidas estão contidas na biblioteca para fornecer à análise qualitativa dos valores gerados. As medidas de erros disponibilizadas na biblioteca são apresentadas na Tabela 1.

Tabela 1: Operações para análise de erro disponíveis

Operação	Assinatura Método
Diâmetro	diam()
Erro Absoluto (EA)	absError()
Erro Relativo (ER)	relativeError()

Tabela 2: Indicadores Estatísticos Descritivos intervalares disponibilizados pela biblioteca

	Operação	Assinatura Método
ME	Média	iAverage()
MD	Mediana	iMedian()
AT	Amplitude Total	iRange()
VA	Variância	iVariance()
DP	Desvio Padrão	iSDeviation()
CO	Covariância	icoVariance()
CV	Coeficiente de Variação	icoefVariance()
CC	Coeficiente de Correlação	icoefCorrelation()

4. RESULTADOS

Todos os dados utilizados para a implementação foram obtidos através do banco de dados do IBGE, voltado a oferecer dados oficiais oriundos de pesquisas sobre a sociedade publicamente. Todos os resultados tanto reais quanto intervalares foram obtidos através da implementação de todas as expressões dos indicadores estatísticos descritivos no ambiente de programação intervalar IntPy.

4.1 INCIDÊNCIAS DE AIDS NO ESTADO DO RS NO PERÍODO DE 1990 ATÉ 2008

Considerando a quantidade de incidências de AIDS anuais no estado do Rio Grande do Sul durante os anos de 1990 até 2008, obteve-se os seguintes dados {5, 12, 7, 47, 7, 79, 10, 59, 12, 46, 14, 42, 17, 55, 21, 78, 28, 62, 24}. Como os indicadores de Covariância e Coeficiente de Correlação necessitam de dois objetos de estudos para realizar seu processamento, toma-se como outra fonte de informações as incidências de AIDS no estado de São Paulo durante o mesmo período registrado no estado do Rio Grande do Sul. Assim os dados obtidos referentes ao estado de São Paulo são {16, 68, 21, 27, 25, 97, 27, 48, 28, 19, 31, 53, 34, 1, 34, 13, 36}. A fim de manter a integridade e confiabilidade sobre todos os processos realizados, será aplicado a mesma precisão de 10^{-11} para ambos os estados analisados.

Tabela 3: Indicadores estatísticos calculados para incidência de AIDS no estado do Rio Grande do Sul, durante o o período de 1990 até 2008

	Real	Intervalar
ME	32.8947368421	[32.894736842095234, 32.89473684211529]
MD	24.0	[23.999999999999999, 24.000000000000001]
AT	74.0	[74.0, 74.0]
VA	579.041551247	[579.041551245673, 579.0415512474017]
DP	24.063282221	[24.063282220962147, 24.06328222099807]
CO	181.221606648	[181.221606647611, 181.22160664878797]
CC	0.329395877614	[0.32939587761237493, 0.3293958776154382]
CV	0.731523779518	[0.7315237795170262, 0.7315237795185644]

Tabela 4: Medidas de Qualidade dos Indicadores estatísticos calculados para incidência de AIDS no estado do Rio Grande do Sul

	Erro Absoluto	Erro Relativo
ME	$7.1054273576 \times 10^{-15} < 1.00293107153 \times 10^{-11}$	$2.16004991671 \times 10^{-16} \leq 3.04891045744 \times 10^{-13}$
MD	$0.0 < 1.00008890058 \times 10^{-11}$	$0.0 \leq 4.16703708576 \times 10^{-13}$
AT	$0.0 < 0.0$	$0.0 \leq 0.0$
VA	$0.0 < 8.64361027197 \times 10^{-10}$	$0.0 \leq 1.49274439 \times 10^{-12}$
DP	$3.5527136788 \times 10^{-15} < 1.79607440032 \times 10^{-11}$	$1.47640444316 \times 10^{-16} \leq 7.46396266239 \times 10^{-13}$
CO	$2.84217094304 \times 10^{-14} < 5.88485704611 \times 10^{-10}$	$1.56833999853 \times 10^{-16} \leq 3.24732638396 \times 10^{-12}$
CC	$5.55111512313 \times 10^{-17} < 1.5316359292 \times 10^{-12}$	$1.68524122504 \times 10^{-16} \leq 4.64983332609 \times 10^{-12}$
CV	$0.0 < 7.69107000309 \times 10^{-13}$	$0.0 \leq 1.05137662212 \times 10^{-12}$

Na Tabela 3 todos os intervalos encapsularam o valor real e contém qualidade, uma vez que seu diâmetro é pequeno. Posteriormente analisando a Tabela 4 verifica-se que os resultados mantiveram-se com pouco erro, com valores compreendidos entre 10^{-13} e 10^{-10} .

5. CONCLUSÕES

Operações aritméticas realizadas sem um controle do erro podem conter erros difíceis de serem tratados. Desta maneira, faz-se necessário o uso da aritmética intervalar para manter a robustez no controle do erro numérico.

O objetivo deste trabalho é fornecer uma ferramenta para o cálculo dos indicadores estatísticos descritivos visando a alta exatidão e garantias de resultados com qualidade. A biblioteca foi implementada no ambiente intervalar IntPy, apresentando os indicadores estatísticos descritivos de forma intervalar. Conjuntamente é fornecido medidas de qualidade dos intervalo soluções obtidos. Todo o processo foi guiado a fornecer a comunidade científica uma ferramenta confiável.

6. REFERÊNCIAS BIBLIOGRÁFICAS

CAMPOS, M. A., DIMURO, G. P., ARÁUJO, J. F. F. and DIAS, A. M., **Probabilidade intervalar e cadeias de markov intervalares no maple**, Tendências em Matemática Aplicada e Computacional, vol.3, no.2, pp.53-62, 2002.

LORETO, A. B.; **Complexidade computacional de problemas de estatística descritiva com entradas intervalares**. Dissertação (Doutorado em Computação)-Curso de Pós-graduação em Computação, Universidade Federal do Rio Grande do Sul, 71f, 2006.

MOORE, R. E. **Interval Analysis**. Englewood Cliffs: Prentice Hall, 1966.

RATSCHEK, H.; ROKNE, J. **New Computer Methods for Global Optimization**. Ellis Horwood, 1988.

VARJÃO, F. R. G. **IntPy: Computação científica auto validável em Python**. Dissertação (Mestrado em Ciência da Computação)-Programa de Pós-Graduação em Ciência da Computação, Universidade Federal de Pernambuco, 2011.