A Hybrid Estimation Approach to State Level Estimates in the Survey of Industrial Research & Development

John Slanta, Census Bureau and Jeri M. Mulrow, NSF

Abstract

The Census Bureau annually conducts the company-based Survey of Industrial Research and Development (R&D) in conjunction with the National Science Foundation (NSF). The survey provides national data on the amount of basic and applied research and development conducted within the United States. In addition, state level estimates of industrial R&D are of great interest. It is estimated that less than 2% of all in scope U.S. based companies perform R&D. A new national sample is selected each year and stability across years in the state estimates is a concern given the rare event nature of R&D. This paper explores the use of an estimator that is a hybrid between a design-based and a model-based estimate that also meets constraints required for data publication. This paper will present the estimation and variance estimation methodologies and show results from the survey.

1.0 Introduction

The Survey of Industrial Research and Development (referred to as the R&D survey) is conducted annually and a new sample is selected each year with a large overlap in R&D performers from year to year. The R&D frame consists of all non-farm companies with employment greater than or equal to five employees that are located in the U.S. and operate for profit. The sampling frame is constructed from the Census Bureau's most recently updated Standard Statistical Establishment List (SSEL) file[1]. The frame size varies year to year from about 1.8 million companies to 1.9 million companies, where company is both the sampling unit and, for all but a handful of cases, is the collection unit. We estimate that less than 2% of the companies in the frame have any R&D expenditures, so this can be characterized as a rare event population.

In the past, the sample was controlled at the survey industry level, hereafter referred to as the recode[2] level. A company, while possibly having activity in more than one industry recode, was assigned to a single recode. The company was selected into the sample based on that recode and all data for that company were tabulated in that recode. For sample allocation purposes, probabilities of selection were assigned based on recode CV constraints.

Due to the increased interest in state-level estimates, beginning with the 2002 survey year, state-level CV constraints were added. In contrast to the situation for industry recode, a company receiving the long form of the survey (Form RD1) that had activity in more than one state had its data tabulated in each of the reported states. Multiple probabilities of selection were assigned to each company, one probability for each recode and state in which a company had activity. Thus, each company could have received up to 52 probabilities (1 recode + 50 states + DC). The largest overall probability became the company probability of selection.

In the 2002 survey year there were approximately 2,300 out of 31,000 companies that received the long form. All of these companies were selected with certainty and they accounted for roughly 84% of the U.S. estimate.

Historically, the survey estimates, were computed using the Horvitz-Thompson (HT) estimator. While the HT estimator is design unbiased, it is sensitive to large weights, particularly in the rare event situation of R&D expenditures. In the past, due to the sampling plan, each year we would observe a few companies having large weights with moderately large values of reported R&D. These companies, in the following year, would receive a weight of one and their weighted contribution to the estimate would dramatically decrease. The weights are the inverse of the probability of selection. The probabilities are assigned each year proportionally within recode to how much R&D the company historically reports, or what we think they might do based on the size of the company. If a company with no historical R&D is selected with a large weight and then reports a moderate to large amount of R&D, then that company could receive a probability of one for the following year. This would cause a spike in the time series for the year that these companies would enter the sample. This scenario would repeat itself for different states in different years, but the total estimate of R&D across all industries and states remained fairly stable over time. However, at the more disaggregated industry or state level, the estimates may have exhibited large year-to-year changes.

Thousands of estimates in over forty tables are produced annually from the R&D survey[3]. The estimates must show additivity within and across all tables. That is, the sum of R&D for all industries, or for all states, must add to the total U.S. level. Another example is federal R&D plus

[1] The Standard Statistical Establishment List is being replaced with the Business Registry in Spring 2004.

[2] Recodes may be one or more combinations of North American Industrial Classification System (NAICS) codes.

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau nor the National Science Foundation.

[3] *Research and Development in Industry, Detailed Statistical Tables*, Division of Science Resources Statistics, National Science Foundation.

company R&D must add to total R&D within any particular aggregate.

Research began in early 2001 to address the volatility observed in the state estimates over time. One result of this research was to implement a new sampling plan in 2002 to take better advantage of historic information about R&D expenditures and to incorporate more overlap of R&D performers in the samples from year to year. Another result of the research was to develop a hybrid estimator with desired characteristics to produce state estimates for the 2001 survey. This hybrid estimator would be robust against large weights, and it could still capture a trend or change when one is present. In addition, back estimates were produced for the years 1998-2000 using the new estimator. This paper presents and explains the research and the final formula for the new state estimator.

2.0 Hybrid State Estimator

The new estimator used to produce state estimates from the R&D survey has the following form:

$$\hat{Y}_S = \sum_{h=1}^{L}\sum_{k=1}^{N_h} a_{hk}\, y_{Shk} + \sum_{h=1}^{L}\sum_{k=1}^{N_h} a_{hk}\left(\sum_{I=1}^{N_I} R_{IS}\left(w_{hk}-1\right) y_{Ihk}\right) \tag{2.1}$$

where

$$R_{IS} = \frac{\sum_{k=1}^{N}\left(1-\pi_k\right) X_{ISk}}{\sum_{k=1}^{N}\left(1-\pi_k\right) X_{Ik}} \tag{2.2}$$

and

N = population size
N_h = population size of stratum h
N_I = number of independent non-aggregate industry publication tabulations
L = Number of sampling strata

$$N = \sum_{h=1}^{L} N_h$$

y_{Shk} = reported or imputed R&D in state S of k^{th} company in stratum h
y_{Ihk} = reported or imputed R&D in industry I of k^{th} company in stratum h
w_{hk} = weight of k^{th} company in stratum h, = reciprocal of probability of selection
a_{hk} = one (1) if k^{th} sampling unit in stratum h is selected and zero (0) otherwise
X_{ISk} = payroll in industry I and state S of k^{th} company, available from the frame
X_{Ik} = payroll in industry I of k^{th} company, available from the frame
π_k = probability of selection of k^{th} company

For a given sample we can also write the estimator in the following form.

$$\hat{Y}_S = \sum_{h=1}^{L}\sum_{i=1}^{n_h} y_{Shi} + \sum_{h=1}^{L}\sum_{i=1}^{n_h}\left(\sum_{I=1}^{N_I} R_{IS}\left(w_{hi}-1\right) y_{Ihi}\right) \tag{2.3}$$

where

n_h is the sample size in stratum h,

Note that:

- the index i need not equal k
- y_{Shk} need not equal y_{Ihk} and X_{IShk} need not equal X_{Ihk}
- $\sum_{S=1}^{51} R_{IS} = 1$, i.e. the sum of the ratios across all states and within an industry sum to one.

Payroll by industry and state is first obtained at the establishment level then rolled up to a company level. Note a company can have payroll in more than one industry or state. The numerator of R_{IS} is the expected value of the payroll of any given state within a given industry from companies that are not selected. The denominator of R_{IS} is the expected value of the payroll of a given industry from companies that are not selected. Companies selected with certainty do not figure in the calculation of R_{IS}.

The estimator itself can be decomposed into two major parts. The first part is the unweighted sum of the reported or imputed R&D in the state of interest. This value is the lower bound of all possible values of the true value given the selected sample. The second part is the portion of the difference between the weighted and unweighted R&D that is allocated to the state.

This decomposition helps to explain how the estimator can be robust against large weights and still can capture changes and trends when one is present. The unweighted sum accounts for 89.7% of the total U.S. estimate for the 2002 survey year. This means about 10%, or less than 20 billion dollars of the 200 billion dollars, of the R&D is allocated to the states. If a weight is high and the R&D is moderate then the weighted portion is spread over 51 states. If a weight is one, or close to one, and a company's reported value changes dramatically from the prior year then the change is going to be reflected in only the state involved.

An additional nice attribute of this estimator is that it may be easily modified to incorporate industry estimates, and it may be expanded to incorporate companies being assigned

multiple industries. Currently, the R&D survey does not take advantage of this feature of the estimator.

3.0 Sample Variance of the Estimator

To facilitate calculation of the sample variance of the state estimator

$$\hat{Y}_S = \sum_{h=1}^{L}\sum_{k=1}^{N_h} a_{hk} y_{Shk} + \sum_{h=1}^{L}\sum_{k=1}^{N_h} a_{hk}\left(\sum_{I=1}^{N_I} R_{IS}\left(w_{hk}-1\right) y_{Ihk}\right) \quad (2.1)$$

note that it can also be expressed in the following form:

$$\hat{Y}_S = \sum_{h=1}^{L}\sum_{k=1}^{N_h} a_{hk} w_{hk} \frac{\left(y_{Shk} + \left[\sum_{I=1}^{N_I} R_{IS}\left(w_{hk}-1\right) y_{Ihk}\right]\right)}{w_{hk}} = \sum_{h=1}^{L}\sum_{k=1}^{N_h} a_{hk} w_{hk} \tilde{y}_{Shk} \quad (3.1)$$

where

$$\tilde{y}_{Shk} = \frac{\left(y_{Shk} + \sum_{I=1}^{N_I} R_{IS}\left(w_{hk}-1\right) y_{Ihk}\right)}{w_{hk}}. \quad (3.2)$$

Note:

$\tilde{y}_{Shk}$ is a constant value for a given *k* and *h* (company and stratum) fixed for a given population.

Now we can simply use the sample variance formula of the HT estimator replacing y_{Shk} with $\tilde{y}_{Shk}$ to estimate the variance for the state estimator.

The stratified Yates-Grundy-Sen[4] sample variance of the HT estimator is:

$$\hat{\sigma}^2\left(\hat{Y}_S\right) = \sum_{h=1}^{L} \hat{\sigma}^2\left(\hat{Y}_{Sh}\right) \quad (3.3)$$

where

$$\hat{\sigma}^2\left(\hat{Y}_{Sh}\right) = \sum_{k=2}^{N_h}\sum_{u=1}^{k-1} a_{hku}\left(\beta_{hku}-1\right)\left(w_{hk}\tilde{y}_{Shk} - w_{hu}\tilde{y}_{Shu}\right)^2 \quad (3.4)$$

and

$$\beta_{hku} = \frac{\pi_{hk}\pi_{hu}}{\pi_{hku}} \quad (3.5)$$

Under Tillé sampling[5] this can be closely approximated to:

$$\hat{\sigma}^2\left(\hat{Y}_{Sh}\right) \approx \sum_{i=1}^{n_h} \gamma_{hi}\left[i\left(\sum_{j=1}^{i}\left(w_{hj}\tilde{y}_{Shj}\right)^2\right) - \left(\sum_{j=1}^{i} w_{hj}\tilde{y}_{Shj}\right)^2 \right] \quad (3.6)$$

where

$$\gamma_{hi} = \begin{cases} 0 & \text{if } i = 1 \\ \beta_{hi1} - \beta_{h(i+1)1} & \text{if } 1 < i < n \\ \beta_{hi1} - 1 & \text{if } i = n \end{cases}$$

If Tillé sampling is used and the probabilities of selection are skewed then, with a few exceptions, the following identity will hold true,

$$\beta_{hij} = \beta_{hi1} \text{ for } j < i$$

and the sampling units are sorted by ascending order of π_{hi} within sampling stratum.

4.0 Research

Two other types of estimation techniques were investigated prior to settling on the use of the hybrid estimator described above.

1) In the past, windsorizing[6] had been used to adjust state estimates when necessary. This methodology required (1) identification of outlier estimates, (2) development of multiple weights, one for state R&D and another one for industry R&D, and (3) a remainder field needed to balance the tables in order to maintain additivity. The procedures for determining the outliers and weights was somewhat ad hoc and varied from year to year depending upon the observed data making it difficult to standardize the procedure. In addition, the remaining field, called Undistributed Funds was not useful analytically.

[4] Cochran (1977)

[5] Slanta and Fagan (1997), Tillé (1996)

[6] Lee (1995)

2) Composite estimators were researched in depth, but two significant problems arose in applying them. The first problem encountered was in deriving a stable estimate for the Mean Squared Error (MSE) of the synthetic estimate. The MSE is an integral part of the composite estimate. If, by chance, the difference between the HT and synthetic estimates are less than one standard error of that difference, then an unbiased estimate of the MSE will be negative even though the true MSE is by definition nonnegative.

The second major problem with the composite estimator was the state estimates would not add to the total R&D at the U.S. level. To try to compensate for this, raking of the composite estimates was tried, but this resulted in many of the state estimates having more variability than the original HT estimator. Using this technique resulted in estimates for the sample variance that were biased.

The composite estimators studied had the form:

$$\hat{Y}_C = W\hat{Y}_{HT} + (1-W)\hat{Y}_S \qquad (4.1)$$

where

$\hat{Y}_S$ is the synthetic estimate and

$$W = \frac{\hat{mse}(\hat{Y}_S) - \hat{\text{cov}}(\hat{Y}_{HT}, \hat{Y}_S)}{\hat{mse}(\hat{Y}_S) + \hat{\sigma}^2(\hat{Y}_{HT}) - 2\hat{\text{cov}}(\hat{Y}_{HT}, \hat{Y}_S)} \qquad (4.2)$$

Note that the mean square error is an integral part of the estimator itself.

As a result of the research, the hybrid estimator described earlier was implemented. This new estimator has many advantages. (1) Unlike the composite estimator, it yields unbiased estimates of sample variance. (2) It maintains additivity within and between the tables no matter what level of aggregation is used, which was an important requirement to the data users. (3) The estimator is relatively easy to implement and (4) the procedures can be standardized from year to year.

5.0 Results of the Research

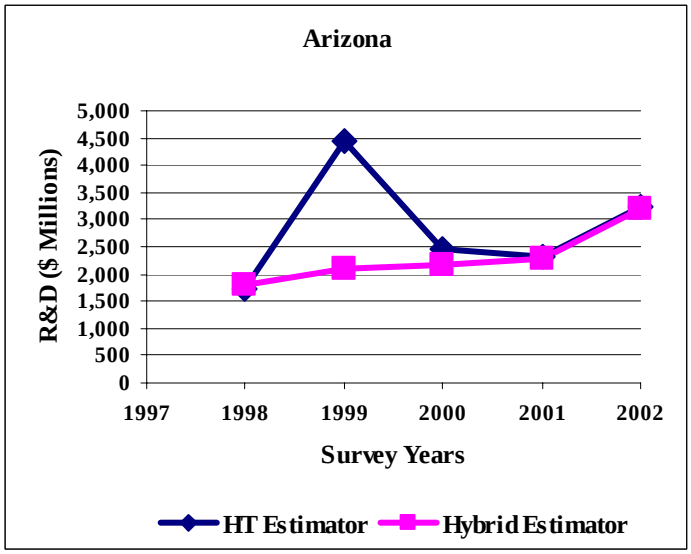
The Survey of Industrial R&D was established by the National Science Foundation Act of 1950, as amended and has been conducted annually since 1953. The Bureau of the Census has conducted the survey in conjunction with NSF since 1957. Data users are not only interested in the current year estimates produced from the survey, but they are also interested in the estimates over time, particularly changes over time. The new sample design implemented in 2002 and alluded to earlier in this paper provides one step towards the process of improving year-to-year estimates. The implementation of the new hybrid state estimator is another step in this process.

Trend estimates using the new hybrid estimator for five years for the 50 states plus the District of Columbia are shown in Table 1. The estimates shown are for total R&D performed in the state. The estimates are rounded to the nearest million, and due to the rounding the sum of states plus undistributed for the years 1998, 2001, and 2002 do not equal the rounded U.S. total. One goal of the research was to mitigate the influence of large weights due to the sampling procedures and stabilize the state estimates from year to year. It can be seen in Table 1 that the state level estimates are relatively smooth showing generally upward trends for the years 1998 to 2000 with some flattening after that. Changes are observable from year to year as expected.

To better illustrate how the new estimator helps to mitigate the effect of large weight cases but allows actual change, the graphs in Figures 1, 2 and 3 show a comparison for three states between the HT estimator and the hybrid estimator. Note that these states are all very small R&D performers where it is expected that a sample estimate might be heavily influenced by one or two observations with large weights.

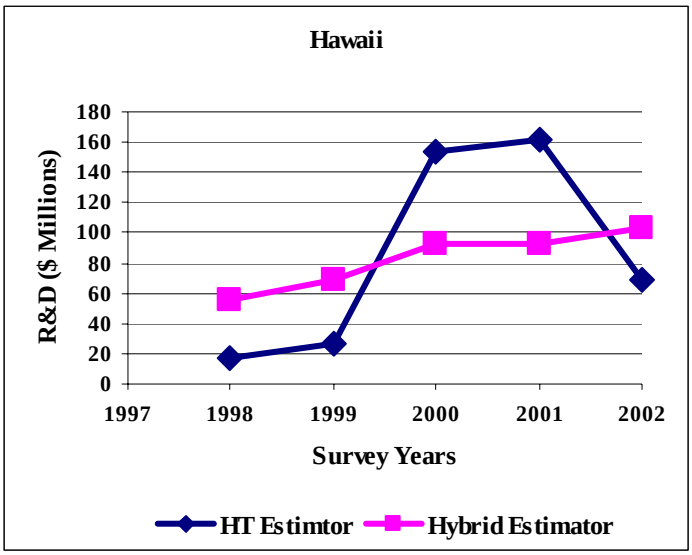
In the cases of Arizona and Hawaii shown in Figures 1 and 2, the new hybrid estimator helps to smooth the year-to-year estimates while the HT estimator is heavily influenced by a small number of large weight cases. In the case of Arizona, the spike in the 1999 HT estimate is due to one new company with a large weight in the sample reporting R&D in that year. The 2000 HT estimate does not show nor maintain this level because that company became a certainty case in 2000 and no new weighted sampled companies reported such a large amount of R&D that year.

Figure 1 – Comparison of HT and Hybrid Estimator Showing the Influence of Large Weights for Arizona



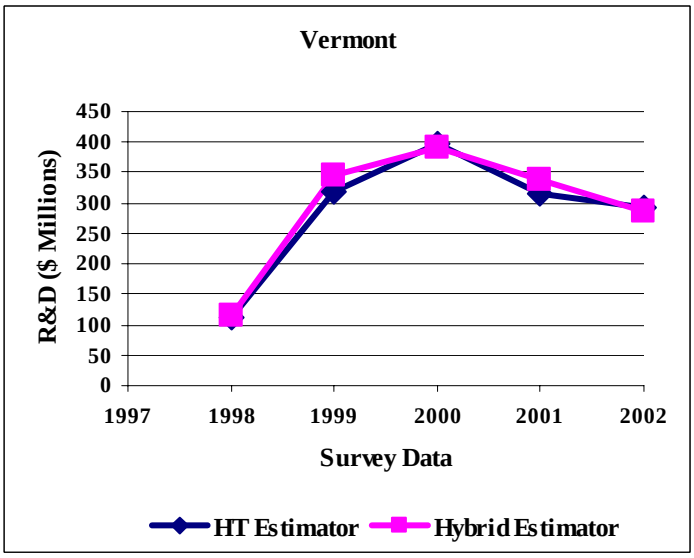
Hawaii is a different case. The level of R&D in Hawaii is extremely low. Thus companies that report even low levels of R&D generally have large sample weights which increases the relative variability of the HT estimator. The hybrid estimator is less influenced as can be seen.

Figure 2 – Comparison of HT and Hybrid Estimator Showing the Influence of Large Weights for Hawaii



In the case of Vermont shown in Figure 3, the year-to-year changes are due to actual observed changes in large influential certainty companies, not due to large weight cases. Thus the hybrid estimator captures change when it is observed and helps to smooth change due to the sampling procedures.

Figure 3 – Comparison of HT and Hybrid Estimator Showing Actual Observed Year-to-Year Changes - Vermont



One concern with the hybrid estimator is that it is not designed unbiased. What is the cost of year-to-year stability in terms of bias? As a start in answering this question, the difference between the HT estimator and the hybrid estimator was computed for the 2002 survey estimates. Recall that the HT estimator is unbiased. The results of this comparison for 2002 data are shown in the last two columns of Table 1.

It should be noted that a bias of one percent of the estimate may be statistically significant, i.e. the difference between the HT and hybrid estimates is significantly different than zero at some predefined level of confidence, but may not be practically significant to a user.

The hybrid estimates for very small R&D performing states are more heavily impacted by the nature of the estimator than are those for large R&D performing states that are more impacted by the observed data. On average, ninety percent of the contribution to the estimates is due to reported data, which are excluded from the modeling portion of the estimate.

As a data producer, it is important to provide the data user with information about limitations to the data including the potential for bias in the estimates.

6.0 Next Steps

The new hybrid estimator demonstrates many promising advantages to the HT estimator for state estimates of R&D in the Survey of Industrial R&D. Further research into the impact of the bias on the estimates is needed. A first look at this was taken as part of this paper but additional work should be conducted in this area.

It was alluded to in Section 2 that the estimator could be expanded to the industry estimates. As the survey moves forward, this may be an extremely valuable feature of the estimator. There have been discussions about providing estimates at the line of business level. Further research into this application would be helpful.

Table 1 - New Hybrid Estimates for Total R&D by State for 1998 to 2002 and a Comparison of HT and Hybrid for 2002

	Hybrid					HT	Std. Error Difference
Survey Year	1998	1999	2000	2001	2002	2002	2002
Alabama	845	824	821	924	846	714	97
Alaska	37	82	48	69	51	26	7
Arizona	1,801	2,111	2,182	2,288	3,201	3,224	97
Arkansas	213	327	400	256	225	114	28
California	32,856	37,006	43,564	40,909	39,664	45,794	2,858
Colorado	3,180	3,268	3,157	3,101	2,823	2,812	110
Connecticut	3,346	4,147	4,132	5,522	6,077	6,173	184
Delaware	1,356	1,295	1,468	1,334	1,219	1,186	10
District of Columbia	598	269	196	248	194	123	13
Florida	3,265	3,489	3,782	3,805	3,707	3,246	243
Georgia	1,617	1,907	2,159	1,924	2,107	1,931	181
Hawaii	55	69	93	93	103	69	9
Idaho	1,103	1,239	1,363	887	992	1,014	61
Illinois	7,318	8,108	8,393	8,376	7,616	7,277	278
Indiana	2,922	2,865	2,888	3,458	3,572	3,823	497
Iowa	750	731	762	818	753	716	47
Kansas	1,384	1,449	1,327	1,299	1,427	1,330	35
Kentucky	606	778	762	645	656	520	42
Louisiana	377	518	364	318	248	99	50
Maine	137	208	255	249	250	254	35
Maryland	1,905	2,022	2,239	3,684	3,800	3,512	94
Massachusetts	10,367	9,629	10,614	11,229	10,279	10,140	184
Michigan	12,554	16,881	17,489	13,889	13,565	13,264	197
Minnesota	3,367	3,697	3,971	4,388	4,460	4,279	71
Mississippi	183	225	242	229	224	143	23
Missouri	1,505	1,666	1,978	1,742	1,592	1,686	294
Montana	63	92	78	70	66	50	7
Nebraska	195	218	335	307	342	275	22
Nevada	476	492	433	287	339	335	62
New Hampshire	1,138	1,157	722	1,346	1,153	1,125	36
New Jersey	11,107	10,149	10,580	9,973	11,566	11,016	143
New Mexico	1,450	1,352	1,203	228	331	372	69
New York	10,283	12,268	11,631	11,014	9,234	8,874	353
North Carolina	3,483	3,636	4,328	4,184	3,443	3,175	123
North Dakota	46	95	83	79	154	138	7
Ohio	5,742	6,535	6,245	6,948	6,230	5,887	162
Oklahoma	369	563	463	580	412	333	49
Oregon	1,345	1,409	1,533	4,964	2,320	2,184	43
Pennsylvania	7,393	7,478	8,483	9,094	7,064	6,784	234
Rhode Island	1,332	1,317	1,167	1,148	1,121	1,044	19
South Carolina	996	924	1,093	936	1,054	949	55
South Dakota	40	57	89	87	53	30	5

	Hybrid					HT	Std. Error Difference
Survey Year	**1998**	**1999**	**2000**	**2001**	**2002**	**2002**	**2002**
Tennessee	**2,440**	**2,207**	**1,644**	**1,485**	**1,289**	**1,085**	**86**
Texas	**8,984**	**8,670**	**10,081**	**9,939**	**10,744**	**10,123**	**319**
Utah	**1,119**	**1,029**	**1,063**	**1,072**	**1,116**	**1,110**	**47**
Vermont	**114**	**346**	**389**	**338**	**286**	**293**	**17**
Virginia	**2,540**	**2,665**	**2,683**	**2,971**	**2,920**	**2,427**	**120**
Washington	**7,072**	**7,095**	**8,261**	**8,690**	**8,579**	**8,623**	**189**
West Virginia	**335**	**352**	**334**	**227**	**264**	**240**	**27**
Wisconsin	**1,929**	**2,196**	**2,415**	**2,485**	**2,649**	**2,449**	**63**
Wyoming	**20**	**65**	**37**	**28**	**21**	**11**	**3**
Undistributed Funds	**5,521**	**5,647**	**9,517**	**8,337**	**8,406**	**8,406**	**0**
Total U.S.	**169,180**	**182,824**	**199,539**	**198,505**	**190,809**	**190,809**	**0**

7.0 References

Cochran, William (1977), Sampling Techniques, John Wiley & Sons, p. 261

Lee, Hyunshik (1995), Outliers in Business Surveys, Business Survey Methods, John Wiley & Sons, pp. 503-526

Slanta and Fagan (1997), A Modified Approach to Sample Selection and Variance Estimation with Probability Proportionate to Size and Fixed Sample Size, MCD Working Paper Number: Census/MCD/WP-97/02

Tillé, Yves, (1996) An Elimination Procedure for Unequal Probability Sampling Without Replacement, *Biometrika*, **83**, 1, pp. 238-241