Fall 2021, Zewail City of Science and Technology, 6th
October City, 12588 Giza, Egypt

STATISTICAL ANALYSIS OF TEXT FILES

Mariam Wagdy 201801585

STATISTICAL ANALYSIS OF TEXT FILES

DEFINITIONS

MEAN

Mean is the first central moment and is defined by

$$Mean = \sum_{i=1}^{26} xf(x)$$

Since the situation deals with discrete random variable "x", where $x = 1,2,3, \ldots .26$

And $f(x)$ is the PMF of x

VARIANCE

Variance is the second central moment and is defined by

$$Variance = \varepsilon\{(x-\mu)^2\} = \sum_{i=1}^{26}(x-\mu)^2 f(x) = \sum_{i=1}^{26} x^2 f(x) - \mu_x^2$$

SKEWNESS AND KURTOSIS

Skewness is a measure of symmetry and is the third central moment. Kurtosis is a measure of whether the data is heavy-tailed or light-tailed relative to a normal distribution and it is the fourth central moment. (*1.3.5.11. Measures of Skewness and Kurtosis*, 2000).

For random variable Y, the Fisher-Pearson formula is:

$$skewness = \frac{\sum_{i=1}^{N} \frac{\left(Y_i - \overline{Y}\right)^3}{N}}{\sigma^3}$$

Where $\overline{Y}$ is the mean, σ is the standard deviation, and N is the number of data points.

Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right.

$$kurtosis = \frac{\sum_{i=1}^{N} \frac{\left(Y_i - \overline{Y}\right)^4}{N}}{\sigma^4}$$

CODE

Code is divided into a **main** part and a function **LetterCount.**

```
function LetterCount = count(filename,A)
fid = fopen(filename,'rt');
if fid < 0
    LetterCount = -1;
    return
end
if fid >0 && ischar(A)
    r=0;
    readline = fgets(fid);
    while ischar(readline)
        r = r + count(readline,A)+count(readline,upper(A));
        readline = fgets(fid);
    end
    LetterCount = r;
else
    LetterCount = -1;
end
fclose(fid);
```

```
clc;
clear;
s=input('insert file name','s');
AZ='a':'z';
n=1:26;
for k=1:26
  num(k,1)=LetterCount(s,AZ(k));
end
total=sum(num);
%%disp(num);
%%disp("......");
%%disp(total);
freq=100*num./total;
%%disp("......");
%%disp(freq);
aspacez={'a   ','b   ','c   ','d   ','e   ','f   ','g   ','h   ','i   ','j
aspacez=transpose(aspacez);
X = cellstr(aspacez);
X = categorical(X);
%%X = reordercats(X,aspacez);
bar(X,freq);
```

```
22 -    title('Probability in Percentage of Each Letter');
23
24      %%%%%%%%%%%%%%max 5%%%%%%%%%%%%%%%
25 -    [maxfreq,index]=maxk(freq,5);
26 -    disp(transpose(maxfreq));
27 -    for i=1:5
28 -        maxletter(i)=aspacez(index(i));
29 -    end
30 -    disp(maxletter);
31 -    Y = cellstr(maxletter);
32 -    Y = categorical(Y);
33 -    Y = reordercats(Y);
34 -    b = bar(Y,maxfreq);
35 -    xtips1 = b(1).XEndPoints;
36 -    ytips1 = b(1).YEndPoints;
37 -    labels1 = string(b(1).YData);
38 -    text(xtips1,ytips1,labels1,'HorizontalAlignment','center',...
39          'VerticalAlignment','bottom')
40 -    title('Probability in Percentage of the Most Repeated 5 Letters')
41
42      %%%%%%%%%%%%%%%%%%%PDF CDF%%%%%%%%%%%%%%%%%%
43 -    bar(n,freq);
44 -    title('PDF of letters');
45      %%f=sum(freq);
46 -      cumulative(1)=freq(1);
47 -    for i=2:length(freq)
48 -        cumulative(i)=cumulative(i-1)+freq(i);
49 -    end
50      %%disp(cumulative);
51 -    bar(n,cumulative);
52 -    title('CDF of letters');
53      freq=freq/100;
54 -      mean=0;
55 -      expx2=0;
56 -      expx3=0;
57 -    for i=1:26
58 -        mean=mean+freq(i)*n(i);
59 -        expx2=expx2+ (n(i)^(2)*freq(i));
60 -        expx3=expx3+ (n(i)^(3)* freq(i));
61 -    end
62 -    mean=mean/100;
63 -    disp("mean is");
64 -    disp(mean);
65 -    var=expx2-(mean*mean);
66 -    var=var/100;
```

```
67 -    disp("Variance is");
68 -    disp(var);
69 -    skew=(expx3-3*mean*var-mean^(3))/var^(3/2);
70 -    disp('Skewness is');
71 -    disp(skew);
72 -    kurt=0;
73 -    for i=1:26
74 -        kurt=kurt+((n(i)-mean)/var^(1/2))^(4)*freq(i);
75 -    end
76 -    disp('Kurtosis is');
77 -    disp(kurt);
```

RESULTS

The letter count was compared to the word document result. The code was applied to two documents **a** and **b.**

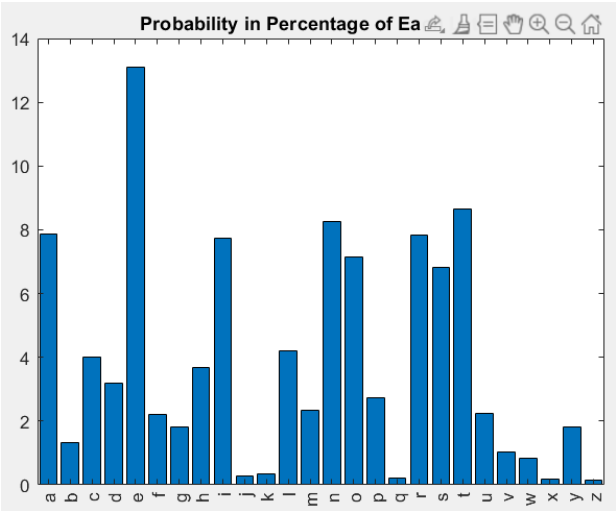

FIGURE 2 THE PROBABILITY OF EACH LETTER IN FILE A

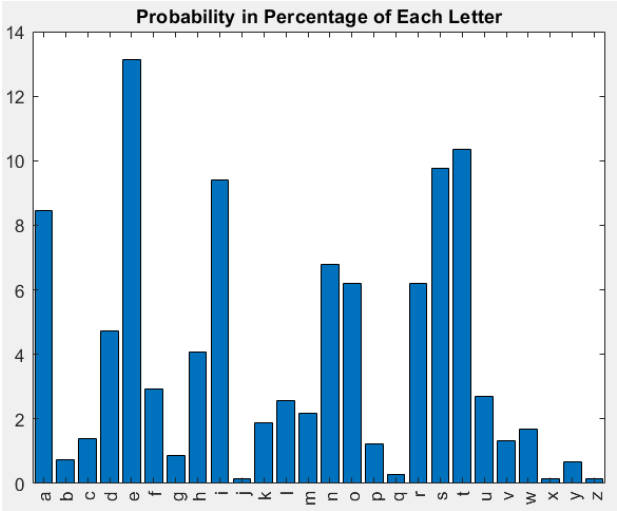

FIGURE 2 THE PROBABILITY OF EACH LETTER IN FILE B

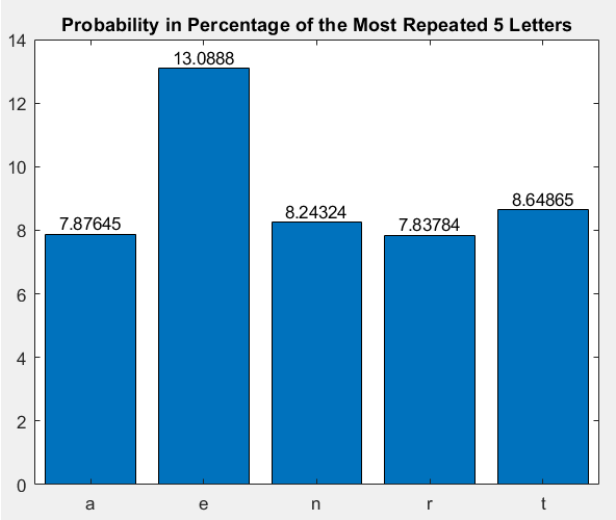

FIGURE 4 MOST REPEATED LETTER IN FILE A

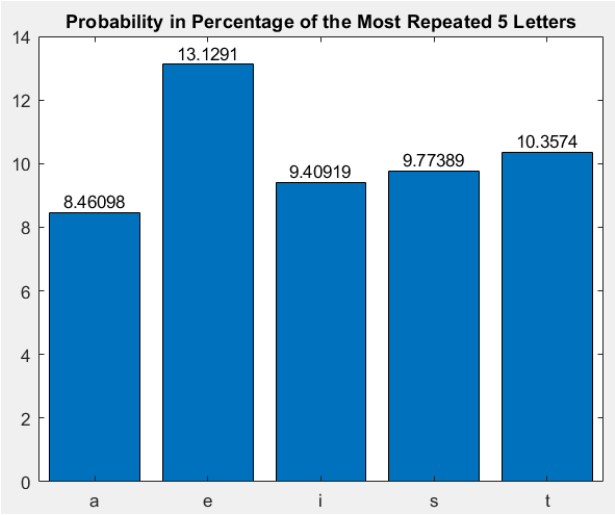

FIGURE 4 MOST REPEATED LETTER IN FILE B

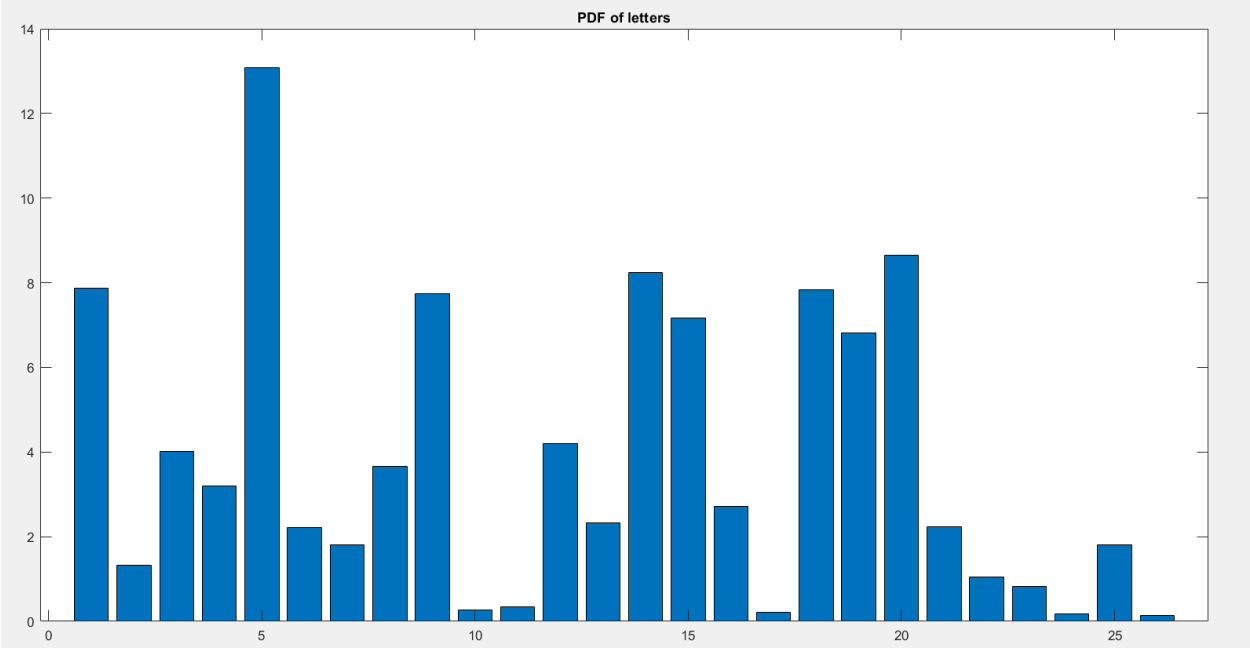

FIGURE 5 PDF OF FILE A

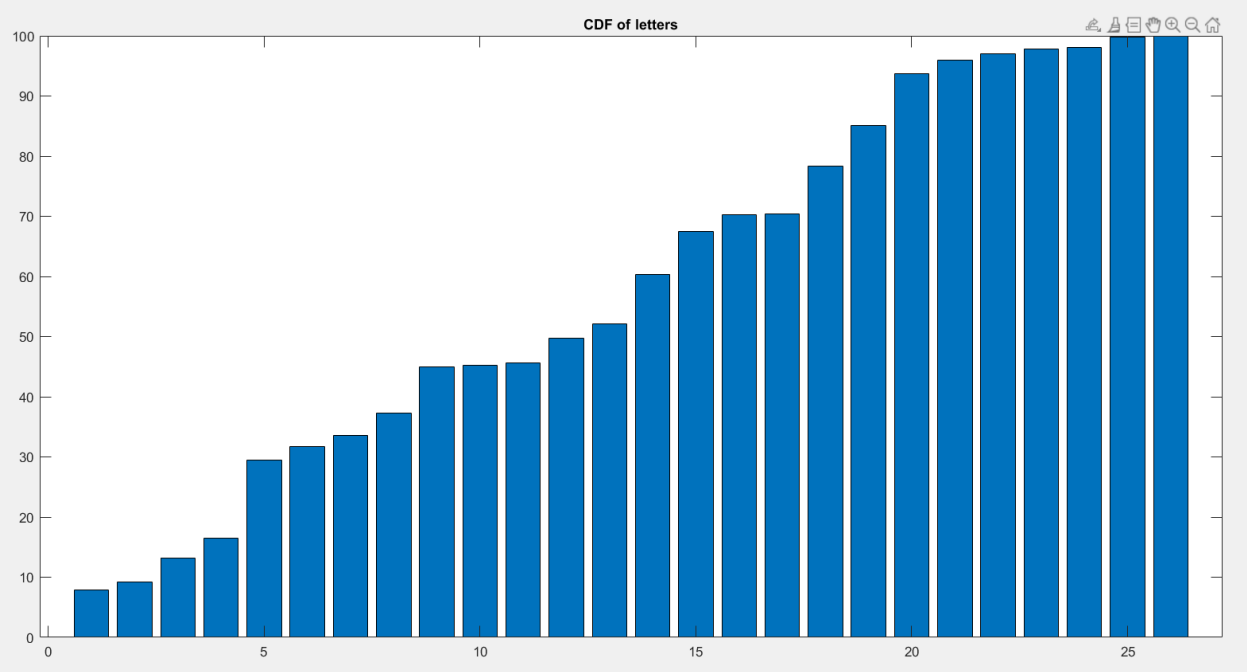

FIGURE 7 CDF OF FILE A

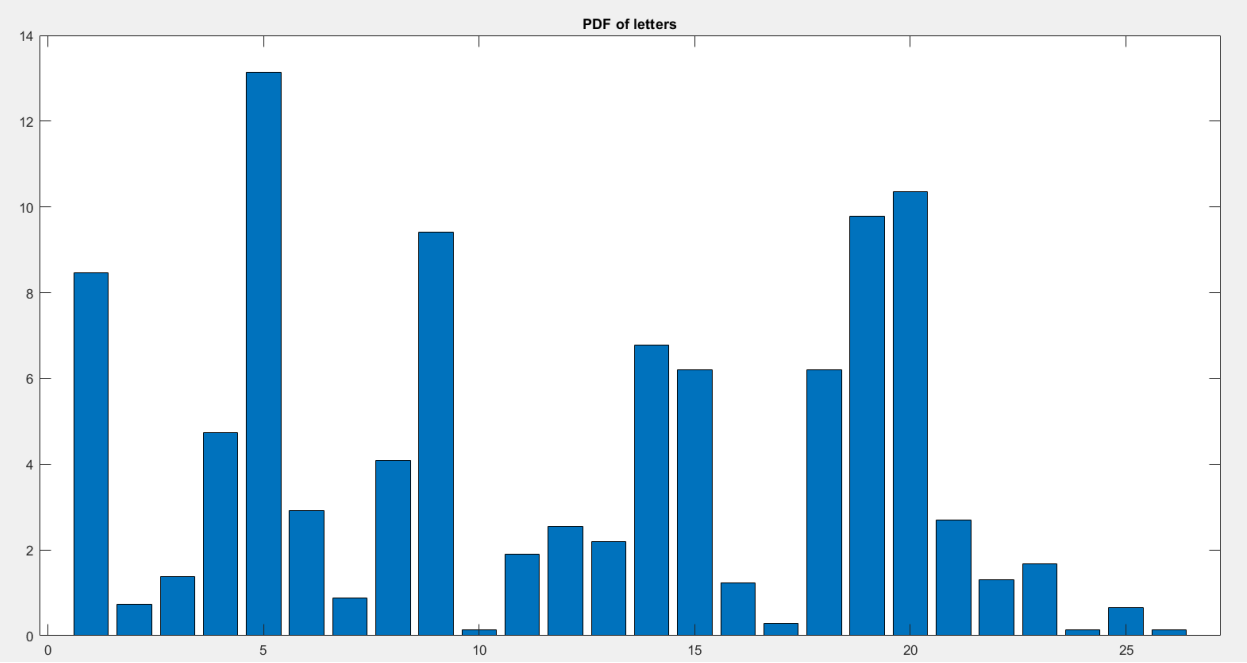

FIGURE 6 PDF OF FILE B

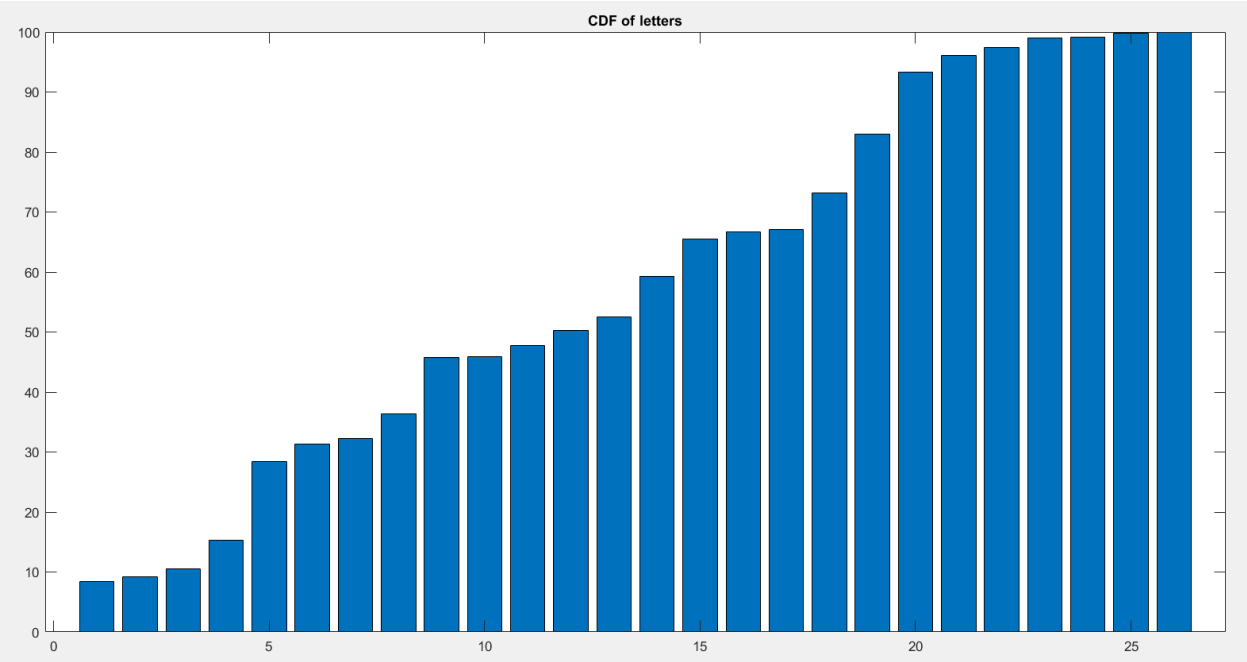

FIGURE 8 CDF OF FILE B

File	a	b
No. of letters	**5180**	**1371**
Mean	**11.69**	**11.8614**
Variance	**45.66**	**46.2390**
Skewness	**-0.0132**	**-0.0422**
Kurtosis	**1.7429**	**1.6552**

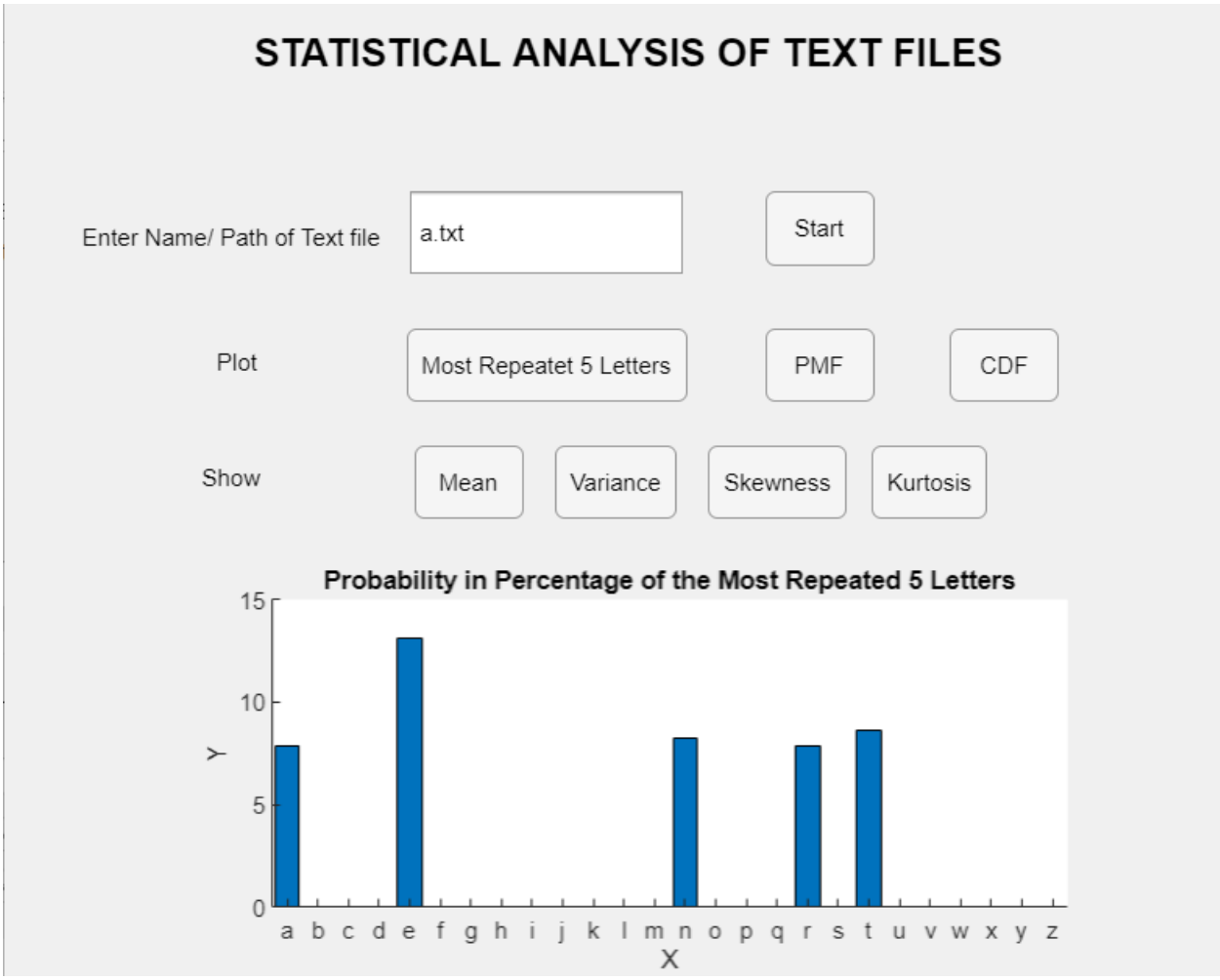

FIGURE 10 SCREENSHOT FROM THE GUI, WHEN THE “MOST REPEATED 5 LETTERS” BUTTON WAS CLICKED.

The GUI also prints the next figure for more clarity.

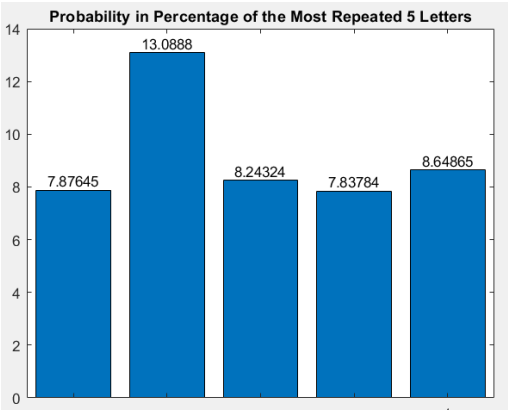

FIGURE 9 SECOND OUTPUT WHEN THE "MOST REPEATED 5 LETTERS" BUTTON IS PRESSED

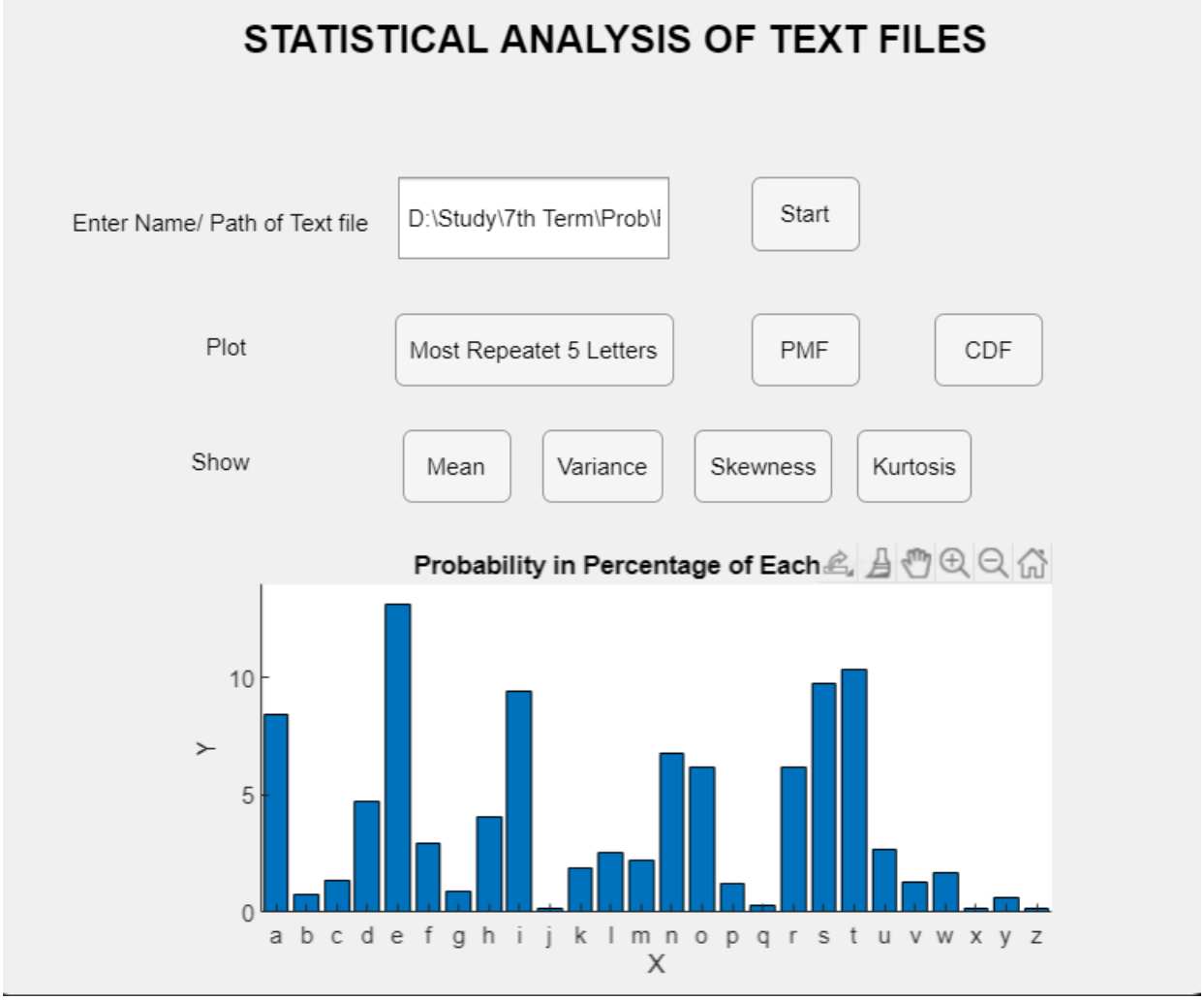

FIGURE 11 SCREENSHOT FROM THE GUI, WHEN THE “START” BUTTON WAS CLICKED.

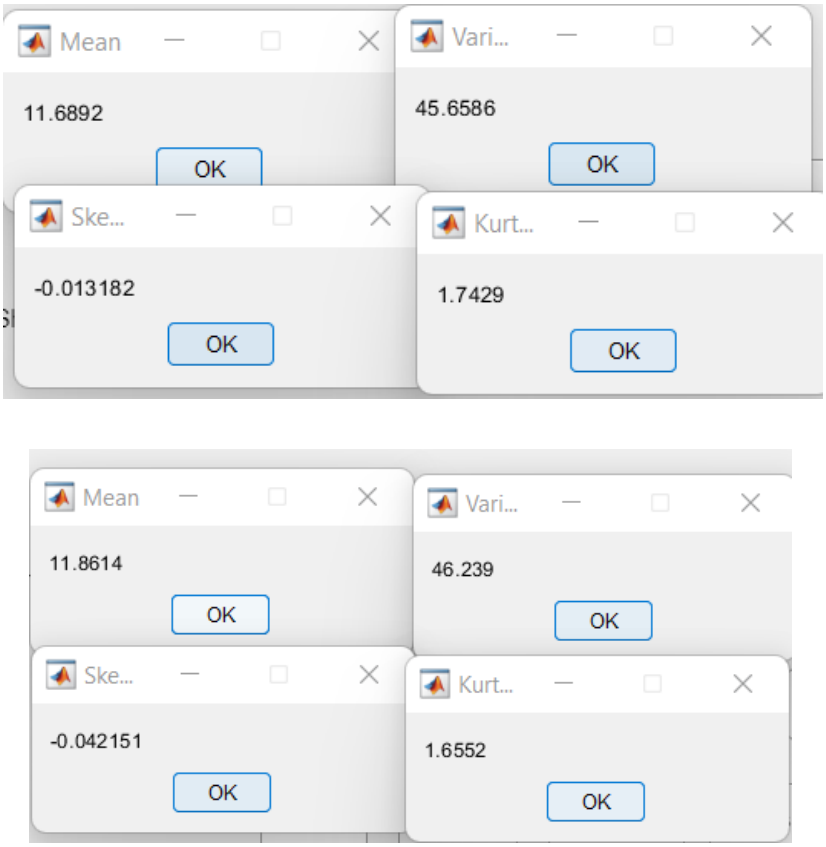

FIGURE 12 OUTPUTS OF GUI. TOP: FILE A. BOTTOM: FILE B

REFERENCES:

1.3.5.11. Measures of Skewness and Kurtosis. (2000). Engineering Statistics Handbook. https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm