PERSON DUPLICATION IN THE CENSUS 2000 DRESS REHEARSAL

John Jones and Danny Childers
U. S. Census Bureau, Washington DC 20233[1]

Keywords: Integrated Coverage Measurement, Dual System Estimation

1. Introduction

Census 2000 procedures were rehearsed in three sites during 1998: Sacramento, California; the Menominee Indian Reservation in Wisconsin; and the Columbia, South Carolina area. In each location, after the Census was taken, an independent enumeration of sampled block clusters was performed for the purpose of census coverage measurement. During the Dress Rehearsal, this process was called Integrated Coverage Measurement (ICM). The people and housing units contained in this independent enumeration is known as the P-sample. People and housing units from the census that are counted in the sampled block clusters are called the E-sample. Both the P-sample and the E-sample contain within sample person and housing unit duplication. This duplication is examined with emphasis on E-sample person duplication.

The data used is derived from the ICM matching at all three sites of the Dress Rehearsal; there are E-sample duplicates at each site and each duplicate has exactly one individual that it duplicates. Individuals that have duplicates are called primary persons. Duplicates are identified by clerical match code during matching and subsequent analyst verification. Person duplication is analyzed by the post-strata gender, race, age, and housing tenure to see if duplication occurs more regularly in one group than in another group and to see if we can accurately predict duplication by examining personal characteristics. We attempt to associate duplication with gender, race, age, and housing tenure and search for significant relationships.

Section 2 discusses the methods used to analyze the data. Section 3 examines the frequency of duplication within various post-strata while Section 4 discusses the percentage agreement between primary persons and duplicates on these same post-strata. Section 5 examines the randomness of E-sample duplication while Section 6 states the final conclusions of the paper.

2. Methodology

The database used consists of all census persons at each site before subsampling. Person duplication at each site is analyzed by four post-strata: gender, age, race, and housing tenure. At each site, there are individual records with missing values of post-strata so that each post-strata variable has a level called missing. The age variable has four non-missing levels (ages 0-17, ages 18-29, ages 30-49, and ages 50 and over). The race variable has two non-missing levels in Menominee and South Carolina. At these sites the levels are White and Other where all nonWhites have been collapsed into the Other category. In Menominee the Other category is primarily American Indian while in South Carolina this Other category is primarily African-American. Sacramento has four non-missing racial categories: White, Black, Asian, and Other where the Other category is primarily Hispanic.

The frequency of duplication is determined by taking the ratio of the number of persons duplicated to the total number of people in sample, both in total and for each level of post strata. These ratios are computed for each site and then converted to percentages. Standard errors of these percentages are calculated using the software package VPLX, which uses replication methods to calculate variances of estimates derived from complex surveys as described in Fay (1990). Once these frequencies and their standard errors are determined, within post-strata comparisons are made at each site to check for significant differences in the frequency of duplication.

These comparisons are made using critical values of t-statistics. These critical values are determined using a multiple comparison of means technique with a Bonferroni adjustment, as described in Hocking (1986). The technique allows the overall type 1 error probability to be .10 for a family of tests at a given site. For example,

[1] John Jones and Danny Childers are mathematical statisticians in the Decennial Statistical Studies Division of the U. S. Census Bureau. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official U. S. Census Bureau publications. Research results and conclusions expressed are those of the authors and do not necessarily indicate concurrence by the U. S. Census Bureau. It is released to inform interested parties of current research and to encourage discussion.

the age variable at a given site has four non-missing levels so that pairwise comparison of these levels results in six (4 choose 2) different comparisons. The Bonferroni adjustment reduces the significance level of each individual test so that the overall type 1 error for the entire family of tests is .10. Critical values of t are based on this reduced significance level. Absolute values of observed t statistics are reported and compared with critical values of t. Only comparisons of non-missing levels of post-strata variables are deemed to be important.

Duplicate persons are identified by name and often live in duplicate housing units. To investigate the agreement of duplicate persons with those persons being duplicated (otherwise known as primary persons) on gender, race, age, and housing tenure; we use the subset of the E-sample database consisting of primary persons and their associated duplicates. A new database is formed with each individual record consisting of the primary name, gender, age, race, and tenure linked with the duplicate name, gender, race, and tenure. Some primary persons have more than one duplicate; when this happens a separate record is created for each primary-duplicate pair.

To analyze the randomness of duplication, we make a comparison of percentages within levels of post-strata. For every level of post-strata, we compare its percentage occurrence among the duplicates with its percentage occurrence among the nonduplicates and check for significant differences. Again, standard errors of these percentages are calculated via VPLX. If there are significant differences, then the characteristic (level of post-strata) in question is said to be more (or less) likely to occur among duplicates than among nonduplicates. The characteristic is then not a random occurrence. The critical value calculation and subsequent hypothesis testing proceed in a manner analagous to that used to analyze duplication frequency.

3. The Frequency of E-sample Duplication and a Description of the Duplicates

Table 1 gives the unweighted percentage of duplicates that are in each site and the associated standard errors of these percentages. Menominee has the highest percentage of duplicates but it has the smallest population.

Table 1: Percentage of Duplicates in E-sample

Site	Percentage	Standard error
SC	1.26	0.13
Sac	1.02	0.10
Men	4.22	0.83

Table 2 gives the unweighted percentage of males and females that are duplicated at each site. The percentage that is missing gender is also given. (Standard errors are in parentheses). For each site, only the male-female comparison is made and the critical value for each male-female comparison is 1.65. Neither Menominee (t=0.49), South Carolina (t=0.57), nor Sacramento (t=0.85) exhibit significant gender difference in duplication frequency. While Menominee has person records that are missing gender, none of these individuals happen to be duplicated.

Table 2: Percentage of Gender that are Duplicated

	Male	Female	Missing
SC	1.31(0.14)	1.37(0.14)	0.22(0.10)
Sac	1.11(0.12)	1.03(0.12)	0.32(0.14)
Men	5.21(1.43)	6.05(1.14)	0.00(0.00)

Table 3 gives the duplicate percentage of each age group. There are six comparisons made at each site, implying that the critical value of t at each site is 2.39. Neither South Carolina nor Menominee exhibit significant age difference in duplication frequency. Note that while there are person records with missing age in Menominee, none of these individuals happen to be duplicated. The only significant difference is between persons aged 30-49 and persons aged 50 and over in Sacramento (t=2.5).

Table 3: Percentage of Age Grouping that are Duplicated:

Age	SC	Sac	Men
0-17	1.23(0.18)	1.10(0.19)	6.16(1.68)
18-29	1.44(0.24)	1.20(0.20)	5.88(2.22)
30-49	1.37(0.15)	1.20(0.15)	5.36(1.19)
50+	1.45(0.18)	0.84(0.12)	5.40(1.62)
Missing	0.18(0.08)	0.32(0.12)	0.00(0.00)

Table 4 gives duplicate percentage of race in both Menominee and South Carolina while Table 5 gives the duplicate percentage of race in Sacramento (Standard errors are in parentheses). In Menominee and South Carolina only one comparison is made so that the critical value of t is 1.65. In Sacramento six racial comparisons are made so that the critical value of t is 2.39. At each site the observed t value for each comparison is less than the critical value, meaning that there are no racial differences in duplication frequency .

Table 4: Percentage of Race that are Duplicated: South Carolina and Menominee

Site	White	Other	Missing
SC	1.45(0.18)	1.19(0.16)	0.29(0.10)
Men	6.08(2.26)	10.81(7.26)	3.58(0.97)

Table 5: Percentage of Race that are Duplicated: Sacramento

White	Black	Asian	Other	Missing
1.16 (0.15)	1.09 (0.22)	0.78 (0.21)	1.17 (0.20)	0.44 (0.15)

Table 6 gives the percentage of housing unit owners and renters that are duplicates at each site. There is one comparison made at each site so that the critical value of t is 1.65. Menominee (t=0.20) and South Carolina (t=1.18) have no significant difference in duplication frequency between owners and renters. In Sacramento (t=2.90), there is strong evidence that more duplication occurs among renters than among owners. There are person records with missing housing tenure at each site, however, none of those persons are duplicated.

Table 6: Percentage of Housing Tenure that are Duplicated

	Owner	Renter	Missing
SC	1.00(0.21)	1.13(0.29)	0.00(0.00)
Sac	0.80(0.11)	1.37(0.18)	0.00(0.00)
Men	5.92(1.71)	4.69(2.98)	0.00(0.00)

4. Comparison of Characteristics for the Linked Primary-Duplicate Pair

Next we describe the duplicates by examining the extent of agreement on the post-strata between the duplicates and those persons who are duplicated (primary persons). Here, we use the database of linked primary-duplicate pairs. Table 7 gives the percentage agreement between primary persons and duplicate persons at each site. With the exception of age and race in Sacramento and housing tenure in South Carolina, there is at least 89% agreement between primary persons and duplicate persons on these variables.

Table 7: Percentage Agreement between Primary persons and Duplicates on Post-Strata

Post Strata	SC	Sac	Men
Gender	92.9	92.9	97.0
Age	91.1	88.6	94.0
Race	95.8	85.9	89.6
Tenure	78.0	89.9	92.5

Tables 8, 9, and 10 give cross-classifications of age in Sacramento, race in Sacramento, and tenure in South Carolina for the linked primary-duplicate database. These are the three site-variable combinations which have less than 90% percentage agreement. The rows are levels of post-strata for the primary person while the columns are levels of post-strata for the duplicates. Each table shows that the numerous missing data fields contribute most heavily to the low agreement percentage.

Table 8: Cross Classification by Age in Sacramento

	0-17	18-29	30-49	50+	Miss
0-17	103	0	4	0	1
18-29	1	62	4	1	2
30-49	2	4	121	6	4
50+	2	2	5	66	5
Miss	0	0	0	2	0

Table 9: Cross Classification by Race in Sacramento

	White	Black	Asian	Other	Miss
White	174	1	1	6	2
Black	4	47	0	2	0
Asian	0	0	35	4	5
Other	4	0	1	85	12
Miss	4	0	2	3	5

Table 10: Cross Classification by Tenure in South Carolina

	Owner	Renter	Missing
Owner	246	13	65
Renter	14	130	14

5. **Randomness of Duplication**

To learn which characteristics that duplicates are likely to have, we compare the percentage of the occurrence of that characteristic (level of post-strata) among the duplicate population with that same percentage among the nonduplicate population and check for significant differences.

Table 11 compares the percentage of duplicates that are male and female with the percentage of nonduplicates that are male and female. There are persons with missing gender that are duplicates and persons with missing gender that are nonduplicates but they are not the subject of this study. At each site there are two comparisons made: the percentage of duplicates that are male with the percentage of nonduplicates that are male and the percentage of duplicates that are female with the percentage of nonduplicates that are female. The critical value of the t-statistic is 1.96. For females, the observed value of t is less than 1.96 in absolute value at each site, meaning that females occur equally among duplicates and nonduplicates. For males, Menominee (t=1.50) and South Carolina (t= 0.81) exhibit no significant difference. However, in Sacramento (t=2.00) a significantly larger percentage of duplicates than nonduplicates are male. This does not imply that a significantly smaller percentage of duplicates than nonduplicates are female because of the existence of missing gender fields.

Table 11: Gender Percentage of Duplicates and Nonduplicates

	Female		Male	
	%dup	%ndup	%dup	%ndup
SC	50.81 (4.52)	52.29 (0.49)	42.74 (4.43)	46.27 (0.52)
Sac	47.37 (2.52)	48.58 (0.33)	49.74 (2.40)	44.84 (0.30)
Men	52.24 (6.41)	37.84 (2.64)	46.27 (6.52)	35.81 (2.38)

Table 12 gives the percentage of duplicates and nonduplicates that are in each age category. At each site there are four comparisons made, each comparison is made at an alpha level of .025 and the corresponding critical t value is 2.23. Menominee and Sacramento have no significant differences while South Carolina has a significantly higher percentage of duplicates than nonduplicates over the age of 50 (t=5.10).

Table 12: Age group Percentage of Duplicates and Nonduplicates

	Ages 0-17		Ages 18-29	
	%dup	%ndup	%dup	%ndup
SC	22.58 (4.32)	25.27 (1.05)	8.87 (3.27)	14.19 (1.01)
Sac	26.32 (3.75)	24.90 (0.75)	17.11 (2.74)	14.98 (0.46)
Men	31.34 (7.12)	23.13 (3.69)	11.94 (4.54)	8.41 (1.52)
	Ages 30-49		Ages 50+	
	%dup	%ndup	%dup	%ndup
SC	23.39 (4.33)	29.53 (0.64)	37.90 (6.33)	28.22 (1.25)
Sac	32.11 (2.39)	28.45 (0.48)	18.68 (2.55)	23.45 (0.87)
Men	22.39 (4.06)	17.41 (2.11)	32.84 (6.36)	24.11 (2.40)

Table 13 gives the percentage of duplicates and nonduplicates in each racial category for Menominee and South Carolina. At these sites there are two comparisons made so that the critical value of t is 1.96.

Table 14 gives the percentage of duplicates and nonduplicates in each racial category for Sacramento. Here, there are four comparisons made so that the critical value of t is 2.23. There are no significant differences in the occurrence of each race among duplicates and nonduplicates at each site.

Table 13: Racial Percentage of Duplicates and Nonduplicates in South Carolina and Menominee

	Other		White	
	% dup	% ndup	%dup	%ndup
SC	33.87 (8.27)	41.09 (3.09)	64.52 (8.37)	57.21 (3.08)
Men	2.99 (3.21)	2.30 (0.68)	32.84 (9.81)	18.00 (4.23)

Table 14: Racial Percentage of Duplicates and Nonduplicates in Sacramento

	%dup	%ndup
White	46.05 (4.41)	39.32 (1.60)
Black	12.63 (3.15)	13.17 (0.68)
Asian	10.00 (3.03)	15.04 (0.99)
Other	23.16 (3.28)	22.28 (0.87)

Table 15 gives the percentage of duplicates and nonduplicates that are owners and renters. There are two comparisons in each site so that the critical value of t is 1.96. In Menominee and South Carolina there are no significant differences for either owners or renters. In Sacramento, there are significant differences for both owners and renters (t=2.41 and t=3.44 respectively). A significantly lower percentage of owners and a significantly higher percentage of renters are duplicates at this site.

Table 15: Tenure Percentage of Duplicates and Nonduplicates

	Owners		Renters	
	%dup	%ndup	%dup	%ndup
SC	64.52 (4.74)	59.54 (2.41)	35.48 (4.74)	34.56 (2.43)
Sac	40.53 (4.83)	51.45 (1.82)	59.47 (4.83)	43.68 (1.71)
Men	76.12 (10.96)	53.22 (5.52)	23.88 (10.96)	23.35 (8.48)

6. Conclusions

Duplicates are identified by name, and they generally agree on race, gender, age, and housing tenure. Data capture problems in the form of missing data fields prevent more substantial agreement between primaries and duplicates, although there are examples of disagreement on post-strata.

Duplication occurs among both genders, all races, all age strata, and with both owners and renters. There is a significantly higher percentage of persons aged 30-49 than persons aged 50 and over that are duplicated in Sacramento. However, this does not occur at the other sites. There is a significantly higher percentage of renters that are duplicated than owners that are duplicated in Sacramento. Again, there are no significant tenure differences at the other sites. Because significant differences do not occur at each site for a given variable, we cannot conclude that there is always more duplication in one level of post-strata than in another level of post-strata.

Similarly, there are examples of significant differences in the percentage occurrence of a characteristic between duplicates and nonduplicates. In Sacramento, a relatively higher percentage of duplicates than nonduplicates are renters and a relatively lower percentage of duplicates than nonduplicates are owners. Also, Sacramento has a relatively higher percentage of duplicates that are male than nonduplicates that are male. South Carolina has a relatively higher percentage of duplicates than nonduplicates over the age of fifty. However, these differences do not repeat themselves at all sites. Therefore, we cannot conclude that there are significant differences in the percentage occurrence of a characteristic (level of post-strata) among duplicates and that same percentage among nonduplicates. The significant differences that do occur are site-specific.

It appears that there is a limited amount to be learned about duplication from examining post-strata alone.We need to investigate the relationship of duplication to census operations to learn about the causes and consequences of duplication.

7. References

Fay, Robert (1990) "VPLX: Variance Estimates for Complex Samples," *Proceedings of the Section on Survey Research Methods,* American Statistical Association.

Hocking, RR (1986) *Methods and Applications of Linear Models: Regression and the Analysis of Variance* (New York: John Wiley and sons), pp 108-9.