PATTERNS OF HEAPING IN THE REPORTING OF NUMERICAL DATA

By Stanley H. Turner, University of Pennsylvania

When a person is asked to volunteer numerical data, he either gives a precise answer or he does not. That is, he bases his response either on precise information or a somewhat hazy estimate. This paper tries to explain a certain type of pattern that emerges from data based on numerical estimates.

The central idea of this paper is that this pattern that emerges is related to the number system used by the estimator. To put it simply: the way we count, influences the way we estimate. That is, when a person estimates, he should do so in convenient units provided for him by the number system. Specifically, he should tend to over-report digits which are multiples of the divisors of the base of the number system and under-report digits which are not multiples of the divisors of the base of the number system.

As an example, consider the reporting of age. We may be unsure of our age or we may be asked to estimate the ages of other persons. Which digits are we more likely to report? This paper is concerned only with the ending digit of age since it is assumed that the decade of age is known accurately. Therefore, which ending digits of age are we more likely to report?

The most familiar way of counting is with the base ten. The divisors of the base ten are ten, five and two. The hypothesis states that estimates should heap at multiples of these three divisors; but more than that, it states that the most heaping should occur at ages ending in multiples of ten, the next largest at multiples of five, and the next largest at multiples of two. Figure I shows the rank order of heaping for part of the ending digits of age.

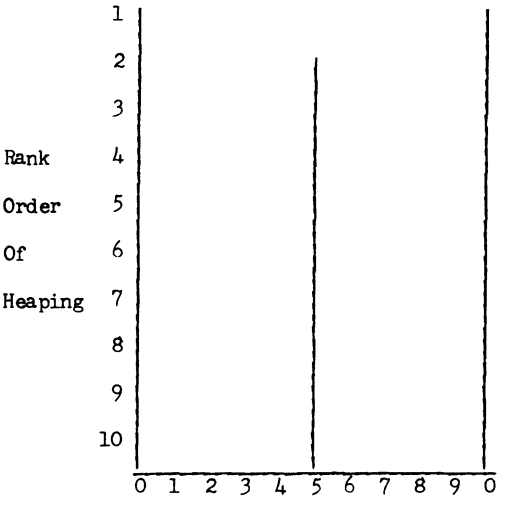

Figure I

Thus far, ages ending in multiples of ten, that is, ages ending in zero, are supposed to receive the most heaping. Ages ending in five, the next largest divisor, are supposed to receive the next largest amount of heaping. Only a single zero is needed in Figure I, but two are shown for symmetry and clarity.

Multiples of the next largest divisor, two, should all come next. That is, ages ending in two, four, six and eight should all receive the next largest amount of heaping. But notice that four and six are right next to five, which is supposed to attract a good deal of heaping. Furthermore, the other two even digits, two and eight, are not next to either zero or five. Therefore, two and eight should be free to attract more heaping than four or six. This line of reasoning implies the following additions to the expected pattern of heaping:

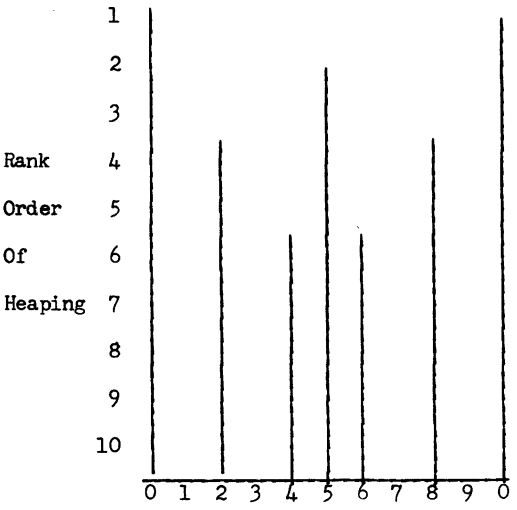

Figure II

All that remains is to fit in the remaining odd digits - one, three, seven and nine. Notice that the digits one and nine are between digits which are ranked as attracting a large amount of heaping. This should put one and nine at a disadvantage compared to three and seven. This enables the ranking to be completed as follows:

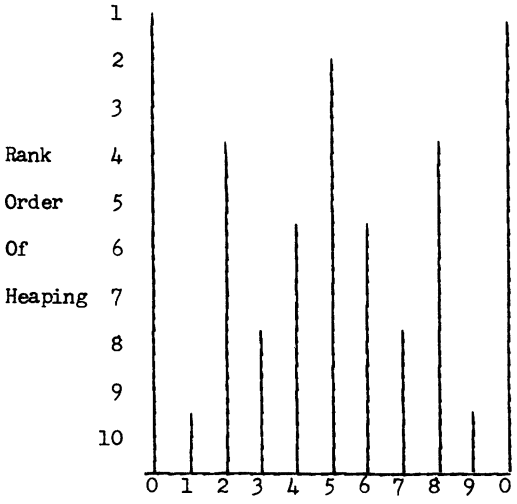

Figure III

From Figure III, the complete rank order prediction for all ending digits of age can be made:

Ending Digit Of Age	Predicted Rank Order
0	1.0
1	9.5
2	3.5
3	7.5
4	5.5
5	2.0
6	5.5
7	7.5
8	3.5
9	9.5

Eight decennial censuses of the United States population, covering the period from 1880 to 1950 were used to test the above expected rank orders. The number of people whose ages ended in each digit in each census was determined. However, a correction suggested by R. J. Myers was needed. If heaping were estimated by adding together all people whose ages ended in one, two, three, etc., a bias would be introduced. Consider the group of people whose ages end in one and those whose ages end in two. The former group is younger and therefore usually more numerous. That is to say, the sum of those aged 10 + 20 + 30 + 40 + 50 + 60 + 70 + 80 + 90 is usually greater than the sum of those age 11 + 21 + 31 + 41 + 51 + 61 + 71 + 81 + 91. In general, starting with any age tends to overstate the heaping at that age. Myers suggested that this bias could be removed by starting at each digit in turn and averaging the results.[1]

Table 1. shows the rank order of heaping for all ending digits of age during the entire period from 1880 to 1950.

Table 1.

RANK ORDER OF HEAPING FOR ENDING DIGITS OF AGE, BOTH SEXES, U. S. CENSUSES FROM 1880 - 1950*

Digit	1880	1890	1900	1910	1920	1930	1940	1950
0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
2	4.5	4.0	4.0	4.0	4.0	4.0	4.0	3.0
3	7.0	6.0	8.5	8.0	8.0	8.0	9.0	8.5
4	6.0	7.0	6.0	6.5	8.0	6.5	7.0	6.5
5	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
6	4.5	5.0	7.0	5.0	5.0	6.5	6.0	6.5
7	8.0	8.0	8.5	9.0	8.0	9.0	8.0	8.5
8	3.0	3.0	3.0	3.0	3.0	3.0	3.0	4.0
9	9.0	9.0	5.0	6.5	6.0	5.0	5.0	5.0

*Columns 1880 - 1930 are based on data taken from Robert J. Myers (See Bibliography)

Table I shows that zero was the most frequently reported ending digit of age during the entire period. Ending digit five was next most frequently reported and digit one was the least reported of all the ending digits of age. The other digits changed their rank orders somewhat. It is interesting to note that the digit nine was the least stable in its rank order. It ranked ninth in 1880 and 1890 and then rose to fifth rank in 1900. None of the other digits displayed such variability.

It might be helpful to analyze these data separately for males and females since the sexes are known, or at least reputed, to differ in their willingness to report their ages. Indeed, the censuses do provide a breakdown of age reporting for males and females. But such figures can be easily misinterpreted. Remember that the census enumerator does not ask each and every person his or her own age. Rather one person is commonly asked to report the ages of perhaps several other persons. Since females may be reasonably expected to be home more frequently when the enumerator calls, then many of the figures listed in the census as male's ages are actually reported or estimated by females. In view of this observation, the decision was made not to analyze the ages of each sex separately.

Instead, an average rank order of each of the ending digits of age was computed. This was done in order to compare the observed rank order derived from the census data to the expected rank order derived from the hypothesis.

The results are shown in Table 2. The difference between the expected and the observed rank orders is small except for the digit nine. A statistical test showed that the overall pattern of heaping conformed quite closely to the predicted pattern. (Spearman's Rank Correlation Coefficient $r' = 0.96$).

Additional work is being done to test the hypothesis against census materials in other countries. All countries tested so far give similar results.

Table 2.

AVERAGE RANK ORDER OF HEAPING FOR ENDING DIGITS OF AGE, BOTH SEXES, U. S. CENSUSES FROM 1880 - 1950*

Ending Digit of Age	Predicted Rank Order	Observed Rank Order	Difference
0	1.0	1.0	0.0
1	9.5	10.0	0.5
2	3.5	3.9	0.4
3	7.5	7.9	0.4
4	5.5	6.7	1.2
5	2.0	2.0	0.0
6	5.5	5.7	0.2
7	7.5	8.4	0.9
8	3.5	3.1	0.4
9	9.5	6.3	3.2

*The average rank order of each ending digit was the sum of its rank order from 1880 - 1950 divided by eight, the number of censuses. All values were taken from Table 1.

Footnotes:

1. For further discussion of this technique of measuring heaping consult the excellent study by R. J. Myers, "Errors and Bias in the Reporting of Ages in Census Data." in the Handbook of Statistical Methods for Demographers, A. J. Jaffe, U.S. Government Printing Office. Washington, 1951.

Heaping as defined by Myers is equal to:

Heaping At Age Ending In Digit	Number of Persons At Age		Number of Persons At Age
0	=(10 + 20 + 30 + ... + 90)	x 1 +	(20 + 30 + 40 + ... + 100) x 9
1	=(11 + 21 + 31 + ... + 91)	x 2 +	(21 + 31 + 41 + ... + 101) x 8
2	=(12 + 22 + 32 + ... + 92)	x 3 +	(22 + 32 + 42 + ... + 102) x 7
.	.	.	. .
.	.	.	. .
.	.	.	. .
9	=(19 + 29 + 39 + ... + 99)	x10 +	(29 + 39 + 49 + ... + 109) x 0