Statistical Analysis Of Text Files

Course Instructor : Samy S. Soliman

Nada Ismail, 202-001-387

December 17, 2022

1 Introduction

This report will discuss project's statiscal analysis, results and the Matlab code.

2 Statiscal analysis

let x be an array containing the encoded values (1 to 61) and f(x) be the probability of x.

The equation of mean is:

$$m = \sum_{x=1}^{61} x f(x) \tag{1}$$

The equation of variance is:

$$\sigma^2 = \sum_{x=1}^{61} (x-m)^2 f(x) \tag{2}$$

The equation of skewness is:

$$\frac{1}{\sigma^3} \sum_{x=1}^{61} (x-m)^3 f(x) \tag{3}$$

The equation of kurtosis is:

$$\frac{1}{\sigma^4} \sum_{i=1}^{61} (x-m)^4 f(x) \tag{4}$$

3 Code

3.1 Data analysis functions

3.1.1 filter

this function filters out the unwanted charcters like: \$,&,%. isstrprop() method is used to return an array of ones and zeroes that shows which indexes in the original data are alphanumeric.

```
        function filtered_data = filter(app,data)
            filtered_data=[];
            j=1;% counter for the return array
            isAlphaNum=isstrprop(data,"alphanum");%returns boolean array of alphanum characters in the data array
            for i=1:length(data)
                if isAlphaNum(i)
                        filtered_data(j)= data(i);
                        j=j+1;
                end
            end
        end
```

3.1.2 get_encoded_data

This function returns a 61x1 array where the index-1 is the encoded value and the each element is number of times the corresponding character appeared in the file.

```
        function results = get_encoded_data(app,data)%return array where the index is the encoded value and the value of each element is the # of occurences
            char_vals=['0','1','2','3','4','5','6','7','8','9','a','A','b','B','c','C','d','D','e','E','f','F','g','G','h','H','i','I','j','J','k','K','l','L','m','M','n','N','o','O','p','P','q','Q','r','R','s','S','t','T','u','U','v','V','w','W','x','X','y','Y','z','Z'];
            results=zeros(1,62);
            for i=1:length(data)
                for j=1:length(char_vals)
                    if data(i)==char_vals(j)
                           results(j)=results(j)+1;
                    end
                end
            end
        end
```

3.1.3 get_sorted_array_of_chars_based_on_occurenceArray

This function Returns a sorted array of character based where the most common character is the first.

```
        function  [sorted_char_array, sorted_array] = get_sorted_array_of_chars_based_on_occurenceArray(app,occurence_array)
            sorted_char_array=[’0’,’1’,’2’,’3’,’4’,’5’,’6’,’7’,’8’,’9’,’a’,’A’,’b’,’B’,’c’,’C’,’d’,’D’,’e’,’E’,’f’,’F’,’g’,’G’,’h’,’H’,’i’,’I’,’j’,’J’,’k’,’K’,’l’,’L’,’m’,’M’,’n’,’N’,’o’,’O’,’p’,’P’,’q’,’Q’,’r’,’R’,’s’,’S’,’t’,’T’,’u’,’U’,’v’,’V’,’w’,’W’,’x’,’X’,’y’,’Y’,’z’,’Z’];
            sorted_array=occurence_array;
            [sorted_array,sort_Index] = sort(sorted_array,’descend’);
            sorted_char_array = sorted_char_array(sort_Index);
        end
```

3.1.4 get_statistics

This function return 4x1 array where the elements are the mean, variance, skewness, and kurtosis.

```
        function [m,v,s,k] = get_statistics(app,data)
            f_x=data./sum(data);
            x=1:62;
            x=x-1;
            m = sum(x.*f_x);
            v = sum(((x-m).^2).*f_x);
            %data_p3=sum((data-m).^3);
            s = sum(((x-m).^3).*f_x)/(v^(3/2));
            %data_p4=sum((data-m).^4);
            k = sum(((x-m).^4).*f_x)/(v^2);
        end
```

3.1.5 get_f_x

This function returns the PMF.

```
        function f_x = get_f_x(app,data)
            f_x=data./sum(data);
        end
```

3.1.6 get_F_x

This function returns the CDF.

```
        function F_x = get_F_x(app,f_x)
            F_x=[];
            for i=1:length(f_x)
                F_x(i)=sum(f_x(1:i));
            end
        end
```

3.2 GUI callback functions

3.2.1 ImportfileButtonPushed

This function promts the user to chose the text file, passes its content to filter() then get occurence array and does the stastistical analysis and plot.

```
               [filename,path]=uigetfile('*.txt');
       app.PathEditField.Value=path;
       app.data=fileread(filename);
       app.PathEditField.Value=path;
       %app.draftTextArea.Value=data;
       app.filter_data=filter(app,app.data);
       app.filter_data_characters=char(app.filter_data);
       app.occurences_of_characters=get_encoded_data(app,app.filter_data_characters);

       [m,v,s,k]=get_statistics(app,app.occurences_of_characters);
       app.ThemeanisTextArea.Value=string(m);
       app.ThevarianceisTextArea.Value=string(v);
       app.TheskewdnessisTextArea.Value=string(s);
       app.ThekurtosisisTextArea.Value=string(k);

       %shift ticks to start at origin and then plot

       app.UIAxes_3.XLim=[1 63];
       app.UIAxes_3.XTickLabel = {' '; '0'; '1'; '2'; '3'; '4'; '5'; '6'; '7'; '8'; '9'; '10'; '11'; '12'; '13'; '14'; '15'; '16'; '17'; '18'; '19'; '20'; '21';'22'; '23'; '24'; '25'; '26'; '27'; '28'; '29'; '30'; '31'; '32'; '33'; '34'; '35'; '36'; '37'; '38'; '39'; '40'; '41'; '42'; '43'; '44'; '45'; '46'; '47'; '48'; '49'; '50'; '51'; '52'; '53'; '54'; '55'; '56'; '57'; '58'; '59'; '60'; '61';'62'};
       app.UIAxes_3.YLim=[0 1];
       app.f_x=get_f_x(app,app.occurences_of_characters);
       stairs(app.UIAxes_3,app.f_x);

       app.UIAxes_2.XLim=[1 63];
       app.UIAxes_2.XTickLabel = {' '; '0'; '1'; '2'; '3'; '4'; '5'; '6'; '7'; '8'; '9'; '10'; '11'; '12'; '13'; '14'; '15'; '16'; '17'; '18'; '19'; '20'; '21';'22'; '23'; '24'; '25'; '26'; '27'; '28'; '29'; '30'; '31'; '32'; '33'; '34'; '35'; '36'; '37'; '38'; '39'; '40'; '41'; '42'; '43'; '44'; '45'; '46'; '47'; '48'; '49'; '50'; '51'; '52'; '53'; '54'; '55'; '56'; '57'; '58'; '59'; '60'; '61';'62'};
       app.UIAxes_2.YLim=[0 1];
       app.F_x=get_F_x(app,app.f_x);
       stairs(app.UIAxes_2,app.F_x);
```

3.2.2 NumberEditFieldValueChanged

This function is triggered when the user enters the most repeated character that he wants. The user cannot choose more than number of diffrent characters in the file.

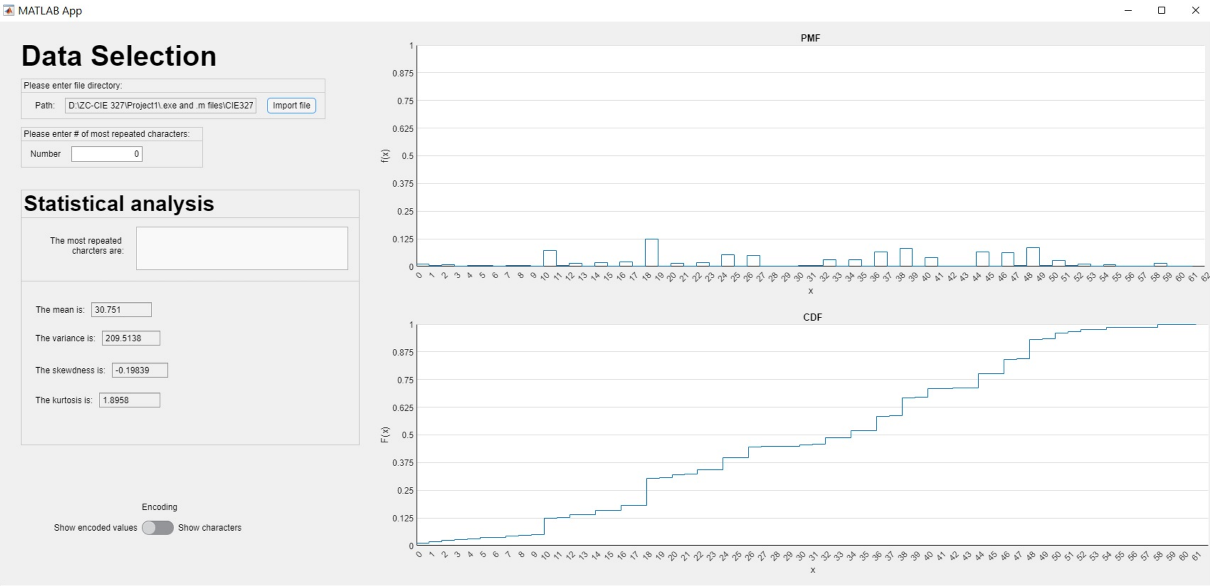
```
    function NumberEditFieldValueChanged(app, event)
        limit= length(nonzeros(app.occurences_of_characters)); %do not allow user to ask for more characters than there are in the document if it only has a's and b's then he cannot askk for 3 most repeated
        if (app.NumberEditField.Value>limit)
            app.NumberEditField.Value=limit;
            msgbox(sprintf('There are %d diffrent alphanumeric charcters in the document,\n you cannot chose more than that',limit));
        end
        [uncutmsg,x]=get_sorted_array_of_chars_based_on_occurenceArray(app,app.occurences_of_characters);
        msg=[];
        for i=1:app.NumberEditField.Value
            msg(i)=uncutmsg(i);
        end
        app.ThemostrepeatedcharctersareTextArea.Value=char(msg);

    end
```

3.2.3 EncodingSwitchValueChanged

this function gives the user the option to see the numbers or the characters corresponding to those numbers.

```
    if app.EncodingSwitch.Value(1:length('Show characters'))=='Show characters'
            app.UIAxes_3.XTickLabel = {' '; '0'; '1'; '2'; '3'; '4'; '5'; '6'; '7'; '8'; '9'; 'a';'A';'b';'B';'c';'C';'d';'D';'e';'E';'f';'F';'g';'G';'h';'H';'i';'I';'j';'J';'k';'K';'l';'L';'m';'M';'n';'N';'o';'O';'p';'P';'q';'Q';'r';'R';'s';'S';'t';'T';'u';'U';'v';'V';'w';'W';'x';'X';'y';'Y';'z';'Z'};
            app.UIAxes_2.XTickLabel = {' '; '0'; '1'; '2'; '3'; '4'; '5'; '6'; '7'; '8'; '9'; 'a';'A';'b';'B';'c';'C';'d';'D';'e';'E';'f';'F';'g';'G';'h';'H';'i';'I';'j';'J';'k';'K';'l';'L';'m';'M';'n';'N';'o';'O';'p';'P';'q';'Q';'r';'R';'s';'S';'t';'T';'u';'U';'v';'V';'w';'W';'x';'X';'y';'Y';'z';'Z'};
        else
            app.UIAxes_3.XTickLabel = {' '; '0'; '1'; '2'; '3'; '4'; '5'; '6'; '7'; '8'; '9'; '10'; '11'; '12'; '13'; '14'; '15'; '16'; '17'; '18'; '19'; '20'; '21';'22'; '23'; '24'; '25'; '26'; '27'; '28'; '29'; '30'; '31'; '32'; '33'; '34'; '35'; '36'; '37'; '38'; '39'; '40'; '41'; '42'; '43'; '44'; '45'; '46'; '47'; '48'; '49'; '50'; '51'; '52'; '53'; '54'; '55'; '56'; '57'; '58'; '59'; '60'; '61';'62'};
            app.UIAxes_2.XTickLabel = {' '; '0'; '1'; '2'; '3'; '4'; '5'; '6'; '7'; '8'; '9'; '10'; '11'; '12'; '13'; '14'; '15'; '16'; '17'; '18'; '19'; '20'; '21';'22'; '23'; '24'; '25'; '26'; '27'; '28'; '29'; '30'; '31'; '32'; '33'; '34'; '35'; '36'; '37'; '38'; '39'; '40'; '41'; '42'; '43'; '44'; '45'; '46'; '47'; '48'; '49'; '50'; '51'; '52'; '53'; '54'; '55'; '56'; '57'; '58'; '59'; '60'; '61';'62'};
        end
```

4 Results



MATLAB App

Data Selection

Please enter file directory:

Path: D:\ZC-CIE 327\Project1\.exe and .m files\CIE327

Please enter # of most repeated characters:

Number: 0

Statistical analysis

The most repeated charcters are:

The mean is: 30.751

The variance is: 209.5138

The skewdness is: -0.19839

The kurtosis is: 1.8958

Figure 1: A screenshot of the results using the provided sample text