SOME MATHEMATICAL AND STATISTICAL ASPECTS OF THE TRANSFORMED TAXPAYER IDENTIFICATION NUMBER: A SAMPLE SELECTION TOOL USED AT IRS

James M. Harte, Internal Revenue Service

Taxpayer identification numbers (TIN's), namely social security numbers and employer identification numbers, are involved in the selection of Statistics of Income (SOI) samples at the Internal Revenue Service (IRS). Formerly, these numbers were used directly to determine inclusion or exclusion in a sample. In recent years, the TIN has been transformed using a multiplicative linear congruence. With the transformed TIN, it is a very simple matter to determine the inclusion or exclusion of returns or other administrative documents which contributes to the operational simplicity and ease of use of the method.

It is the purpose of the present paper to discuss the transform and the associated sample rule so that others may assess its value for their circumstances and be aided in its use. The method would be useful to those who wish to sample based on an identification number, like the TIN, which may contain information correlated with the characteristics to be estimated, or have other non-random properties. To a large extent this is a "how to" paper.

The paper is organized into six sections. The current use of TINs and the motives for switching to the transform are described in Section 1. In Section 2 the transform is defined and placed in the setting of elementary number theory. An example is presented. The type of decision rule used for sampling is defined. In Section 3, some statistical properties and operational aspects are presented. For testing computer programs, it is desirable to use an inverse transform. The next section shows a convenient method for deriving the inverse. This is followed by section 5, which develops a self-inverse transform. The final section is devoted to a comparison with a technique proposed by Sunter that was called to the author's attention after the first draft of this paper was written.

1. SAMPLING WITH TAXPAYER IDENTIFICATION NUMBERS

In processing tax returns and other documents received from the public, information is entered into a computer system by the IRS. This information is available to the Statistics of Income program to classify returns and create efficient stratified sample designs. Within a sample stratum, SOI is able to use the taxpayer identification number for the sample decision, i.e., including or excluding a return. In particular, for the 1968 tax year and subsequently, the Employer Identification Number (EIN) has been used for corporation tax returns [1]. This EIN is also the account number for each corporation, which means that the number is universal, distinctive, and highly accurate. The objective of having good estimates of year-to-year change is important and, by sampling on the EIN, it was possible to have many of the entities reselected, reducing the variance of estimates of year-to-year change.

The EINs may be considered as nine-digit numbers with a zero permitted in the highest order (left-most) position. For computational purposes, EINs may be considered a subset of the integers between ten million and one billion. The two high-order digits identify a district office of the IRS.

Sampling for corporations (and other returns) involved specifying, in advance, a list of low-order EIN digits for each sample stratum. If an EIN of a return belonging to a stratum had one of the specified combinations in the specified places, then the return was selected for the sample. Otherwise, it was not selected. For instance, if the EIN were 12-3456789 (fictitious) then the return would be selected if 78 had been specified for positions 7 and 8 in the EIN. The right-most digit was not used.

But there are problems with the EIN, actually and putatively. A contractor consultant to the SOI program, noted that there were too many multiples of 5, which is why the lowest order digit was not used for sampling. Consequently, runs of 10 consecutive EINs were designated for the sample. The contractor expressed concern about intra-cluster correlation if similar entities had been issued consecutive numbers such as corporate subsidiaries or trusts. Such correlation would diminish the precision of sample estimates. EINs had been issued a long time ago, and it is difficult to trace the historical patterns of issuance.

The contractor suggested that instead of using the EIN, a function of it be computed. This function would yield an integer which was typically larger than the EIN. Low-order digits of this large integer would be used for the sample decision [2].

The sample decision rule used with the transform would be simpler. For instance, if the last three (low-order) digits of the transform were less than 100, then the return might be selected for the sample in a stratum when a sampling rate of 10 percent was desired. This type of rule simplifies computer programming in the first instance, and makes it easier to change sampling rates after programs have been written, or after they have begun to select returns from the revenue processing pipeline.

For individuals, the TIN is typically the social security number. It does not have the problem of heaping at multiples of 5 that the EIN has, but Jabine [3] points out that there is information in the number. Thus, some care in its use for sampling is necessary, as well as some restriction on the usable digits. For SOI 1981 individual income tax returns, selection was based on combinations of low-order digits of the SSN. For SOI 1982 and subsequently, the transformed SSN is used, along with the earlier method to preserve linkage for individuals in the Continuous Work History Sample. The situation for the estate tax return is similar.

2. THE TRANSFORM AND ELEMENTARY NUMBER THEORY

The consultant suggested that we multiply the TIN by a specified large integer and divide the resulting product by a specified larger prime (unfactorable) integer. The integer remainder produced by this process is the transform. Suppose the multiplying integer is c, and the dividing integer is n, then we produce the integers q and T such that for a given EIN, say E,

(1) cE=qn + T, with "T" between zero and "n".

In the discrete land of integers where one counts by n's, this has been called the division algorithm. We shall restrict our attention to positive numbers.

It is useful to put this computation into the slightly more general framework of modular arithmetic. We adopt the notation and concepts attributed to Gauss [4]. Let "m" be a fixed positive integer, then the integer, "a", is defined congruent to the integer b mod m, if and only if a-b=km, where k is an integer. We write a = b mod m. The integers are thus classified into "m" mutually exclusive and exhaustive congruence classes, where "a" and "b" belong to the same class. The binary congruence relation has properties similar to the binary equals relations; namely, they are both equivalence relations [5]. More importantly, the operations of addition and multiplication may be performed on these classes using any integer in a class to represent the class and each element [6]. This means that equations may be formulated using congruence classes such as: ax = b mod n and the solution for the unknown, if any, sought. Having formulated a relationship in congruence notation permits the application of well-known elementary theorems and corollaries without proof. Apropos is the following:

The congruence has solutions for "x" if, and only if, the greatest common divisor (g.c.d.) of "a" and "m" divides b. If so, and the g.c.d. = "d", then there are "d" solutions. We use the notation (a,m) = d.

We formulate (1) as a congruence:

(2) cE = T mod n.

Since n is a prime number and c is an integer less than n with (c,n) = 1, each E maps onto a unique class T mod n. By considering only the smallest positive element in the class (least positive residue), we map the congruence class onto a unique integer solution. In terms of the least positive residues, we may think of the congruence mapping the set of integers from 1 to n onto itself.

If we wish to map the set of EINs, one requirement is that n be at least as large as the number of EINs. In practice, it is much larger. To avoid disclosure of information about individuals and other entities, values for c and n used in SOI sampling will not be given. The examples used will have features paralleling our practice. It is germane to note that the value of n most often used is a prime number larger than 10 billion. The result is that most TINs map onto integers larger than any TIN.

An example: we chose n to be one less than the thirty-first power of 2 (i.e., 2,147,483,647), and c = 204,954,811, whose prime factors are 28,661 and 7,151 [7]. The ratio of n to c is approximately 10.48. This means that transforms of consecutive numbers show cycles of increasing values of length 10 or 11 as can be seen in Figure 1. During sample selection, the transforms are computed straightforwardly on a mainframe computer. We also perform the computation to verify output from the mainframe computer program. Spread-sheet usage is more important to SOI in computing, with the inverse transform, critical values for testing sample selection programs, as will be seen in Section 4.

3. STATISTICAL PROPERTIES AND OPERATIONAL CONSIDERATIONS

There is some art in choosing the multiplying constant "c". Since, in the example, n is a prime number, any positive number less than n would have the property of having a unique inverse. The choice c = 1 is obviously bad, since it leaves the TIN unchanged. The choice c = n - 1 is equally bad because, the result is the TIN subtracted from n. We have some freedom in selecting c and should select it so

Figure 1.--Transform of Consecutive Identification Numbers

c=204,954,811 n=2,147,483,647

Key Number	Transform
621,435,547	100,000,000
621,435,548	304,954,811
621,435,549	509,909,622
621,435,550	714,864,433
621,435,551	919,819,244
621,435,552	1,124,774,055
621,435,553	1,329,728,866
621,435,554	1,534,683,677
621,435,555	1,739,638,488
621,435,556	1,944,593,299
621,435,557	2,064,463
621,435,558	207,019,274
621,435,559	411,974,085
621,435,560	616,928,896
621,435,561	821,883,707
621,435,562	1,026,838,518
621,435,563	1,231,793,329
621,435,564	1,436,748,140
621,435,565	1,641,702,951
621,435,566	1,846,657,762
621,435,567	2,051,612,573
621,435,568	109,083,737

that intra-cluster correlation, that may come from selecting entities with consecutive TIN's, be eliminated. This would depend on the low-order digits of c and the sample decision rule. Suppose, for instance, the rule for selecting a return in a sample is "select when the last three digits of the transform are less than 100." Under the rule, three selections could have been made based on identification numbers 1, 6, and 13 in Figure 1.

Incrementing the identification number (e.g., TIN) by 1 corresponds to adding c to the transform, except when the resultant transform would be larger than n. In that case, it corresponds to decrementing by c - n (Figure 1). If the last three digits of a transform are less than 100, then the last three digits of the transform of the next key would be in the ranges 811 - 900 or 189 - 288. Conditional probabilities of selecting the next TIN are given in Figure 2. The conditional probability of selecting the TIN + 2 is positive as the figure shows. The rationale for computing these probabilities is given with the figure. If c had been chosen with the last three digits equal to zero, the conditional probabilities most often equal 1. If c also were a relatively small number, the transform would not be practically superior to sampling directly using the key number (e.g., EIN).

Figure 2.--Conditional Probability of Selecting Units With Identification Numbers k+1 or k+2, When Unit k Has Been Selected

c=204,954,811 n=2,147,483,647

Selection Probability p/1 Unit k	Conditional p/21 Unit k+1	Probabilities p/31 Unit k+2
0.025	0.000	0.000
0.050	0.000	0.095
0.075	0.000	0.127
0.100	0.000	0.143
0.125	0.000	0.153
0.150	0.000	0.159
0.175	0.006	0.164
0.200	0.067	0.167
0.025	0.171	0.170
0.250	0.254	0.172
0.275	0.321	0.174
0.300	0.378	0.175

NOTE: The objective of the conditional probability computations is to give an idea of proportions over repeated sampling. It is assumed that sample selection comes from a uniform distribution of the transform ending digits eligible for the sample.

The first column gives probability values which may be stipulated for a sample stratum. For the computation of the second and third columns, it is assumed that k+1 or k+2 have been issued to a unit within the scope of the survey, and which belongs to the same sample stratum as unit k. When p/1 is less than .165, p/21 = 0. When p/i is between .165 and .189, then for inclusion of k + 1 in the sample, T must be at least equal to n-c. Then p/21 = c/n((p/1-.164) - p/1). When p/i exceeds .189, the term (n-c)/n((p/1-.189) - p/1) must be added to the computation of the tabulated values. When p/1 is less than .025, p/31 = 0. Otherwise, for the tabulated values, T is equal to or greater than n-2c and p/31 = 2c/n((p/1-.025) - p/1). It can be seen that p/31 is sensitive to the value of c/n, and reducing it could lower the levels considerably.

4. THE INVERSE TRANSFORM

In practice, we select the "c" and "n" values so that we know that they are relatively prime, and thus the existence of the inverse is guaranteed. The fact that they are relatively prime may be verified by using the Euclidean algorithm attributed to the ancient geometer (circa 375 B.C.) [8].

In Figure 3 we show the Euclidean algorithm for the example as computed on an electronic spreadsheet. The intermediate values are needed to develop the inverse transform. The inverse transform is used to develop TINs whose transforms have specified low-order digits at or below the threshold value controlling the sample decision.

Figure 3.--Euclidean Algorithm and Inverse Algorithm

Euclidean Dividends/ Divisors	Euclidean Quotients	Recursive Sequence y's	
2,147,483,647		1	given
204,954,811	10	10	quotient
97,935,537	2	21	= 1+10*2
9,083,737	10	220	= 10+21*10
7,098,167	1	241	= 21+220*1
1,985,570	3	943	= 220+241*3
1,141,457	1	1,184	etc.
844,113	1	2,127	
297,344	2	5,438	
249,425	1	7,565	
47,919	5	43,263	
9,830	4	180,617	
8,599	1	223,880	
1,231	6	1,523,897	
1,213	1	1,747,777	
18	67	118,624,956	
7	2	238,997,689	
4	1	357,622,645	
3	1	596,620,334	INVERSE
1	3	2,147,483,647	'n'
0			

NOTE: The inverse is the next to last term in the sequence whenever the number of divisions is odd. Otherwise, subtract this term from 'n' for the inverse.

Firstly, we develop a formal view of the inverse transform. The transform, when computed directly, is:

(3) cE = T mod n where "c", "E", and "n" are known and we select the least positive residue to represent the class T mod n.

If E is unknown and T is stipulated, we write:

(4) cX = T mod n.

Now, the congruence:

(5) cX = 1 mod n

has a unique solution which we call I, such that

(6) c(I) = 1 mod n.

Multiplying (3) by I gives: Icx = I T mod n.

Thus, x = I T mod n is the formal solution, and the problem is reduced to solving (5).

Making a retrograde step, we note that (5) is equivalent to the Diophantine equation cx + bn = 1. In solving the equation for x and n, a specific solution pair x' and n' must be found. Then the solutions are: x = x'+ bt and y = n' + ct, where t may be any integer.

The Euclidean algorithm may be used to determine the particular pair x' and n' as well as to determine (c,n). It is a sequence of nested division algorithms (the symbol "/" precedes a subscript):

(7) n = c * q/0 + r/0,

r/0 not less than zero but less than c

c = r/0 * q/1 + r/1,

r/1 not less than zero but less than r/0

r/0 = r/1 * q/2 + r/2,

r/2 not less than zero but less than r/1

...

r/(k-3) = r/(k-2) * q/(k-1) + r/k-1,

r/(k-1) not less than zero but less than

r/k-2,

r/(k-2) = r/(k-1) * q/k + r/k, r/k = 0.

The algorithm terminates with the zero remainder and r/(k-1) is the greatest common divisor of n and c.

Let us re-label: c = r/-1. Stewart [9] described how to develop recursively solutions to the sequence of equations:

(8) (the i power of -1) * (r/i) = n * (x/i) -

c * (y/i).

To initiate the recursive computation where i ranges from -1 to k, set x/(-1) = 0; y/(-1) = 1; x/0 = 1; y/0 = q/0.
Then,

(9) x/(i+1) = x/(i-1) + (x/i) * (q/i+1) and,

(10) y/(i+1) = Y/(i-1) + (y/i) * (q/i+1).

The r/i and q/i in (8), (9) and (10) are from the Euclidean algorithm. The penultimate equation is what is needed, and, since r/(k-1) = 1, we write:

(11) (the (k-1) power of -1) x1 = n *

(x/(k-1)) - c * (y/(k-1)).

Consequently, $\pm$ y/(i-1) = 1 mod n.

If the number of divisions in the Euclidean algorithm is odd, then y/(k-1) is the required inverse. Otherwise, n - y/(k-1) is the required inverse. Summarizing:

(12) I = y/(k-1) if k + 1 is odd

(13) I = n - (y/k-1) if k + 1 is even.

It is interesting to note that x/k = c and y/k = n. Since the x/i, y/i sequences are increasing, the maximum values are known. This also provides a verification of the accuracy of the computation. Of course, only the y/i values need be computed. All this is illustrated in Figure 3. A theorem by Lame (1845 A.D.) guarantees that the maximum number of divisions in the Euclidean algorithm is not greater than 5 times the number of digits in "c", the first divisor [9]. In the Figure 3 example, no more than 45 divisions were expected and, in fact, only 21 were needed. A sharper limit is given by Knuth [10].

To illustrate the use of the inverse, suppose from a stratum it was desired to select a 10 percent sample. The sample rule is as follows: compute the transform of the EIN, inspect the last three digits and compare to 100. If less, include in the sample. Otherwise, exclude. For instance in the Figure 3 example, I = 596,620,334. A proposed value of the transform might be 1,111,111,100. With "I" playing the role of the multiplying constant, the TIN computed for testing is 514,333,883. Similarly, the transform value 1,111,110,099 yields the test value 297,833,415. The values "100" and "099" are threshold values for the sample decision.

Before the inverse computation was developed, test values were developed by starting with a value of the direct transform for some TIN and then generating (by computer) the transform of the consecutive TINs by adding "c" to the previous value of the transform and subtracting "n" if the result were larger than n. The desired values were found by inspection, if the list were long enough. Another solution was performing the direct transform on a file of actual EINs, sorting on the low-order digits of the transform and printing out a long listing for reference purposes.

With the development of the inverse transform, more economic computation was possible. A candidate test value, with the specified low-order digit, was used with the inverse transform. The result was frequently an invalid TIN, usually too large, because the transform computation and its inverse maps values all over the range of values 1 to "n". In this case, various BASIC and FORTRAN programs were written which added a power of ten, say 1,000, to the candidate value and recomputed until a valid result was obtained. This is a good approach, especially when the testing is part of the revenue processing system and many extra constraints on the validity of the TIN must be met. The test values can be easily produced on a micro-computer.

The usual constraint is merely to have a number in the range from 10 million to 1 billion. This can be done non-iteratively or with only two iterations on an electronic spreadsheet. The inverse of the candidate number is computed and used to look up an adjustment factor in a table. The adjustment factor is subtracted from the candidate number. The computation is successfully repeated with the new number.

5. THE SELF INVERSE

It is convenient, but not necessarily important, to use a transform that is a self inverse, that is with c = I.

The problem of finding the inverse was characterized in (5) as solving the congruence cx = 1 mod n with "c" and "n" known. For the self inverse, c is not known and the relevant congruence is (** means exponentiation):

(14) x ** 2 = 1 mod n or (x-1) (x+1) = 0 mod n.

This looks like the equation x ** 2 - 1 = 0, whose solutions are +1 and -1. The congruence (14) does in fact have solutions +1 and -1. It is obvious that c = 1 is a useless solution. The solution c = -1 is equivalent to c = n - 1, which is equivalent to subtracting E from n and, thus, also useless. Are there more than two solutions? We note that the congruence: x ** 2 = 1 mod 8 has four solutions represented by the least positive residues 1, 3, 5 and 7. (The square of an odd number is one more than a multiple of 8.)

Can n be a prime number? Let F(x) be a polynomial of degree t with integer coefficients. A theorem of LaGrange [11] establishes that the congruence:

(15) F(x) = 0 mod, p, with p prime, cannot have more solutions than its degree t. Thus, F(x) = x ** 2 - 1 = 0 mod p has just the two solutions.

Thus, we must consider composite n. To avoid needless complications, we solve the problem for a special and very useful case. We require n to be an odd number and the product of prime numbers to the first power only.

Thus,

(16) n = (p/1) * (p/2)...*(p/k) with p/i not equal 2.

We may write:

x**2 - 1 = (x-1) (x+1) = 0 mod n which implies
= 0 mod (p/i) for each i.

By the fundamental lemma of arithmetic, if a prime number divides a product, it divides at least one of the factors. Thus, either:

x = 1 mod p/i or x = -1 mod p/i for each i.

We invoke the Chinese Remainder Theorem (CRT) [12] which was stated and proved by the mathematician Shu Shu Chiu Chang (1247 A.D.) and was known in a special form by Sun Tsu (prior to 500 A.D.) [13]. Let m/i be moduli relatively prime in pairs and associate with each a residue a/i. Let m be the product of the m/i. Then there is exactly one congruence class x mod m whose elements also satisfy the relationships x = a/i mod m/i. (The class x mod m is a subclass of each class a/i mod m/i.)

For each combination x = + 1 mod p/i, CRT guarantees a distinct solution. Thus, there are 2**k solutions where k is the number of primes in the factorization of m.

To compute x we use x = sum of (M/i * x/ix a/i)

where (M/i) * (m/i) = m and (M/i) * (x/i) = 1 mod m/i.

The latter congruences may be solved for x/i using the method described in section 3.

An example is shown as Figure 4.

6. EVALUATION

In a recent paper Sunter advocated use of sampling from administrative record systems based upon identification numbers. Such sampling would create panels in the successive cross sectional samples [14]. The samples in various SOI surveys have this characteristic although the improvement of cross-sectional estimates of change over time was the advantage sought rather than the analysis of panel entities.

Sunter also uses a function of the identification number with the same type of rule used to make the sample decision of inclusion or exclusion. The identifying number is transformed by a multiplicative quadratic congruence and the result scaled to produce an integer in the range [000,999]. Let c and n be constants and E the identification number. S is defined as the least positive residue satisfying the congruence:

cE**2 = S mod n.

Then T (Sunter) = [1,000 * (S/n)]

where the square brackets represent the integer part of the enclosed quantity.

The quadratic congruence does not have a unique inverse (mapping is two on one) and

Figure 4.--Solve X**2 = 1 mod 1,001

1,001 = 7*11*13	1,001 = mI*MI	MI*XI = 1mod 1,001
m1 = 7	M1 = 143	X1 = 5
m2 = 11	M2 = 91	X2 = 4
m3 = 13	M3 = 77	X3 = 12
M1*X1 = 715	M2*X2 = 364	M3*X3 = 924

	715	364	924	Residue	Least Positive Residue	Complement
Factors:	1	1	1	2003	1	1000
	1	1	-1	155	155	846
	1	-1	1	1275	274	727
	1	-1	-1	-573	428	573

The eight solutions are displayed in the right columns. The values just above the line are multiplied by the row factors and summed to a residue. This is converted to a least positive residue by adding or subtracting multiples of 1,001 as necessary. The complement and the least positive residue add to 1,001. The complement is the solution produced with each factor in the row multiplied by -1.

Note, for instance, that the solution 155 maps multiples of 13 onto themselves, and multiples of 77 onto their complements. This is characteristic of these solutions, but is unimportant for numbers with only large prime factors, the practical situation.

consequently is more difficult to work with when inverse values are needed. The implications for sampling are more laborious to evaluate for a given set of parameters as Sunter's paper demonstrates for the two sets he recommends.

However, there is a difference in sampling strategy behind the two methods. Under the method proposed in this paper the selection of entities with consecutive identification numbers is prevented or discouraged in most cases. This is to reduce intra-cluster correlation. The alternative method apparently makes such selections independently (with suitable choice of parameters).

The SOI program is also currently using a transform to select a sample of documents not having permanent identification numbers. The serial number of a statistical edit sheet is used. The method provides ease of control and use. The sample selection decision is unrelated to the processing history of the document. In this case, a self-inverse version of the transform was specified.

ACKNOWLEDGMENTS

Thanks are extended to Wendy Alvey and Beth Kilss for editorial assistance and preparation of graphic illustrations; to Mike Bush for electronic spreadsheet computations; and to Darlene Reynolds, Susan Rhodes, Nancy Dutton, and Bettye Jamerson for their typing efforts.

NOTES AND REFERENCES

[1] Jones, H. W. and McMahon, Paul. (1984) "Sampling Corporation Income Tax Returns for Statistics of Income, 1951 to Present," Statistics of Income and Related Administrative Record Research: 1984, Internal Revenue Service, pp. 269-274.

[2] Houseman, E. E. (1976) "Methods of Selecting Tax Returns," prepared at Westat, Inc. in a report commissioned for the Internal Revenue Service. He also cites a memorandum by B. Tepping of the Census Bureau mentioning the heaping of EIN's at multiples of 5.

[3] Jabine, T. B. (1985) "Properties of the Social Security Number Relevant to Its Use in Record Linkages," Record Linkage Techniques - 1985, Internal Revenue Service, pp. 213-225.

[4] Stewart, B. M. (1952) Theory of Numbers Chapter 17. The MacMillan Company.

[5] An equivalence relation is determinative, reflexive, symmetric and transitive.

[6] Addition and multiplication are associative and commutative. Subtraction is inverse to addition. Multiplication does not generally have an inverse.

[7] Selby, S. M. (1971) Standard Mathematical Tables, Nineteeth Edition, Primes 1 to 100,000, pp. 686 et seq., Chemical Rubber Co.

[8] Knuth, D. E. (1981) "Seminumerical Algorithms," Art of Computer Programming, Second Edition, Vol 2, Addison-Wesley Publishing Company, p. 318.

[9] Stewart, B. M., op.cit., pp. 75-77.

[10] Knuth, D. E., op.cit, pp. 343.

[11] Stewart, B. M., op.cit., pp. 141-142.

[12] Stewart, B. M., op.cit, p. 130.

[13] Knuth, D. E., op.cit., pp. 270-271.

[14] Sunter, A. B. (1986) "Implicit Longitudinal Sampling from Administrative Files: A Useful Technique," Journal of Official Statistics, Vol. 2, No. 2 1986, Statistics Sweden, Statistics Sweden.