ISSN No:-2456-2165

Using Demographic Techniques to Detect Errors in Age Based Data

Adeyemo, S.O.
Mathematics and Statistics Department
Federal Polytechnic, Nekede,
Owerri, Imo State, Nigeria

Abstract:- Demographic data are usually classified by age and sex, as they both are very important variables. National plans for the provision of such need as housing, food, education, health, e.t.c depend on the relevant socio – demographic statistics classified by age and sex. The importance of accurate age-sex data in demographic analysis cannot be over emphasized as there are errors associated with age and sex. As age data tend to be more inaccurate than sex data, there is need for investigation and evaluation of the quality of the data collected on age. This paper is purposely prepared to evaluate the accuracy of age reporting using the demographic and health survey 2018 for Nigeria using demographic techniques. The Whipples' and Meyer's indices as well as the United Nations Age-Sex accuracy index were determined. The result of the work has shown very accurate age data reporting for all except male at age ending with "0" in Nigeria. For the Myer's index, the most preferred final digits are '5' and '0', while the most avoided final digit by both sexes is '1'. Furthermore, the calculated age-sex accuracy index is 57.61 that qualified the age data is deficient and requires massive adjustment before they could be meaningfully used.

Keywords:- *Digit preference, Whipple's index, Myer's index, Age, UN age-sex accuracy Index*

I. INTRODUCTION

Demographic analysis remains an important tool for evaluating census data, particularly in countries where independent sources of data, such as vital registration and sample surveys, are lacking or where a Post-Enumeration Survey (PES) is not conducted. Demographic data, which is the bed rock for any meaningful analysis, are usually classified by age and sex. Other parameters are often analyzed and results interpreted in relation to age. And as such, age is an important demographic variable. If any meaningful decision is to be made at any level, correct age or error free age data must be used because many decisions are age-sensitive and misrepresentation of age may lead to inappropriate action. For example, (in medicine) a screening mammography is recommended to start at age 50 and patient's inaccurate age may result in having either an unnecessary or delayed test.

The importance of accurate age-sex data in demographic analysis cannot be overemphasized. National development plans for the provision of such need as housing, food, education, health, employment, manpower, and etc, depend on the relevant socio-demographic statistics classified by age and sex. Most population analysis like fertility and mortality are either age-sex dependent or age-sex selective (Lerche, 1983 and Kpedekpo, 1982).

The accuracy of age data collected by house-to house surveys varies in different set-ups and depends on numerous factors. This is clearly indicated in studies which describe the age-related data of census from different countries (Pardeshi, 2010). A number of errors and misstatements have been noted with respect to age-related data. Misstatement of age is a common example of content error in census and surveys. Of these irregularities, age heaping and age shifting are common phenomena. Age data frequently display excess frequencies at round or attractive ages, such as even numbers and multiples of 5 leading to age heaping. Age heaping is considered to be a measure of data quality and consistency.

In demographic studies, age misreporting is a common phenomenon (Shryock, 1976). The most common phenomenon among the irregularities is the age heaping. Age data frequently displays excess frequencies at round or attractive ages, such as even numbers and multiples of 5 leading to age heaping. Age heaping is considered to be a measure of data quality and consistency (Pardeshi, 2010).

In developing countries, the deficiencies among other problems are as the results of individual ignorance about certain personal details, also digits preference and sometimes open hostility to some types of inquiry due to ignorance (Yusuf, 2012). In Nigeria, age heaping is one of the irregularities in census/survey reporting of age. Since age misreporting is a common phenomenon in demographic studies, while the quality of data in age-sex distribution is very important in medical studies, innovative methods in data collection along with demographic techniques for evaluation of the age and sex data should be applied to ensure accuracy of the age data. The demographic evaluation techniques include Whipple's index, Myer's index and age-sex accuracy index. The Whipple's index (index of concentration), invented by the American demographer George Chandler Whipple (1866–1924), indicates the extent to which age data show systematic heaping on certain ages as a result of digit preference or rounding. Myer's blended index of digit preference is used for evaluating single-year age-sex data by giving the extend of digit preference for all the digits 0, 1, 2, …,9 (Kpedekpo, 1982). Pardeshi (2010) has used Whipple's index and identified a large age heaping at ages ending with terminal digits '0' and '5', on a data collected during a community survey in the Yavatmal district, Maharashtra, India. Shirley et al., (2004) have used Myer's blended index and age-sex accuracy index to evaluate age and sex data from the Census population of Provinces and Territories of Canada. The result has shown no preference or avoidance of any year of birth, and the accuracy of the age-sex population data for almost all the provinces in Canada. The Quality of age data in patients from developing countries was evaluated using Whipple's and Myer's indices (Denic et. al, 2003).

Despite the risk of misclassification of age data in medical studies, analysis on the prevalence and magnitude of age misreporting in medical studies is rarely available in Nigeria. Yusuf (2012), discovered that the age data collected for the outpatients in General Hospital Dutsin-ma were very rough in quality for both male and female outpatients. There was age heaping at ages ending with terminal digits '0' and '5', indicating a preference in reporting such ages. The result has shown that, about 86 percent of male outpatients and 88 percent of female outpatients reported ages with incorrect final digits. The evaluation of the outpatients' age data using the demographic techniques has finally qualified the data inaccurate as the results of systematic heaping and digit preference.

This study explored the Whipple's index, Myer's index and UN age-sex accuracy index to evaluate the accuracy of age and sex based data from Nigeria (DHS, 2018).

II. MATERIALS AND METHODS

A. Methods

The data used was from the Demographic and Health Survey (DHS), 2018 for Nigeria. The data presented according to sex and age in single years was used for both the Whipple's and Myer's indices. The data was then interpolated for five-year for the UN age-sex accuracy index.

B. Statistical Analysis

Age heaping and digit preference were measured by calculating Whipple's index and Myers' blended index. Age accuracy index was also calculated.

The analyses performed are:

- *Whipple's Index:*

Whipple's index is applicable where age is reported in single-years. It gives the relative preference for digits '0' and '5' while reporting age in the interval 23 and 62 years (Kpedekpo, 1982). It is computed as

$$Whipple'sIndex = \frac{\sum(P_{25}+P_{30}+P_{35}+...+P_{60})}{1/5\sum(P_{25}+P_{30}+P_{35}+...+P_{60})} \times 100$$

(1)

To evaluate for ages ending with '0', i.e. 30, 40, 50 and 60, the index is calculated as

$$Whipple'sIndex = \frac{\sum(P_{30}+P_{40}+...+P_{60})}{1/5\sum(P_{25}+P_{30}+P_{35}+...+P_{60})} \times 100$$

(2)

To evaluate for ages ending with '5', i.e. 25, 35, 45 and 55, the index is calculated as

$$Whipple'sIndex = \frac{\sum(P_{25}+P_{35}+P_{45}+P_{55})}{1/5\sum(P_{25}+P_{30}+P_{35}+...+P_{60})} \times 100$$

(3)

If there is no heaping at age reporting ending with '0' and '5', the index will have a value of 100. If there is complete heaping, the index will have a value of 500. Therefore, the value of Whipple's Index (WI) for each end-digit lies between zero (when no person is reported at ages ending with the end-digit) and 500 when every person was reported at ages with the end-digit

Whipple's Index	Quality of data	Deviation from
<105	Very Accurate	5%
105 – 110	Relatively Accurate	5 – 9.99%
110 - 125	Ok	10 – 24.99%
125 - 175	Bad	25 – 74.99%
> 175	Very Bad	≥ 75%

Table 1: The united nations' recommendation for measuring age heaping as identified by whipple's index

- *Myer's Blended Index*

This index is used for evaluating single-year age-sex data. It gives the extent of digit preference for all digits 0, 1, 2, 3,…, 9. It can be used to report errors for all ages 10 – 89 years (Kpedekpo, 1982). The underlined assumption of this method is that in the absence of systematic irregularities in the reporting of age, the blended sum at each terminal digit should be approximately equal to 10% of the total blended population. If the sum at any given digit exceeds 10% of the total blended population, it indicates over selection of ages ending in that digit (digit preference). On the other hand a negative deviation (or sum that is less than 10% of the total blended population) indicates under-selection of the ages ending in that digits (digit avoidance). If age heaping is non-existent, the index would be approximately 0 (Kpedekpo, 1982). The procedure for computation is as follows:

- Sum all the population ending in each terminal digit over the whole range for the ages 10 – 89.
- Sum all the population ending in each terminal digit over the whole range for the ages 20 – 89.
- Multiply the sums of ages at each terminal digit in (1) above by co-efficient, 1,2,3,4,5,6,7,8,9,10.
- Multiply the sums of ages at each terminal digit in (2) above by co-efficient 9,8,7,6,5,4,3,2,1,0.
- Add the product of (3) and (4) above to obtain the blended sum at each terminal digit.
- Add up the blended sum in (5) above.

- Find the percentage of the blended sum at each terminal digit to the total of the blended sum.
- Find the deviation of the percentage distribution from 10.

➢ *Ages-Sex Accuracy Index (Joint Score)*

This index measures the level of quality of age-sex population data. It employs the age ratios and the sex ratios simultaneously (Kpedekpo, 1982), and computed as:

$$JS = ARSM + ARSF + 3(SRS)$$

JS = Joint Score

ARSM = Age Ratio Score for Male

ARSF = Age Ratio Score for Female

The United Nation (UN) scaling for estimating the reliability of the data is:

- Reliable if JS < 20
- Usable with adjustment if $20 \leq JS \leq 40$
- Deficient and requires massive adjustment before use and interpreted with care and caution if $40 \leq JS \leq 60$ and
- Grossly erroneous and risky to utilize for any inference if $JS > 60$

➢ *Age Ratios*

Age ratio is usually defined as the ratio of the population in the given age group to one half of the population in the two adjacent groups.

Mathematically,

Let ${}_5P_x$ be the age group from age x to age $x + 5$, ${}_5P_{x-5}$ and ${}_5P_{x+5}$ be the preceding and the following age groups respectively, then,

$$Age\ ratio = \frac{{}_5P_x}{1/2\,({}_5P_{x-5} + {}_5P_{x+5})} \times 100$$

(5)

The discrepancy at each computed age group when compared with the expected value (which is usually 100) is a measure of net age misreporting. An age accuracy index is derived by taking the absolute average deviation from 100 of the age ratios and summing over all the age groups. (The overall measure of the accuracy of an age distribution, called an age accuracy index.

➢ *Sex Ratios*

The sex ratios or age specific sex ratios (number of males per 100 females in each age group) is defined as the ratio of the population of males in the given age group to the population of females in that given age group.

Mathematically,

Let ${}_5P_x^m$ be male aged x to age $x+5$ and ${}_5P_x^f$ be female aged x to age $x+5$

$$Age\ specific\ sexratio = \frac{{}_5P_x^m}{{}_5P_x^f} \times 100$$

(6)

When the age specific sex ratio is summed over successive absolute differences between one age group and the next one, the average of the summation is called the sex accuracy index (Kpedekpo, 1982)

III. RESULT

The data consisting of the age and sex distributions of Nigeria was obtained from the Demographic and Health Survey (DHS), 2018. Various demographic techniques discussed were employed to evaluate and discuss the accuracy of the age and sex data.

A. *Whipple's Index*

The total female population in the age group '23–62' was 35512. The total male population in the age group '23–62' was 32023. Among them, the population reporting age ending in '0' was 6149 (female), 6063 (Male) and those reporting age with the terminal digit of '5' were 7393 (female), 6211 (male). Thus, Whipple's index for age with terminal digit '0', was 86.58 (female) and 94.67 (male). Whipple's index for age with terminal digit '5', was 104.09 (female) and 96.98 (male).

Table 2 shows the Whipple's indices and their corresponding interpretations

DIGIT END	FEMALE	MALE
0	86.58	94.67
	(Highly Accurate)	(Highly Accurate)
5	104.09	96.98
	(Highly Accurate)	(Highly Accurate)

Table 2: Whipple's indices and their corresponding interpretation

B. *Myer's Blended Index*

Myer's blended index of digit preference is used for evaluating single-year age-sex data by giving the extend of digit preference for all the digits 0, 1, 2, … , 9. The computation for Myer's index

Terminal Digits	Sum of ages 10 - 69	coefficients	Product	Sum of ages 20 - 69	coefficient
0	11951	1	11951	8682	9
1	4477	2	8954	2734	8
2	7288	3	21864	4616	7
3	5681	4	22724	3409	6
4	4208	5	21040	2696	5
5	10048	6	60288	7949	4
6	4453	7	31171	2850	3
7	4667	8	37336	3039	2
8	6184	9	55656	4071	1
9	3352	10	33520	2076	0

Table 3: (a) myer's blended index (female)

Terminal Digits	Product	Blended Sum	% Distribution	Deviation from 10	Remark
0	78138	90089	17.28	7.28	Preference
1	21872	30826	5.91	-4.09	Avoidance
2	32312	54176	10.39	0.39	Preference
3	20454	43178	8.28	-1.72	Avoidance
4	13480	34520	6.62	-3.38	Avoidance
5	31796	92084	17.67	7.67	Preference
6	8550	39721	7.62	-2.38	Avoidance
7	6078	43414	8.33	-1.67	Avoidance
8	4071	59727	11.46	1.46	Preference
9	0	33520	6.43	-3.57	Avoidance
		521255	100.00		

Table 4: myer's blended index (female) (….contd)

The result (from table III) revealed that there is over selection of ages ending with digits '0' and '5' with the respective preferences of 17.28% and 17.67%. However, the ages ending with '1' have the highest avoidance, followed by ages ending with '9' and '4'.

Terminal Digits	Sum of ages 10 – 69	coefficients	Product	Sum of ages 20 - 69	coefficient
0	11093	1	11093	7885	9
1	4250	2	8500	2431	8
2	6843	3	20529	4195	7
3	5159	4	20636	2999	6
4	4270	5	21350	2445	5
5	9127	6	54762	6909	4
6	4319	7	30233	2806	3
7	4367	8	34936	2805	2
8	5448	9	49032	3554	1
9	2827	10	28270	1826	0

Table 5: myer's blended index (male)

Terminal Digits	Product	Blended Sum	% Distribution	Deviation from 10	
0	70965	82058	17.29	7.29	Preference
1	19448	27948	5.89	-4.11	Avoidance
2	29365	49894	10.51	0.51	Preference
3	17994	38630	8.14	-1.86	Avoidance
4	12225	33575	7.08	-2.92	Avoidance
5	27636	82398	17.36	7.36	Preference
6	8418	38651	8.14	-1.86	Avoidance
7	5610	40546	8.54	-1.46	Avoidance
8	3554	52586	11.08	1.08	Preference
9	0	28270	5.96	-4.04	Avoidance
		474556	100.00		

Table 6: MYER'S BLENDED INDEX (MALE) (…..contd)

Table IV shows the computation of male population from Nigeria DHS, 2018. The result has also shown the over selection of ages ending with digits '0' and '5' with the respective preferences of 17.29% and 17.36%. Similarly, the ages ending with '1' have the highest avoidance, followed by ages ending with '9'.

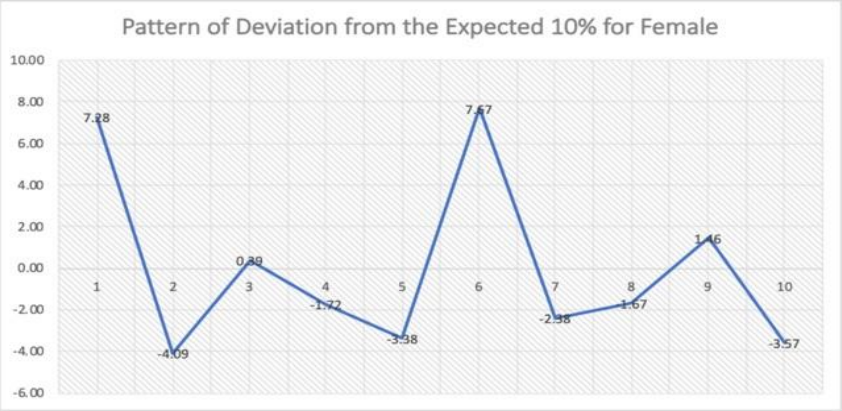

Fig. 1: Pattern of Deviation from the Expected 10% for Female

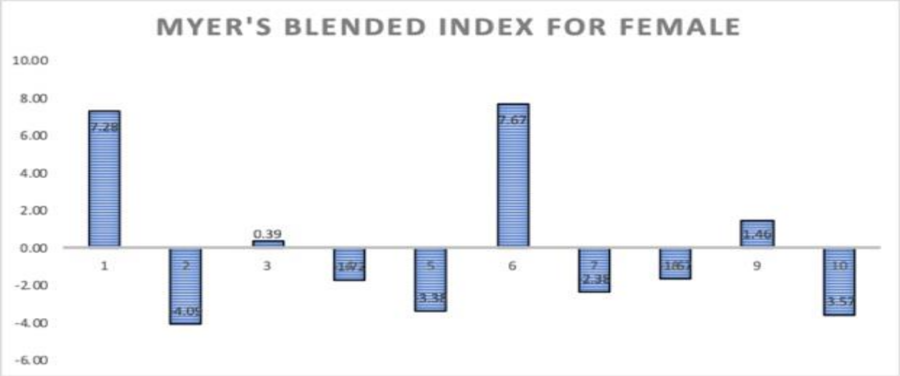

Fig. 2: Myer's Indices for Female

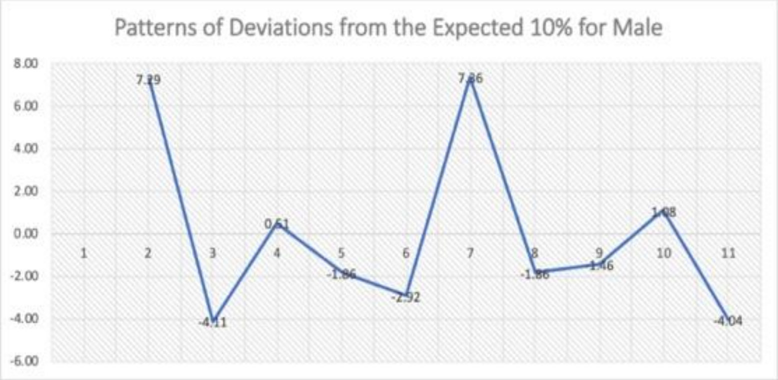

Fig. 3: Pattern of Deviation from the Expected 10% for Male

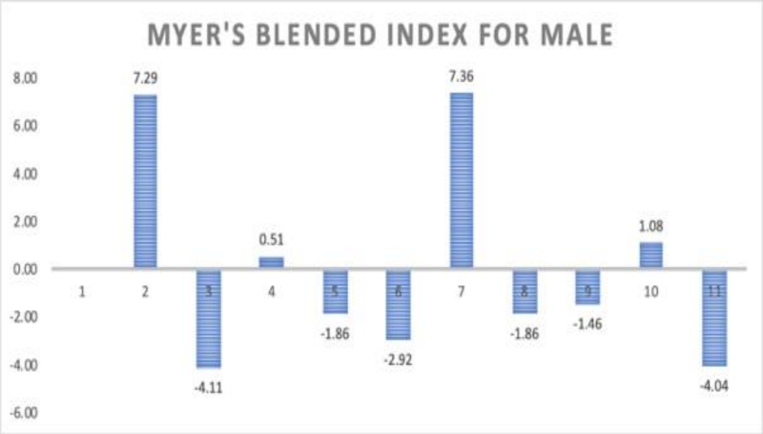

Fig. 4: Myer's Indices for Male

Figure 1 - 4 describe the deviations of the percentage of the blended population from 10 among each of the terminal digits. The most preferred terminal digits while reporting ages were '0' and '5' for both male and female population.

C. Age-Sex Accuracy Index (Joint Score)

1	**2**	**3**	**4**
Age (in 5 yrs)	**Female**	**Age Ratio**	**Deviation from 100**
0 - 4	15647		
5 - 9	15407	113.64	13.64
10 - 14	11468	95.07	-4.93
15 - 19	8719	93.67	-6.33
20 - 24	7149	86.57	-13.43
25 - 29	7798	115.07	15.07
30 - 34	6404	95.72	-4.28
35 - 39	5583	108.18	8.18
40 - 44	3918	88.31	-11.69
45 - 49	3290	93.81	-6.19
50 - 54	3096	114.20	14.20
55 - 59	2132	91.38	-8.62
60 - 64	1570	94.75	-5.25
65 - 69	1182		
Total (irrespective of sign)			**111.81**
Mean			**9.3175**
UN Joint Score			
5	**6**	**7**	**8**
Age (in 5 yrs)	**Male**	**Age Ratio**	**Deviation from 100**
0 - 4	16305		
5 - 9	15935	113.96	13.96
10 - 14	11660	96.67	-3.33
15 - 19	8188	96.90	-3.10
20 - 24	5240	77.12	-22.88
25 - 29	5401	100.95	0.95
30 - 34	5460	101.24	1.24
35 - 39	5385	109.80	9.80
40 - 44	4349	97.91	-2.09
45 - 49	3499	99.07	-0.93
50 - 54	2715	97.14	-2.86
55 - 59	2091	85.24	-14.76
60 - 64	2191	121.22	21.22
65 - 69	1524		
Total (irrespective of sign)			**97.12**
Mean			**8.09**
UN Joint Score			

9	10	11
Age (in 5 yrs)	**Sex ratio**	**First difference**
0 - 4	95.96	-0.72
5 - 9	96.69	-1.67
10 - 14	98.35	-8.13
15 - 19	106.49	-29.95
20 - 24	136.43	-7.95
25 - 29	144.38	27.09
30 - 34	117.29	13.61
35 - 39	103.68	13.59
40 - 44	90.09	-3.94
45 - 49	94.03	-20.01
50 - 54	114.03	12.07
55 - 59	101.96	30.30
60 - 64	71.66	-5.90
65 - 69	77.56	
Total (irrespective of sign)		**174.21**
Mean		**13.40**
UN Joint Score	**57.61**	

Table 7: Results of age ratios, sex ratios and joint score

The Joint Score for Nigeria is 57.61 The implication of this is that the data is deficient and requires massive adjustment.

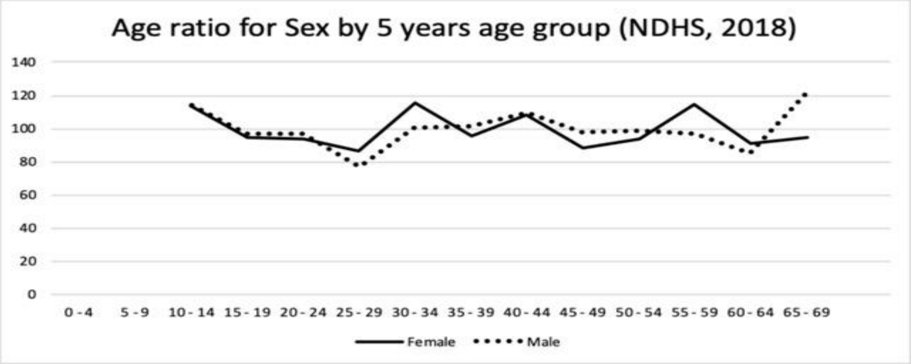

Fig. 5: Pattern of Age ratio

Figure 5 shows the patterns in the reporting of age ratio for sex by 5-years age group in Nigeria using the NDHS, 2018

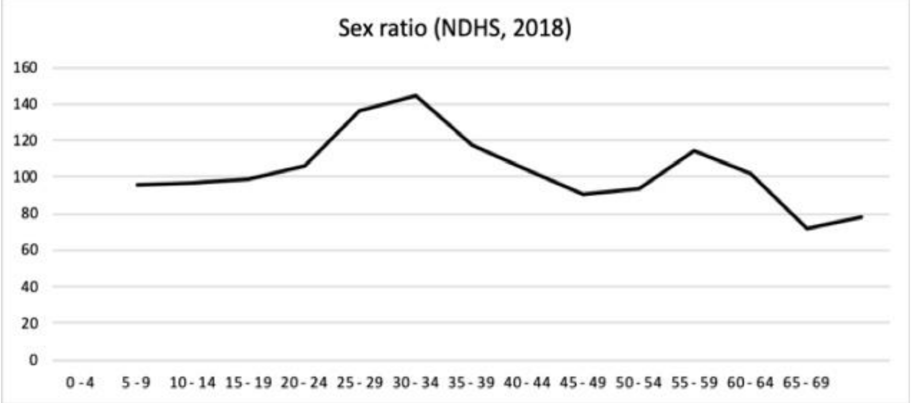

Fig. 6: Pattern of Sex ratio

Figure 6 shows the patterns in the reporting of sex ratio in Nigeria using the NDHS, 2018

IV. CONCLUSION

The age data collected from the Demographic and Health Survey, 2018 for Nigeria. The data for the separate age end-digit were highly accurate quality. There was age heaping at ages ending with terminal digits '0' and '5', indicating a preference in reporting such ages. The evaluation of the age data using the demographic techniques has finally qualified the data inaccurate as the results of systematic heaping and digit preference. The UN Joint Score revealed that the data from the data is deficient and requires massive adjustment before they could be meaningfully used, and any interpretation with the data should be with care and caution. It is hereby recommended that the data be adjusted to remove the irregularities

ACKNOWLEDGEMENT

All authors and contributors are duly acknowledged

REFERENCES

[1.] C.O. Lerche, Social and Economic Statistics for Africa, (2nd Edition), Longman group, London, 1983

[2.] G.M.K. Kpedekpo, Essentials of Demographic Analysis for Africa, Heinemann Educational Books Inc., New Hemisphere, 1982

[3.] G.S. Pardeshi, "Age heaping and accuracy of age data collected during a community survey in the Yavatmal district, Maharashtra," Indian J Community Med, 35(3), 391-395, 2010

[4.] B, Yusuf, "Error detection in outpatients' age data using demographic techniques," International Journal of Pure and Applied Sciences and Technology. ISSN 2229 – 6107, 2012

[5.] H. S. Shyrock and J. S. Siegel, with E. D. Stockwll, The Methods and Materials of Demography, Condensed Edition: 187, San Diego: Academic Press, 1976

[6.] Y. Aida, A.P. Mohamad and A. Aliraza, "Digit preference in Iranian age data", Italian Journal of Public Health. Volume 9, Number 1, 2012

[7.] J.F. Polly and I. Johannes, Introduction to Biostatistics, (2nd Edition) 1999

[8.] Demographic and Health Survey (2018)