STATISTICAL ANALYSIS OF LETTER FREQUENCY IN SWAHILI LANGUAGE

By

Esther A. Kahamba

Dissertation Submitted in Partial Fulfilment of the Requirements for the Degree of

Master of Science in Mathematics of the University of Dodoma

The University of Dodoma

October, 2012

CERTIFICATION

The undersigned certifies that he has read and hereby recommends for the acceptance by the University of Dodoma a dissertation entitled **"Statistical analysis of letter frequency in Swahili language"** in partial fulfilments for the degree of Master of Science in Mathematics of University of Dodoma.

.....

Dr.Soud Mohamed

(SUPERVISOR)

.

Date.....

DECLARATION

AND

COPYRIGHT

I Esther Kahamba, declare that this dissertation is my own original work and that it has not been presented and will not be presented to any other University for a similar or any other degree award.

Signature.....

No part of this dissertation may be reproduced, stored in any retrieval system, or transmitted in any form or by any means without prior written permission of the author or the University of Dodoma.

ACKNOWLEDGEMENT

I would like to thank my supervisor Dr. Soud Mohammed for offering suggestions for corrections and additions to this dissertation. He offered me some books and his enormous amounts of time and energy toward the accomplishment of the dissertation. Indeed his contribution is of paramount important.

I thanksful Mr. Ochanga Frank for typing and setting mathematical symbols for this project and May God bless him with his family.

I also extend my heartfelt appreciation to all the people who offered valuable assistance when preparing this dissertation.

Last and not least, I would like to thank my family for supporting me morally and financially in all time of the study.

(Esther A.Kahamba)

DEDICATION

I wish to dedicate my work to my beloved husband Mr. Valentine, my children, my friend Mr. Alberto Gwaha, Mr. Ochanga Frank Mr. Lucas Ngwema and all my colleague students at University of Dodoma.

ABSTRACT

The sudy of a language is very important for the cryptanalysis of substitution and/ or permutation ciphers. In that type of ciphers one letter is substituted by another one, or its order is changed with the order of another letter also from the text. In either case the “personality”of the alphabets remains intact, hidden inside a different vest, but intact anyway.

If it is true that the morden block ciphers hide those characteristics, given the fact that they operate at bit level, we think that it is still important to have at hand such a tool for our own laguage, we can think more as an education tool, in order to present and/ or study the classical ciphers. We also have one more tool in our cryptanalyst toolbox.

In this dissertation we present the statistical analysis of letters frequencies in Swahili language; we have discussed the mostly frequent letter in the Arabic language as well as in the English language. We also discussed the Swahili alphabets and analyse some Swahili articles and present the frequencies of the letters as used in the Swahili language. Lastly, we present the use of $t-test$ and provide the level of confident interval for each letter.

TABLE OF CONTENT

CERTIFICATION i
DECLARATION AND COPYRIGHT ii
ACKNOWLEDGEMENT iii
DEDICATION iv
ABSTRACT v

CHAPTER ONE 1
1.0 Introduction 1

CHAPTER TWO: FREQUENCY ANALYSIS OF LETTERS 4
2.1 Frequency Analysis of Arabic Language 6
2.3 Frequency Analysis of English Language 10

CHAPTER THREE: FREQUENCY ANALYSIS OF LETTERS IN SWAHILI TEXTS 19
3.1 Introduction 19
3.2 Structure of Swahili language 19
3.3 Frequency Analysis of Letters in Swahili articles 20
3.3.1 Data source for frequency analysis of letters in Swahili texts 20
3.3.2 Methodology and analysis 21
3.3.3 Stefan Trost Character frequency: Swahili 21
3.3.4 Letter Frequencies in Swahili articles 22

CHAPTER FOUR: STATISTICAL ANALYSIS OF LETTER FREQUENCY IN SWAHILI LANGUAGE 26
4.1 Introduction 26
4.2 Definitions and Examples 27
4.2.1 Mean 27
4.2.2 Standard deviation 27
4.2.3 Universe and Population 28
4.2.4 Statistics and Parameters 28

4.2.5 Confidence level and significance level 28
4.2.6 Sampling distribution 29
4.2.7 Sampling error 30
4.2.8 Frequency table 30
4.2.9 Probability 30
4.2.10 Random sample 30
4.3 Student's t-test 30
4.4 Confidence Intervals (C.I) 35
4.5 Results and Conclusion. 36
REFERENCES 38
APPENDIX 41
SAMPLE COLLECTIONS 41

LIST OF TABLES

Table 3.1 Swahili alphabets. 19

Table 3.2 The Relative frequency table sorted according to the letters in percentage. 22

Table 3.3 The frequency distribution table showing the frequencies and 24

relative frequencies of the Swahilialphabets sorted by letters(left) and by frequencies (right) 24

Table 4.1 The mean, standard deviation, 95% interval population mean μ and

$t - test$ values for each letter. 34

Table 4.2 The mean, standard deviation, standard error and 95% confidence interval for mean for each letter. 36

LIST OF FIGURE

Figure 2.3 A histogram of Arabic letter frequency distribution data sorted by frequency.....10

Figure 2.5 A histogram of English letter frequency distibtion.....18

Figure 8 A histogram of Swahli letter frequency distribution table showing the relative frequency in percentage 25

LIST OF CHART

Chart 3.1 Letter frequencies data as observed from the Swahili articles. 23

Chart 4.1 Letter frequencies, means and standard deviations as observed from the samples studied. 33

LIST OF APPENDENCES

APPENDIX SAMPLE COLLECTIONS.....38

CHAPTER ONE

1.0 Introduction

Statistics can be called the body of analytical and computational methods by which characteristics of a population are inferred through observations made in a representative sample from that population. It refers to a collection of methods used to process large amounts of data and report overall trends. Statistical analysis is particularly useful when dealing with noisy data. Statistical analysis provides ways to objectively report on how unusual an event is based on historical data. The relative frequencies of the alphabets, diagrams, trigrams, the first and last letters of a word, the average length of words and the frequencies of the "small" words are all characteristics of a given language [2, 15]. The behavior of the letters and words reflects the way people use its own language and characterize that language in a unique way.

Frequency lists can be listings of words or of letters. Letter frequencies typically are used in cryptography. One of the simplest codes is a substitution cipher, where each letter is replaced with another letter or symbol. For example, the message "attack at dawn" might be encoded as "zoozhlzoazqp." The benefit of substitution ciphers is that they don't require a code book, but the weakness is that they can be cracked by comparing the frequency of letters and letter combinations within the message to a frequency list of common usage [22].

Historically, the first known recorded explanation of frequency analysis (indeed, of any kind of cryptanalysis) was given in the 9th century by Al-Kindi, an Arabpolymath, in A Manuscript on Deciphering Cryptographic Messages [8]. It has

been suggested that close textual study of the Qur'an first brought to light that Arabic has a characteristic letter frequency [21]. Its use spread, and similar systems were widely used in European states by the time of the Renaissance. By 1474 Cicco Simonetta had written a manual on deciphering encryptions of Latin and Italian text [9]. Arabic Letter Frequency and a detailed study of letter and word frequency analysis of the entire book of Qur'an are provided by Intellaren Articles [10] that is the article that introduces an unprecedented study on the frequency distribution.

Several schemes were invented by cryptographers to defeat this weakness in simple substitution encryptions. These included:

- Use of *homophones* — several alternatives to the most common letters in otherwise monoalphabetic substitution ciphers (for examples, for English, both X and Y ciphertext might mean plaintext E).
- *Polyalphabeticsubstitution*, that is, the use of several alphabets — chosen in assorted, more or less devious, ways (Leone Alberti seems to have been the first to propose this); and
- *Polygraphic substitution*, schemes where pairs or triplets of plaintext letters are treated as units for substitution, rather than single letters (for example, the Playfaircipher invented by CharlesWheatstone in the mid-19th century).

A disadvantage of all these attempts to defeat frequency counting attacks is that it increases complication of enciphering and deciphering, leading to mistakes. Famously, a British Foreign Secretary is said to have rejected the Playfair cipher because, even if school boys could cope successfully as Wheatstone and Playfair had shown, 'our attachés could never learn it!'.

During World War II (WWII), both the British and the Americans recruited codebreakers by placing crossword puzzles in major newspapers and running contests for who could solve them faster. Several of the ciphers used by the Axis

powers [21] were breakable using frequency analysis (for example, some of the consular ciphers used by the Japanese). Mechanical methods of letter counting and statistical analysis (generally IBM card type machinery) were first used in WWII, possibly by the US Army's SIS.

Today, the hard work of letter counting and analysis has been replaced by computer software, which can carry out such analysis in seconds. With modern computing power, classical ciphers are unlikely to provide any real protection for confidential data [21].

As we know that scientists rarely observe entire populations, hence sampling and statistical inference are essential. Since the objective of statistical methods is to make the process of scientific research as efficient and productive as possible, the purpose of this dissertation is to find the mostly frequent alphabet in the use of Swahili language and to provide the level of confident interval for each alphabet.

This dissertation is organized as follow: first in Chapter Two, we discuss the frequency analysis of Arabic language as well as that of English language. Next, in chapter three, we discuss the alphabets used in Swahili language, and we make some considerations about the Swahili sample articles used as a base for the study of the frequencies analysis (see appendix), data summary are presented using charts, frequency distribution table and a histogram. We also consider Stefan Trost Character frequency (Swahili) analysis. In chapter four, we discuss the statistical analysis of letter frequency in Swahili language, a strong emphasis is put on the use of simple $t-test$ and we provide the level of confident interval for each letter. Lastly is our conclusion.

CHAPTER TWO

FREQUENCY ANALYSIS OF LETTERS

Almost any item, feature, characteristics, events, etc, that can be reliable identified can be analyzed qualitatively and/or qualitatively depending on the purposes of being analyzed. Organizations, people and decision makers depend on the identified analyzed and collected information (data) for further decisions. The data can be analyzed, presented and summarized qualitatively and/or qualitatively to meet the required demand (purposes) [3, 23]. This is what is called statistics. In statistics we basically deal with frequencies and frequency analysis when collecting, analyzing and interpreting the data. Frequency is the number of opportunities a certain data (event) happened to occur [3, 6]. For example, a survey on HIV/AIDS disease was found that 100 people were affected then the responses were recorded according to their ages as follows:-

Age (years)	17-23	24-30	31-37	38-44	45-51	52-58	59-65	66-72
No. of People (%)	6	17	20	18	25	10	3	1

Table 2.1 The frequency distribution table showing the number of people affected agewise

From the table above, there are two rows, first row indicates years of people in group, the second row indicates the number of people being affected for each group, and from the table, and the most affected group is of age between 45-51.

Frequency analysis is a statistical procedures involved in interpreting the data (information) in relation to the occurrences of that data for future use.

Frequency analysis administratively is a statistical procedures involved in interpreting the past record of events such as floods, droughts, food storage, rainfall, etc. to occurrences that event in the future.

In cryptanalysis frequency analysis is the study of the frequency of letters or group of letter in a cipher text (the transformed message to secrete form), and is used as an aid to breaking classical cipher.

Frequency analysis involves the process of colleting data (information), analysis of the data collected and interpreting those data for a certain demand.

Frequency analysis of letter in any given language as in content analysis or textual analysis is the methodology for studying the structure of that language and finding out how many times each letter (alphabet) occurs.

Using a definition of context analysis by Harold Lasswell, we can say frequency analysis of letters is the way of making inferences by objectively and systematically identifying specified characteristics or letter (character) used in any given language text.

On this dissertation "character" is used as a single letter used in a word formation (formulation) and all the data given were quantitatively analyzed. Therefore all the data in letter frequency analysis were summarized quantitatively, thus by counting how many times letters or combination of letters occurs on any given text.

In this chapter we go through on the frequency analysis of Arabic language, were the Arabic alphabet frequency analysis documents were studied. We also study the frequency analysis of English language the general observation were concluded using frequency distribution tables, relative frequency distribution tables and graphs.

2.1 Frequency Analysis of Arabic Language

Chiefly, the Arabic alphabet consists of 28 primary letters; these are letters 1 to 28 in Table 2.2. However, when scripting in Arabic, the eight modified letters listed in positions 29 to 36 in the same table are used just the same. If these 8 modified forms are lumped back into the primary list based on shape or phonetic similarity, the outcome then is as shown in Table 2.3. For accurate frequency analysis, each of the 36 letters of Table 2.2 gets its frequency counted independently. The ordering of the alphabet shown in the tables is more logical than is used by the Unicode standard.

Hand written letters in Arabic manuscript

وتبين البحوث أنوتحليلاالشخصياتفي لغات مختلفةتعتمد علىمحتوى الموضوعأوالمؤلف.

The figure given below shows the full set of Arabic characters includes about ten diacritics, that set can be produced using the Arabic letter key board interlard.

ذ	د	خ	ح	ج	ث	ت	ب	ء	إ	أ	ا
ق	ف	غ	ع	ظ	ط	ض	ص	ش	س	ز	ر
ه	ن	م	لا	لأ	ي	ى	ھ	ن	م	ل	ك

Figure 2.1 Arabic characters that can be produced using the Arabic letter keyboard interllark.

If these eight modified letters are changed back into its primary letter list, then the out come is as shown in Table 2.2.

No	Letter	No	Letter	No	Letter
1	ا	13	ش	25	ن
2	ب	14	ص	26	ه
3	ت	15	ض	27	و
4	ث	16	ط	28	ي
5	ج	17	ظ	29	أ
6	ح	18	ع	30	إ
7	خ	19	غ	31	آ
8	د	20	ف	32	ء
9	ذ	21	ق	33	ة
10	ر	22	ك	34	ؤ
11	ز	23	ل	35	ى
12	س	24	م	36	ئ

Table 2.2 The Arabic alphabet showing the 28 primary letters and 29 to 36 modified letters

No	Letter	No	Letter	No	Letter
1	اأإآء	13	ش	25	ن
2	ب	14	ص	26	ه
3	ت,ة	15	ض	27	وؤ
4	ث	16	ط	28	يىئى
5	ج	17	ظ	29	أ
6	ح	18	ع	30	إ
7	خ	19	غ	31	آ
8	د	20	ف	32	ء
9	ذ	21	ق	33	ة
10	ر	22	ك	34	ؤ
11	ز	23	ل	35	ى
12	س	24	م	36	ئ

Table 2.3 The Arabic alphabet, with modified letter lumped on to their primary forms

Frequency analysis of Arabic letters (characters) concerned only with finding the frequency of the Arabic alphabet by counting the Arabic letters shown in table 2.2. The study made through a number of famous Arabic sources such as.

- The first seven volumes of the seriesالبدايةوالنهاية (The beginning and the end) of IbnKathir).
- The book of الرحيقالمختوم(The sealed Nectar, the life of Prophet Mohammad (Peace be upon him) of Almubarakfur)
- The book of تحفةالعروسين(The master piece of the Brides for Al-Shuri)

All together with a total of 5,122,132 letters come up with the following observation as they were presented below using two figures that show the letter frequency distribution for the counted letter.

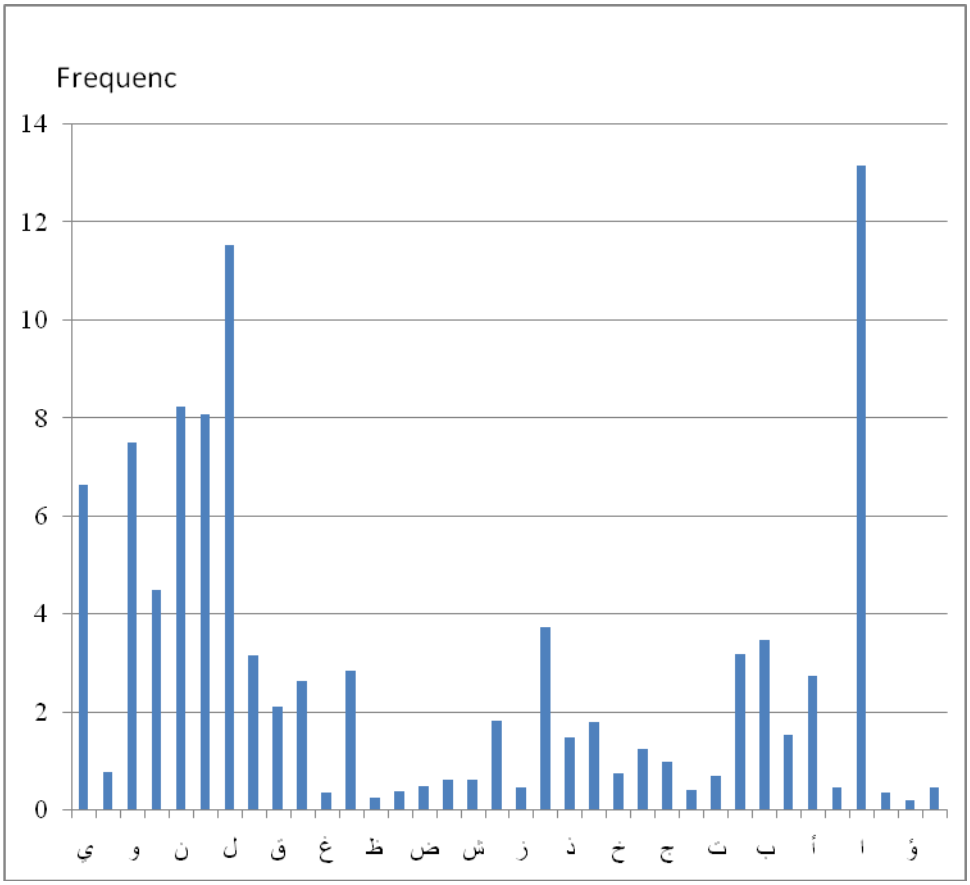


Figure 2.2 A histogram of Arabic letter frequency distribution data sorted by letter

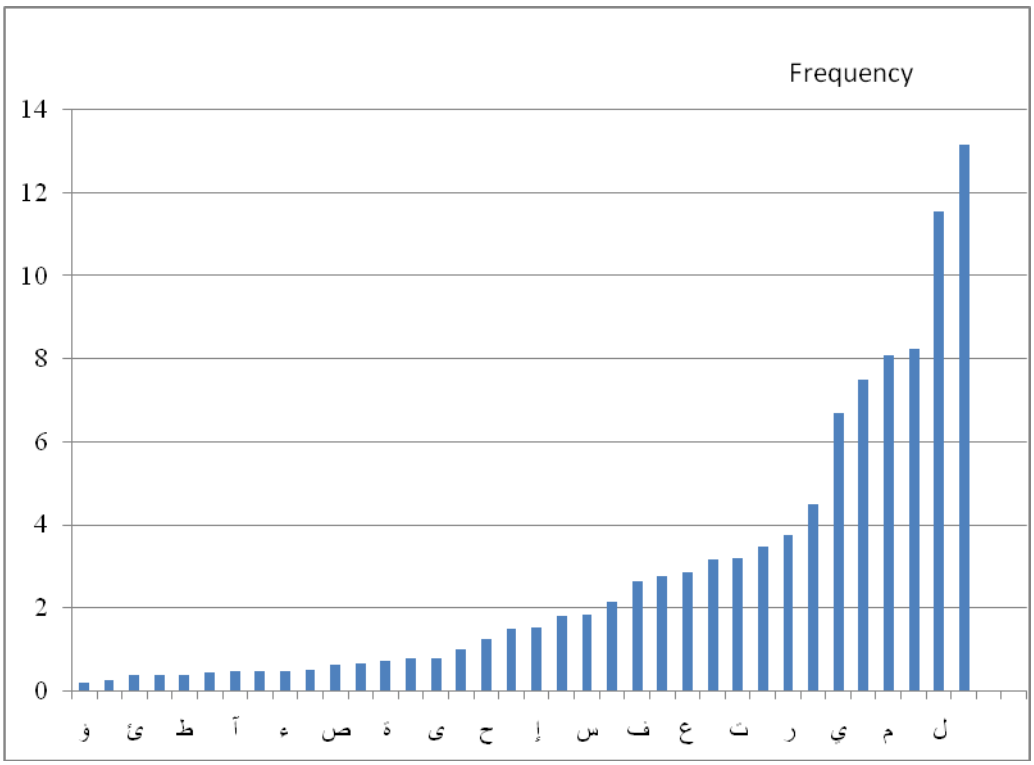


Figure 2.3 A histogram of Arabic letter frequency distribution data sorted by frequency

From the figure above, then the most frequency letter is ‘A’ written as (ا) in Arabic letter and the least frequent letter is ﻭٛ (wow with dot above).

2.3 Frequency Analysis of English Language

English language has its own structure like other languages used in communication. That structure shows that there are 26 English alphabets as shown in the table below.

No.	Letter	No.	Letter
1	A	14	N
2	B	15	O
3	C	16	P
4	D	17	Q
5	E	18	R
6	F	19	S
7	G	20	T
8	H	21	U
9	I	22	V
10	J	23	W
11	K	24	X
12	L	25	Y
13	M	26	Z

Table 2.4 The English alphabet showing 1 to 26 letters used in any English text.

Through different study in frequency analysis of English text it revealed that in any text (English) there is no exact letter frequency distribution which underlies a given English text due to the writers' content. For example, Herbet S.Zim [7], in his classic introduction cryptography text "codes and secret writing" gives the English letter frequency as "ETAON RISH LFCMU GYPWB VKXJQZ" the most common letter pairs as "TH, HE, AN, RE, ER,IN, ON, AT, ND, ST, ES, EN, OF, TE, ED, OR, TI, AS, and TO" and the most common doubled letters as "LL, EE, SS, OO, TT, FF, RR, NN, PP, CC".

Linotype machine assumed the letter order, from most frequency letters to least common, to be **etaoinshrdlucmfwypvbgkjqxz** based on the experience and custom of manual compositors. Thus, E, T, A and O are the most common letters, while Z, Q and X are rare. Likewise, the most common pairs of letter (bigrams or digraphs) are TH, ER, ON, and AN and SS, EE, TT, and FF are the most common repeats. From it, we can conclude that the non sense phrase that represents the twelve most frequency letters in typical English language text is "ETAON SHRDLU". Modern International Morse code encodes the most frequent letters with the shortest symbols; arranging the Morse alphabet into groups of letters that require equal amounts of time to transmit, and then sorting these groups in increasing order, yields **e it san hurdmwgvlfbkopjxczyq,** showing that the most frequent letter is E and the least frequent letter is Q. Similar ideas are used in modern data-compression techniques such as Huffman coding.

Another study done by the Black Chamber [16], gives an approximate ordering of English letter by frequency of use as "ETOIN SHRDL, UCMFG YPWB KXJQZ" (relative frequencies of letters, 2009) and the study molded on the text taken from the website

"http://www.nytimes.com/2011/06/19/world/middleeast/19syria.html?ref=world"

shows that the most frequency letter is E, T and O, with the nonsense phrase "ETAION SHRDL CUMWFGY PBVK JXQZ"

Kenton Born and Dr. David Gu in their study they analysis that, English has a distribution of characters that greatly affects the randomness of the language. Patterns and characters to form made it easier to predict what the following bits of information would be, as more information was collected. Due to such slight differences; it can be concluded that each given English text analysis is differently

from one another due to writers' content (subject). **Thus they differ with differences on the subjects'content. That is letter frequencies, like word frequencies, lied to very both by writer and by subject.** Since accurate average of letter frequencies can only be close by analyzing a large amount of text representative to be sure with a very high degree of accuracy, Kristan Lawson (1998) collects 16 different reliable analyses of letter frequencies each of 1000 letters. By doing so he used the information found in all of the true frequency of occurrence of all letter in the English language (texts).

Below is the chart showing the totals of all the original studies and his averaged result as used by Kristan Lawson.

Letters	Total	Average	Rank
E	1972.6	123.3	1
T	1452	90.8	2
A	1262.2	78.9	3
O	1222.3	76.4	4
N	1135.1	70.9	5
I	1134.4	70.9	6
R	1088.1	68	7
S	997.4	62.3	8
H	798	49.9	9
D	627.2	39.2	10
L	608.1	38	11
C	496.3	31	12
U	456.9	28.6	13
M	433.3	27.1	14
F	385	24.1	15
P	367.4	23	16
Y	316.2	19.8	17
G	291.8	18.2	18
W	270.3	16.9	19
B	229.7	14.4	20
V	171.6	10.7	21
K	93	5.8	22
X	63.7	4	23
J	53.5	3.3	24
Q	43.9	2.7	25
Z	34	2.1	26

Chart 2.1 Frequency Distribution Analysis, showing the total number of each letter is occurrences in the English language (text) for 1000 letters.

Using the above chart (information), Kristan comes out with the final totals of frequencies of the letters of the English alphabet expressed as percentages, that is, the number of times each letter occurs out of random sampling of 100 letters.

No	1	2	3	4	5	6	7	8	9	10	11	12	13	**Total**
Letter	E	T	A	O	N	I	R	S	H	D	L	C	U	
%	12.33	9.08	7.89	7.64	7.09	7.09	6.80	6.23	4.99	3.92	3.80	3.10	2.86	
No	14	15	16	17	19	19	20	21	22	23	24	25	26	
Letter	M	F	P	Y	G	W	B	V	K	X	J	Q	Z	
%	2.71	2.41	2.30	1.98	1.82	1.69	1.44	1.07	0.58	040	0.33	0.27	0.21	**100%**

Chart 2.2 Frequency Distribution Analysis showing the final totals frequencies of the letters in percentages

Using Chart 2.2 above, we have the following table of English frequency with two main columns, the first column showing the frequency distribution analysis in percentages sorted by letter and the second column showing the frequency distribution analysis in percentages sorted by frequency.

Sorted by Letter		Sorted by Frequency	
letter	**Frequency (%)**	**Letter**	**Frequency (%)**
A	7.89	E	12.33
B	1.44	T	9.08
C	3.10	A	7.89
D	3.92	O	7.64
E	12.33	N	7.09
F	2.41	I	7.09
G	1.82	R	6.80
H	4.99	S	6.23
I	7.09	H	4.99
J	0.33	D	3.92
K	0.58	L	3.80
L	3.80	C	3.10
M	2.71	U	2.86
N	7.09	M	2.71
O	7.64	F	2.41
P	2.30	P	2.30
Q	0.27	Y	1.98
R	6.80	G	1.82
S	6.23	W	1.69
T	9.08	B	1.44
U	2.86	V	1.07
V	1.07	K	0.58
W	1.69	X	0.40
X	0.40	J	0.33
Y	1.98	Q	0.27
Z	0.21	Z	0.21
Total	**100%**	**Total**	**100%**

Table 2.5 English letter frequency Distribution table showing the frequency distribution analysis in percentage sorted by letter and by frequency

If we make a histogram graph of the data given in Table 2.5 above, we can have the following two figures, one showing English letter frequency histogram for the data sorted by letter frequency and the second showing English letter frequency histogram for the data sorted by frequency.

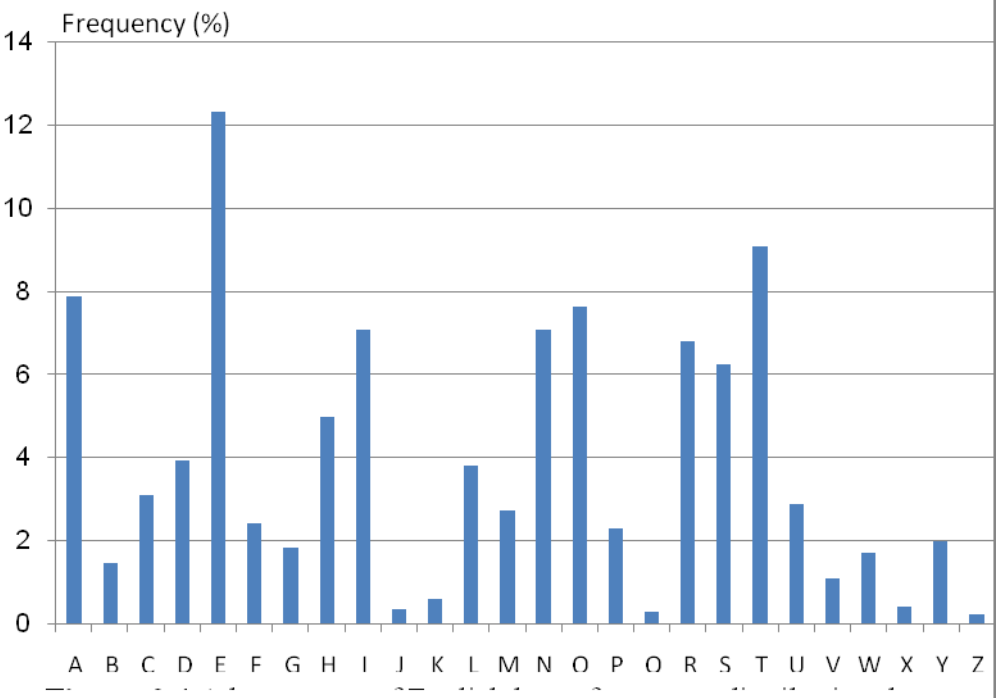
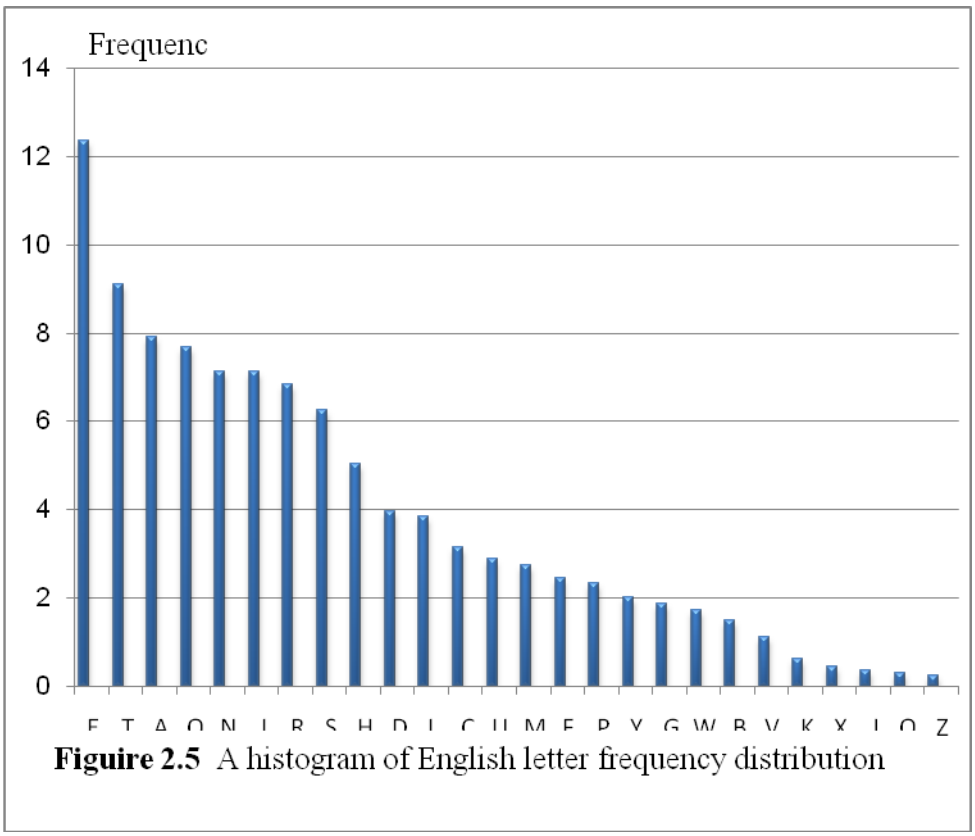


Figure 2.4 A hostogram of English letter frequency distribution data

Figure 2.4 A histogram of English letter frequency distribution data



Figuire 2.5 A histogram of English letter frequency distribution

Figure 2.5 A histogram of English letter frequency distibtion

From the data given above, Kristan concluded that for any given English text, the number of each letter to occur out of a random sampling of 100 letters is as it indicated above. Thus for example, the letter E accounts for 12.33% of all letters in any English text where as Z accounts for 0.21%.

Since many studies do not give statistical analysis of the individual letters, but simply present them ranked in order of frequency, using Kristan results we can conclude that for each single character we have an approximate ordering of English letters by frequency of use as “ETAONI RSHDLC UMFPYG WBVKXJ QZ”, and hence the most frequent letter being “E” as well as the least frequency letter being “Z”

CHAPTER THREE

FREQUENCY ANALYSIS OF LETTERS IN SWAHILI TEXTS

3.1 Introduction

In this chapter we discuss the frequency analysis of letters in Swahili texts. Swahili languageis one of natural language (Bantu language) of East Africa. We focus only on single alphabet (letters) from Tanzania Swahili newspapers. In Section 3.2, we discuss the structure of Swahili language and Section 3.3; we discuss frequency analysis of letters in Swahili articles.

3.2 Structure of Swahili language

Swahili language like other languages has its own structure. It is constructed by twenty four (24) alphabets, either forming a word using a combination of a single alphabet and digraphs or trigraphs, such as "anakwenda shule". A combination of words forms a sentence in which later forms a paragraph (text).

Below is a table that shows twenty four (24) alphabets (letters) used in Swahili language

No.	1	2	3	4	5	6	7	8	9	10	11	12
Letters	A	B	C	D	E	F	G	H	I	J	K	L
No.	13	14	15	16	17	18	19	20	21	22	23	24
Letters	M	N	O	P	R	S	T	U	V	W	Y	Z

Table 3.1 Swahili alphabets

Among those 24 letters, five of them ***"a, e, i, o, u"*** are called primary letters (vowels) and the rest 19 are consonants. Using those vowels and any consonants, we can form a word (words) and then a passage (text) as well, for example;

"Sense ya watu na makazi ni zoezi maalum lenye lengo la kupata idadi ya watu wote waliopo nchini ambapo taarifa zao za kijamii na kiuchumi hukusanywa. Sensa pia hukusanya hali ya makazi ya watu hao"

From the text above we have 35 words, each formed using a consonant and a vowel or as a combination of vowels and consonants, for example in the word; ***"sensa"*** we have ***s*** + ***e***+ ***n*** + ***s*** + ***a.*** Thus the Swahili language is not like Arabic language where words are written in a cursive style that is letters are joined together in word. The Swahili language is like the English language, but the only difference is that in Swahili we use only 24 letters, letters Q and ***X*** is not used and that a word is formed as a combination of consonants and vowels.

Again as in English language, in Swahili language there are repeated (doubled) alphabets such as ***"aa, ee, ii, oo, uu and mm"***, digrams (digraphs) letters that often go together, always formed by combining a consonant letter and a vowel such as ***"ba, ka, me, go"***, trigraphs (a combination of three letters), such as ***"sha, chi, kwa, mwa"***, and a combination of four (quadrigraphs) letters such as ***"shwa, chwe, nywa"***.

In this dissertation we will only deal with single letters and we will base with writen Swahili articles.

3.3 Frequency Analysis of Letters in Swahili articles

3.3.1 Data source for frequency analysis of letters in Swahili texts

Frequency analysis of letters in Swahili language concerned only with finding the frequency of each alphabet (character) show in Table 3.1 above. The study was made by collecting ten Swahili articles together with a total of 10000 letters. (See the appendix)

As we have seen from above that almost all Swahili words are made up with a combination of consonants and vowels letters. Susan Choge [17] argues that vowels have high frequency in use than consonants. Also from Wikipedia [21], we confirm that in any given string of written language, certain letters and combinations of letters tend to differ. To prove those statements on this section we consider only ten samples of Swahili texts with a total of 10000 letters from different authors and contents.

From the studies made on English language [7], it was observed that, in any given language (even in Swahili), no exact letter frequency distribution underlies that language because of writters' interest and the subject matter.

3.3.2 Methodology and analysis

For collecting, organizing and analysing data for Swahili letter frequency analysis, we employ the method of counting how many times each letter occurs by using a computer (the Microsoft word software),where by we copy those Swahili text in a blank word document and then using a "find" feature of word to see the frequencies of each letter. Then the results were recorded by using frequencies and relative frequencies tables. The relative frequency was obtained by using the following formula;

$$\textbf{\textit{Relative frequency}} = \frac{letter\,frequency}{total\,no.\,of\,letters} \times 100 \qquad \textbf{(3.1)}$$

3.3.3 Stefan Trost Character frequency: Swahili

It was observed from Stefan Trost that in Swahili Character frequency, the letters A, I and K are the most common letters. This was observed by using some Swahili texts with together 2,068,457 characters (321,456 words) the texts consist of a good mix of different literacy contents. Below were the Stefan Trost results.

letters	Frequencies (%)
A	20.15
I	11.05
K	7.19
N	6.44
U	6.41
E	5.86
M	4.74
W	4.22
O	3.84
T	3.54
L	3.49
S	3.29
H	2.92
Y	2.75
R	2.43
D	1.85
B	1.84
G	1.56
Z	1.23
J	1.20
C	1.15
P	1.11
F	0.87
V	0.73

Table 3.2 The Relative frequency table sorted according to the letters in percentage

3.3.4 Letter Frequencies in Swahili articles

To make our study more accurately, we combined all frequencies observed for all ten studies made above and we find the total average and the relative frequencies for each letter and then rank those averaged totals to find the most occuring letter as shown in the Table 3.3 below.

Then results were presented using a distribution table and histograms.

l	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Total
A	209	227	249	225	211	220	197	210	244	217	2209
B	21	31	31	9	12	12	16	21	16	27	196
C	9	4	4	8	21	7	3	21	14	5	96
D	13	12	13	25	22	19	26	22	13	15	179
E	40	44	47	40	56	38	44	58	47	51	465
F	8	5	14	13	20	7	1	7	14	4	93
G	17	21	11	15	11	17	42	12	6	9	162
H	33	38	32	28	36	29	17	46	32	32	322
I	131	111	106	134	107	139	101	110	106	102	1147
J	12	4	3	15	15	18	9	11	13	13	113
K	61	56	59	74	58	71	45	60	76	71	631
L	36	35	38	32	33	49	31	24	38	29	345
M	42	60	33	49	33	34	56	45	38	48	438
N	69	64	63	60	77	53	92	64	54	68	664
O	43	50	45	47	53	31	47	32	34	31	413
P	16	10	9	13	12	14	5	5	12	10	108
R	7	23	21	8	25	23	25	15	12	11	172
S	23	40	29	23	32	39	23	35	29	23	296
T	41	27	26	41	26	31	28	18	35	31	304
U	70	47	64	57	48	55	92	71	59	76	637
V	2	9	1	2	10	6	13	2	11	16	72
W	55	31	49	33	48	39	39	53	54	60	461
Y	27	36	19	22	17	23	27	27	23	37	258
Z	17	15	34	27	17	26	21	30	20	14	219
	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	10000

Chart 3.1 Letter frequencies data as observed from the Swahili articles

From the above chart we constructed the frequency distribution tables given below, the left table is sorted according to the letters, and the right table is according to the frequency of the letters. The Relative frequencies (Rf) were calculated using formula (3.1) above.

letters	*frequencies*	*Rf*	*letters*	*frequencies*	*Rf*
A	2209	22.09	A	2209	22.09
B	196	1.96	I	1147	11.47
C	96	0.96	N	664	6.64
D	179	1.79	U	637	6.37
E	465	4.65	K	631	6.31
F	93	0.93	E	465	4.65
G	162	1.62	W	461	4.61
H	322	3.22	M	438	4.38
I	1147	11.47	O	413	4.13
J	113	1.13	L	345	3.45
K	631	6.31	H	322	3.22
L	345	3.45	T	304	3.04
M	438	4.38	S	296	2.96
N	664	6.64	Y	258	2.58
O	413	4.13	Z	219	2.19
P	108	1.08	B	196	1.96
R	172	1.72	D	179	1.79
S	296	2.96	R	172	1.72
T	304	3.04	G	162	1.62
U	637	6.37	J	113	1.13
V	72	0.72	P	108	1.08
W	461	4.61	C	96	0.96
Y	258	2.58	F	93	0.93
Z	219	2.19	V	72	0.72
Total	10000	100	*Total*	10000	100

Table 3.3 The frequency distribution table showing the frequencies and Relative frequencies of the Swahili alphabets sorted by letters (left) and by frequencies (right)

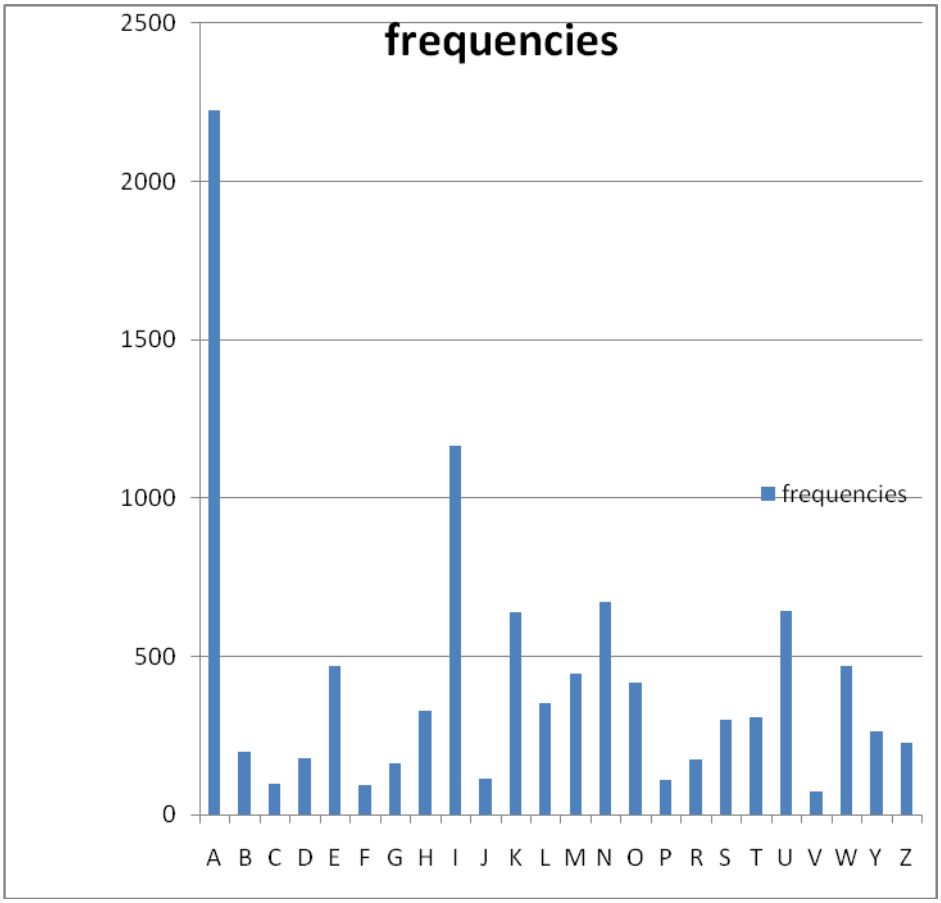


Figure 3.1 A histogram of Swahli letter frequency distribution table showing the relative frequency in percentage

CHAPTER FOUR

STATISTICAL ANALYSIS OF LETTER FREQUENCY IN SWAHILI LANGUAGE

4.1 Introduction

Statistics is the study of the collection, organization analysis, interpretation and presentation of data. It represents the body of methods by which characteristics of a population are inferred through observations made in a representative sample from that population [23]. /lStatistical analysis refers to a collection of methods used to process large amounts of data and report overall trends. Statistical analysis is particularly useful when dealing with noisy data. Statistical analysis provides ways to objectively report on how unusual an event is based on historical data [3]. Since it is not easy to observe the entire populations, sampling and statistical inference are essential so as to make the process of scientific observation as efficient and productive as possible. In this chapter therefore, we discuss some general principles used for the data analysis, interpretation and presentation of the results, the emphasis being on the choice of appropriate student statistical models and methods of statistical inference.

4.2 Definitions and Examples

4.2.1 Mean

This is defined as the value which we get by dividing the total values of various given items in the series by the total number of items, sometimes is also known as arithmetic average, denoted as $\bar{x}$.

For example, if you have given the following values $x_1, x_2, \ldots x_n$, then its mean can be worked out as

Mean $(\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n}$

Where $\bar{x}$ = the symbols used for mean

$\sum$ =Symbols for summarition

x_i= values of the ith item, $i = 1,2,\ldots,n$

n = total number of items.

4.2.2 Standard deviation

That is the most widely used measure of dispersion of a series of data and is commonly denoted by the symbol$\boldsymbol{\sigma}$. It is defined as the square root of the average of squares of deviations, when such deviations for the values of individual items in a series are obtained from the arithmetic average (mean). Its formula is given below;

Standard deviation $(\boldsymbol{\sigma}) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}}$

4.2.3 Universe and Population

From statistical point of view, the term universe refers to the total of the iterms or units in any field of inquiry where as the term population refers the total of iterms about which information is desired.

The population or universe can be finite or infinite. The population is said to be finite if it consists of a fixed number of elements so that it is possible to enumerate it in its totality, where by the symbol N is generally used to undicate how many elements are there. An infinite population is the population in which it is theoretically impossible to observe all the elements; that is; we can not have any idea about the total number of items.

4.2.4 Statistics and Parameters

A statistic is a characteristic of a sample whereas a parameter is a characteristic of a population. For example, when we work out certain measures such as mean, median or mode from sample, then there are called statistics for they describe the characteristics of a sample. But when such measures describe the characteristics of a population, they are known as parameters. For example, the population mean is a parameter, whereas the sample mean is a statistic.

4.2.5 Confidence level and significance level

The confidence level or reliability is the expected percentage of times that the actual value will fall within the stated precision limits. Thus, if we take a confidence level of 95%, then that there are 95 chances in 100 (or.95 in 1) that the sample results represent the true condition of the population within a specified precision range against 5 chances in 100(or.05 in 1) that is does not. Precision is the range within

which the answer may vary and still be acceptable, confidence level indicates the likelihood that the answer will fall within that range, and the significance level indicates the likelihood that the answer will fall outside that range. We can always remember that if the confidence level is 95%, then the significance level will be (100 – 95) i.e., 5%; if the confidence level is 99%, the significance level is (100-99) i.e., 1%, and so on.

4.2.6 Sampling distribution

We are often concerned with sampling distribution in sampling analysis. If we take certain number of samples and for each sample compute various statistical measures such as mean, standard deviation, etc., then we can find that each sample may give its own value for the statistic under consideration. All such values of a particular statistic, say mean, together with their relative frequencies will constitute the sampling distribution of the particular statistic, say mean.

Accordingly, we can have sampling distribution of mean, or the sampling distribution of standard deviation or the sampling distribution of any other statistical measure. It may be noted that each item in a sampling distribution is a particular statistic of a sample. The sampling distribution tends quite closer to the normal distribution if the number of samples is large. The significance of sampling distribution follows from the fact that the mean of a sampling distribution is the same as the mean of the universe. Thus, the mean of the sampling distribution can be taken as the mean of the universe.

4.2.7 Sampling error

A sample survey represents the study of small portion of the population, and as such there would arise some certain amount of inaccuracy in the information collected. Therefore sampling errors are those errors which arise on account of sampling and they generally happen to be random variations (in case of random sampling) in the sample estimates around the true population values.

4.2.8 Frequency table

The table that shows for an ungrouped discrete variable, the values that a variable can take together with the counts of each value in the sample is known as frequency table.

4.2.9 Probability

Probability is an expressed degree of belief or a limit relative frequency of an event over an infinite sample. The set of values taken by a random variable together with their probabilities is known as probability distribution.

4.2.10 Random sample

A sample in which every member of the population has the same chance of appearing such that the members of the sample are chosen independently is known as random sample. Stratified random sample is the random samples that are taken from the strata of a population.

4.3 Student's t-test

Student's t-test, in statistics, a method of testing hypotheses about the mean of a small sample drawn from a normally distributed population when the population standard deviation is unknown. It tell us if there is a significant difference between

two sets of data, or if the average of a set of data differs significantly from a predicted value.The *t* distribution is a family of curves in which the number of degrees of freedom (the number of independent observations in the sample minus one) specifies a particular curve. As the sample size (and thus the degrees of freedom) increases, the *t* distribution approaches the bell shape of the standard normal distribution. In practice, for tests involving the mean of a sample of size greater than 30, the normal distribution is usually applied.

In this dissertation, the method used for data analysis was the student's t-test. We decide to use this method because it is used for a small size sample, that is, $n \leq 30$. Before calulating student's t- test (t), we need to find the sample mean ($\bar{x}$) and sample standard deviation (s). The methods used to calculate sample mean ($\bar{x}$) and sample standard deviation(s) are;

$$Mean\ (\bar{x}) = \frac{\sum x_i}{n} = \frac{x_1+x_2+\cdots+x_n}{n} \tag{4.1}$$

$$standarddeviation(s) = \sqrt{\frac{1}{n-1}(\sum_{i=1}^{n}(x_i - \bar{x})^2)} \tag{4.2}$$

respectively,

Where n = sample size

$\sum$ = Symbol for summation

$\bar{x}$ = Sample mean

x_i = Individual value

$n-1$ = The number of degree of freedom.

The formula used to calculate student's $t-test$ values is as follows;

$$t = \frac{x-\mu}{s}\sqrt{n} \tag{4.3}$$

With$(n-1)$ degree of freedom, and s= standard deviation of the sample.

The level of significance initially chosen was 0.05 (5%) at 23 degree of freedom, which produces t-test critical range of ± 1.83, at this level of significance, we have the t-test calculated using formula (4.3), whereas the population mean μ is obtain by using the interval estimation as follows:

$$\bar{x} \pm t_{\propto} SE(x), \qquad (4.4)$$

Where$SE(x)$ is the standard error of mean and is define as;

$$SE(x) = {}^{s}/_{\sqrt{n-1}} \qquad (4.5)$$

Where s is the sample standard deviation.

The value of $t_{\propto}$ at $100(1-\propto)\%$level is obtained from the $t-table.$

The following table shows the 95% interval population mean μ and $t-test$ values for the letter frequency analysis as they occur in observed samples.

l	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Total	Mean(x)	STDEV
A	209	227	249	225	211	220	197	210	244	217	2209	220.09	16.06
B	21	31	34	9	12	12	16	21	16	27	199	19.9	7.96
C	9	4	4	8	21	7	3	21	14	5	96	9.6	6.80
D	13	12	13	25	22	19	26	22	13	15	179	18	5.55
E	40	44	47	40	56	38	44	59	47	51	465	47	6.77
F	8	5	14	13	20	7	1	7	14	4	93	9.3	5.77
G	17	21	11	15	11	17	42	12	6	9	162	16.2	10.05
H	33	38	32	28	36	29	17	47	32	32	322	32.6	7.45
I	134	111	106	134	107	139	101	111	106	105	1165	116.5	14.24
J	12	4	3	15	15	18	9	11	13	13	113	11.3	4.79
K	61	56	59	74	58	71	45	60	76	71	631	63.9	9.69
L	36	35	38	32	33	49	31	24	38	29	345	35	6.65
M	42	60	33	49	33	34	56	45	38	48	438	44.5	9.54
N	69	64	63	60	77	53	92	64	54	68	664	67.1	11.42
O	43	50	45	47	53	31	47	32	34	31	413	41.6	8.47
P	16	10	9	13	12	14	5	5	12	10	108	10.8	3.55
R	7	23	21	8	25	23	25	15	12	11	172	17.2	7.32
S	23	40	29	23	32	39	23	35	29	23	296	29.9	6.72
T	41	27	26	41	26	31	28	18	35	31	304	30.9	7.12
U	70	47	64	57	48	55	92	71	59	76	637	64.3	13.81
V	2	9	1	2	10	6	13	2	11	16	72	7.2	5.35
W	55	31	49	33	48	39	39	53	54	60	461	47	9.97
Y	27	36	19	22	17	23	27	27	23	37	258	26.1	6.56
Z	17	15	34	27	17	26	21	30	20	14	219	25.5	6.44
	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	10000	1000	

Chart 4.1 Letter frequencies, means and standard deviations as observed from the samples studied

$t-test$ Values for 24 alphabets as used in the investigation of the analysis of letter frequency in Swahili language.

Table 4.1 the mean, standard deviation, 95% interval population mean μ and $t-test$ values for each letter

l	Mean(x)	STDEV(s)	95% interval population mean, μ ($\pm$)	t-test values ($\pm$)
A	220.09	16.06	1	0.1969
B	19.6	7.96	0.49	0.1947
C	9.6	6.8	0.42	0.1953
D	17.9	5.55	0.34	0.1955
E	46.5	6.77	0.42	0.1962
F	9.3	5.77	0.36	0.1973
G	16.2	10.05	0.62	0.1951
H	32.2	7.45	0.46	0.1953
I	11.47	14.24	0.26	0.0577
J	11.3	4.79	0.30	0.1981
K	63.1	9.69	0.60	0.1958
L	34.5	6.65	0.41	0.1950
M	43.8	9.54	0.59	0.1956
N	66.4	11.42	0.71	0.1966
O	41.3	8.47	0.52	0.1941
P	10.8	3.55	0.22	0.1960
R	17.2	7.32	0.45	0.1944
S	29.6	6.72	0.42	0.1976
T	30.4	7.12	0.44	0.1954
U	63.7	13.81	0.86	0.1969
V	7.2	5.35	0.33	0.1951
W	46.1	9.97	0.62	0.0825
Y	25.8	6.56	0.41	0.1976
Z	21.9	6.44	0.40	0.1964

4.4 Confidence Intervals (C.I)

In statistics it is important to measure how certain we can be in the results of an analysis or experiment. One way of computing the measure of confidence in statistical results is by the use of the confidence interval.Confidence intervals express the range within which a result for the whole population would occur for a particular proportion of times an experiment or test was repeated among a sample of the population. Confidence interval is a standard way of articulate the statistical accuracy of an experiment based assessment. If an assess has a high error level, the equivalent confidence interval will be ample, and the less confidence we can have that the experiment results depict the situation among the whole population When quoting confidence, it is common to refer to the 95% confidence interval around an experiment assessment or test result, even though other confidence intervals can be reported (e.g. 90%, 95%, even 99%). We can be rationally confident that the range comprises the true value for the population as a whole where a 95% confidence interval is reported. The characteristics of a normal distribution describe the measure of a confidence interval. The confidence interval is the mean ±1.96 multiplied by the standard deviation when the statistic is distributed normally.

Using chart 3.1 above, then the mean and standard deviation, a 95% confidence interval were calculated. The calculated statistics and the corresponding confident interval (C.I) were included in the following table.

l	n	Mean	STDEV	C.I($\pm$)	95% C.I for Mean	
					Lower Bound	Upper Bound
A	10	220.09	16.06	1	219.9	22.19
B	10	19.6	7.96	0.49	19.11	20.09
C	10	9.6	6.80	0.42	9.18	10.02
D	10	17.9	5.55	0.34	17.56	18.24
E	10	46.5	6.77	0.42	46.08	46.92
F	10	9.3	5.77	0.36	8.94	9.66
G	10	16.2	10.05	0.62	15.58	16.82
H	10	32.2	7.45	0.46	31.74	32.66
I	10	114.7	14.24	0.26	114.44	114.96
J	10	11.3	4.79	0.30	11.00	1.16
K	10	63.1	9.69	0.60	62.50	6.37
L	10	34.5	6.65	0.41	34.09	34.91
M	10	43.8	9.54	0.59	43.21	44.39
N	10	66.4	11.42	0.71	65.69	67.11
O	10	41.3	8.47	0.52	40.78	41.82
P	10	10.8	3.55	0.22	10.58	11.02
R	10	17.2	7.32	0.45	1675	17.65
S	10	29.6	6.72	0.42	29.18	30.02
T	10	30.4	7.12	0.44	29.96	30.84
U	10	63.7	13.81	0.86	62.84	64.56
V	10	7.2	5.35	0.33	6.87	7.53
W	10	46.1	9.97	0.62	45.84	46.72
Y	10	25.8	6.56	0.41	25.39	26.21
Z	10	21.9	6.44	0.40	21.5	22.30

Table 4.2 The mean, standard deviation, standard error and 95% confidence interval for mean for each letter

4.5 Results and Conclusion

From Charts 4.1, we have the information about each letter/alphabet frequency (number of occuarrance in each alphabet), means and standard deviations, and we

have the mean, standard deviation, 95% interval population mean, μ and $t-test$ values for each letter in Table 4.1. The main thing we are interested in is the $t-test$ values column. Since the $t-test$ value for each letter is less than $t-critical$ values (based on $'t'$ distribution table), then there is no significant difference observed samples and the expected population (the sample mean and the interval population mean for the 24 alphabets).

Table 4.2 gives us the mean, standard deviation, standard error and 95% confidence interval for mean for each letter. With the help of this table, one may find the intervals estimates about the population parameter in probabilistic terms. For example, if we take, a letter A with mean equal to 220.09 and its standard deviation to be equal to 16.06 then the 95% Confidence interval estimate of μ is from 219.9 to 221.9, it means that there is a 95% chance that the population mean is within 219.9 to 221.9 intervals. In other words, this means that if we were to take a complete count of all letters in the entire population, the chance are 95 to 5 that we would find the population mean lies between 220.9 to 221.9. (In case we want to change the degree of confidence in the interval estimate, the same can be done using the table of areas under the normal curve).

Generally, we can say that there is a 95% chance of occuarance for each letter lies between those intervals.

Using Stefan Trost Character frequency (Swahili) analysis and Table 3.3, we can conclude that for any given Swahli article, the number of each letter to occur out of a random sampling of 1000 letters is as indicated in Figure 3.1 and Table 3.2, showing that the top five letters are A, I, N, K and U, and the last five letters are J, P, C, F, and V, A being the most frequent letter whereas V is the least frequent letter.

REFERENCES

[1]. Abraham Sinkov, "Elementary Cryptanalysis; A Mathematical Approval", the mathematical Association of America, 1966 ISBN 0883856220

[2]. Alfred J. Menezes, Paul C. van Oorschot and Scott A. Vanstone. "Handbook of Applied Cryptoraphy." CRC Press, 5th edition, 2001.

[3]. C.R.Kothari. "Research Methodology, Methods and Techniques," 2nd revised edition 2004, New Age International(P)Limited,Ansari Road, Daryaganj New Delhi.

[4]. Daniel L. Joverhoeven Haute Ecole de BruxellesInstitutSupériecr de TraducteursetInterprétes rue Joseph Hazard B-1180 Brussels dep of Germanic Languages.

[5]. DavidL. Hoover, "Quantitative Analysis and Literary Studies," in Susan Schreibman and Ray Siemens, eds., A Companion to Digital Literary Studies, (Oxford: Blackwell, 2008), Chapter 28,
http://www.digitalhumanities.org/companionDLS/

[6]. Frequency and Regression Analysis. Chapter in H.P. Ritzema (ed. 1994), Drainage principles and Applization, Publ. 16, page 175-224 (ILRI) Wageninjen. The Netherlands ISBN 90-70754-33-9

[7].Herbert S. Zim. "Codes and Secret Writing" (abridged edition) Scholastic Book Services, fourth printing 1962. Copyright 1948 Herbert S. Zim, Kenton Brown and Dr. David Gustatfin

[8]. Ibrahim A. Al-Kadi "The origins of cryptology: The Arab contributions", Cryptologia, 16(2) (April 1992) pp. 97–126.

[9]. Kahn, David L. (1996). "The codebreakers: the story of secret writing." New York: Scribner. ISBN 0-684-83130-9.

[10]. Kristan Lawson (1998) .www.atlas-games.com/letterhead. Letter frequency in letter head.

[11]. Madi, Mohsen M. (2010). "Quran Suras Statistics".Intellaren Articles.Retrieved 16 January 2011.

[12]. Michael Collins “A new Statistical parsed based on texical dependencies.”In proceeding of the 34-Annual meeting of the Association of computational Linguistics Sanifu Cruz, CA. 1996.page184-191.

[13]. Neal Koblitz, “A course in Number Theory and Cryptograph,” 2nd Edition 1948 page 54-55.

[14]. Nordic Journal of African Studies 12(3):405-423 (2003) “Semantic Analysis of Kiswahili words using the self organizing help.”WanjilenNg’ang’a University of Heinsinki Finland.

[15]. Richard Spillman. “Classical and Contemporary Cryptology.”Prentice Hall. 2005.

[16]. Singh Simon. "The Black Chamber: Hints and Tips". Retrieved 26 October 2010.

[17]. Susan Choge. “Understanding Kiswahili Vowels, Journal of Pan Africa Studies” (M.Phil) MasindeMulilo University of Science and Technology, Kakamega Kenya.

[18]. W. Diffie, M. E. Hellman, “Privacy and Authentication: An Introduction to Cryptography” in Proc. IEEE, Vol. 63(3) Mar 1979, page 397-427.

[19]. White Field Diffiet and Martin Hellnon, “New directions in Cryptograph IEEE Tran Theory,”VOl. IT-22 no.6, Nov. 1976 page 644-654.

[20]. www.richkni.comuk/php/cryptal/freq.php. Decrypting Text-code breaking software

[21]. www.wikipedia, the free encyclopedia “frequency analysis”

[22]. www.wisegeek.com/ what is a frequency list htm?

[23]. en.wikipedia.org/wiki/statistics,the free encyclopedia,"statistics."

[24]."http://www.nytimes.com/2011/06/19/world/middleeast/19syria.html?ref=world

"

APPENDIX

A. SAMPLE COLLECTIONS

Sample 1

Halmashauri za shauriwa kuunda chombo kusimamia maendeleo

WAZIRI Mkuu mstaafu Bw. David Msuya, amezishauri Halmashauri za Wilaya,mkoani Kilimanjaro, kuunda chombo ambacho kitasimamia utekelezaji wa shughuli za maendeleo ya wananchi.

Bw. Msuya alitoa ushauri huo mwishoni mwa wiki wakati akizungumza na wajumbe wa Kamati ya Fedha ya Baraza la Madiwani wa Halmashauri ya Wilaya ya Mwanga pamoja na kuweka mikakati ya kusukuma maendeleo wilayani humo.

Alisema kazi kubwa ya chombo hicho ni kubuni miradi mbalimbali ya maendeleo na kutafuta njia ya kupatafedha za utekelezaji ili kuharakisha maendeleo ya wananchi. Bw. Msuya ambaye ni Mwenyekiti wa Mfuko wa Maendeleo mkoani Kilimanjaro (KDF), alisema vyombo hivyo vitafanya kazi karibu na mfuko huo ili kuhakikisha miradi husika inatekelezwa.

“Fedha za miradi zimekuwa zikitumwa, kuna watu ambao wanapaswa kutekeleza miradi hii lakini tunahitaji watu wengine ambao watakuwa wanaifuatilia kwa karibu ili isikwame,” alisema.

Awali Mkuu wa Wilaya hiyo, Bw. Athuman Mdoe, wamepata mwekezaji kutoka nchini India ambaye anataka kujenga kiwanda cha kutengeneza juisi na tayari ekari 3,000 zimetengwa kwa ajili ya kilimo cha miwa ya kiwanda hicho.

Alisema kiwanda

Sample 2

Mbwa mwitu 40 wauawa kwa moto Hifadhi ya Serengeti

Zaidi ya mbwa mwitu 40 wameuawa kwa kuchomwa moto ndani ya mapango katika vijiji vya Kertalo na Orkiu mpakani mwa Hifadhi ya Taifa ya Serengeti Wilaya ya Ngorongoro mkoani Arusha.

Habari kutoka Loliondo zinaeleza kuwa mbwa mwitu hao waliuawa na watu wasiojulikana ambao wanadaiwa kujaza kuni ndani ya mapango wanayoishi wanyama hao nakuwasha moto.

Imeelezwa kuwa baada ya kuwasha moto, wauaji waliyafunika mapango hayo kwa mawe na kuwateketeza wanyama hao wanaotajwa kuwa katika hatari ya kutoweka duniani.

Mkuu wa Wilaya ya Ngorongoro, Elius Wawa Lali, alithibitisha kutokea kwa tukio hilo na kueleza uchunguzi wa kubaini wahusika unaendelea.

“Ni kweli tukio hilo limetokea na kwa sasa vyombo vya ulinzi na usalama vinalifanyia uchunguzi na ukikamilika tutawapa taarifa kamili," alisema. Wakizungumza na waandishi wa habari, baadhi ya wananchi wa vijiji hivyo walisema siku chache kabla ya mbwa mwitu hao kuuawa, watu hao waliua ndama wanne na mbuzi 157 mali ya wanakijiji. Mbaaryo Papalai alisema kuwa mbwa mwitu hao wamekuwa wakitishia usalama wa mifugo yao na kila walipo ua mbuzi hakuna fidia yeyote au kifutama chozi wanachopata kutoka serikalini. “Mi…

Sample 3

…alo limekuwa maarufu hadi leo; “Hivi mwendesha farasi ni lazima azaliwe kuwa farasi kwanza ili ajue kupanda farasi?” Uchezaji huu wa Sacchi umeigwa na baadhi ya makocha kama Zdenek Zeman na makocha wote ambao hufuata mfumo wa Mtaliano huyu wanajulikana sasa kwa jina la *Sacchistas*.

Tangu Sacchi, watu wakaanza kubadili mawazo yao kuhusu ukocha wampira wa miguu. Kwamba silazima mtu acheze mpira ndiyo ajekuwa kocha mzuri. Baada ya Sacchi, mtu wa pili ambaye amebadili mawazo ya wengi kuhusu soka ni Jose Mourinho. Kwa baadhi ya watu, huyu huenda ndiye kocha bora zaidi wa soka kwa sasa.

Hakuwahi kucheza soka la ushindani. Mwenyewe amewahi kukiri kwamba hakucheza soka kwa vile hakuwa na shida nyumbani kwao. Kwenye sikukuu yake ya kuzaliwa kutimiza miaka 18, baba yake mzazi ambaye alikuwa mwanasoka, alimnunulia zawadi ya gari aina ya Mercedes Benz.

“Wachezaji wengi wanacheza soka kwa sababu wanashida ya maisha. Miye nilikuwa Napata kila kitu kwa wazazi wangu. Nimeanza kuendesha benz bila ajira nikiwa na miaka 18. Ningetaka nini tena kingine duniani?” aliwahi kuhoji.

Sifa kubwa ya Mourinho nikutengeneza timu zenye nidhamu ya kitimu ya hali ya juu. Hutengeneza wachezaji wenye roho ngumu na wasiokubali kushindwa.

Sample 4

…kutoka Uganda, Kenya, Rwanda, Burundi, India na Comoro na wengine kutoka Mikoa ya Manyara, Arusha, Morogoro, Dar es Salaam, Mbeya na Mwanza ndiyo wanaoongoza kununua vitunguu vya Singida kwa wingi katika soko lililotengwa. Bw. Rugangila aliyataja, maeneo yanayolima vitunguu kwa wingi Wilaya ya Iramba kuwa ni Mwanga, Nduguti, Gumanga, Ilunda na maeneo mengine ni Iguguno, Ulemo, Msingi, Singa.Maeneo yanayolima vitunguu wilayani Singida ni Mtinko, Merya, Kinyeto, Ilongero, Ikhanoda na mengine ni Mdida, Mwanyonye, Msange na Ngamu. Ilieleza wakulima wa vitunguu huanza kutayarisha mashamba yao mwezi Novemba hadi mwezi Aprili, tayari vitunguu huanza kuvunwa na kuuzwa Bw. Rugangila alitaja mbegu za vitunguu zinazolimwa sana Singida kuwa ni Red Bombay, Red Globe, Red Creole na Tropicana na amewahimiza wakulima kutumia mbegu bora ili waweze kupata mazao mengi na bei nzuri ilikukidhi mahitaji ya soko.

Hata hivyo amewahimiza wakulima kutumia mashamba darasa. Mkoa wa Singida katika miaka ya nyuma ulikuwa maarufu sana kwa uzalishaji wa vitunguu lakini sasa, uzalishaji wake umeanza kupungua kutokana na uhaba wa mvua, kukosa dawa za kuua wadudu, ukungu na utaalamu mdogo juu ya upandaji, uvunaji na utunzaji wake.

Sample 5

Ni waliostaafu na pili wale ambao wana umri wa kuanzia miaka 45 nakuendelea. Isingewezekana kwa mtu mwenye umri wa miaka 33 kupewa nafasi hiyo. Huo ni umri wa kucheza na watoto wako baada ya kustaafu soka au kufanya kazi nyingine utakazopewa na timu yako.

Hata hivyo, miaka 20 na ushee iliyopita, akaja mtu anaitwa Arrigo Sacchi (tamka Saki). Mtoto wa muuza viatu ambaye hakucheza soka lolote kueleweka katika ujana wake. Watu wakacheka sana alipoteuliwa kuwa kocha wa AC Milan lakini akawashangaza kwa kutwaa mataji mfululizo.

Lakini, kubwa zaidi ni kuwa alileta vitu vipya katika soka ambavyo havikuwapo kabla yake. Yeye ndiye mtu aliyegawanya uwanja wa soka katika vipande alivyoviita zones. Mpaka leo hii kuna ukabaji unafahamika kwa jina la zonal marking (ukabajiwanafasi).

Falsafa ya Sacchi ilikuwa kwamba timu yake ni lazima icheze kwa kuziba nafasi na si wachezaji watimu husika. Timu ya Sacchi ilikuwa inatembea kwa pamoja. Inashambulia na kujilinda. Na akatwaa mataji ya kutosha. Kwa vile hakuwa na historia ya kucheza kandanda kabla, waandishi wa habari wakawa wanamdhihaki. Akawapa jibu ambalo limekuwa maarufu hadi leo; “Hivi mwendesha farasi ni lazima azaliwe kuwa farasi kwanza ili ajue kupanda farasi?”

Sample 7

MAPOTO SEKTA YA UTALII YAPAA JUU

Mapato yatokanayo na sekta ya utalii nchini yameongezeka kutoka dola za kimarekani milioni 739.06 hadi kufikia dola za kimarekani milioni 1,353.29. Mkuu wa mkoa wa Mwanza, Evarist Ndikilo alisema mapato hayo yameongezeka kati ya mwaka 2000 na mwaka jana ampapo mchango wa utalii katika pato la Taifa umeongezeka na kufikia asilimia 17 huku ukichangia zaidi ya asilimia 25 ya fedha za kigeni.

Akizungumza wakati wa ufunguzi wa wiki ya utalii inayofanyika kitaifa mkoani hapa, Ndikilo alisema Tanzania imekuwa ikinufaika na sekta ya utalii kwa kuwa na ongezeko kutoka 501,669 mwaka 2000 hadi kufikia watalii 867,994 mwaka jana. Akizungumzia changamoto zinazohitaji uawjibikaji wa wadau wa sekta ya utalii, alisema ni pamoja na huduma za mikopo na fedha ambazo sekta za fedha na mitaji zinapaswa kusaidia utoaji wa elimu ya mikopo ili kuwezesha kuanzisha na kuendeleza miradi ya utalii kwa ufanisi.

“Kuanzisha miradi ya utalii huhitaji mtaji na ufahamu wa kuendesha biashara ya utalii hivyo ipo haja kwa benki kama NMB, CRDB na NBC kuwasaidia wajasiriamali wadogo ambao wanamchango mkubwa katika kusaidia utalii,” alisema Ndikilo. Aidha Ndikilo alitaja changamoto nyingine kuwa ni uwekezaji katika huduma za utal

Sample 8

ZIKIWA zimebaki siku 10, kabla masalia ya mwili wa aliyekuwa Kadinali wa kwanza mweusi barani Afrika, Laurean Rugambwa, kuzikwa upya kwenye Kanisa Katoliki la kwanza mkoani Kagera, uongozi wa kanisa hilo umetoa onyo kali kwa wafanyabiashara walioanza kunufaika na tukio hilo.

Hatua hiyo imekuja siku chache baada ya wafanyabiashara katika Manispaa ya Bukoba, kutengeneza fulana zenye nembo ya kanisa hilo na picha ya Kadinali Rugambwa, ambazo zinauzwa sh. 12,000 bila ridhaa ya uongozi wa kanisa hilo. Akizungumza na gazeti hili, Katibu wa Askofu wa Kanisa Katoliki Jimbo la Bukoba,

Padri Deodatus Rwehumbiza, alisema waumini ambao watanunua fulana zilizotengenezwa na wanyabiashara hao bila Baraka za kanisa hilo, watakuwa wamefanya makosa.

“Upo utaratibu ambao umeandaliwa na kanisa kwa kukaa katika maeneo maalumu yaliyotengwa, waumini ambao watavaa Fulani zinazouzwa na wafanyabiashara hawatapewa kipaumbele kwasababu si zale za tukio lililopo mbele yetu. “Zipo sare ambazo zimeandaliwa ili waumini ambao watahudhuria tukio hili waweze kufahamika kirahisi, wanaponunua sare zetu wanachangia ufanisi wa sherehe hizi,” alisema. Padri Rwehumbiza alitoa ufafanuzi huo baada ya gazeti hili kutaka kujua kama waumini hao watakwenda kanisani siku hiyo

Sample 9

“Katika bara la Afrika tatizo la msongamano magerezani ni kuwepo kwa mahabusu wengi na hiyo inatokana nasababu mbalimbali ambazo zinajulikana kama kutokumalizika mapema mashauri yao yaliyoko mahakamani” alisema Mboya. “Zaidi ya nusu ya walioko magerezani si wafungwa ni mahabusu hiyo LSRP walijitahidi kupambana na tatizo hilo la msongamano wa mahabusu nchini.” Katibu wa Wizara ya Katiba na Sheria, alisema kuwa lengo la LSRP ni kuhakikisha sekta ya sheria inakuwa bora” “Sekta ya sharia bado inakabiliwa na changamoto nyingi hasahasa jeshi la magereza. Kitendo cha kutoa magari haya ni kwa ajili ya kukabiliana na changamoto hizo,” alisema Mbonde. Alisema mahabusu wanahaki ya kupaa kila kitu kama binadamu wengine na kwamba kukaa kwenye magari mazuri ni haki ya mahabusu. Mbonde alisema magari hayo yaliyotolewa yanathamani ya Sh.967 milioni na kwamba ni muhimu ya katumika kwa kazi iliyokusudiwa na sivinginevyo.

“Vifaa vinavyotolewa vitumike kwa kazi iliyokusudiwa. Magari haya pia yafanyiwe matengenezo kwa kufuata utaratibu na sivinginevyo,” alisema Mbonde. Alisema program hiyo ya maboresho ipo kwa Mikoa ya Dar es Salaam na Pwani na kwamba hadi hivi sasa yameshatolewa magari 22 yakubeba mahabusu

Sample 10

Tulifundishwa pia (na inabidi tuwafundishe wana wetu) kwamba haki haigawanyiki; ama iko kila mahali ama haipo popote. Haiwezekani kutetea haki kwa ajili ya sehemu ya watu na kuikataa kwa ajili ya sehemu nyingine ya watu, Wawe ni watu wajaii au nchi moja au jamii mbalimbali au nchi mbalimbali.

Ndiyo maana mambo yalikuwa yakifanyika Guatemala, au CHILE, au El Salvador, au Timor Leste au Afrika Kusini au Msumbiji,…kokote kulikotokea hali ya kusigina haki za watu, Watanzania tulipaza sauti kuwaunga mkono wanyonge na kuwakemea waonezi, hata pale walipokuwa ni wakubwa.

Hii ilijenga taswira ya Tanzania kama nchi ndogo yenye Taifa kubwa. Hadi leo wapo watu wengi katika Afrika na duniani wanaoitazama Tanzania kwa macho hayo ya miaka mingi iliyopita, wasijue kwamba tulikwisha kuiacha misingi hiyo. Ndiyo maana wanashtushwa na matukio ya siku hizi wanapo ambiwa kwamba Tanzania nayo inaweza kuvumilia polisi wake kuwafanyia unyama raia wake.

Kila siku linapotangazwa tukio jingine la aina hii tunapokea maswali mengi ambayo tunashindwa kuyajibu. Tunashindwa kujibu juu ya upuuzi wa “kikombe cha

Loliondo”kilichopigiwa mbiu na wakuu wa serikali waliotaka tuamini kwamba tunaweza tukazifunga Muhimbili, Bugando, KC