ACTUARIAL NOTE

NUMBER 62
DECEMBER 1969

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE
SOCIAL SECURITY ADMINISTRATION

AN ANALYSIS OF SOCIAL SECURITY NUMBERS IN THE SMI ACTUARIAL SAMPLE

by Terence Hawkes and Ronald Harris
Office of the Actuary

This Actuarial Note resulted from an investigation into the completeness of the Actuarial Sample of Supplementary Medical Insurance (SMI) benefit payment records. The Office of the Actuary selects a 0.1% sample of all payment records on the basis of the last three digits of the Social Security Numbers. Since the sample is a 0.1% one, any bias in the sample is magnified a thousandfold in the universe figures.

Social Security Numbers are nine-digit numbers, consisting of three parts—area, group, and serial. The first three digits of a Number identify the geographical area of issuance; the following two digits divide each area into groups; and the last four digits comprise the serial portion of the Number. Social Security Numbers are issued at various points throughout the country in serial number sequence.

Payment records with Social Security Numbers ending in 595 are chosen on a continuing basis for the Actuarial Sample. Offhand, this method of selection would appear to produce exactly a 0.1% sample (1 in 1,000). From a purely theoretical standpoint, as a result of the method of issuing Social Security Numbers, this is not true, because no numbers are issued whose last four digits are 0000. Consequently in the Actuarial Sample 10 cases are selected out of 9,999 possibilities, which in theory produces a 0.10001% sample.

Experience has revealed fewer Social Security Numbers in the Actuarial Sample than was expected, and as a result questions about the completeness of the 0.1% Actuarial Sample have been raised. In an effort to examine the problem in detail, a distribution of the last three digits of the Social Security Numbers of beneficiaries who at any time had elected to enroll for SMI was obtained from the Master Beneficiary Record System.

This data includes beneficiaries who are presently enrolled in SMI, deceased beneficiaries, and beneficiaries who have terminated enrollment. Also, certain beneficiaries have been counted in the data more than once. These irregularities seem to be relatively small, and their presence is believed to be inconsequential. Little reason can be found to point to a particular bias that any of these inclusions and irregularities may produce in the distribution. This note will present an analysis of the distribution of Social Security Numbers and a comparison between the size of the Actuarial Sample and an exact 0.1% sample.

DISTRIBUTION OF SOCIAL SECURITY NUMBERS BY TERMINAL DIGIT

The total number of Social Security Numbers under observation suggests that a reliable estimate of the distribution by the terminal digit could be obtained by using the binomial distribution.[1] Table 1 compares the distribution of Social Security Numbers by terminal digit under the assumption of a binomial distribution with the actual numbers of cases. Graph 1 locates the actual numbers of cases relative to the mean and standard deviation measurements of this binomial distribution. Theoretically one expects to find 68% of all points in the interval of the mean plus or minus one standard deviation; similarly, 95% are expected in a two standard deviation interval.[2]

The points in the graph representing the actual numbers of cases fluctuate about the theoretical mean, as one would expect. Any pattern of fluctuation, however, appears to be the result of random assignment. A more detailed breakdown, by terminal digit within third-last digit groups, supports this conclusion. In the more detailed breakdown, different random patterns occur in each third-last digit grouping, but each appears to be binomially distributed. Rather sizable fluctuations are characteristic of both breakdowns, although no one particular digit is consistently outside of the two standard deviation interval.

DISTRIBUTION OF SOCIAL SECURITY NUMBERS BY THIRD-LAST AND SECOND-LAST DIGITS

In the past, Social Security Numbers have been issued in serial order sequence, which suggests that third-last and second-last digits may not be equally distributed about a single mean. In other words, the probabilities of different digits may not be equal. One might reasonably expect to observe a larger number of cases with low third-last digits than with high third-last digits. Also, this pattern would be expected to appear in the distribution of second-last digits, although less pronounced. This distribution by second-last digit, in addition, should display some of the random characteristics found in the terminal digit distribution.

An analysis of the distributions of Social Security Numbers by third-last and by second-last digits reveals that the digits are related to the corresponding numbers of cases in a systematically decreasing manner. Shown below are the correlation coefficients for three distributions of Social Security Numbers—by third-last digit, by second-last digit, and by third-and-second-last digits.

Distribution	*Correlation Coefficient*
Third-last Digit	r−0.98
Second-last Digit	r−0.84
Third-and-second-last Digits	r−0.85

A correlation coefficient (r) for a set of paired variables (in this case third-last or second-last digit and the corresponding number of cases) provides a measure of the linear relationship between the two variables.[3] A correlation coefficient of $r=\pm1$ indicates a perfect linear relationship between two variables; and, conversely, the value $r=0$ indicates no linear correlation. Negative values obtained for r indicate that the variables are inversely related. For example, the distribution of third-last digits shows that as the size of the digit increases from 0 to 9 the corresponding number of cases observed decreases. The values calculated for r indicate that the linear correlation between digits and corresponding numbers of cases is relatively high. This close relationship strongly supports the initial hypothesis that third-last and, to a lesser extent, second-last digits are distributed in a linearly decreasing manner, rather than equally or randomly.

Table 2 presents a comparison between the distribution of Social Security Numbers by third-last digit under the assumption of a linearly decreasing function and the actual numbers of cases. The expected numbers of cases under this assumption are derived from a linear regression line which fits the given data to a first degree equation.[4] Table 3 displays similar data for the second-last digit.

The scatter diagram in Graph 2 shows the distribution of Social Security Numbers in the data by third-and-second-last digits—i.e., XXX–XX–XOOX through XXX–XX–X99X. The line shown in the diagram is the linear regression line approximation to the actual data. This graph, together with Tables 2 and 3, clearly demonstrates that the Social Security Numbers under observation are distributed in a linearly decreasing manner.

CONCLUSION

The third-last digits of the Social Security Numbers of beneficiaries who had enrolled for SMI are distributed in a linearly decreasing fashion. A similar decreasing function is characteristic of the second-last digits, although to a lesser degree. Terminal digits appear to be distributed in a random fashion.

The Actuarial Sample is selected on the basis of the last three digits of a Social Security Number. Approximate figures for the universe may be obtained merely by adding three zeroes to the sample figures. This approximation assumes that Numbers ending in 595 are exactly 1/1,000 of the universe. However, the absence of serial number 0000 and the linearly decreasing nature of serial numbers violate this assumption (the former factor producing a higher proportion than 1 in 1,000, but the latter factor more than offsetting this). Shown below is a comparison of the relative sizes of an exact 0.1% sample, an expected sample (under the linear decrement assumption), and the observed Actuarial Sample that was drawn from the universe of 19,205,977 Social Security Numbers used in this analysis.

The data shows a bias error of 0.2% (due to the somewhat smaller likelihood than 1 in 1,000 of Number 595 occurring[5]) and a sampling error of 0.5% in the Actuarial Sample. In the future, one can expect that the size of the sampling error will vary.

TYPE OF SAMPLE	*NUMBER IN SAMPLE*	*PERCENT OF TOTAL*	*INFLATION FACTOR*
Exact 0.1%	19,206	0.1000%	1.000
Expected	19,178	0.0998	1.002
Actuarial	19,071	0.0993	1.007

[1] In the binomial distribution, the mean (μ) and the standard deviation (σ) are based on the following formulae: $\mu = n \cdot p$ and $\sigma = \sqrt{n \cdot p \cdot q}$ where p = proportion expected, q = 1–p, and n = number of cases observed.

[2] The 68% and 95% inclusion factors are derived by means of a normal approximation to the binomial distribution.

[3] The formula used to calculate these correlation coefficients is $r = [\Sigma X \cdot Y - n \cdot \overline{X}\overline{Y}] \div \sqrt{[\Sigma(X^2) - n \cdot (\overline{X})^2] \cdot [\Sigma(Y^2) - n \cdot (\overline{Y})^2]}$; the symbols are defined as X = third-last (or second-last) digit, Y = actual number of cases, n = number of (X, Y) pairs, $\overline{X} = {}^1/_n \Sigma X$, and $\overline{Y} = {}^1/_n \Sigma Y$.

[4] The formulation of these linear regression lines assumes that the given data are approximately linearly related and employs the method of least squares. The lines derived are of the form $Y' = b(X - \overline{X}) + \overline{Y}$; the symbols are defined as X = third-last (or second-last) digit, Y' = expected number of cases, Y = actual number of cases, $b = [\Sigma X \cdot Y - n \cdot \overline{X} \cdot \overline{Y}] \div [\Sigma(X)^2 - n \cdot (\overline{X})^2]$, n = number of (X, Y) pairs, $\overline{X} = {}^1/_n \cdot \Sigma X$, and $\overline{Y} = {}^1/_n \cdot \Sigma Y$.

[5] If the Social Security Number for the Actuarial Sample had been 495, instead of 595, it is likely that the bias error inherent in the Sample would have been virtually nonexistent.

TABLE 1

Distribution of Social Security Numbers by Terminal Digit

Terminal	Number	
Digit	Expected (μ)	Actual
0	1,918,869	1,920,987
1	1,920,790	1,917,696
2	1,920,790	1,919,723
3	1,920,790	1,920,984
4	1,920,790	1,921,634
5	1,920,790	1,921,600
6	1,920,790	1,922,452
7	1,920,790	1,918,735
8	1,920,790	1,920,563
9	1,920,790	1,921,603
Total	19,205,979	19,205,977

GRAPH 1

Number of Social Security Numbers in Terminal Digit Groups

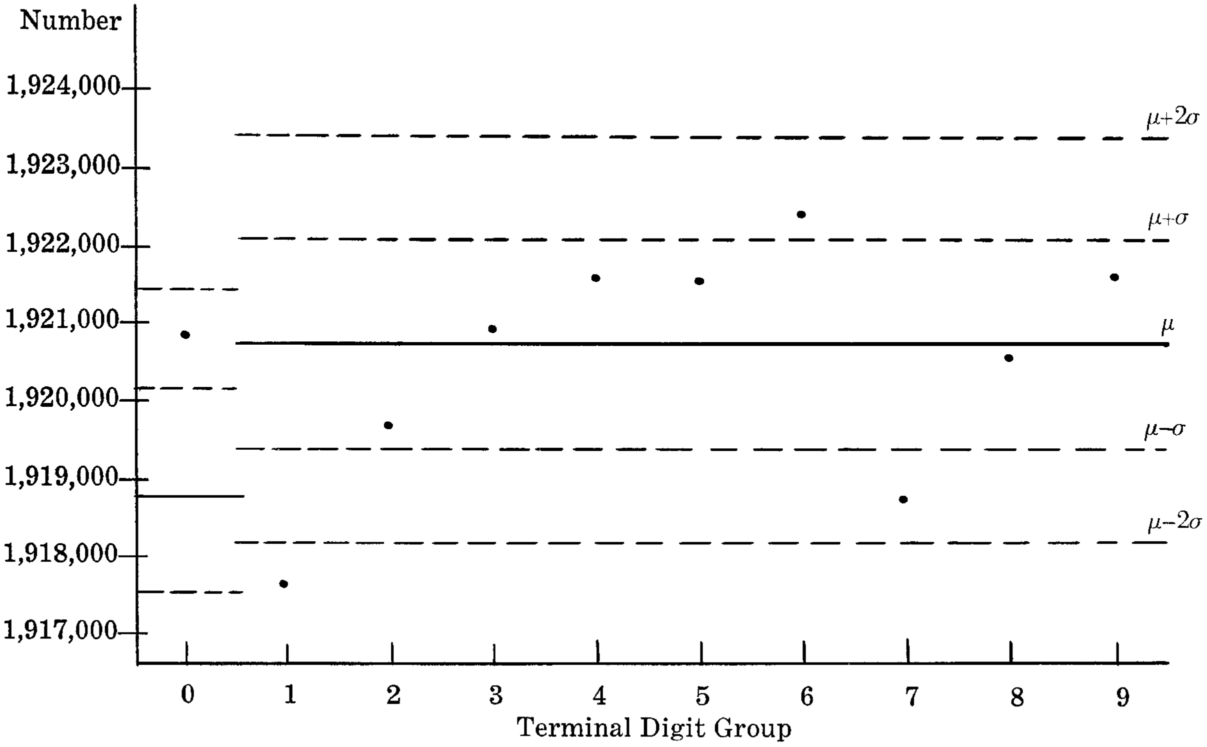


TABLE 2

Distribution of Social Security Numbers by Third-Last Digit

Third-last Digit	Number Expected[1]	Number Actual[2]
0	19,348	19,357
1	19,317	19,334
2	19,286	19,267
3	19,255	19,262
4	19,224	19,184
5	19,193	19,204
6	19,162	19,158
7	19,131	19,143
8	19,100	19,106
9	19,069	19,066
Total	192,085	192,081

TABLE 3

Distribution of Social Security Numbers by Second-Last Digit

Second-Last Digit	Number Expected[3]	Number Actual[2]
0	19,271	19,332
1	19,257	19,245
2	19,243	19,224
3	19,229	19,205
4	19,215	19,182
5	19,201	19,197
6	19,187	19,181
7	19,173	19,179
8	19,159	19,164
9	19,145	19,169
Total	192,080	192,078

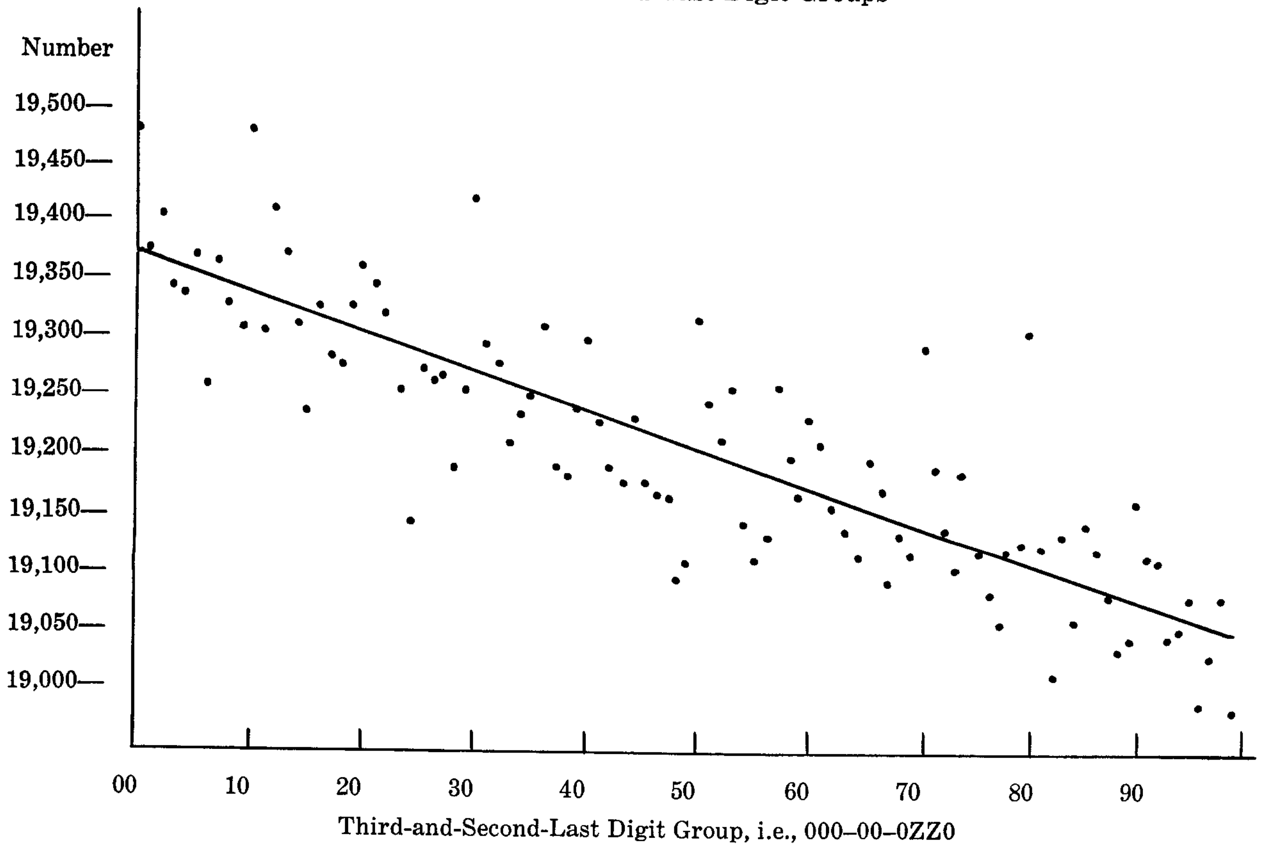
[1] The linear regression line employed to calculate these values is Y = 19,348 − 31X.

[2] These figures are the actual numbers of cases divided by 100, except the figures for third-last and second-last digits of 0. These values are the actual numbers of cases divided by 99.9, since the serial 0000 is not issued.

[3] The linear regression line employed to calculate these values is Y = 19,271 − 14X.

GRAPH 2

Number of Social Security Numbers in Third-and-Second-Last Digit Groups


Number
19,500—
19,450—
19,400—
19,350—
19,300—
19,250—
19,200—
19,150—
19,100—
19,050—
19,000—
Third-and-Second-Last Digit Group, i.e., 000–00–0ZZ0