Cálculo Numérico

Introdução ao Cálculo Numérico e Erros

Ivo Calado

Instituto Federal de Educação, Ciência e Tecnologia de Alagoas

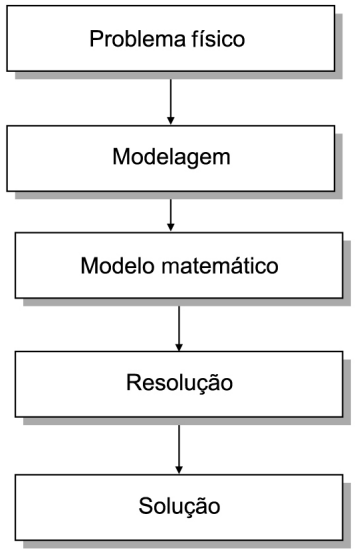
23 de Fevereiro de 2016

Roteiro

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Fases na resolução de problemas físicos

- O processo de resolução de problemas físicos, em geral, é composto por 5 fases



- A disciplina de Cálculo Numérico tem como enfoque a **Resolução** do modelo matemático construído

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Etapas do processo de resolução e objetivos da disciplina

- Definir se o problema tem solução
- Caso possua solução, definir se a solução é única ou não
- Definir uma solução aproximada do problema, **exclusivamente** através de métodos númericos concebidos na área de **Análise Numérica**

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Etapas do processo de resolução e objetivos da disciplina

- Definir se o problema tem solução
- Caso possua solução, definir se a solução é única ou não
- Definir uma solução aproximada do problema, **exclusivamente** através de métodos númericos concebidos na área de **Análise Numérica**

Afinal, qual o objetivo da disciplina?

Propiciar ao estudante o conhecimento de processos numéricos convebidos pela análise numérica

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Exemplo prático da utilização de Cálculo Numérico I

Como obter uma solução para o seguinte problema?

$$x^2 - 5x + 6 = 0$$

$$x = \frac{-b \pm \sqrt{\Delta}}{2a} \; .: \Delta = b^2 - 4ac$$

Mas como obter uma solução para o seguinte problema?

$$x^6 - 20x^5 - 110x^4 + 50x^3 - 5x^2 + 70x - 100 = 0$$

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Problema numérico x Problema não-numérico

- Problema numérico é aquele cujos dados de entrada e saída são conjuntos numéricos finitos

$$x^6 - 20x^5 - 110x^4 + 50x^3 - 5x^2 + 70x - 100 = 0$$

- Problema não-numérico é aquele onde tanto os dados de entrada quanto os de saída não se apresentam como uma quantidade finita de números reais

$$\begin{cases} \dfrac{d^2y}{dx^2} = x^2 + y^2, x \in (0,5) \\ y(0) = 0 \\ y(5) = 1 \end{cases}$$

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

O que é um método numérico?

O que é um método numérico?

É um conjunto de procedimentos utilizados para resolver um problema numérico

A escolha do método mais eficiente para resolver um problema numérico deve envolver os seguintes aspectos:

- precisão desejada para os resultados
- capacidade do método em conduzir aos resultados desejados (velocidade de convergência)
- esforço computacional despendido (tempo de processamento, economia de memória necessária para a resolução)

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Iteração ou Aproximação Sucessiva

- Uma das ideias fundamentais do Cálculo Numérico é a de iteração ou aproximação sucessiva
- Grande parte dos métodos numéricos tem como característica a iteratividade
- Um método iterativo se caracteriza por envolver os seguintes itens:
  - Tentativa inicial
  - Equação de recorrência
  - Teste de parada

Como enumeramos?

- Todo o processo de contagem é baseado em sistemas de numeração
- São caracterizados por serem posicionais (i.e., são formados por somas de potências)

Como enumeramos?

- Todo o processo de contagem é baseado em sistemas de numeração
- São caracterizados por serem posicionais (i.e., são formados por somas de potências)

Decimal

$$(20{,}36)_{10} = 2 \text{ X } 10^1 + 0 \text{ X } 10^0 + 3 \text{ x } 10^{-1} + 6 \text{ x } 10^{-2}$$

Binário

$$(101{,}11)_2 = 1 \text{ X } 2^2 + 0 \text{ X } 2^1 + 1 \text{ X } 2^0 + 1 \text{ X } 2^{-1} + 1 \text{ X } 2^{-1}$$

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Como enumeramos?

- Todo o processo de contagem é baseado em sistemas de numeração
- São caracterizados por serem posicionais (i.e., são formados por somas de potências)

Decimal

$(20{,}36)_{10} = 2 \text{ X } 10^1 + 0 \text{ X } 10^0 + 3 \text{ x } 10^{-1} + 6 \text{ x } 10^{-2}$

Binário

$(101{,}11)_2 = 1 \text{ X } 2^2 + 0 \text{ X } 2^1 + 1 \text{ X } 2^0 + 1 \text{ X } 2^{-1} + 1 \text{ X } 2^{-1}$

Hexadecimal

$(\text{ABC,D})_{16} = \text{A}(10) \text{ X } 16^2 + \text{B}(11) \text{ X } 16^1 + \text{C}(12) \text{ X } 16^0 + \text{D}(13) \text{ X } 16^{-1}$

Representação em binário e decimal

- Todo sistema de numeração posicional de base β admite apenas digitos 0, 1, ..., $\beta - 1$
- Números racionais e irracionais ocorrem em ambas as representações e muitas vezes um número com uma sequência finita de casas em uma representação possui uma representação infinita na outra.

$$\frac{1}{10} = (0{,}1)_{10} = (0{,}0001100110011...)_2 = \frac{1}{16} + \frac{1}{32} + \frac{0}{64} + \frac{0}{128} + \frac{1}{256} + ...$$

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Conversão entre representações

Para converter um número real x de decimal para binário deve-se converter separadamente as partes inteira ($ip(x)$) e fracionária ($fp(x)$) de x:

1. Converte-se $ip(x)$ dividindo-se sucessivamente o número por 2 e obtendo o último quociente e os restos das sucessivas divisões
2. Converte-se $fp(x)$ multiplicando-se sucessivamente o número por 2. O número à esquerda da virgula será um dos dígitos da representação binária. Se o dígito for 1, subtrai esse valor e continua com o processo.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Bits, Bytes e palavras

Definições

- Bit: menor unidade de informação em um computador (0 e 1)
- Byte: conjunto formado por 8 bits
- Palavra: conjunto formado por 1 ou mais bytes

- A leitura e o salvamento de dados são feito em termos de palavras e não de bytes!
- Tamanho típicos de palavras atualmente são 32 bits e 64 bits

A memória de um computador está organizado como um conjunto de palavras.

Representação de números inteiros

- Atualmente os números inteiros são armazenados em palavras de 32 bits
- Considerando-se apenas números positivos, tem-se 2^{32} números possíveis de serem representados
- Para representar números negativos fazemos uso do bit mais significativo para indicar o sinal

Exemplo, considerando uma palavra de 8 bits

10 = 00001010
-10 = 10001010

- Reduz-se o intervalo de representação de números
- O zero passa a ter duas representações: +0 e -0
- Faz-se necessário um hardware para processar a subtração

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Representação via complemento-de-2

- Utiliza-se apenas o hardware da somas
- a subtração passa a ser definida como a soma de um número positivo com um negativo

10 - (+10) => 10 + (-10)

Tipos

1. Representação *racional*
2. Representação em *ponto-fixo*
3. Representação em *ponto-flutuante*

Representação racional

- Utiliza a representação racional de dois inteiros (numerador e denominador)
- Pode-se representar de maneira exata as frações
- Presente nos sistemas MAPLE, MATHEMATICA e DERIVE
- Não há implementação em *hardware* disponível, apenas em *software*. Como consequência, aumenta-se o tempo de proessamento

Representação em ponto fixo

- O ponto binário ocupa uma posição fixa. Uma parte dos bits representa a parte inteira e a outre parte representa a parte decimal
- A palavra do computador é dividida em três campos

1. s, sinal do número
2. e, digitos a esquerda do ponto binário
3. d, digitos a direita do ponto binário

Exemplo de representação do número 11,75 numa palavra de 32 bits

|1|00000000001011|11000000000000000|

Problemas na representação de ponto-fixo

- Duas representações do zero
- Intervalo bastante pequeno para representar números

Representação em ponto-flutuante

- Os números são representados em notação científica

Representação binária de um número em ponto-flutuante

$$x = \pm M \times 2^{\pm E}$$

$$F = (\beta, |M|, |E|)$$

- β representa a base trabalhada
- M representa a mantissa (23 bits float, 52 bits double)
- E representa o expoente (8 bits float, 11 bits double)

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Convertendo número decimal em ponto-flutuante

1. Transformar o número em notação científica de base 2 ($1,x * 2^y$)
2. Calcular a mantissa baseado na parte fracionária
3. Definir o sinal do número (0 => positivo, 1 => negativo)
4. Definir o expoente (E + 127 para o float) ou (E + 1023 para o double)
5. Montar a representação binária

Convertendo número em ponto flutuante em decimal

1. Extrair os campos
2. Calcular o expoente real (exp - 127)
3. Calcular o valor decimal da mantissa
4. Montar a representação binária

Exercício I

Desenvolver um programa em C++ que possibilite a conversão de numeros entre os sistemas binário-decimal. Inicialmente, o usuário deverá escolher o tipo de conversão que poderá ser um dos seguintes:

- Decimal para binário
- Binário para decimal

O sistema deve levar em consideração a possibilidade do número passado ser fracionário.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Exercício II

Converter:

- 2,4
- 0,25
- -2,64
- 4,66
- -1562,234

Exemplo

Considere, por exemplo, uma máquina que opera no sistema:
$F = f(\beta, t, m, M) = f(10, 3, -5, 5)$
Nesse sistema os números serão representados na seguinte forma:

$$0.d_1\, d_2\, d_3 \cdot 10^e, \qquad 0 \leq d_j \leq 9, \qquad d_1 \neq 0, \qquad E \in [-5, 5]$$

O menor número, em valor absoluto, representado nessa máquina é:

$$m = 0,100 \cdot 10^{-5} = 10^{-6}$$

e o maior número, em valor absoluto, é:

$$M = 0,999 \cdot 10^5 = 99900$$

Considere o conjunto dos números reais $\mathbb{R}$ e o seguinte conjunto:

$$G = \{x \in \mathbb{R} | m \leq |x| \leq M\}$$

Dado um número real x, várias situações poderão ocorrer:

Considere o conjunto dos números reais $\mathbb{R}$ e o seguinte conjunto:

$$G = \{x \in \mathbb{R} | m \leq |x| \leq M\}$$

Dado um número real x, várias situações poderão ocorrer:

Caso 1) $x \in G$: O número será representado exatamente, truncado ou arredondado. Exemplo: $x = 235,8 = 0,2358 \cdot 10^3$ será representado por $0,235 \cdot 10^3$ ou $0,236 \cdot 10^3$.

Considere o conjunto dos números reais $\mathbb{R}$ e o seguinte conjunto:

$$G = \{x \in \mathbb{R} | m \leq |x| \leq M\}$$

Dado um número real x, várias situações poderão ocorrer:

Caso 1) $x \in G$: O número será representado exatamente, truncado ou arredondado. Exemplo: $x = 235,8 = 0,2358 \cdot 10^3$ será representado por $0,235 \cdot 10^3$ ou $0,236 \cdot 10^3$.

Caso 2) $|x| < m$: Exemplo: $x = 0,345 \cdot 10^{-7}$. O número não pode ser representado nesta máquina porque o expoente é menor que -5, ocorrendo *underflow*.

Considere o conjunto dos números reais $\mathbb{R}$ e o seguinte conjunto:

$$G = \{x \in \mathbb{R} | m \leq |x| \leq M\}$$

Dado um número real x, várias situações poderão ocorrer:

Caso 1) $x \in G$: O número será representado exatamente, truncado ou arredondado. Exemplo: $x = 235,8 = 0,2358 \cdot 10^3$ será representado por $0,235 \cdot 10^3$ ou $0,236 \cdot 10^3$.

Caso 2) $|x| < m$: Exemplo: $x = 0,345 \cdot 10^{-7}$. O número não pode ser representado nesta máquina porque o expoente é menor que -5, ocorrendo *underflow*.

Caso 3) $|x| > M$: Exemplo: $x = 0,872 \cdot 10^9$. O número não pode ser representado nesta máquina porque o expoente é maior que 5, ocorrendo *overflow*.

Erros de arredondamento

Um problema que pode surgir ao se representar valores decimais na forma binária está ligado ao fato de não haver tal representação finita. Este tipo de problema dá origem aos chamados **Erros de Arredondamento**.

Erros de arredondamento

Um problema que pode surgir ao se representar valores decimais na forma binária está ligado ao fato de não haver tal representação finita. Este tipo de problema dá origem aos chamados **Erros de Arredondamento**.

Exemplo

$$0,1_{10} = 0,000110011001100...{}_2$$

Erros de arredondamento

Um problema que pode surgir ao se representar valores decimais na forma binária está ligado ao fato de não haver tal representação finita. Este tipo de problema dá origem aos chamados **Erros de Arredondamento**.

Exemplo

$$0,1_{10} = 0,000110011001100...{}_2$$

O valor decimal $0,1$ tem como representação binária um número com infinitos dígitos, logo, ao se representar $0,1_{10}$ no sistema binário com 16 bits comete-se um erro, pois

0	1	1	0	0	1	1	0	0	1	1	1	0	0	1	1

que na forma decimal é igual a $0,099976_{10}$.

Erros de truncamento

Os **Erros de Truncamento** são provenientes da utilização de processos que deveriam ser infinitos ou muito grandes para a determinação de um valor e que, por razões práticas, são truncados.

Erros de truncamento

Os **Erros de Truncamento** são provenientes da utilização de processos que deveriam ser infinitos ou muito grandes para a determinação de um valor e que, por razões práticas, são truncados.

Exemplo

O cálculo da função $sen(X)$ em um programa computacional pode ser obtido pela seguinte série infinita:

$$sen(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$$

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Erros de truncamento

Os **Erros de Truncamento** são provenientes da utilização de processos que deveriam ser infinitos ou muito grandes para a determinação de um valor e que, por razões práticas, são truncados.

Exemplo

O cálculo da função $sen(X)$ em um programa computacional pode ser obtido pela seguinte série infinita:

$$sen(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$$

A solução adotada é a de interromper os cálculos quando uma determinada precisão é atingida.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Erros de truncamento

Os **Erros de Truncamento** são provenientes da utilização de processos que deveriam ser infinitos ou muito grandes para a determinação de um valor e que, por razões práticas, são truncados.

Exemplo

O cálculo da função $sen(X)$ em um programa computacional pode ser obtido pela seguinte série infinita:

$$sen(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$$

A solução adotada é a de interromper os cálculos quando uma determinada precisão é atingida.
De maneira geral, pode-se dizer que o erro de truncamento pode ser diminuído até chegar a ficar da ordem do erro de arredondamento; a partir deste ponto, não faz sentido diminuir-se mais, pois o erro de

Erros Absolutos e relativos

- Define-se *erro absoluto* como a diferença entre o valor exato de um número x e seu valor aproximado $\bar{x}$: $EA_x = x - \bar{x}$
- Em geral, apenas o valor de $\bar{x}$ é conhecido, tornando impossível obter o valor exato do erro absoluto.

Erros Absolutos e relativos

- Define-se *erro absoluto* como a diferença entre o valor exato de um número x e seu valor aproximado $\bar{x}$: $EA_x = x - \bar{x}$
- Em geral, apenas o valor de $\bar{x}$ é conhecido, tornando impossível obter o valor exato do erro absoluto. Nestes casos, obtem-se um limitante superior ou uma estimativa para o módulo do erro absoluto

Exemplo

Sabendo-se que $\pi \in (3,14; 3,15)$ tomaremos para π um valor dentro deste intervalo e teremos $|EA_\pi| = |\pi - \bar{\pi}| < 0,01$

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Erro relativo

Exemplo 2

- Considere x representado por $\bar{x} = 2112{,}9$ com $|EA_x| < 0,1$ (x $\in$ (2112,8; 2113))
- Considere y representado por $\bar{x} = 5{,}3$ com $|EA_y| < 0,1$ (y $\in$ (5,2; 5,4))

Tanto para x quanto para y os valores são representados com o mesmo erro absoluto. Porém, um é tão preciso quanto o outro?

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Erro relativo

Exemplo 2

- Considere x representado por $\bar{x} = 2112{,}9$ com $|EA_x| < 0,1$ ($x \in (2112{,}8; 2113)$)
- Considere y representado por $\bar{x} = 5{,}3$ com $|EA_y| < 0,1$ ($y \in (5{,}2; 5{,}4)$)

Tanto para x quanto para y os valores são representados com o mesmo erro absoluto. Porém, um é tão preciso quanto o outro?

O erro de 1 cm na medição entre dois planetas é o mesmo que o erro de 1 cm na medição entre letras vizinhas num texto?

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Erro relativo

Definição

O erro relativo ER_x é definido por:

$$\mathrm{ER}_x = \frac{|EA_x|}{|\bar{x}|} = \frac{x - \bar{x}}{\bar{x}}$$

Quais seriam os erros relativos dos exemplos anteriores?

$$\mathrm{ER}_x = \frac{|EA_x|}{|\bar{x}|} < \frac{0,1}{2112,9} \approx 4,7 * 10^{-5}$$

$$\mathrm{ER}_x = \frac{|EA_x|}{|\bar{x}|} < \frac{0,1}{5,3} \approx 0,02$$

Consequências do truncamento/arredondamento

Será mostrado abaixo, através de um exemplo, como os erros de arredondamento e, ou, truncamento podem influenciar o desenvolvimento de um cálculo.

Consequências do truncamento/arredondamento

Será mostrado abaixo, através de um exemplo, como os erros de arredondamento e, ou, truncamento podem influenciar o desenvolvimento de um cálculo.

Supondo-se que as operações abaixo sejam processadas em uma máquina com 4 dígitos significativos e fazendo-se

$$x_1 = 0,3491 \cdot 10^4$$

$$x_2 = 0,2345 \cdot 10^0$$

tem-se

$$(x_2 + x_1) - x_1 = (0,2345 \cdot 10^0 + 0,3491 \cdot 10^4) - 0,3491 \cdot 10^4$$

$$= 0,3491 \cdot 10^4 - 0,3491 \cdot 10^4 = 0,0000$$

Continuação...

$$x_2 + (x_1 - x_1) = 0,2345 \cdot 10^0 + (0,3491 \cdot 10^4 - 0,3491 \cdot 10^4)$$

$$= 0,2345 \cdot 10^0 + 0,0000$$

$$= 0,2345 \cdot 10^0$$

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Continuação...

$$x_2 + (x_1 - x_1) = 0,2345 \cdot 10^0 + (0,3491 \cdot 10^4 - 0,3491 \cdot 10^4)$$

$$= 0,2345 \cdot 10^0 + 0,0000$$

$$= 0,2345 \cdot 10^0$$

Os dois resultados são diferentes, quando não deveriam ser, pois a adição é uma operação distributiva. A causa desta diferença foi o arredondamento feito na adição $(x_2 + x_1)$, cujo resultado tem 8 dígitos. Como a máquina só armazena 4 dígitos, os menos significativos foram desprezados.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Continuação...

$$x_2 + (x_1 - x_1) = 0,2345 \cdot 10^0 + (0,3491 \cdot 10^4 - 0,3491 \cdot 10^4)$$

$$= 0,2345 \cdot 10^0 + 0,0000$$

$$= 0,2345 \cdot 10^0$$

Os dois resultados são diferentes, quando não deveriam ser, pois a adição é uma operação distributiva. A causa desta diferença foi o arredondamento feito na adição $(x_2 + x_1)$, cujo resultado tem 8 dígitos. Como a máquina só armazena 4 dígitos, os menos significativos foram desprezados.

Ao se utilizar máquinas de calcular deve-se estar atento a essas particularidades causadas pelo erro de arredondamento, não só na adição, mas também nas outras operações.

Exemplo 2

Considerando um sistema de aritmética de ponto flutuante de 4 digitos, na base 10, e com acumulador de precisão dupla. Dados x $= 0{,}937 * 10^4$ e y $= 0{,}1272 * 10^2$, obter x + y com truncamento e arrendondamento

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Exemplo 2

Considerando um sistema de aritmética de ponto flutuante de 4 digitos, na base 10, e com acumulador de precisão dupla. Dados x $= 0{,}937 * 10^4$ e y $= 0{,}1272 * 10^2$, obter x + y com truncamento e arrendondamento

Alinha-se os pontos decimais dos valores acima

$x + x = 0{,}937 * 10^4$ e $y = 0{,}001272 * 10^4 = 0{,}938272 * 10^4$

Então,

$$x + y = (0{,}937 + 0{,}001271) * 10^4$$

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Exemplo 3

Considerando um sitema de aritmética de ponto flutuante de 4 digitos, na base 10, e com acumulador de precisão dupla. Sejam x e y anteriores, obter xy

Exemplo 3

Considerando um sitema de aritmética de ponto flutuante de 4 digitos, na base 10, e com acumulador de precisão dupla. Sejam x e y anteriores, obter xy

$$xy = (0{,}937 * 10^4) * (0{,}1272 * 10^2) = (0{,}937 * 0{,}1272) * 10^6 =$$
$$= 0{,}1191864 * 10^6$$

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS

Exercício 1

Calcular o valor de $sen(\pi/6)$ pela série $sen(x) = \sum_{i=0}^{n} \frac{(-1)^i \cdot x^{2i+1}}{(2i+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$ utilizando $n = \{1, 2, \ldots, 5\}$. Apresente em uma tabela os valores de $sen(x)$ exato e aproximados com os respectivos valores de n.

Exercício 2

Implementar um programa computacional para calcular o valor de $sen(\pi/6)$ pela série, imprimir os valores de $n = 1, 2, \ldots, 20$, do valor aproximado de $sen(\pi/6)$, do valor exato $sen(\pi/6) = 0.5$ e do erro (valor exato - valor aproximado), utilizando ao menos 10 algarismos significativos.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA ALAGOAS