ESTATÍSTICA DESCRITIVA PARA ANÁLISE DE DADOS (PRODUÇÃO DE SOJA) VIA PROGRAMA ITERATIVO

DESCRIPTIVE STATISTICS FOR DATA ANALYSIS (SOY PRODUCTION) VIA ITERATIVE PROGRAM

Alfredo BONINI NETO[11]*

Fernando Ferrari PUTTI[2]

Ricardo César Gonçalves SANT'ANA[2]

Nelson Aparecido BONINI JUNIOR[3]

RESUMO

Neste trabalho é apresentado o desenvolvimento de uma interface gráfica, ou seja, um programa iterativo, para o estudo da estatística descritiva em dados de produção de soja no Brasil. O objetivo principal é fazer um programa simples e de fácil entendimento e que possa ser utilizado em diversas áreas de aplicação, inclusive na área de engenharia de Biossistemas. As análises feitas são: média aritmética, mediana, variância, desvio padrão e coeficiente de variação percentual. Estas análises têm por objetivo verificar se os dados estudados apresentam altos níveis de dispersão. O programa foi desenvolvido no ambiente Matlab que fornece todos os comandos possíveis para o desenvolvimento.

Palavras-chave: Estatística Descritiva, Programa Iterativo, Média, Desvio Padrão.

ABSTRACT

This paper presents the development of a graphical interface, i.e., an iterative program for the study of descriptive statistical on the data of soy production in Brazil. The main goal is make a simple program and easy to understand and that can be used in various application areas, including in the area of Biosystems Engineering. These analyzes are: arithmetic mean, median, variance, standard deviation and coefficient of variation percentage. These analyzes are intended to verify if the data studied show high or low levels of dispersion. The program was developed in Matlab environment that provides all possible commands for development.

Keywords: Descriptive Statistics, Iterative Program, Average, Standard Deviation

[1]Departamento de Matemática - Faculdades de Dracena/UNIFADRA

* alfredoboninineto@hotmail.com

[2] UNESP, Univ Estadual Paulista, Campus de Tupã.

[3] ETEC – Escola Técnica Estadual

INTRODUÇÃO

A característica de um conjunto de dados mais comumente investigada é o seu centro ou o ponto ao redor do qual as observações tendem a se agrupar. A medida de tendência central mais frequentemente utilizada é a média aritmética, calculada com a soma de todas as observações de um conjunto de dados e divisão do resultado pelo número total de medidas. A média aritmética é normalmente representada pela letra $\overline{x}$ e sua fórmula de cálculo é:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

Uma medida de tendência central que não é sensível ao valor de cada medida é a mediana, que pode ser usada como uma medida resumo para as observações ordinais, assim como para dados discretos e contínuos. Uma lista de observações é ordenada da menor até a maior, metade dos valores são maiores ou iguais à mediana, enquanto a outra metade é menor ou igual a ela. Consequentemente, se uns conjuntos de dados contêm um total de n observações, no qual n é ímpar, a mediana é o valor do meio ou a $(n+1)/2$ -ésima medida; se n for par, a medida é usualmente tomada como a média dos dois valores mais centrais do intervalo, a $(n/2)$ -ésima e $[(n/2)+1]$ -ésima observações, Pagano e Gauvreau (2004).

Outra análise importante a ser estudada neste trabalho é a medida de dispersão de um conjunto de dados. A variância quantifica a variabilidade ou o espalhamento ao redor da média das medidas. Mais explicitamente, a variância é calculada ao se subtrair a média de um conjunto de valores de cada uma das observações, elevar ao quadrado esses desvios, soma-los de dividir a soma pelo número de observações do conjunto de dados menos 1 (amostra). Representar a variância por s^2.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} \tag{2}$$

O desvio padrão de um conjunto de dados é a raiz quadrada positiva da variância.

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}} \tag{3}$$

Caso o conjunto de dados for uma população, deve-se trocar (n-1) por n.

O coeficiente de variação percentual (CV) é uma medida de dispersão relativa, pois permite comparar a dispersão de diferentes distribuições (com diferentes médias e desvios padrões). Quanto menor o coeficiente de variação percentual, mais os dados estão concentrados em torno da média, pois o desvio padrão é

pequeno em relação à média, Stevenson (2001).

$$CV = \frac{s}{\bar{x}}.100 \quad (4)$$

Segundo Pimentel Gomes (1985) estudando os coeficientes de variação obtidos nos ensaios agrícolas, classifica-os da seguinte forma:
CV<10% baixa dispersão de dados;
$10 \leq CV<20$ média ou moderada dispersão de dados;
$20 \leq CV<30$ alta dispersão de dados;
$CV \geq 30$ elevada dispersão de dados.

O SOFTWARE MATLAB

O Matlab é uma linguagem de programação apropriada ao desenvolvimento de aplicativos de natureza técnica. Para isso, possui facilidades de computação, programação e baixo custo, dentro de um ambiente amigável e de fácil aprendizado (Huang, Zhang, 2000). Com o Matlab é possível resolver problemas computacionais mais rápido do que com linguagens de programação tradicionais, como C, C + + e Fortran (Mathworks, 2009). O Matlab foi desenvolvido no início da década de 80 por Cleve Moler, no Departamento de Ciência da Computação da Universidade do Novo México, EUA. As versões posteriores ao Matlab 4.0, foram desenvolvidas na firma comercial MathWorks Inc., que detêm os direitos de autores destas implementações. O Matlab foi originalmente desenvolvido para prover um acesso amigável ao tratamento de vetores e matrizes. Atualmente o Matlab dispõe de uma biblioteca bastante abrangente de funções matemáticas, geração de gráficos e manipulação de dados que auxiliam muito o trabalho do programador.

Características dos Recursos Gráficos do Matlab

Existem muitos comandos para criação da interface gráfica no Matlab, citaremos alguns dos comandos. Podemos criar uma janela através da função Figure e formatar essa janela através de seus parâmetros (alguns destes parâmetros serão mostrados a seguir). A figura 1 mostra uma janela feita com a função Figure e seus parâmetros devidamente configurados.

```
dx=0.2850;
dy=0.2200;
pos = [(1-dx)*0.5,  (1-dy)*0.5, dx, dy];
h0 = figure('Color',[  0.800  0.800  0.800], ...
        'Units','normalized',  ...
        'MenuBar','none', ...
        'NumberTitle','off',  ...
        'Position',pos, ...
        'Resize','off',  ...
        'name','');
```

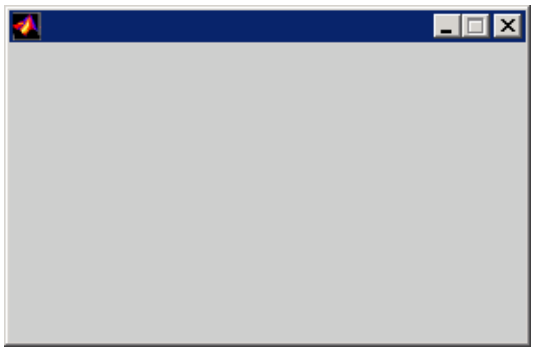

Figura 1: Exemplo da função *Figure.*

Parâmetros da Função Figure:

Color: Representa a cor de fundo da janela. É um vetor com os componentes RGB. Exemplo: a seqüência Color [0 0 0] equivale a cor preta e a seqüência Color[1 1 1] equivale a cor branca.

Units: É uma unidade usada para posicionar o controle. A posição e tamanho de um controle dentro da janela, que são feitos através de coordenadas como: Normalized (máximo e mínimo da janela correspondendo a 0 e 1) e Pixels(pontos gráficos).

MenuBar: Se o valor dessa propriedade for 'none' nenhum menu é mostrado na janela. Se for 'Figure' a janela terá o menu padrão de figuras.

NumberTitle: Se o valor dessa propriedade for 'on' aparecerá o nome e o número da janela. Se for 'off' a barra de título aparece em branco.

Position: Especifica a posição e tamanho da janela através das propriedades: [left, bottom, width, height].

Resize: Se estiver em 'on' a janela pode ter seu tamanho alterado. Se tiver em 'off' o tamanho da janela não pode ser alterado.

Name: Dá um nome para a janela. O valor desta propriedade deve ser uma string.

Controlando os Controles

No *Matlab* existe uma maneira muito prática de se programar a resposta de um controle ao usuário. Por exemplo, ao apertar-se um botão queremos que seja plotado um gráfico, ou fechar a janela que se está operando. Os controles também servem para retornar algum valor para o usuário de maneira mais amigável.

O Comando Uicontrol

O *Uicontrol* é um comando de controle para a janela que está ativa. Para criar os controles, deve-se configurar apropriadamente seus parâmetros. A figura 2 apresenta o exemplo de *BackgroundColor*.

```
h0 = figure('Color',[ 0.800 0.800 0.800], ...
            'Units','normalized', ...
            'MenuBar','none', ...
            'NumberTitle','off', ...
            'Position',pos, ...
            'Resize','off', ...
            'name','');
h1 = uicontrol('Parent',h0, ...
           'Units','normalized', ...
           'BackgroundColor',[ 1 1 1], ...
           'ForegroundColor',[0.000 0.000 0.502],...
           'HorizontalAlignment','center', ...
           'Position',[0.0787 0.7678 0.8287 0.1547], ...
           'String','Sistemas:', ...
           'FontSize',17,...
```

```
'Fontname','Arial',...
'Style','text', ...
'Tag','StaticText1');
```

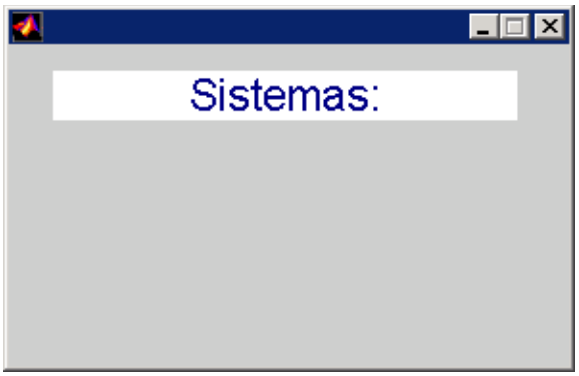

Figura 2: Exemplo do *BackgroundColor*.

O PROGRAMA DESENVOLVIDO E RESULTADOS

Utilizaremos os dados de produção de soja no Brasil durante os anos de 1990 a 2010 para aplicar ao programa desenvolvido (análise da tendência central e dispersão de dados). A tabela 1 e 2 apresentam os dados de produção de soja nos anos de 1990/91 a 1999/00 e 2000/01 a 2009/10, Conab, 2010. Foram computados os valores em mil toneladas durante os últimos 20 anos, de 1990 a 2010.

Tabeal 1: SOJA – BRASIL - Safras 1990/91 a 1999/00 em mil toneladas

REGIÃO/UF	1990/91	1991/92	1992/93	1993/94	1994/95	1995/96	1996/97	1997/98	1998/99	1999/00
NORTE	11,5	19,4	36,4	59,0	45,5	14,2	28,6	99,8	123,2	177,0
NORDESTE	564,3	520,3	682,1	1.018,4	1.267,8	921,9	1.300,1	1.561,1	1.609,8	2.064,0
CENTRO-OESTE	6.667,0	7.313,2	8.484,2	9.907,0	10.084,7	8.846,4	10.438,1	12.889,9	13.356,1	15.467,6
SUDESTE	1.930,4	1.910,7	2.314,3	2.499,4	2.365,9	2.274,5	2.498,4	2.495,5	2.757,0	2.569,7
SUL	6.221,3	9.655,0	11.525,1	11.575,4	12.170,2	11.132,7	11.894,8	14.323,6	12.918,9	12.611,7
NORTE/NORDESTE	575,8	539,7	718,5	1.077,4	1.313,3	936,1	1.328,7	1.660,9	1.733,0	2.241,0
CENTRO-SUL	14.818,7	18.878,9	22.323,6	23.981,8	24.620,8	22.253,6	24.831,3	29.709,0	29.032,0	30.649,0
BRASIL	15.394,5	19.418,6	23.042,1	25.059,2	25.934,1	23.189,7	26.160,0	31.369,9	30.765,0	32.890,0

Fonte: Conab

Tabela 2: SOJA – BRASIL - Safras 2000/01 a 2009/10 em mil toneladas

REGIÃO/UF	2000/01	2001/02	2002/03	2003/04	2004/05	2005/06	2006/07	2007/08	2008/09	2009/10
NORTE	216,6	367,4	557,5	913,7	1.419,9	1.255,2	1.079,9	1.472,4	1.414,0	1.691,7
NORDESTE	2.075,9	2.124,6	2.519,3	3.538,9	3.953,1	3.560,9	3.867,2	4.829,8	4.161,9	5.309,5
CENTRO-OESTE	17.001,9	20.533,4	23.532,5	24.613,1	28.973,5	27.824,7	26.494,8	29.114,0	29.134,9	31.586,7
SUDESTE	2.873,9	3.519,8	4.067,6	4.474,4	4.752,0	4.137,1	4.005,4	3.983,4	4.057,6	4.457,6
SUL	16.263,5	15.684,8	21.340,6	16.252,6	13.206,2	18.249,2	22.944,5	20.618,1	18.397,1	25.642,7
NORTE/NORDESTE	2.292,5	2.492,0	3.076,8	4.452,6	5.373,0	4.816,1	4.947,1	6.302,2	5.575,9	7.001,2
CENTRO-SUL	36.139,3	39.738,0	48.940,7	45.340,1	46.931,6	50.211,0	53.444,7	53.715,5	51.589,6	61.687,0
BRASIL	38.431,8	42.230,0	52.017,5	49.792,7	52.304,6	55.027,1	58.391,8	60.017,7	57.165,5	68.688,2

Fonte: Conab

Para calcular a dispersão dos dados foram utilizadas as fórmulas das equações (2), (3) e (4). O objetivo é de analisar a dispersão de dados de produção de soja no Brasil, ou seja, verificar se os dados apresentam uma variabilidade muito grande em relação a média. A figura 3 apresenta a tela inicial do programa. Ao clicar no botão "Análise de tendência central" outra tela é aberta (figura 4), esta tela representa a figura de entrada de dados do exercício para análise de tendência central.

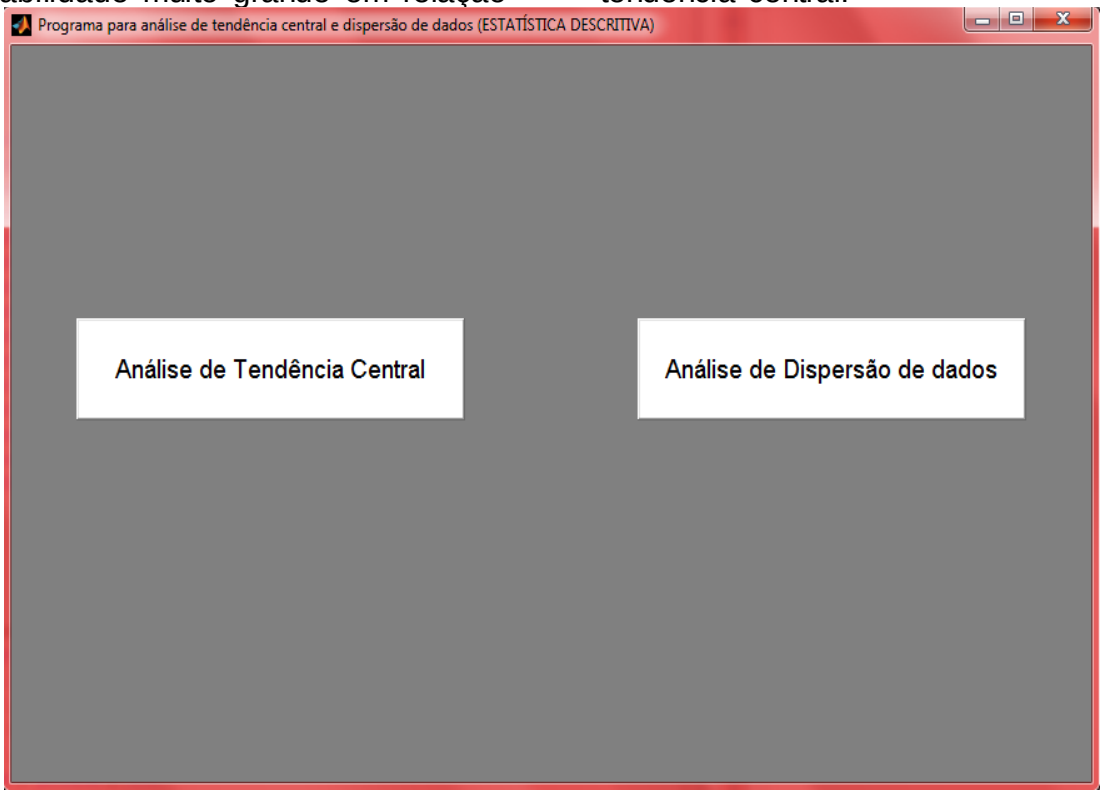

Figura 3: Tela inicial do programa.

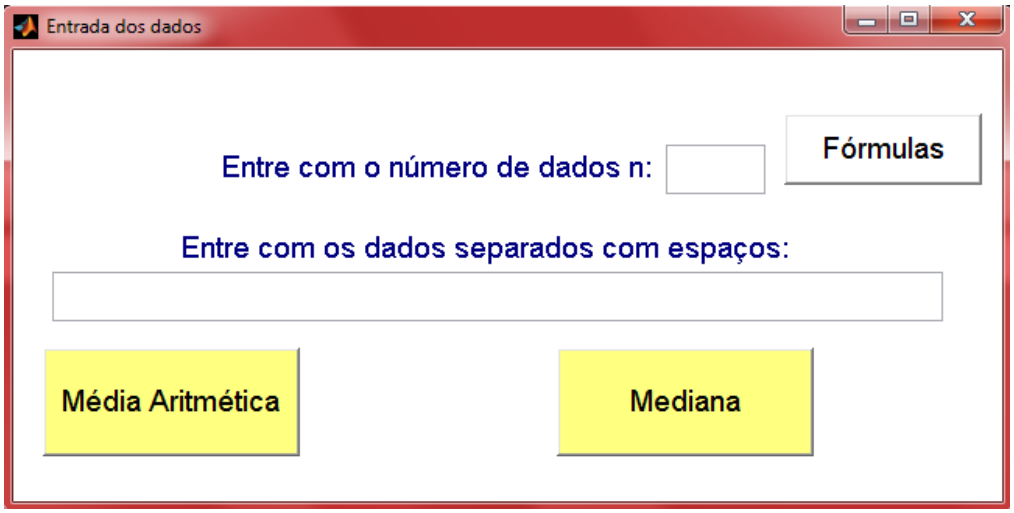

Figura 4: Tela de entrada de dados.

Apresenta-se na figura 5, a mesma figura 4 com os valores de entrada de dados para a aplicação apresentada, foram 10 pontos (n=10), correspondente aos anos de 1990/91 a 1999/00. Ao clicar no botão "Média Aritmética" o programa calcula via equação (1) a média dos dados inseridos na interface. Para os dados de produção de soja entre 1990/91 e 1999/00, o valor da média foi de 25322,31 mil toneladas. Caso o usuário queira saber o valor da mediana, basta clicar no botão "Mediana". Para os dados de soja o valor é de 25496,7.

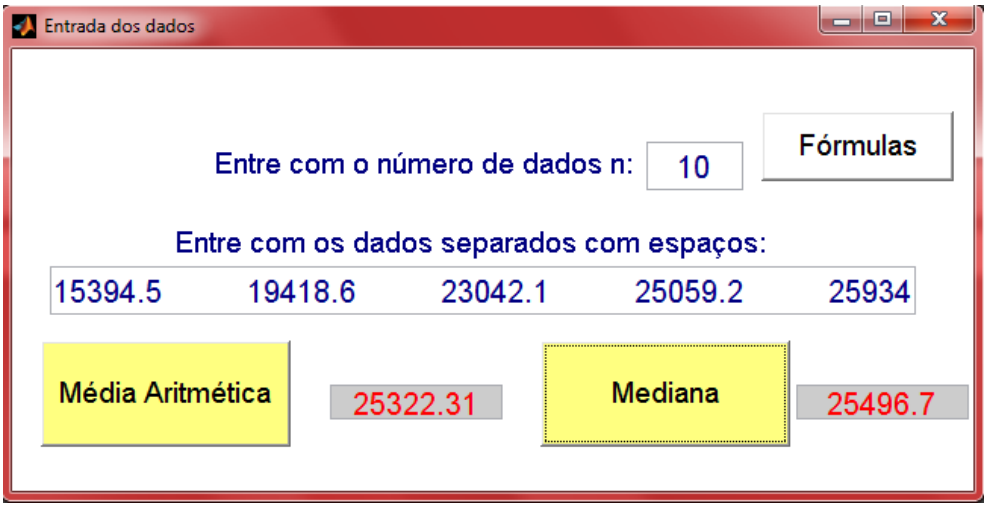

Figura 5: Tela de entrada e saída de dados.

Para análise de dispersão de dados, utilizando-se a figura 3 novamente, basta o usuário clicar no botão "Análise de Dispersão de dados" que aparecerá dois outros botões "Populacional" e "Amostral" para que o usuário escolha se os dados a serem analisados serão de uma população ou amostra, figura 6.

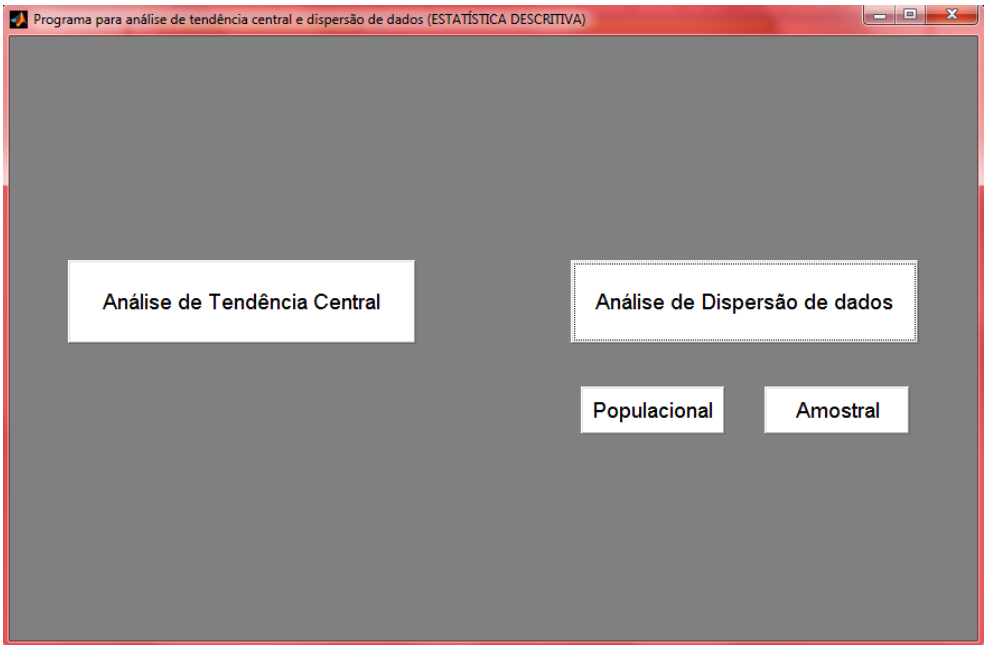

Figura 6: Tela inicial do programa para escolha populacional ou amostral dos dados.

Como os dados apresentados neste artigo são amostrais, clica-se no botão “Amostral” e uma nova tela se abre para inserção dos dados, figura 7. Os dados inseridos são de produção de soja no Brasil entre os anos de 1990/91 a 1999/00 e 2000/01 a 2009/10 e para calcular a variância desses dados basta clicar no botão “Variância”, da mesma forma para o Desvio Padrão e CV. Os resultados podem ser vistos na figura 8.

Os valores encontrados foram obtidos utilizando as fórmulas (1), (2), (3) e (4). Observa-se dos resultados obtidos que o CV foi de 21,54% para os anos de 1990/91 a 1999/00 e 16,32% para os anos de 2000/01 a 2009/10, que segundo Pimentel Gomes (1985) se $20 \leq CV < 30$ há uma alta dispersão de dados e se $10 \leq CV < 20$ há uma média ou moderada dispersão de dados.

Figura 7: Tela de entrada e saída de dados, análise de dispersão.

Outra parte importante do programa é não deixar que o usuário cometa erros despercebidos, por exemplo, caso o usuário digite um numero n diferente da quantidade de dados apresentado, o valor da média será diferente do valor real e outras análises também. Então, para não ocorrer este erro o programa avisará que os números são diferentes, ver figura 8.

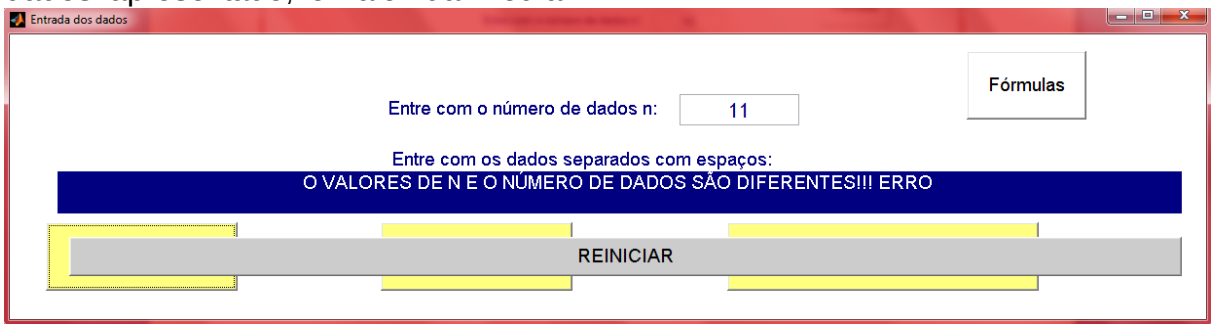

Figura 8: Mensagem de erro caso n seja diferente do número de dados inseridos.

A figura 9 apresenta as fórmulas utilizadas via o programa iterativo proposto clicando no botão “Fórmulas”.

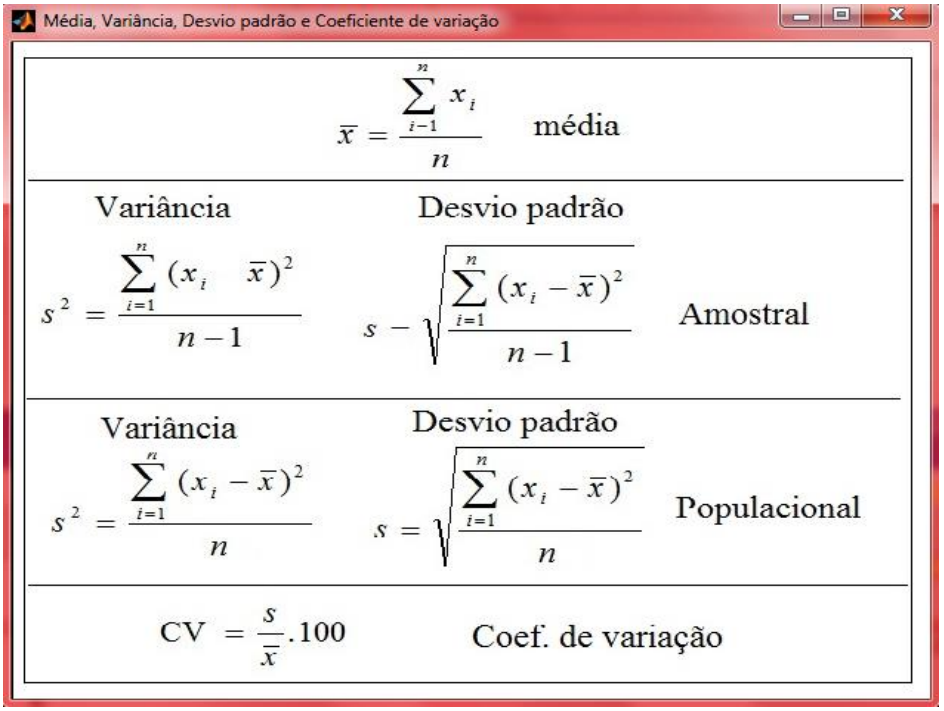

Figura 9: Fórmulas utilizadas para desenvolver o programa iterativo.

CONCLUSÃO

Neste trabalho foi apresentado um método iterativo via interface gráfica para cálculos da estatística descritiva, a análise de tendência central e a dispersão de dados. Foi utilizado o software Matlab para criação da interface gráfica tornando o programa mais iterativo para o usuário. Os dados utilizados para o teste foi de produção de soja nos anos entre 1990/91 até 2009/10.

REFERÊNCIAS

CONAB – Companhia Nacional de Abastecimento. Disponível em: < http://www.conab.gov.br>. Acesso em: 15/09/2010.

PAGANO, M.; GAUVREAU, K. Princípios de Bioestatística. 2. ed. São Paulo: Pioneira Thompson Learning, 2004.

PIMENTEL-GOMES, Curso de Estatística Experimental. Piracicaba-SP. ESALQ/USP, 1985.

STEVENSON, W. J. Estatística Aplicada à Administração, Harbra, 2001.

HUANG, G.M.; ZHANG, H. A. New Education Matlab Software for Teaching Power Analysis that Involves the Slack Bus Concept and Allocation Issues. Power Engineering Society Winter Meeting. IEEE, v. 2, 23-27, p. 1150- 1158, Jan. 2000.

MATHWORKS. Disponível em: < http://www.mathworks.com>. Acesso em: 15/03/2010.