A Study of Mass Imputation in Small Area Estimation[1]

Nancy Robbins and Richard A. Moore, Jr.
US Census Bureau, Washington DC, 20233-8700

Key Words: Mass Imputation, Small Area estimation

Introduction

The 2002 Survey of Business Owners (SBO), formerly known as the Survey of Minority Business Owners (SMOBE), is a survey conducted every five years in conjunction with the Economic Census. It publishes information on the aggregate number, receipts, payroll, and employment of minority-owned sole proprietorships, partnerships, and corporations. It is the most comprehensive survey source of basic economic statistics on businesses owned by people of Black, Hispanic, Asian, Hawaiian-Pacific Islander, or American Indian-Native Alaskan ancestry. (Hawaiian-Pacific Islanders identified as a category distinct from Asians was begun with the 2002 SBO). The sample is designed to produce reliable estimates by state (50 states plus the District of Columbia) and by industry (2002 SBO at the 3-digit North American Industrial Classification System (NAICS) level, formerly SMOBE at the 2-digit Standard Industrial Classification (SIC) level). The universe includes all private non-farm businesses operating in the U.S. The survey is designed to yield coefficients of variation (CVs) for the estimates of less than 20 percent for the number of firms and less than 25 percent for sales, employment, and payroll (the higher CV being due to more variability in these variables).

Requests, however, are often received for estimates at more detailed ethnic, geographic (particularly county or metropolitan statistical area (MSA)) and/or industrial levels. Direct survey estimators for these smaller areas fail because the sample is not designed to be representative at this level of detail. Direct small area estimates are usually inaccurate, misleading, and sometimes even unreasonable. They are also likely to produce estimates with standard errors large enough to render the estimate meaningless. This paper discusses two methods of small area estimation under study to provide better estimates at these finer levels of detail. This study used 1997 SMOBE data for three states (Delaware, New Jersey and Illinois as representative of small, medium and large states, respectively) in the analysis.

Current Process

For the 1997 SMOBE, a probabilistic sample of about 2.5 million businesses was selected from a universe of nearly 21 million companies using a stratified random sampling scheme, with sampling rates varying between strata in order to meet the prescribed CV targets. The selected cases were canvassed for race, Hispanic ethnicity, and gender. Since it is known that minority-owned businesses constitute a relatively small percentage (about 10 percent) of the total universe, it is desirable to pre-identify as many "potentially" minority-owned businesses as possible, thus holding down the sample size and cost while retaining relatively accurate estimates. A frame construction plan was developed which used several different sources of race information. For the 1997 SMOBE, fourteen sources were used to infer a minority race/ethnicity/gender to any business exhibiting an indication of being minority or female owned. This race/ethnicity/gender inference was assigned to each record as well as a probability that the inference was correct. Since some of the sources were extremely good predictors of the race of the owner (e.g., the firm's response to the previous SMOBE), while others provided only marginal evidence of minority-ownership (e.g., inferences based on the distribution of minority-owned businesses in the previous SMOBE), the distribution of the probabilities was bimodal. The minority inference along with state and industry classification was used in the stratification.

Following stratification, firms with exceptionally higher receipts than the other cases in their stratum were identified and designated to be self-representing or "selected with certainty". It was deemed that these cases were so large that their responses would significantly influence the estimates of receipts, payroll and employment. In addition, the magnitude of the receipts was large enough that directly expanding the response by the inverse of the sampling weight would greatly increase the variance of these variables, unless that sampling weight was 1.00.

All selected cases were mailed a questionnaire and approximately 82 percent of mailed cases responded. All respondents were assigned a set of tabulation weights, i.e., the sampling weights adjusted for non-

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.

responses. Summing up the tabulation-weighted variables over the responding cases yielded the direct survey estimates for the number of firms, total receipts, number of paid employees, and annual payroll. Variances of the estimates were calculated using the method of random groups [Wolter, 1985]. The firms selected with non-certainty were randomly divided into k=10 groups. Weights were adjusted and domain estimates were calculated for each group 1 to k, and the variance of each of these k independent estimates was calculated and further divided by k (using the principle of the Central Limit Theorem) to obtain a variance for the full sample domain.

For small area estimates derived from 1997 SMOBE responses, direct estimation is possible, and the estimates for disjoint small areas or groups add up to the larger area total. The variance estimates do not have this property, however, due to the method of random groups. Another disadvantage is that the sample is representative only at the stratum level; biases for estimates and variances occur when small areas are either over-represented or under-represented in the sample.

Historically, it has been found that each minority category owns no more than 10 percent of the businesses in any specific geographic or industrial cell. This proportion can be higher for some industries in metropolitan areas, but it rarely exceeds 50 percent. We compared the direct small area estimates with the universe; if the estimate was greater than half of the universe, it was flagged as a potential problem. We first established a baseline at the state by 2-digit SIC level where a comparison of the estimates for number of firms and receipts uncovered **no** cells where over half the businesses were owned by any single minority group. Then we repeated this comparison at the state by county by 4-digit SIC level. The results (Table 1) show that there are a handful of cells flagged in all categories. This gives some indication that a small percentage of small area cells may be overestimated.

Table 1 - Number of Small Areas for which Direct Estimation Estimate is greater than 50% of Universe, Total by State by County by 4-digit SIC (Substrata)

State	Number of Substrata	Variable	Race				
			Black	Asian	Indian	Hspanic	Total
Delaware	1,754	Firms	18	15	2	8	43
		Receipts	10	12	3	7	32
Illiinois	33,464	Firms	87	114	60	97	358
		Receipts	161	124	50	64	398
New Jersey	13,762	Firms	102	184	10	154	450
		Receipts	25	157	10	91	282
Total	48,980	Firms	207	313	72	259	851
		Receipts	195	293	63	161	712

Small Area Estimation

Small area estimation literature suggests a wide range of estimation techniques. We focused on two methods, both of which employ creating a complete rectangular file for tabulation. The first method is based on a binomial model, while the second uses mass imputation. It is important to note that the race/ethnicity fields are the response variables we seek from the sample; the receipts, employment and payroll information are all available for the universe from administrative records.

Binomial Model

The Binomial model uses the assignment of the probability of belonging to a specific race/ethnicity to each unit in the universe which was referred to in the sample design; the assigned probability of responding to a particular race/ethnicity is binomial. For respondents this probability becomes a one or zero, while non-respondents and non-selected cases keep their assigned probability between zero and one based on the source of the inference. In this method the sum of the probabilities are raked to the estimated total number of minority-owned firms in the strata, i.e.,

Estimate = $\sum_{i=1}^{N} p_i$ where p_i = raked probability;

for the other variables, Estimate = $\sum_{i=1}^{N} p_i * variable_i$.

The variance is calculated using the binomial variance formula: Since the assigned probability is the estimated chance of responding to a particular race/ethnicity, the variance of this estimate should approximate the variance of a set of independent binomial trials, namely, for firm counts the

Variance = $\sum_{i=1}^{N} p_i * q$, where p_i = the raked probability and $q_i = 1 - p_i$. For the receipts, payroll, and employment variables,

Variance= $\sum_{i=1}^{N} p_i * q_i * variable^2$.

In practice, we found that this variance is severely underestimated. We found we had to inflate the variance by a factor of

$$\frac{N-n}{n} * \left[1 + \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} p_i q_i} \right] ,$$

where $e_i = p_i - \bar{p}$ and $\bar{p}$ = *average probability over all N cases*.

The probability of responding to a particular race/ethnicity is a bimodal distribution, and for bimodal distributions the summation in the inflation factor can be substantial. (The mathematics of this inflation factor is in the Appendix).

Mass Imputation

Mass imputation imputes the missing parts of the non-sampled primary units to create a complete set of responses for every unit in the universe. It is implicitly equivalent to an expansion estimator with variable weights [Kovar and Whitridge, 1995]. In this study, the responses of those cases that are selected for sampling and respond are naturally used. For those cases that are selected but do not respond and for those not selected, the donor-based method of imputation is employed; that is, a respondent with similar known characteristics as the non-respondent is identified and its responses are donated to the non-respondent. The random group is donated as well so that the method of random groups is used to calculate variances.

In SMOBE, each response was assigned a tabulation weight, usually a non-integer. For the mass imputation process, these weights were rounded to integers in order to preserve the published marginal firm count totals. Thus a respondent with integer weight n would donate its set of responses to n-1 non-responding or non-selected cases within its stratum and with similar characteristics according to the administrative data. In theory, the strata were the state by 2-digit SIC blocks; in practice, these strata were too large and required excessive computer time to match records. (Illinois used over twenty-two hours). The strata were modified to include MSA and the race and sex of the business owner as provided by information from the Social Security Administration. Within these strata, matches were made based on agreement with frame, legal form of organization (LFO), whether the business was an employer, and number of employees. A second iteration of matches based on strata of state and 2-digit SIC was then performed. Not all donors were matched to recipients after these two iterations, but a third iteration did not significantly change the results and so was dropped.

Results

One of the difficulties with any technique of small area estimation is how best to evaluate the results. Since the true race/ethnicities are unknown, it is not possible to calculate an error term with which to assess the various results. Comparisons of the estimates at the lower levels of detail are an inconclusive proof of one method's superiority over another. One comparison, however, that we required to be met was that the estimates of the small area methods by race/ethnicity by state by 2-digit SIC should be close to the direct estimates of the design-based sample, since the survey was specifically designed to be valid at these levels. The following chart (Table 2) compares the three methods at this race/ethnicity by state by SIC level. One can easily see that both estimators perform about equally well for the number of firms; the mass imputation estimator outperforms the binomial model noticeably better for all race/ethnic groups and all states in estimating total receipts.

Table 2 - Percentage of Estimates at Race/Ethnicity by State by 2-digit SIC which are within 95% Confidence Interval of Design Estimates

Race/Ethnicity	State	Number of 2-digit SIC	Firms		Receipts	
			Binomial	Imputation	Binomial	Imputation
Black	Delaware	47	100.0%	100.0%	59.6%	85.1%
	New Jersey	60	100.0%	98.3%	43.3%	73.3%
	Illinois	58	100.0%	96.6%	50.0%	63.8%
Asian	Delaware	37	97.4%	97.4%	60.5%	86.8%
	New Jersey	63	100.0%	98.4%	61.9%	88.9%
	Illinois	62	100.0%	100.0%	54.8%	83.9%
American Indian	Delaware	23	100.0%	100.0%	60.9%	87.0%
	New Jersey	37	100.0%	100.0%	40.5%	86.5%
	Illinois	55	100.0%	100.0%	40.0%	78.2%
Hispanic	Delaware	34	100.0%	100.0%	58.8%	76.5%
	New Jersey	65	96.9%	96.9%	50.8%	73.8%
	Illinois	63	100.0%	98.4%	47.6%	79.4%

The binomial method consistently overestimates the volume of receipts. One reason may be that the probabilities were calculated independently of the receipts variable; if they are not independent, this problem could result. When we compare the estimates of the three methods for employment size ranges, we can see that although the firm count distributions for the methods are approximately the same, the binomial model overestimates the receipts for many of these cells. For the lower employment ranges, the receipts estimates of the binomial are within tolerance of the design-based estimates, based on the corresponding CV's of the design-based estimates. However, at the higher ranges,

the estimates begin to differ more severely. Table 3 illustrates this result for New Jersey; the other states exhibit a similar problem. This leads to the conjecture that while raking of probabilities may provide reasonably accurate estimates for receipts, payroll, and employment of small businesses, it leads to severe overestimation when used on large businesses.

The above results tend to suggest that the mass imputation method outperforms the binomial model when estimating receipts. The most important question remains: How well do the models estimate firm counts for the small areas of interest, such as county? Delaware has only three counties and thus is the easiest state to compare. The following graphs compare the distributions of number of firms for each race/ethnicity in Delaware (Table 4). Since the survey is designed at the state level, we do not necessarily expect that the design-based county distributions are correct, but it is interesting to note any large divergences. The distributions of the number of black-owned businesses are quite close for all three estimation procedures, while the distributions for American Indian-owned firms are not; but it should be noted that the variability of the state-level estimates is much higher in Delaware for the American Indian-owned firms than for the other three groups: a 29 percent relative standard error for American Indian-owned firms versus 6-13 percent RSE for the others. The distributions for Asian-owned and Hispanic-owned firms show some variability; both small area methods for New Castle and Sussex counties are closer to each other than they are to the design-based estimate.

Note that for Asian-owned and Hispanic-owned businesses, the small area estimation methods for the number of firms using the binomial model and mass imputation are more closely distributed in both New Castle and Sussex counties. For Asian-owned firms, the small area estimates are nearly 10% lower in New Castle and 10% higher in Sussex, and Hispanic-owned firms show a similar pattern.

It is also interesting to note that the estimates of these small area models are more precise than the design-based estimates. Virtually all estimates for the number of firms have a CV less than or equal to the direct estimates; the binomial estimate for American Indians firms in Kent county is the only exception. And for receipts, the mass imputation estimate is more precise for all groups except Sussex county Asian-owned businesses. These results also hold for the other states as the following chart shows (Table 5). Note that once again, with respect to receipts, the binomial model lags significantly behind mass imputation.

Table 5 - Percentage of Estimates with a CV less than or equal to that of the Direct Estimate by Race/Ethnicity by State by County

Race/Ethnicity	State	Number of Counties	Firms		Receipts	
			Binomial	Imputation	Binomial	Imputation
Black	Delaware	3	100%	100%	100%	100%
	New Jersey	21	95%	100%	43%	67%
	Illinois	19	89%	100%	26%	84%
Asian	Delaware	3	100%	100%	0%	67%
	New Jersey	21	100%	100%	43%	81%
	Illinois	19	89%	95%	32%	84%
American Indian	Delaware	3	67%	100%	0%	100%
	New Jersey	21	76%	86%	62%	67%
	Illinois	19	63%	74%	42%	68%
Hispanic	Delaware	3	100%	100%	0%	100%
	New Jersey	21	95%	100%	43%	86%
	Illinois	19	95%	100%	21%	79%

Conclusion

The SBO publishes many small area design-based estimates; for the 1997 SMOBE, details by race/ethnicity for county and larger MSA, employment size and LFO were available. Even finer levels of refinement are often provided upon request. It is important, therefore, that these estimates be as accurate and precise as possible. The mass imputation method of small area estimation that is under review seems the most promising. Both methods often yield lower CVs than the design-based estimates. While the binomial method is more efficient in terms of computer time, the mass imputation method leads to more accurate estimates, particularly receipts. This method has not been adopted and further research into the validation of this method needs to continue.

References

Cochran, W. G. (1977). Sampling Techniques. John Wiley and Sons, New York, NY.

Ghosh, M. and Rao, J.N.K. (1994). "Small Area Estimation: An Appraisal". Statistical Science, Vol. 9, No. 1, 55-93.

Kovar, John G. and Whitridge, Patricia J., "Imputation of Business Survey Data" from Cox, Brenda, Binder, David, Chinnappa, B. N., Christiansen, Anders, Colledge, Michael, Kott, Phillip, (1995). Business Survey Methods. John Wiley and Sons, New York, NY.

Moore, R., Caldwell, C., Detlefsen, R. (1998). "Changing the Sample Design to Meet User Needs – The Survey of Minority-Owned Business Enterprises – Past, Present, and Future", Proceedings of the Section on Survey Research Methodology. American Statistical Society, 493-498.

Moore, R. and Williams, A. (1998). "Using Administrative Data to Enhance the Sampling Frame for the 1997 Survey of Minority-Owned Business Enterprises", Proceedings of the Section on Survey Research Methodology. American Statistical Society, 499-504.

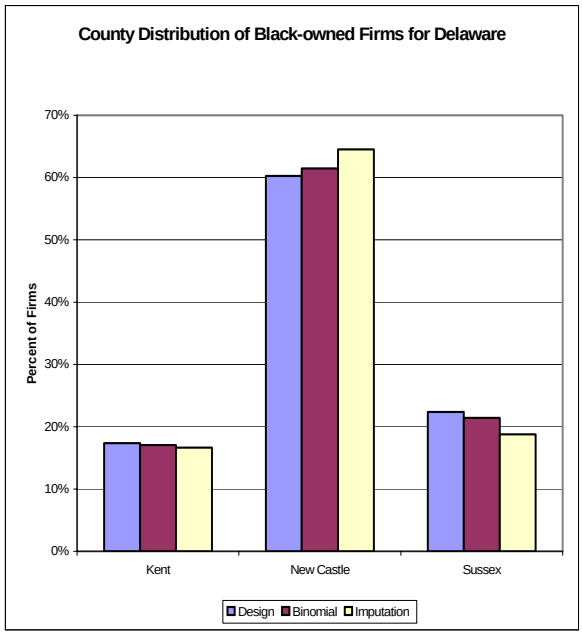
Statistical Policy Working Paper 21, "Indirect Estimators in Federal Programs" (1993).

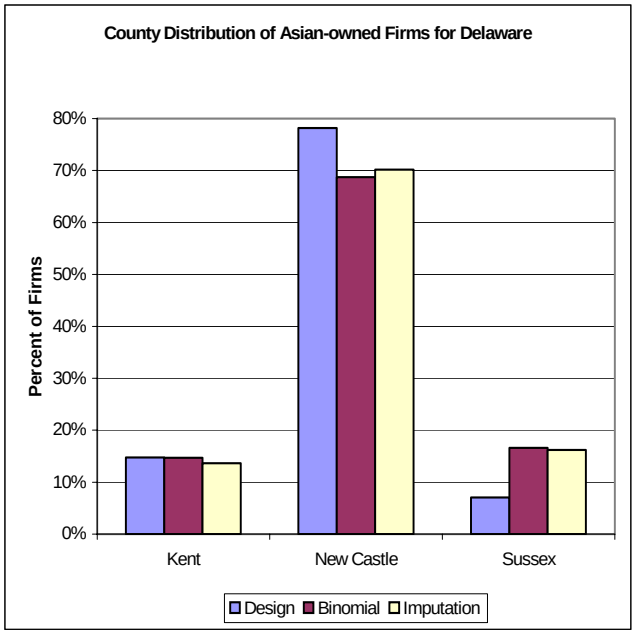
Wolter, K. (1985). Introduction to Variance Estimation. Springer-Verlag, New York, NY.

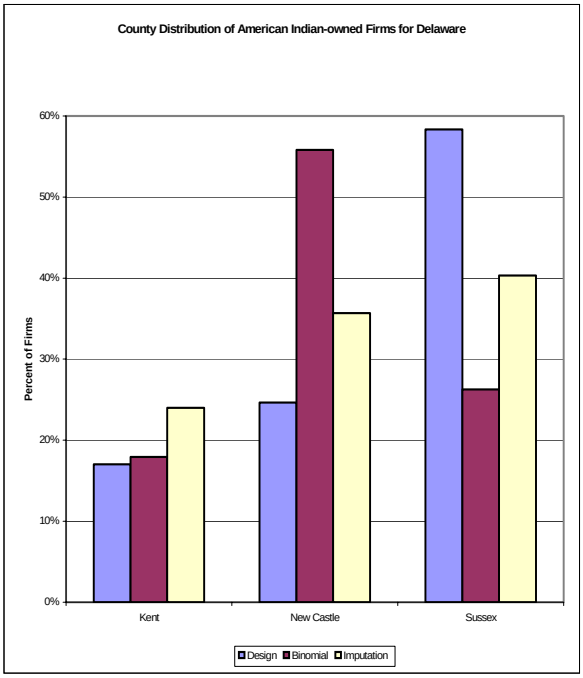
Table 3 - New Jersey Comparisons of Estimates of Employment Size
Note: The "D" in the table represents a cell whose value has been suppressed in accordance with Title 13 disclosure rules.

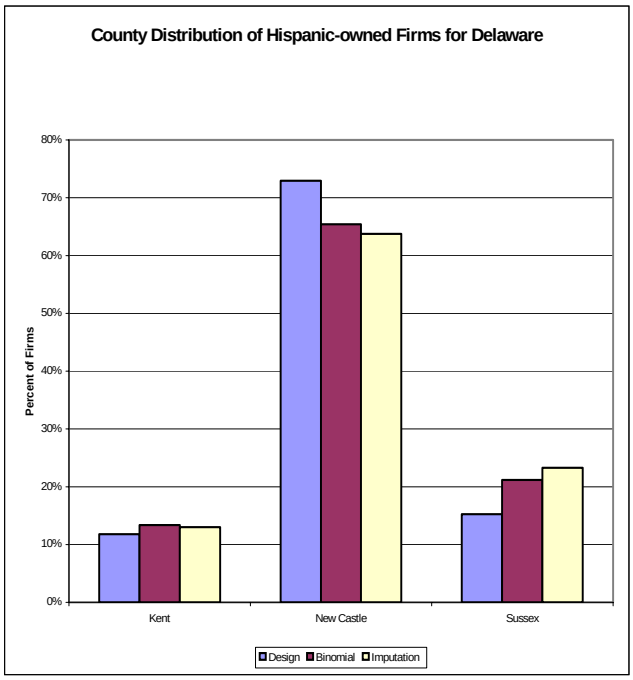
Number of Employees	Number of Firms						Receipts					
	Design-Based		Binomial		Imputation		Design-Based		Binomial		Imputation	
Black-owned	**Estimate**	**CV**	**Estimate**	**CV**	**Estimate**	**CV**	**Estimate**	**CV**	**Estimate**	**CV**	**Estimate**	**CV**
0	23,620	29	23,814	3	22,612	29	634,772	28	701,044	14	862,754	27
1 - 4	2,078	17	1,720	11	1,805	17	378,193	19	462,731	7	459,664	18
5 - 9	487	23	490	22	448	23	291,416	34	390,368	10	345,602	25
10 - 19	169	36	242	32	115	40	210,324	46	361,118	13	160,575	46
20 - 49	77	27	140	41	57	28	209,830	47	437,505	13	211,654	42
50 - 99	41	39	42	66	26	29	233,053	30	328,430	18	328,510	46
100 - 249	22	0	31	62	22	0	162,313	0	321,371	20	162,313	0
250 - 499	6	0	9	110	6	0	28,207	0	126,860	24	28,207	0
500 - 999	4	0	4	144	4	0	12,381	0	65,558	99	12,381	0
1000 - 2499			1	539					D	40		
2500 - 4999			1	652					D	90		
5000 - 9999			1	637					D	29		
10,000 or more			1	451					D	15		
	26,504	3	26,496		25,095	3	2,160,489	8	3,194,985		2,571,660	12
	Design-Based		Binomial		Imputation		Design-Based		Binomial		Imputation	
Asian-owned	**Estimate**	**CV**	**Estimate**	**CV**	**Estimate**	**CV**	**Estimate**	**CV**	**Estimate**	**CV**	**Estimate**	**CV**
0	28,893	28	32,293	2	28,391	28	2,199,823	33	1,694,361	6	1,964,432	30
1 - 4	8,861	17	5,807	6	8,165	17	3,398,309	19	2,619,728	5	3,153,863	18
5 - 9	2,249	24	1,700	12	2,279	25	2,526,851	24	2,050,618	10	2,463,759	26
10 - 19	788	26	889	17	796	32	2,012,886	29	2,060,461	12	1,861,483	26
20 - 49	493	23	511	21	440	24	2,436,558	15	3,132,348	9	2,171,654	12
50 - 99	70	20	142	37	70	24	987,791	7	1,674,434	19	1,143,892	18
100 - 249	47	12	82	46	53	18	982,418	3	1,456,877	18	989,958	4
250 - 499	19	16	24	65	19	17	820,837	2	940,118	14	816,415	2
500 - 999	12	0	15	80	12	0	923,215	0	1,100,527	15	923,215	0
1000 - 2499	8	0	9	111	8	0	407,972	0	654,427	39	407,972	0
2500 - 4999	3	0	4	175	3	0	D	0	D	167	D	0
5000 - 9999	1	0	2	194	1	0	D	0	D	100	D	0
10,000 or more			1	546					D	47		
	41,444	3	41,479		40,237	3	16,696,660	6	17,383,899		15,896,643	5
	Design-Based		Binomial		Imputation		Design-Based		Binomial		Imputation	
Indian-owned	**Estimate**	**CV**	**Estimate**	**CV**	**Estimate**	**CV**	**Estimate**	**CV**	**Estimate**	**CV**	**Estimate**	**CV**
0	1,994	14	2,143	9	1,927	19	65,504	19	87,697	14	93,034	31
1 - 4	282	25	192	18	254	31	91,183	26	85,806	9	124,072	36
5 - 9	49	45	51	36	60	49	18,892	46	52,606	22	27,806	57
10 - 19	30	42	32	42	29	48	32,128	38	62,123	19	85,419	69
20 - 49	14	64	17	53	12	57	91,106	55	102,138	14	82,759	50
50 - 99	1	0	3	123	1	0	D	0	60,410	31	D	0
100 - 249	4	68	3	85	4	62	D	65	29,492	86	D	70
250 - 499			0	793					D	321		
500 - 999			0	1000					D	220		
1000 - 2499			0	944					D	402		
2500 - 4999			0	2000					D	509		
5000 - 9999			0	1000					D	140		
10,000 or more			0	781					D	141		
	2,374		2,441		2,287		298,813	6	480,272		413,090	5
	Design-Based		Binomial		Imputation		Design-Based		Binomial		Imputation	
Hispanic-owned	**Estimate**	**CV**	**Estimate**	**CV**	**Estimate**	**CV**	**Estimate**	**CV**	**Estimate**	**CV**	**Estimate**	**CV**
0	30,015	29	31,022	3	29,546	29	1,389,744	43	1,129,543	4	1,506,704	30
1 - 4	4,270	19	3,347	9	3,632	20	1,009,944	21	1,029,854	4	1,123,074	19
5 - 9	1,072	26	944	18	968	30	566,889	26	754,931	7	555,865	24
10 - 19	411	26	479	25	277	34	613,590	26	721,808	9	477,978	30
20 - 49	283	28	286	29	204	27	999,142	13	1,346,089	7	1,031,028	15
50 - 99	44	36	66	51	30	9	243,992	7	471,309	16	241,576	5
100 - 249	19	29	30	74	16	15	256,808	10	474,552	27	269,721	12
250 - 499	6	0	7	142	6	0	D	0	70,928	148	D	0
500 - 999	1	0	2	294	1	0	D	0	D	497	D	0
1000 - 2499			2	334					D	655		
2500 - 4999			1	664					D	150		
5000 - 9999	1	0	2	285	1	0	D	0	D	141	D	0
10,000 or more			1	838					D	73		
	36,122	2	36,189		34,681	5	5,080,109	8	5,999,014		5,205,946	5

Table 4 - Comparison of Distribution of Number of Firms by County by Race/Ethnicity for Delaware









Appendix: Developing a Generalized Variance Formula for the Binomial

Main Result:

The $\sum_{i=1}^{N} p_i * q_i$ underestimates the actual sample variance by a factor of approximately (N–n)/n. There is also an additional underestimation due to the dispersion of the $\{p_i\}$. When the $\{p_i\}$ are bimodal, the dispersion factor may be substantial.

Small Area estimation Variance Inflation Factor = $\frac{N-n}{n} * (1 + DispersionFactor)$

$$= (AvgWt - 1) * (1 + DispersionFactor)$$

Design-Based Variance:

1. Design-based (assuming equal weighting)
   N=Number of units in stratum (state by 2-digit SIC)
   n=number of respondents in stratum (state by 2-digit SIC)

   Unit Var Mean = $\frac{\overline{p} * \overline{q}}{n}$

   FPC = $\left(1 - \frac{n}{N}\right)$

   $$StratumVar_{Design} = N^2 * FPC * UnitVarMean$$

   $$= \frac{N-n}{n} * N * \overline{p} * \overline{q}$$

2. Proposed Binomial Variance:

   $p_i = \overline{p} - e_i$

   $$StratumVar1_{Bin} = \sum_{i=1}^{N} p_i * q_i$$

   $$= \sum_{i=1}^{N} (\overline{p} - e_i) * (\overline{q} + e_i) \qquad (2.1)$$

   $$= (N * \overline{p} * \overline{q}) - (N * Var(e_i))$$

   Multiplying each term in the above equation by a factor of (N–n)/n and rearranging the terms gives:

   $$\frac{(N-n)}{n} * StratumVar_{Bin} + \frac{(N-n)}{n}(N * Var(e_i)) = StratumVar_{Design}$$

If the second term of the left-hand side is negligible, then we need only adjust our original guess for the variance of the binomial estimator by a factor of (N–n)/n. So our new variance estimate for the binomial is:

$$StratumVar2_{Bin} = \frac{(N-n)}{n} * \sum_{i=1}^{N} p_i * q_i \qquad (2.2)$$

In our situation, the likelihoods are bimodal. Hence the variance of the error terms are not negligible and the second term must be included, thus

$$StratumVar3_{Bin} = \frac{(N-n)}{n} * \left[\sum_{i=1}^{N} (p_i * q_i) + N * Var(e_i^2)\right] \qquad (2.3)$$

$$= \left[\sum_{i=1}^{N} p_i * q_i\right] * \frac{(N-n)}{n} * (1 + DispersionFactor)$$