A Measure on "Digit Non-Heaping"

Barun Kumar Mukhopadhyay

Abstract

Objective of my paper is to focus on two research contributions one presented at the Social Statistics Section, American Statistical Association in 1958, and the current work proposed to be presented at the Annual Meeting of the Population Association of America in 2015. Former one purported hypothetical ranking of ten digits of ages, 0-9, when there was data deficiency 70 years back irrespective of residence. Developed countries improved over time whereas developing world remains almost same. An alternative approach only for the former countries uses ranking of ten digits from life table population and compares it with the same for raw data from censuses/surveys. Analysis reveals almost a one to one matching between the two rankings—life table and actual data, using Spearman rank correlation coefficient—which proclaims a new nomenclature, "digit non-heaping". The earlier vis-a-vis proposed work, if cited in literature on population analysis, students from developing countries may be benefited.

Keywords: "digit non-heaping", digits, life table, Spearman rank correlation coefficient

A Measure on "Digit Non-Heaping"

Barun Kumar Mukhopadhyay

Introduction

Turner presented a paper at the Social Statistics Section, American Statistical Association in 1958. He proposed a hypothesis that people reporting their age (numerical data) with highest preferences in order of ten, five and two as ending digits in censuses in western countries in his time. Out of the remaining odd digits, one, three, seven and nine as one and nine were between the most attracted digit '0', hence they were at a most disadvantaged position compared to three and seven surrounding even digits, '2', '4', '6' and '8'. He then had shown graphically the pattern of heaping for all the digits. He precisely gave a table of hypothetical ranking of all the ten digits from '0' to '9' according to his observation above. He then computed the rank of each ending digit of age and compared the observed rank order to the expected rank from his hypothesis using Spearman rank difference correlation coefficient (ρ). His methodology has been cited in many literatures (e.g., Stockwell, 1966 ; Stockwell and Wicks, 1974 ; Ewbank, 1981; and Jowett and Li, 2004, Mukhopadhyay, 2006).

As the quality of age data in censuses of developed countries improved significantly after a long period of almost 7 decades the methodology proposed by him might not be so suitable in the present time. Nevertheless, his technique is quite legitimate in the developing regions still with inferior quality of age reporting having much digit preferences. It is also true that as there is Myers' technique, (1940) readily available especially in software packages on population analysis, nobody bothers about other methods (King, 1915; Bachi, 1951; Carrier, 1959; Ramachandran, 1965; Mukhopadhyay, 1987, 1986 and etc.). An alternative to Turner uses ranking of ten digits from life table population and compares it with the same for raw data from censuses/surveys has been proposed in the present paper.

The technique of ranking is based on the life table stationary populations according to ten digits from '0' to'9' within a broad age band say, 10-69 depleted only by mortality with a fixed birth cohort being taken an ideal age distribution monotonically declining, i.e., $\sum P_i > \sum P_{i+1}$ without any digit preference error. The rankings of the raw single year population figures for the developed countries have been done according to the ten digits '0' to '9' within 10 to 69 years of age usually there being very little international net migrations and having no current effect of fertility in the middle ages, vulnerable for heaping. And then the Spearman rank difference correlation coefficient (ρ) has been calculated as done by Turner. In the

present paper, the new methodology based on life table comparison has been applied only on the population of developed countries. In order to test the methodology a few countries has been taken as a case study. However, only the country India has been chosen to represent the developing world where Turner's method has been applied to solidify his technique still is valid for data with much digit preferences. As the new technique is not at all valid for them hence is avoided in the paper the aim of which is to prescribe the new technique which is an alternative to Turner only for the developed world.

Analysis

The broad age band, 10-69 years is arbitrarily taken as middle age band 23-62 taken by Whipple (1920) to consider forty points for his index of concentration at digits '0' and '5'. Moreover digit heaping is usually studied in the middle ages of the population. However, the extreme two age bands, 0-9 years and 70+ years, early young and older population respectively possess different kind of error which are usually avoided to measure the quantity of error of digit preferences.

The general format of the final procedure to find out the comparison between the observed and the hypothetical (Turner) ranking is made using Spearman's rank difference correlation coefficient (ρ). It may be expected that the ρ value for the highly erroneous data pertaining to the illiterate masses of developing countries would be highly significant (p value < 0.00).

Table 1: Format showing the rankings of 10 digits of ages

Digit	Population of age band 10-69		Observed ranking		Turner's ranking	
	Male	Female	Male	Female	Male	Female
0	X	Y	Z	P	1	1
1	X	Y	Z	P	9.5	9.5
2	X	Y	Z	P	3.5	3.5
3	X	Y	Z	P	7.5	7.5
4	X	Y	Z	P	4.5	4.5
5	X	Y	Z	P	2	2
6	X	Y	Z	P	4.5	4.5
7	X	Y	Z	P	7.5	7.5
8	X	Y	Z	P	3.5	3.5
9	X	Y	Z	P	9.5	9.5

The similar format is also required for the ranking of the present paper where life table stationary population is used to find out the comparison between the observed and the life table ranking is made using Spearman's rank difference correlation coefficient (ρ). It may be expected that the ρ value for the highly good quality data pertaining to the developed countries would also be highly significant (p value < 0.00) that is what is exactly given a new nomenclature of "digit non-heaping". Further in so far as index

of “Non Heaping” in age data among developed countries may also be compared using the new ‘ρ’ values calculated among any desired group under study.

Table 2: Format showing the rankings of 10 digits of ages

Digit	Population of age band 10-69		Observed ranking		New ranking	
	Male	Female	Male	Female	Male	Female
0	X	Y	Z	P	1	1
1	X	Y	Z	P	2	2
2	X	Y	Z	P	3	3
3	X	Y	Z	P	4	4
4	X	Y	Z	P	5	5
5	X	Y	Z	P	6	6
6	X	Y	Z	P	7	7
7	X	Y	Z	P	8	8
8	X	Y	Z	P	9	9
9	X	Y	Z	P	10	10

Application of the Turner’s method in India

The method of Turner has been applied in India in the following table. For this the data of demographic yearbook obtained from the online network service as a representative for the developing countries are arranged as par Turner.

Table 3 : Turner’s methodology in Indian census data

Digit	Population(10-69)		Turner’s ranking	Observed rank	
	Male	Female		Male	Female
0	**84322279**	**77737611**	**1.0**	**1**	**1**
1	**15319710**	**12669114**	**9.5**	**8**	**8**
2	**36490822**	**33137339**	**3.5**	**3**	**3**
3	**16887032**	**16007060**	**7.5**	**7**	**7**
4	**18514147**	**17402879**	**5.5**	**6**	**6**
5	**67070638**	**56131229**	**2.0**	**2**	**2**
6	**20739446**	**18490575**	**5.5**	**5**	**5**
7	**13689233**	**11787993**	**7.5**	**9**	**9**
8	**29974746**	**30799679**	**3.5**	**4**	**4**
9	**9251836**	**8202481**	**9.5**	**10**	**10**

Turner’s ρ = 0.96 (M), 0.96 (F).

The Spearman rank-difference correlation coefficient, ρ is found highly significant (p value < 0.00). It shows Turner’s method is truly applicable as the raw ranking with huge peaks and troughs at ‘0’ and ‘5’ and even digits and odd digits respectively tally with the ranking proposed by him in his paper.

The application of proposed new methodology for six developed countries

Before explaining the new methodology, Turner's ranking method is done for a few developed areas where less heaping is commonly found. And age pattern almost follows a progressively declining series. The values are 0.16 (Male) and 0.14 (Female) for France, 0.28 and 0.28 for Ireland, 0.22 and 0.38 for Chile, and 0.24 and 0.27 for Canada. It is clear that there is almost no one to matching of ranking between Turner and the three developed countries. It simply shows that the quality of age reporting in advanced countries has improved much over the 7 decades or so. Now, thinking is that, life table population by assuming there is no effect of fertility in the middle ages, vulnerable for heaping depleted by only death could be matched with the raw single year data of developed countries.

The application of proposed new methodology for six developed countries has been described in the present section. The proposed methodology already mentioned earlier has been applied taking data from the list obtained from the demographic year book online network service as are done similarly for India.

Table 4: Sex-wise new ρs for the developed countries

Developed Countries	New ρ	
	Male	Female
France	0.89	0.82
New Zealand	0.87	0.77
Ireland	1.00	1.00
Canada	0.99	0.98
Australia	0.98	0.98
Chile	0.89	0.82

The ranking of the raw single year population figures for the developed countries have been done according to the ten digits also progressively declining as that of life table age distribution with some variations among them within the broad age band of 10 to 69 years usually there being very little international net migrations and being no current effect of fertility in the middle ages, vulnerable for heaping. And then the Spearman rank difference correlation coefficients (ρ) have been calculated (Table 5) above between the two rankings, separately for male and female populations and found statistically highly significant ($p < 0.00$).

Discussion

Data quality was discussed in the entire countries of the world about 7 decades back as some paper was presented at the Social Statistics Section, American Statistical Association in 1958 by Turner S.H who proposed some hypothetical ranking of ten digits of ages showing huge preferences and disliking of ten

digits. After that long period the same hypothetical ranking is almost entirely null and void in the developed countries improving significantly their age reporting whereas developing world remains almost same as before when his method is quite applicable to them. This was tested by the Indian census data on age having huge peaks and troughs at the preferred digits and disliking digits corresponding to the hypothetical ranking. And the Spearman rank difference correlation coefficient (ρ) is observed to be 0.96 (M) and 0.96 (F) with high level of significance ($p < 0.00$). A similar test was done for three developed countries but with very less value, say for France (0.16 (Male) and 0.14 (Female)) with insignificant ρ values. An alternative ranking in the present paper replacing the Turner's one is based on the life table stationary populations according to ten digits from '0' to'9' depleted only by mortality with a fixed amount of birth (cohort) being taken an ideal age distribution within a middle broad age band (10 to 69) monotonically declining according to ten digits. That is in this case the hypothetical ranking is entirely different from the one proposed by Turner. As the quality in reporting of age improved a lot in the developed regions, the pattern almost same as that of life table indicating almost a one-to-one correspondence. As a result this concept is applied in the present paper as to how these two rankings could be used as a tool to find out the "Digit Non Heaping" as a new nomenclature to measure the high standard of quality of age reporting among the developed countries. Comparison of the observed ranking with the proposed new one has been done using Spearman rank difference correlation coefficient (ρ). It gave the digit accuracy of the country rather than the digit preferences as found in developing countries. Hence the new name, "Digit Non-Heaping" in the developed countries applies. Further in so far as development in terms of "Non Heaping" in age data are concerned in developed countries, a comparison may also be done using the ρ values among, say, most developed western European countries versus the less developed east European countries. The earlier vis-a-vis proposed work, if cited in literature on population analysis, students, researchers may be benefited.

References

BACHI, R. (1951). The Tendency to Round off Age Returns, Measurement and Correction , ***Bulletin of the International Statistical Institute***, 33, part IV**.**

CARRIER, N.H. (1959) . A Note on the Measurement of Digital Preference in Age Recordings, ***Journal of the Institute of Actuaries*** , Oxford.

EWBANK, D.C.(1981). Age Misreporting and Age-selective Underenumeration: Sources, Patterns and Consequences for Demographic Analysis, Report 4, Committee on Population and Demography, ***United States National Academy of Sciences***, Washington.

KING, G. (1915). The New National Life Tables, ***Journal of the Institute of Actuaries***, Vol.49, Oxford.

JOWETT, A. JOHN and LI, YUAN-QING (1992). Age – Heaping: Contrasting Patterns from China, ***Geojournal***, Vol. 28 no. 4.

MUKHOPADHYAY, B.K. (2014). Shortcoming in some Index: Improvisation through Mathematical Modeling, presented at ***SMTDA International Conference***, Lisbon, Portugal, 11-16 June 2014.

MUKHOPADHYAY, B.K. AND MUKHERJEE , B.N. (1988). A study of Digit Preference and Quality of Age Data in Turkish Censuses , ***Genus***, Vol XLIV-n.1-2, Italy.

RAMACHANDRAN, K.V. (1965). An Index to Measure of Digit Preference Error in Age Data , ***Proceedings of the UN's World Population Conference (Summary)***, Belgrade, Yugoslavia.

STOCKWELL, E.G. and WICKS, J.W.(1974). Age Heaping in recent National Censuses, ***Social Biology***, 21:2, pp163-167.

STOCKWELL, E.G. (1966). Patterns of Digit Preference and Avoidance in the Age Statistics of Some Recent National Censuses: a Test of the Turner Hypothesis, ***Eugenics Quarterly***, 13:3.

TURNER, S.H. (1958). Patterns of Heaping in the Reporting of Numerical Data, ***Proceedings of the Social Statistics Section (American Statistical Association)***.