Expected Number of Random Duplications Within or Between Lists

William E. Yancey[*]

Abstract

The U.S. Census Bureau seeks to determine duplicately listed individuals by searching across lists to identify records with the same name and birth date. The question arises of how many of these agreeing records are random agreements, two different people with the same name and birth date. To formally answer this question, we consider first the familiar Birthday Problem and then the more complicated Collision Problem. For each of these problems we exhibit the explicit probability distributions from which we can compute means and variances for some parameter values. We apply this result to voter registration lists for Oregon and Washington to estimate the number of "false matches" that occur across these lists.

Key Words: record linkage, deduplication, Stirling numbers, birthday problem

1. Introduction

One sometimes wishes to identify duplicate records in lists. By *duplicates* we mean two different records that refer to the same entity. In the U.S. Census Bureau, the lists are often files with records that contain values of Census variables for people. If we are looking for duplicate records for the same person at different addresses, then other than some categorical variables for sex and ethnicity, the main personal variables are name and date of birth. We wish to consider the expected number of random agreements within a list or between two lists where two records happen to agree on name and date of birth but do not represent the same person. We consider the problem in the general form of a finite set, or sets, of elements, each of which has been randomly assigned with replacement a tag from a finite set of tags. For our application, we are thinking of the sets as (records of) people with the same name, and the tags are birth dates (within a fixed range). For each of the problems we will present the probability distribution, expected value, and variance. We then discuss an application of these results to voter registration files for Oregon and Washington.

2. Two problems of random agreement

2.1 The Birthday Problem

A standard example or problem in an elementary probability course is to compute the probability for a set A of n people (*e.g.* the students in a class) what is the probability that at least two people have the same birthday. What is sometimes referred to as the "birthday paradox" is the fact that the smallest number n for which this probability is greater than one half is $n = 23$, an evidently, if not paradoxically, small number. However, the discussion usually stops short of specifying the probability distribution or the expected number of duplicates. So to state the problem a little more generally and formally, for a set of n elements, assign a tag each element, where the tags are selected randomly with replacement from a set of D tags. Let $X(\omega)$ be the random variable where ω is a particular assignment of tags and $X(\omega)$ is the number of distinct tags in the assignment. The birthday problem is asking for $D = 365$ to compute the value of

$$\Pr(X < n) = 1 - \Pr(X = n).$$

[*] Statistical Research Division, U.S. Census Bureau

2.2 The Collision Problem

A related problem concerns the number of duplicates across two lists. Here we consider two sets A, B with m, n elements respectively. The $m+n$ elements are each assigned a tag chosen randomly with replacement from a set of D tags. For a given assignment ω of tags, we let the random vector $X(\omega) = (i, j)$ when there are i elements in set A with tags distinct from any tag in B and there are j elements in B which have tags that are not assigned to any element of A. The name for this problem arises in hashing where one is concerned about two distinct records colliding when they are assigned the same address. Presumably, the problem was considered under other names before the advent of computer science.

3. Some Combinatorial Notations

Consideration of these problems naturally involves the *falling factorial power*, which is sometimes denoted $[n]_k$ (*e.g.* in [Constantine [1987]]) where for positive integer k is

$$[n]_k = \underbrace{n(n-1)(n-2)\dots(n-k+1)}_{k \text{ factors}}.$$

For instance, in the above birthday problem, we see that

$$\Pr(X = n) = \frac{[D]_n}{D^n} \tag{1}$$

since we sequentially draw a distinct tag from the remaining unused tags.

The most common combinatorial element is the *binomial coefficient* $\binom{n}{k}$ which has the combinatorial interpretation as the number of subsets with k elements can be chosen from a set of n elements. Perhaps the next most basic combinatorial elements are the Stirling numbers. Following the excellent notational recommendation of Knuth (Graham et al. [1994]), we denote the *Stirling number of the second kind* by $\left\{ {n \atop k} \right\}$, which has the interpretation as the number of ways a set of n elements can be partitioned into k subsets. Similarly the *Stirling number of the first kind* $\left[{n \atop k} \right]$ has the interpretation as the number of ways that n objects can be arranged into k cycles. However, the Stirling numbers of the first kind will be useful in the present context as the coefficients of the polynomials determined by the falling (or rising) factorial powers,

$$[n]_k = \sum_{j=1}^{k} (-1)^{k-j} \left[{k \atop j} \right] n^j.$$

4. Probability Distributions

4.1 The Birthday Problem

For the probability distribution (or mass function) of the birthday problem random variable, we note that for $X = k$, we need to assign exactly k distinct tags to the n elements of the set A, so we can divide A into k subsets and then assign a distinct tag to the elements of each subset. Thus we have

$$\Pr(X = k) = \frac{1}{D^n} \left\{ {n \atop k} \right\} [D]_k\,,$$

which reduces to (1) in the case $k = n$. If we wish to express this in terms of powers of D, we can use the Stirling numbers of the first kind to get

$$\Pr(X = k) = \frac{1}{D^n} \left\{ {n \atop k} \right\} \sum_{j=1}^{k} (-1)^{k-j} \left[{k \atop j} \right] D^j.$$

4.2 The Collision Problem

The combinatorics for the collision problem are considerably more complicated, mainly because we are only interested in duplicated tags between the two sets and not any possible duplicate tags within each of the sets. For sets A with m elements and B with n elements, we describe a procedure for selecting tags so that i elements of A have the same tag as some of j elements of B, and conversely.

1. Select a subset A_i of i elements of A
2. Select a subset B_j of j elements of B
3. For each k, $1 \leq k \leq \min(i, j)$
   (a) Partition A_i into k subsets
   (b) Partition B_j into k subsets
   (c) Assign each of the k subsets of B_j to one of the k subsets of A_i
   (d) Assign k distinct tags to the k pairs of subsets
4. For the remaining complementary subset $A_{m-i} = A \backslash A_i$, partition it into u subsets, $0 \leq u \leq m - i$
5. For the remaining subset $B_{n-j} = B \backslash B_j$, partition it into v subsets, $0 \leq v \leq n - j$
6. Assign $u + v$ distinct tags to these subsets from the $D - k$ remaining tags

Actually since we defined the random vector X in terms of unique tags (non-duplicates), the above procedure produces a tag assignment ω such that $X(\omega) = (m - i, n - j)$. By counting the number of ways each sequential selection can be made, we see that we can write

$$\Pr(X = (m-i, n-j)) = \frac{1}{D^{m+n}} \binom{m}{i} \binom{n}{j} \sum_{k=0}^{\min(i,j)} \genfrac{\{}{\}}{0pt}{}{i}{k} \genfrac{\{}{\}}{0pt}{}{j}{k} k! \, [D]_k \sum_{u=0}^{m-i} \sum_{v=0}^{n-j} \genfrac{\{}{\}}{0pt}{}{m-i}{u} \genfrac{\{}{\}}{0pt}{}{n-j}{v} [D-k]_{u+v} .$$

Aside from combining the two descending factorial powers

$$[D]_k \, [D-k]_{u+v} = [D]_{u+v+k}$$

the above expression does not appear to admit much simplification. Again if we wish, we may replace $[D]_{u+v+k}$ with a polynomial in powers of D using Stirling numbers of the first kind for coefficients.

5. Expected Value and Variance

For the birthday problem random variable, it is possible to compute the expected value directly from the probability distribution by using the recursion formula for Stirling numbers of the second kind

$$\genfrac{\{}{\}}{0pt}{}{n+1}{k} = k \genfrac{\{}{\}}{0pt}{}{n}{k} + \genfrac{\{}{\}}{0pt}{}{n}{k-1}$$

along with the somewhat more obscure identity

$$\genfrac{\{}{\}}{0pt}{}{n+1}{m+1} = \sum_k \binom{n}{k} \genfrac{\{}{\}}{0pt}{}{k}{m}$$

It is also possible to compute the variance using the above identities along with the analogous ones for Stirling numbers of the first kind

$$\begin{bmatrix} n+1 \\ k \end{bmatrix} = n \begin{bmatrix} n \\ k \end{bmatrix} + \begin{bmatrix} n \\ k-1 \end{bmatrix}$$

and

$$\begin{bmatrix} n+1 \\ m+1 \end{bmatrix} = \sum_k \begin{bmatrix} n \\ k \end{bmatrix} \binom{k}{m}$$

(Graham et al. [1994], pp.264–265). However, I was unable to do anything of the sort with the more complicated collision problem random vector probability distribution. Fortunately, it was pointed out to me by Prof. Bikas Kumar Sinha that these calculations are a lot easier using indicator functions.

5.1 The Birthday Problem

For the birthday problem random variable, we want to count the number of distinct tags that are chosen. To do this, we enumerate the tags and for $i = 1, 2, \ldots, D$, we define the indicator function Y_i by $Y_i = 1$ if the i^{th} tag was never chosen for any of the n elements of A, and $Y_i = 0$ otherwise (*i.e.* the i^{th} tag was chosen for one or more of the elements in A). Thus we may express the birthday problem random variable X as

$$X = D - \sum_{i=1}^{D} Y_i$$

since the sum counts the number of distinct tags that were not chosen, so the difference is the number of distinct tags that were chosen. A virtue of indicator functions is that their expected value is just the probability that they equal one. In this case,

$$\mathrm{E}\left[Y_i\right] = \Pr\left(Y_i = 1\right) = \left(\frac{D-1}{D}\right)^n$$

since n independent choices were made from the set of all tags except the i^{th} one. From this we easily have

$$\begin{aligned} \mathrm{E}\left[X\right] &= D - \sum_{i=1}^{D} \mathrm{E}\left[Y_i\right] \\ &= D\left(1 - \left(1 - D^{-1}\right)^n\right) \end{aligned} \qquad (2)$$

since all of the summands are equal. We might note that if we are finding the expected value of the number of duplicates, $Z = n - X$, then viewed as a polynomial in D^{-1}, the leading term is $\binom{n}{2}D^{-1}$,

$$\mathrm{E}\left[Z\right] = \binom{n}{2} D^{-1} \left(1 - \frac{n-3}{3} D^{-1} + O\left(D^{-2}\right)\right).$$

In terms of the traditional birthday problem, we might note that while $n = 23$ is the first value of n for which the probability exceeds one-half that there is at least one duplicate birthday ($\Pr\left(Z\right) > \frac{1}{2}$), the above formula indicates that for $D = 365$, the first value of n for which the expected number of duplicates exceeds 1 is $n = 28$.

To compute the variance, we can find

$$\begin{aligned} \mathrm{E}\left[Y^2\right] &= \mathrm{E}\left[\left(\sum_{i=1}^{D} Y_i\right)^2\right]. \\ &= \sum_{i=1}^{D} \mathrm{E}\left[Y_i^2\right] + \sum_{\substack{(i,j) \\ i \neq j}} \mathrm{E}\left[Y_i Y_j\right]. \end{aligned}$$

Since Y_i is an indicator function, $Y_i^2 = Y_i$ so the first summand is the same as before. Also $Y_iY_j = 1$ if and only if $Y_i = 1$ and $Y_j = 1$, so this happens exactly when neither the i^{th} nor the j^{th} tag is chosen, so $\mathrm{E}\left[Y_iY_j\right] = \left(\frac{D-2}{D}\right)^n$. Thus

$$\begin{aligned}\mathrm{Var}\left[X\right] &= \mathrm{Var}\left[Y\right] \qquad (3)\\ &= \mathrm{E}\left[Y^2\right] - \left(\mathrm{E}\left[Y\right]\right)^2\\ &= D\left(1-D^{-1}\right)^n + D\left(D-1\right)\left(1-2D^{-1}\right)^n - D^2\left(1-D^{-1}\right)^{2n}.\end{aligned}$$

We may observe that some lower order terms cancel so we may write

$$\mathrm{Var}\left[Z\right] = \binom{n}{2}D^{-1}\left(1 - \frac{5n-7}{3}D^{-1} + O\left(D^{-2}\right)\right)$$

so that when D is large relative to n, the variance of the number of duplicates is nearly equal to the expected value of the number of duplicates. For example, in the case where $D = 365$ and $n = 28$, the variance is 0.917 so the standard deviation is 0.958. If we use the two term polynomial estimate, we get the slightly smaller estimated variance 0.910 and standard deviation 0.954.

5.2 The Collision Problem

From the collision problem random vector $X = \left(X^A, X^B\right)$, we work with the component X^A, the other component being symmetric. Thus for a random tag assignment ω to the $m+n$ elements of A and B, $X^A\left(\omega\right) = i$ means that their are i elements of A which have tags that are not assigned to any elements of B. Enumerating the elements of A and the tags, then for each $i = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, D$, we define the indicator function X_{ik} by $X_{ik}\left(\omega\right) = 1$ if the k^{th} tag was not chosen for any element of B, but it was chosen for the i^{th} element of A, and $X_{ik}\left(\omega\right) = 0$ otherwise. We see that

$$X^A = \sum_{i=1}^{m}\sum_{k=1}^{D} X_{ik}$$

and clearly

$$\mathrm{E}\left[X_{ik}\right] = \frac{1}{D}\left(\frac{D-1}{D}\right)^n$$

for all i, k the expected value is

$$\mathrm{E}\left[X^A\right] = mD\,\mathrm{E}\left[X_{ik}\right] = m\left(1-D^{-1}\right)^n, \qquad (4)$$

a formula even a bit simpler than (2). Of course $\mathrm{E}\left[X^B\right]$ is obtained by switching m and n. If we want to know the number duplicates, *i.e.* elements of A whose tags are also tags assigned to elements of B, we want $Y^A = m - X^A$, and

$$\begin{aligned}\mathrm{E}\left[Y^A\right] &= m\left(1-\left(1-D^{-1}\right)^n\right)\\ &= mnD^{-1}\left(1 - \frac{n-1}{2}D^{-1} + O\left(D^{-2}\right)\right).\end{aligned}$$

For the variance,

$$\begin{aligned}\mathrm{E}\left[\left(X^A\right)^2\right] &= \mathrm{E}\left[\left(\sum_{i=1}^{m}\sum_{k=1}^{D} X_{ik}\right)^2\right]\\ &= \sum_{i=1}^{m}\sum_{k=1}^{D}\mathrm{E}\left[X_{ik}^2\right] + \sum_{i\neq j}\sum_{k=1}^{D}\mathrm{E}\left[X_{ik}X_{jk}\right] + \sum_{i=1}^{m}\sum_{k\neq l}\mathrm{E}\left[X_{ik}X_{il}\right] + \sum_{i\neq j}\sum_{k\neq l}\mathrm{E}\left[X_{ik}X_{jl}\right].\end{aligned}$$

As before, for indicator functions $X_{ik}^2 = X_{ik}$, so

$$\sum_{i=1}^{m}\sum_{k=1}^{D} \mathrm{E}\left[X_{ik}^2\right] = m\left(1 - D^{-1}\right)^n.$$

For each $(i, j, k), i \neq j$, $X_{ik}X_{jk} = 1$ means the k^{th} tag was not chosen for any element of B, but was chosen for both the i^{th} and j^{th} elements of A, so

$$\mathrm{E}\left[X_{ik}X_{jk}\right] = \Pr\left(X_{ik}X_{jk} = 1\right) = \frac{1}{D^2}\left(\frac{D-1}{D}\right)^n$$

and since there are $m(m-1)D$ such terms,

$$\sum_{i\neq j}\sum_{k=1}^{D} \mathrm{E}\left[X_{ik}X_{jk}\right] = m(m-1)D^{-1}\left(1 - D^{-1}\right)^n.$$

On the other hand, for $X_{ik}X_{il}$, $k \neq l$, we cannot have two different tags chosen for the same element of A, so $X_{ik}X_{il}$ is always 0, so $\mathrm{E}\left[X_{ik}X_{il}\right] = 0$. For $i \neq j$, $k \neq l$, $X_{ik}X_{jl} = 1$ means that neither the k^{th} nor the l^{th} tags were chosen for any of the elements of B, while one was chosen for the i^{th} element of A and the other was chosen for the j^{th} element, so

$$\mathrm{E}\left[X_{ik}X_{jl}\right] = \Pr\left(X_{ik}X_{jl} = 1\right) = \frac{1}{D^2}\left(\frac{D-2}{D}\right)^n$$

and since there are $m(m-1)D(D-1)$ such terms,

$$\sum_{i\neq j}\sum_{k\neq l} \mathrm{E}\left[X_{ik}X_{jl}\right] = m(m-1)\left(1 - D^{-1}\right)\left(1 - 2D^{-1}\right)^n.$$

Since $\mathrm{Var}\left[Y^A\right] = \mathrm{Var}\left[X^A\right] = \mathrm{E}\left[\left(X^A\right)^2\right] - \left(\mathrm{E}\left[X^A\right]\right)^2$, we have

$$\mathrm{Var}\left[X^A\right] = m\left(1 - D^{-1}\right)^n + m(m-1)D^{-1}\left(1 - D^{-1}\right)^n + m(m-1)\left(1 - D^{-1}\right)\left(1 - 2D^{-1}\right)^n - m^2\left(1 - D^{-1}\right)^{2n}, \quad (5)$$

so as a polynomial in D^{-1},

$$\mathrm{Var}\left[X^A\right] = mnD^{-1}\left(1 - \frac{3n-1}{2}D^{-1} + O\left(D^{-2}\right)\right)$$

in comparison with $\mathrm{E}\left[X^A\right]$(4). However, unlike the coefficients of higher order D^{-1} terms of $\mathrm{E}\left[X^A\right]$ which are polynomials in n, the coefficients of higher order terms of $\mathrm{Var}\left[X^A\right]$ can involve a factor of m as well as polynomials in n.

6. Empirical Studies

In their study related to assessing the extent of double voting, McDonald and Levitt (McDonald and Levitt [2008]) examined the voter registration files for New Jersey by simulating the occurrence of random name and birth date agreements. Here at the U.S. Census Bureau, Dr. William E. Winkler simulated birth date distribution to measure the expected number of random agreements of name and birth date across the voter registration lists of Oregon and Washington. We similarly compute the likely random agreements of name and birth date both within each of the voter registration lists in Oregon and Washington and across the two lists using the analytic formulas for the expected value and variance.

6.1 The Data Sets

We are working with the voter registration files for Oregon and Washington state. The Oregon file contains the records for 2,040,589 voters and the Washington file contains the records for 3,407,596 voters. While these data files in general appear to be quite clean, there are no doubt some errors in them. If we consider the birth years listed in the files, there is a small number of records ($< 0.03\%$) with the birth year missing. In addition, there are some birth years dating from the 19^{th} and early 20^{th} century. One Mr. Brown has a birth year listed as 1763. Some of these records may have birth years that are typographical errors and some of them may be records for people who are no longer voting. We will somewhat arbitrarily restrict the files to records with birth years over a 64 year period, from 1927 to 1990. No doubt some of the records with earlier birth years represent people who are still living and some of them may still be active voters, but we presume that this subset represents the preponderance of the active voters. The Oregon list has 1,941,108 and the Washington list has 3,258,537 records with birth years in this interval. These represent more than 95% of the total records in each state.

For each of these restricted state files, we sort by the full names and count the number of times each of these names occurs. By "full names" we mean last name, first name, and middle initial. We next repeated the name count using only first and last name for "full names." We only consider exact name matches, not attempting to adjust for nicknames or typographical errors. The Oregon list contains $94,923$ names that occur more than once; the Washington list has $183,304$ names that occur more than once. For each problem we compute the mean and variance over the data sets using the analytic formulas that we have derived. For some cases, we also estimate the mean and variance using simulation.

6.2 A Note on the Simulation

When we did our preliminary simulation studies, we used the C library program `rand()`. However, for our more intensive simulations, we decided to use a more robust random number generator. One reason that one is cautioned against using `rand()` is that while the C language requirements specify the form of the interface for the function, they do not specify its implementation, and historically some common random number generator functions have been shown to be poor randomness simulators. A more immediate problem is that the ANSII C specifications require that `rand()` return an integer between 0 and the system-defined constant `RAND_MAX`, which is commonly defined to equal $32,767$, which means that the function returns a positive two-byte signed integer. Our 64 year age interval requires $D = 64 * 365 = 23,360$ distinct birth dates, a range that is covered only one complete time by the output of `rand()`. If we were to extend our age interval to 90 years, then there would not be enough random integers to cover all possible dates. Therefore, we have decided to use another random number generator, the Mersenne twister as implemented in the Gnu Scientific Library (GSL). This generator has passed extensive randomness tests, is efficiently implemented, and returns a long unsigned integer, which produces more that 4 billion possible outcomes.

Furthermore, in an attempt to more accurately estimate some of the low probability values, we have used a varying number of iterations in the simulations. For example, for the birthday problem the probability that a group of 2 people have the same birth date from our D possible birth dates is 0.0000428082, a event that should happen about once in about $23,360$ trials. In order for there to be a chance that the simulations can produce a reasonably accurate average, we used $1,000,000$ iterations. However, this is too large a number of iterations to use over a large range of possible group sizes, so we used $100,000$ and then $10,000$ iterations as the group sizes and the corresponding means increased in value.

6.3 The Birthday Problem

For each state we wish to compute the expected number of names and birth dates that randomly occur more than once. To do this, we partition the list of voters by full names and consider the number of times that full names occur. If a name occurs n times and we are randomly assigning $D = 64 * 365$ birth date tags, then we want the mean of the random variable $X_{n,D}$, the number of duplicate tags in a set of n

	Analytic Mean	Std Dev	Simulated Mean	Sim Mean w/ Year Dist
Oregon	14.50	3.81	14.41	15.68
Washington	35.87	5.99	35.65	38.63

Table 1: Expected Duplicates Including Middle Initial

	Analytic Mean	Std Dev
Oregon	155.84	12.46
Washington	402.80	20.01

Table 2: Expected Duplicates Without Middle Initial

members. Each full name represents and independent random variable, so the total mean (and variance) for the state is the sum of the means (variances) for each full name. If we use the means and variances for for the birthday problem calculated from the analytic probability distribution, we have for Oregon an expected value of 14.50 and a variance of 14.49, resulting in a standard deviation of 3.81. By comparison, if we use Birthday Problem means computed via simulation using a Mersenne twister random number generator, we get an expected value of 14.41, which is about 0.27 standard deviations from the analytic mean. Since for small values of n, the probabilities are quite small and yet they are counted multiple times for many names that occur a small number of times, we used $1,000,000$ draws for $2 \leq n \leq 15$, $100,000$ draws for $16 \leq n \leq 50$, and $10,000$ draws for $n > 50$. As discussed above, these calculations are based on assuming a uniform distribution for the birth date tags. While the birthdays may be more or less uniformly distributed among the population, the birth years of the voters is not, There is a greater concentration of registered voters in their 50's than in their 20's or 70's. If we perform a simulation with the birthday selected uniformly and the birth year selected using the empirical birth year distribution, the expected number of duplicate Oregon voters is 15.68 (Table 1).

For Washington, analytic mean 35.83, variance 35.85, standard deviation 5.99, simulated mean 35.65, simulated year distribution 38.63. As expected, by using the non-uniform birth year distribution, the mean number of random agreements is increased. While the increase is not great, it is a noticeable eight or nine percent, a bit higher in this case than the approximately seven percent reported by McDonald and Levitt (McDonald and Levitt [2008]).

On the other hand, if we look at the voter registration lists using just first and last name without the middle initial (Table 2), we see that the number of expected random duplications increases more than tenfold. This indicates the added distinguishing power of the middle initial, a data field that appears to be fairly consistently reported in these files, but which is often missing or inaccurate in files in general.

In the actual data, Oregon has 47 matching name/birth date pairs in the 64 year range (48 in all), while Washington had 136 matching name/birth date pairs in the 64 year range (139 in all). There were no matching triples. Comparing these numbers to the estimated random agreements, since the number of observed duplicates is many standard deviations from the mean number of random duplicates, we would conclude that both states definitely have some individuals who are registered twice with the same name and birth date. On the other hand, the number of such cases is quite small and a substantial proportion (perhaps about 25%)of the matching pairs are probably random agreements (Table 3).

When we omit the middle initial, we find around five times as many duplications in the file. However, the expected number of random agreements increases by almost as much, so it is unclear if we are actually finding many more people with duplicate registrations (Table 4).

	Actual Dups	Mean	SDs from Mean	Multiple of Mean
Oregon	47	14.05	8.53	3.24
Washington	136	35.87	16.72	3.79

Table 3: Comparison with Observed Duplicates Including Middle Initial

	Actual Dups	Mean	SDs from Mean	Multiple of Mean
Oregon	269	155.84	9.08	1.73
Washington	635	402.80	11.64	1.58

Table 4: Comparison with Observed Duplicates Without Middle Initial

	Analytic Mean	Std Dev	Observed Dups
Oregon	44.687	6.683	7801
Washington	44.690	6.684	7800

Table 5: Across State Expected Duplicates Including Middle Initial

	Analytic Mean	Std Dev	Observed Dups
Oregon	520.926	22.774	9892
Washington	521.354	22.802	9891

Table 6: Across State Expected Duplicates Without Middle Initial

6.4 The Collision Problem

When we consider the expected number of random agreements across the two states, we see again that by omitting the middle initial increases the number of expected random agreements by an order of magnitude. However, we now see a very large number of observed duplications compared to the expected number. In the case of having the middle initial present, the expected random agreements represent only around 0.5% of the observed duplications, indicating the preponderance of these observed duplicates represent voters registered in both states (Table 5). When we drop the middle initial we increase the number of observed duplicates by more than 2000, although a sizable portion of these new duplicates may be random (Table 6).

In any case, one concludes that the individual states do a thorough job of updating their own voter registration records, so that when someone moves within the state, the voting registration is revised without creating a duplicate registration. However, when someone moves from one state to the other, the first state often continues to carry the old registration on the books.

References

Gregory M. Constantine. *Conbinatorial Theory and Statistical Design.* John Wiley and Sons, New York, New York, 1987.

Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics.* Addison-Wesley, Upper Saddle River, NJ, second edition, 1994.

Michael P. McDonald and Justin Levitt. Seeing double voting: An extension of the birthday problem. *Election Law Journal*, 7(2):111–122, 2008.