African Population Studies Vol. 32, No. 1, 2018

Nigeria's recent population censuses: a Benford-theoretic evaluation

Nehemiah A. Ikoba, Emmanuel T. Jolayemi & Olusola O. M. Sanni

Department of Statistics, University of Ilorin,

Ilorin, Nigeria

Email: ikoba.na@unilorin.edu.ng

Abstract

Context: Population censuses in Nigeria have been plagued with under- or over-enumeration, as well as outright manipulation. This paper examines the claim of manipulated results of Nigeria's 1991 and 2006 population censuses.

Data Source & Method: Data on both censuses were obtained from the National Bureau of Statistics and analyzed via fitting Benford's probability distribution. The overall census data, as well as aggregate data for the six geopolitical zones of the country were examined to determine the level of conformity with Benford's distribution, using the Chi-square goodness of fit test.

Findings: The conformity analyses showed that the overall counts differed significantly from Benford's in both censuses. The North-West region had the highest deviation in both censuses, while the North-East and South-West had the lowest deviation in 1991 and 2006 censuses, respectively. Significant conformity was observed in the sizes of the local government areas and the population density for the 2006 census.

Conclusion: Some datasets with built-in minimum and maximum values may still conform to Benford's distribution provided the range of values of the first significant digit span digits 1 to 9. Census results should be scrutinized on the basis of Benford's distribution as an additional check on the quality.

Keywords: Benford's distribution; demography; population census; fraud-detection.

Introduction

Population censuses in Nigeria have been bedeviled with various errors of under- and over-enumeration. Because the results from any census have far-reaching political and socio-economic implications, there are inherent shortcomings that may significantly alter the quality of the results emanating from such exercises especially in developing countries like Nigeria.

While there are a number of techniques used to evaluate the quality of census results, there is the need to further evolve sufficiently simple measures to complement existing methods, and view authentic census results as emanating from a truly random process.

The aim of this study is to evaluate the quality of Nigeria's recent population censuses from the perspective of a random process via fitting Benford's distribution. The objective is to test the randomness of the last two national population censuses held in Nigeria using Benford's probability distribution of the first significant digits. The distribution of the first significant digits of the enumerated population of the local government areas for both the 1991 and 2006 national population censuses are to be compared with Benford's distribution. The geographical sizes of the 774 local government areas and their population density in the 2006 census will also be examined for conformity to Benford's distribution.

In order to examine the level of disparity in the distribution of the first significant digits of the census results within the six geo-political regions of the country, the aggregate results of each region is to be compared with the theoretical distribution in order to establish the region(s) contributing more to the residual of the overall census counts from the theoretical distribution.

Benford's distribution of first significant digits, d_1, as derived by Newcomb in 1881 (Hill, 1998) and Benford (1938) is given by

$$P(D_1 = d_1) = \log_{10}\left(1 + \frac{1}{d_1}\right), \quad d_1 = 1,2,\ldots,9 \qquad (1)$$

The mean, E(D) and variance, Var(D) of the distribution are given by

$$E(D) = \sum_{d=1}^{9} d \log_{10}\left(1 + \frac{1}{d}\right) = 3.44$$

$$Var(D) = \sum_{d=1}^{9} d^2 \log_{10}\left(1 + \frac{1}{d}\right) - \left(E(D)\right)^2 = 6.06$$

Unlike other probability distributions, the mean and variance of Benford's distribution are fixed. The mean is approximately 4, while the median is digit 3 and the mode is digit 1.

The first significant digit phenomenon was first observed by the astronomer and mathematician, Simon Newcomb in 1881 (Hill, 1998) and later by Benford (1938). A surprisingly diverse collection of empirical data obey Benford's law: tables of physical constants, numbers appearing on newspaper front pages, accounting data, scientific calculations, stock market closing figures, accounting, demography, etc (Hill, 1998; Swanson et al, 2003; Ley, 1996).

Benford's distribution is the only probability distribution that is scale-invariant and the only one that is base-invariant (Hill, 1998). The implication of the scale-invariant property of Benford's distribution is that if a dataset conforms to the distribution, any transformation of the dataset should also conform to the distribution (Hill, 1998).

The basic assumptions governing the use of Benford's distribution for any dataset are (Nigrini, 1999):

1. The numbers describe the sizes of similar phenomena (for example, market value of corporations).
2. The numbers do not contain a built-in maximum or minimum value.

The application of Benford's law in ascertaining the randomness of census returns or election results has been quite minimal. Nigrini (1999) established that the distribution of the first significant digits of the human population of 3141 counties in the 1990 United States census showed a good fit for Benford's distribution. Conformity analysis of the 2009 Albanian parliamentary elections results showed strong departure from Benford's law (Berdufi, 2013), with the strong claim that fraud took place in the areas of non-conformity.

Populations of the countries of the world have been shown to follow Benford's distribution (Olofsson, 2015). Human populations typically increase in a fairly steady rate and the change from one digit to the next requires an ever decreasing rate of population change, hence population sizes stay the longest in the lowest categories (digits 1 and 2) and shortest in the highest (digits 7, 8, 9) (Olofsson, 2015).

It is conjectured that the reason that Benford's law is applicable to so many datasets may simply be due to the fact that many popular parametric lifetime models closely follow the law for particular values of their parameters (Leemis et al., 2000). According to Olofsson (2015), any dataset that is large and in some sense irregular is likely to follow Benford's law.

For an extensive discussion of the mathematical properties of Benford distribution, Hill (1998), Rodriguez (2004) and Barrow (2011) provided excellent insights.

To be clear, Benford's law cannot deduce intention, it can only be used to detect unusual or unexpected data. These unusual or unexpected data may or may not have been an intentional product, but the technique of digital analysis is "blind" to the underlying intention and simply highlights possible irregularities (Hickman and Rice, 2010). Digital analysis could help support investigative efforts, but it is not a substitute for a thorough investigation.

Census taking in Nigeria has been a tense political activity mainly due to its perceived constitutional connections with revenue allocation and political representation (Obono and Omoluabi, 2014). Of the five censuses conducted in Nigeria since independence in 1960, two (1962 and 1973) were cancelled outright (Obono and Omoluabi, 2014). There had always been controversies trailing census exercises right from the colonial era till the last census of 2006 (Bamgbose, 2009).

The Local Government system in Nigeria has evolved over the years from the colonial era to the present system of 774 local government areas that has been in existence since 1996. The aim of establishing the local government system was to bring development closer to the grassroots and for ease of administration by the colonial authorities, and since 1971, they were made the third tier of government (Ukiwo, 2006).

The local government areas, in reality do not have limits as to their population sizes, providing sufficient justification for the application of Benford's distribution, as the basic assumptions for the application of the distribution are satisfied.

Data and methods

Census counts/enumerations of the 592 and 774 local government areas of Nigeria for the 1991 and 2006 census, respectively, were analyzed using both the raw counts and the first significant digits. Descriptive analysis using simple measures of location and spread like the mean, median, variance, etc, were used to provide useful insights into the census results and other auxiliary data. Digital analyses of the census counts in the local government areas for both the 1991 and 2006 censuses were done. In

addition, the geographical sizes of the 774 local government areas and the population density (per square kilometer) in the 2006 census were examined for conformity to Benford's distribution. The datasets were extracted from the Nigerian Bureau of Statistics Annual Abstract of Statistics (NBS, 2009). The goodness-of-fit test deployed to test conformity was the chi-square test.

The chi-square goodness-of-fit test compares the actual counts of the census data with the expected counts, which follows the hypothesized distribution (Benford's). The null hypothesis is that the first digits of the data follow Benford's distribution. The statistic is given as

$$X^2 = \sum_{i=1}^{9} \frac{(O_i - E_i)^2}{E_i} \qquad (2)$$

where O_i and E_i represent the observed count and the expected count, respectively of the ith digit.

The decision rule for the test is to reject the null hypothesis at the 5% level of significance (α) if the computed value of the statistic (X^2) is greater than the tabulated value ($\chi^2_{0.05,8} = 15.51$), or if the p-value is less than the level of significance, α. The p-value (or probability value) is the probability of getting a sample statistic in the direction of the alternative hypothesis when the null hypothesis is true. In other words, the p-value represents the probability of a particular sample statistic occurring if the null hypothesis is true (Bluman, 2012).

The Kolmogorov-Smirnov (K-S) test can equally be applied. This was done and identical results were obtained.

Results

Table 1 presents some descriptive statistics of Nigeria's population in both the 1991 and 2006 censuses, as well as the sizes of the 774 local government areas and the population density in the 2006 census. Table 2 provides the regional population totals for both censuses, as well as their inter-censal growth indices, while table 3 gives the digital analysis of the first significant digits (FSD) of the 1991 and 2006 censuses. Table 4 presents the comparison of the cumulative probabilities of the 1991 and 2006 censuses, as well as the 774 local government sizes in the 2006 census, with Benford's law.

Figure 1 presents the plots of the cumulative probabilities of both censuses in comparison with the corresponding Benford probabilities, while figures 2 and 3 show the cumulative probabilities of the six geo-political regions for the 1991 and 2006 censuses, respectively.

Table 1: Summary statistics of Nigeria's 1991 (592 LGAs), 2006 (774 LGAs) censuses, as well as local government sizes (in square kilometers) and population density (per square kilometer).

Statistic	1991 Census	2006 Census	LGA sizes (sq. km)	Population density (per sq. km)
Mean	150,110	181,405	1210.67	1047
Standard Error	4,049	3,666	51.93	140
Std deviation	98,517	101,993	1,444.66	3,884
Range	1,014,140	1,287,930	11,571.06	55,442
Minimum	21,081	31,641	8.71	9
Maximum	1,035,221	1,319,571	11,579.77	55,451
1st Quartile	91,221	120,853	290.84	98
2nd Quartile	129,280	157,794	731.47	219
3rd Quartile	174,897	212763	1,529.38	518

Table 2: Summary of the 1991 and 2006 census results (in millions) for the six regions of Nigeria as well as their geographical sizes and their inter-censal growth indices.

Region	1991 Census	% of population	2006 Census	% of population	Area ('000 sq. km)	growth rate	Population density	% Growth
North-Central	12.55	14.11	20.37	14.51	231.68	3.38	88	62.25
North-East	11.90	13.37	18.98	13.52	289.42	3.26	66	59.52
North-West	22.91	25.75	35.91	25.58	223.15	3.14	161	56.74

South-East	10.77	12.11	16.39	11.68	28.98	2.93	566	52.16
South-South	13.39	15.05	21.04	14.99	85.31	3.15	247	57.13
South-West	17.45	19.61	27.72	19.74	78.51	3.23	353	58.82
Northern Zones	47.37	53.23	75.27	53.60	744.25	3.23	101	58.90
Southern Zones	41.62	46.77	65.16	46.40	192.80	3.13	338	56.55
Overall	88.99	100.00	140.43	100.00	937.05	3.18	150	57.80

Table 3: Result of the Chi-Square tests for 1991 and 2006 Census data and the local government sizes and population density in the 2006 Census.

Year	Region	Rank	X^2	p-value
	North-East	1	16.976	0.0306
	South-West	2	25.538	0.0013
	South-East	3	27.403	0.0006
1991	North-Central	4	31.000	0.0001
	South-South	5	53.492	< 0.0001
	North-West	6	84.236	< 0.0001
	Overall		**188.608**	< 0.0001
	South-West	1	37.076	< 0.0001
	North-Central	2	42.711	< 0.0001
	North-East	3	44.366	< 0.0001
2006	South-South	4	62.012	< 0.0001
	South-East	5	86.092	< 0.0001
	North-West	6	116.157	< 0.0001
	Overall		**388.036**	< 0.0001
	LGA Sizes		**11.429**	0.18
	Population Density		**4.643**	0.79

Table 4: Comparison of the cumulative distribution of first significant digits for 1991, 2006 censuses and local government sizes, with Benford's law.

FSD	$C_p(1991)$	$C_p(2006)$	$C_p(LG\ Sizes)$	$C_p(Benford's)$
1	0.52	0.56	0.33.	0.30
2	0.66	0.78	0.51	0.48
3	0.69	0.84	0.61	0.60
4	0.73	0.86	0.71	0.70
5	0.76	0.88	0.78	0.78
6	0.81	0.90	0.84	0.85
7	0.86	0.93	0.89	0.90
8	0.94	0.97	0.94	0.95
9	1.00	1.00	1.00	1.00

Key: FSD – First Significant Digit; C_p – Cumulative probability

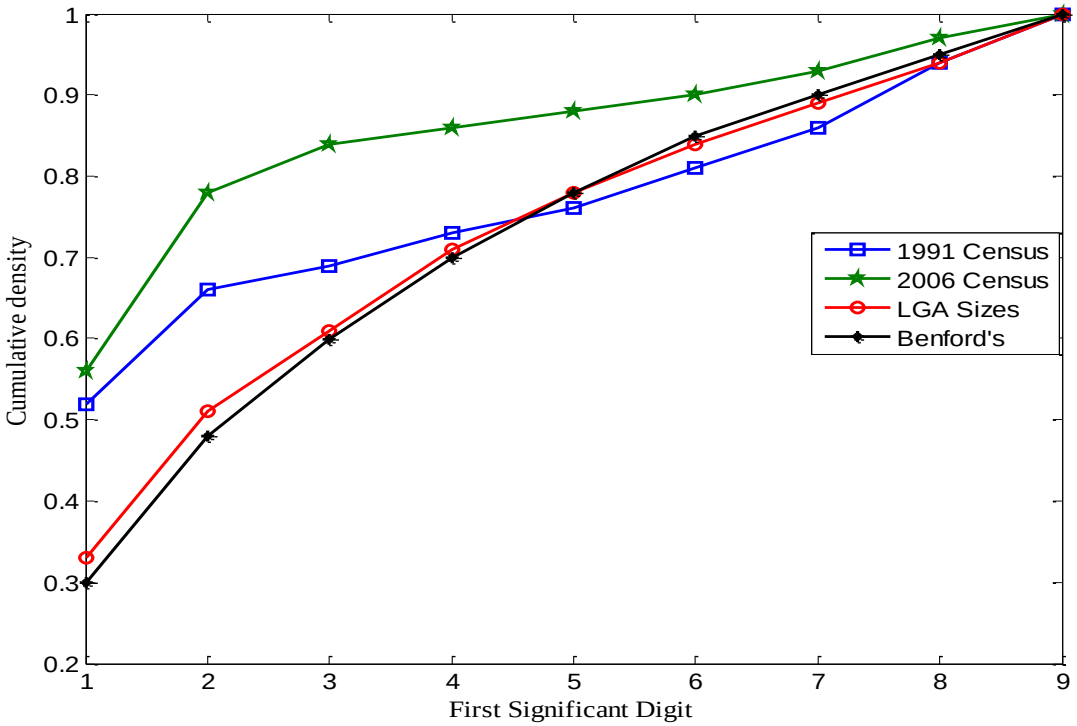

Figure 1: Plots of the cumulative probabilities of Benford's law and the population totals from 1991 and 2006 census, as well as the size of the local government areas for the 2006 census.

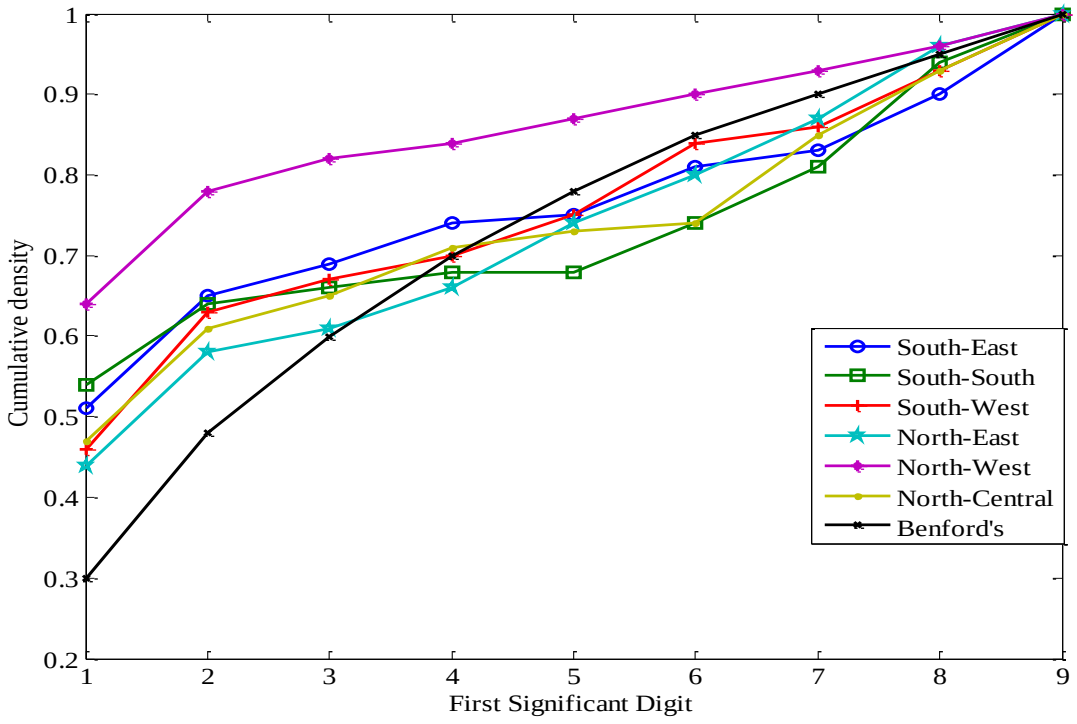

Figure 2: Plots of the cumulative probabilities of Benford's law and the zonal aggregates of the 1991 census.

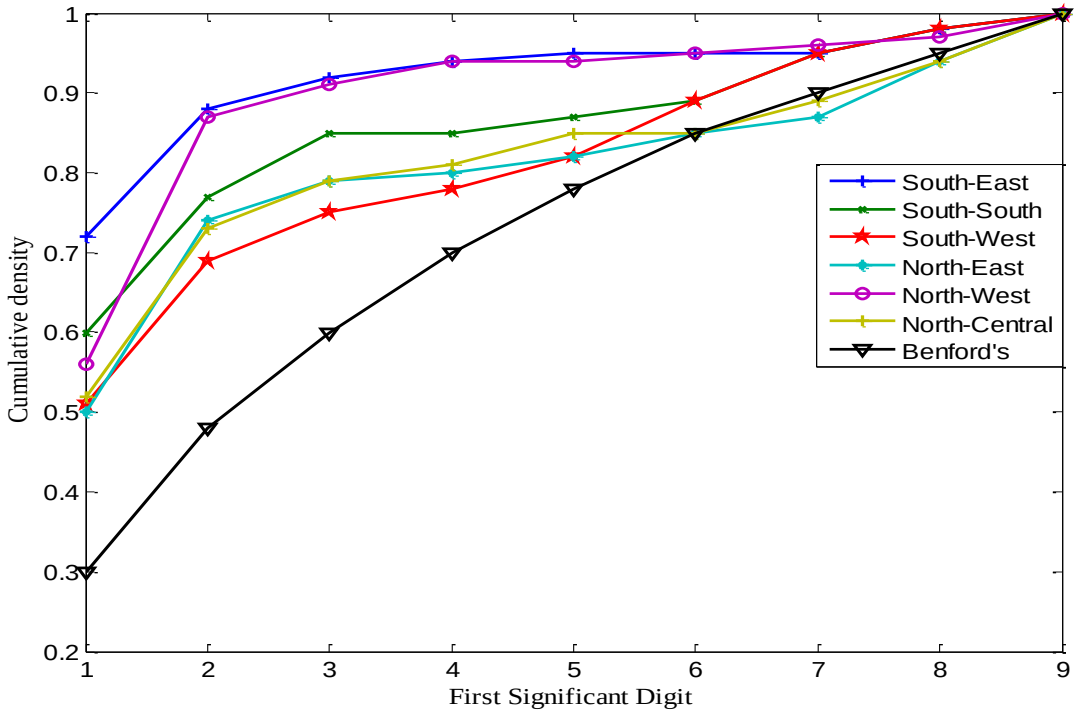

Figure 3: Plot of the cumulative probabilities of Benford's law and the zonal aggregates of the 2006 census.

Discussion

A cursory look at the summary statistics of the data, as presented in table 1 show that the range of values for both censuses were above 1 million, showing sufficient spread in the population of local government areas, spanning all the possible values of Benford's distribution (1-9). From the data in table 1, at least 75% of the first significant digits may be taken by the first two digits 1 and 2, in both censuses.

Table 2 provides great insights into the dynamics and spread of the population over the two censuses in the six regions of Nigeria. All the regions of the country exhibited similar growth rates and similar proportions in both the 1991 and 2006 census. In comparing the northern and southern parts of Nigeria, the northern part accounted for about 53% of the overall population of Nigeria, while the South had about 47% in both censuses.

The regional growth indices provided in table 2 may offer the justification for greater scrutiny of the result of the 2006 census. One of the implications of the growth indices is that the North-Central region exhibited a greater growth than the South-West (which contains Lagos state, the economic hub of the country and Ogun state). Although it may be argued that the North-Central also has the Federal Capital Territory which exhibited the greatest percentage growth (278%) from its 1991 population of about 200,000, as a justification for its growth, there is the need for greater scrutiny of the results. The North-Central and North-East regions, on the basis of the growth information provided in table 2, should be closely scrutinized for possible irregularities in the census results. Similarly, the results of the South-East region could also be closely scrutinized, as these three regions were at the extremes in terms of the growth from the earlier census.

It is also noted that the northern part of Nigeria accounts for about 80% of the land mass of the country, leaving about 20% for the southern part of the country. This disproportionate distribution of the landmass of the country is however not reflected in the population distribution as the north is only marginally more populous than the south in both censuses. Therefore, the southern region of Nigeria has greater population density on the basis of the 2006 census.

It could be seen in table 3 that the sizes of the local government areas (LGAs) and the population density follow Benford's distribution, while the digital analysis of both the 1991 and 2006 census results showed non-conformity with Benford's distribution. A possible reason for this close conformity with Benford's law for the LGA sizes is that the data are factual and truly random to a great extent. As a consequence of the first significant digits of the LGA sizes being Benford distributed, the population density, which is the ratio of the population in the LGA to the size, was also found to conform to Benford's distribution. The North-East region produced the

lowest residual in the 1991 census, while the South-West region had the lowest value in 2006. The other regions in both censuses exhibited great departure from Benford's distribution, as captured by table 3.

When the result of the chi-square tests for the 1991 census data is compared with that of the 2006 census data (table 3), it would seem that there had been a massive shift away from the expected distribution, as the computed X^2 value doubled. The fact that the number of local government areas increased from 592 to 774 may have also contributed to the large value of the test statistic due to the greater variation within the LGAs.

Upon closer scrutiny of the distribution of first significant digits for the 1991 census data, it is seen that digits 6, 7, 8 and 9 had proportions that were very close to that of Benford's distribution but digit 1 exhibited a marked departure from Benford's, which in turn, impacted on the proportions for digits 2, 3, 4 and 5.

Similarly for the 2006 census data, digits 8 and 9 had proportions very close to the corresponding Benford's proportion. With this data, the distribution of the first significant digits was heavily skewed towards 1 and 2, as the two digits accounted for almost 80% of the overall counts.

The fact that the average local government area's population for the 1991 and 2006 censuses were between 100,000 and 200,000 do not indicate that the data will not conform to Benford's distribution. The average or the mean of a dataset is a measure of central tendency and indicates the central value of the data. However, in the case of digital analysis, the interest is not on the actual value of the observations, but on the first significant digits, whose distribution differs from the original data. However, inference could be made about the distribution on the basis of the mean of the distribution of the first significant digits. When this mean is significantly far away from digit 3, the data may not conform to Benford's distribution. The mean LGA size was about 1,211km^2, but the first significant digits of the local government sizes conformed to Benford's distribution. Also, in comparison with the United States 1990 census, the mean county population was 79,182 (having first significant digit 7) but the digit 7 only accounted for 5.51% of the data and the data conformed to Benford's distribution. Hence, the average of a dataset is not an indicator of non-conformity of the dataset to Benford's distribution.

Conclusion

From the evidence of the census data used for this study, it appears that data prone to manipulation may have far too many numbers beginning with 1 and far too few numbers beginning with the digits 4, 5, and 6.

It is conjectured that there may have been possible over-enumeration in the census counts, as well as possible inflation of the results for supposed political benefits. These shortcomings may have been part of the reason for non-conformity of the 1991 and 2006 census data with Benford's distribution.

The mean is not necessarily an indicator of non-conformity of a dataset to Benford's distribution, rather it is the mean of the embedded distribution of the first significant digits of the dataset that could provide information about the conformity. If this mean is significantly different from 3, then the dataset is most likely not to conform to Benford's distribution.

Benford's law could still be applicable to data with an in-built maximum and minimum value but whose range of values span first significant digits 1-9, as shown in the conformity analysis of the local government sizes and the population density of the 2006 census.

Fitting Benford's distribution on population census data may serve as an additional quality control tool to assert the level of authenticity of census results. It is therefore recommended that tests for conformity of census results with Benford's distribution be carried out at all the stages of the enumeration process, in combination with other evaluation methods, as a way of enhancing the quality of the census results.

An area of further study could be the formulation of simple tests of hypotheses relating to the mean, median and variance of Benford's distribution as a means of establishing conformity.

References

Bamgbose, J. A. 2009. Falsification of Population Census Data in a Heterogeneous Nigerian State: The Fourth Republic Example. African Journal of Political Science & International Relations, vol. 3, No. 8, 311-319.

Barrow, J. 2011. Benford's Very Strange law. Lecture presented at Gresham University.

Accessed at: www.gresham.ac.uk/lectures-and-events/benford's-very-strange-law.

Benford, F. 1938. The Law of Anomalous Numbers. Proceedings of the American

Philosophical Society, vol. 78, No. 4, 551-572.

Berdufi, D. 2013. Statistical Detection of Vote Count Fraud: 2009 Albanian Parliamentary Election and Benford's Law. Academic Journal of Interdisciplinary Studies, vol. 2, No. 8, 379-396.

Bluman, A. G. 2012. Elementary Statistics: A Step by Step Approach. McGraw-Hill,

New York.

Hickman, M. J. and Rice, S. K. 2010. Digital Analysis of Crime Statistics: Does Crime Conform to Benford's Law? Journal of Quantitative Criminology, 26 (3), 333-349.

Hill, T. P. 1998. The First Digit Phenomenon. American Scientist, vol. 86. No. 4, 358-363.

Leemis, L. M.; Shmeiser, B. W.; and Evans, D. L. 2000.Survival Distributions Satisfying Benford's Law. The American Statistician, vol. 54, No. 4, 236-241.

Ley, E. 1996. On The Peculiar Distribution of The U.S. Stock Indexes Digits. The American Statistician, vol. 50, No. 4, 311-313.

National Bureau of Statistics 2009. Annual Abstract of Statistics, Published by National Bureau of Statistics (NBS), Nigeria.

Nigrini, M. J. 1999. I've Got Your Number. Journal of Accountancy, 187 (5), 79-83.

Obono, O. and Omoluabi, E. 2014.Technical and Political Aspects of the 2006 Nigerian Population and Housing Census. African Population Studies, 27 (2), 249-262.

Olofsson, P. 2015. Probabilities, The Little Numbers that Rule our Lives. New Jersey: John Wiley and Sons Inc.

Rodriguez, R. J, 2004. First Significant Digit Patterns From Mixtures of Uniform Distributions. The American Statistician, vol. 58, No. 1, 64-71.

Swanson, D.; Cho, M. J. and Eltinge, J. 2003. Detecting Fraudulent or Error-Prone Survey Data Using Benford's Law. Proceedings from 2003 Joint Statistical Meetings – Section on Survey Research Method, 4172-4177.

Ukiwo, U. 2006. Creation of Local Government Areas and Ethnic Conflicts in Nigeria: The Case of Warri, Delta State. Paper Presented at CRISE West Africa Workshop, Accra, Ghana, March 2006

.