

## **Appendix B. Sampling and Estimation Methodologies**

The estimates in this report are based on a stratified simple random sample. The ICT-1 sample consists of 46,200 companies with paid employees (determined by the presence of payroll) in 2005.

The scope of the survey was defined to include all private, nonfarm, domestic companies. Major exclusions from the frame were government-owned operations (including the U.S. Postal Service), foreign-owned operations of domestic companies, establishments located in U.S. Territories, establishments engaged in agricultural production (not agricultural services), and private households.

The 2005 Business Register (BR) was used to develop the 2006 ICT-1 sample frame. The BR is the U.S. Census Bureau's establishment-based database. The database contains records for each physical business entity with payroll located in the United States, including company ownership information and current-year administrative data. In creating the ICT-1 frame, establishment data in the BR file were consolidated to create company-level records. Employment and payroll information was maintained for each six-digit North American Industry Classification System<sup>1</sup> (NAICS) industry in which the company had activity. Next, payroll data for each company-level record were run through an algorithm to assign the company, first to an industry sector (i.e., manufacturing, construction, etc.), then to a subsector (three-digit NAICS code), then to an industry group (four-digit NAICS code), then to an industry (five-digit NAICS code), and finally to an ICT industry code based on the industry. The resulting sample frame contained slightly less than 6.0 million companies.

The 2006 ICT-1 sampling frame consists of a certainty portion and a noncertainty portion. The 17,014 companies with 500 or more employees were selected with certainty. The remaining companies with 1 to 499 employees were then grouped into 135 industry categories. Each industry was then further divided into four strata. Since noncapitalized expenditures data were not available on the sampling frame, 2005 payroll was used as the stratification variable. The stratification methodology resulted in minimizing the sample size subject to a desired level of reliability for each industry. The expected relative standard errors (RSEs) ranged from 1 to 3 percent.

### **ESTIMATION**

Each company selected for the survey has a sample weight which is the inverse of its probability of selection. All sampled companies within the same stratum and industry grouping have the same weight. Weights were increased to adjust for nonresponse. The coverage rate for all companies was 86.4 percent. The coverage rate is calculated by multiplying 100 by the ratio of the noncapital expenditures of all reporting companies weighted by the original sample weights, to the noncapital expenditures of all reporting companies weighted by the adjusted-for- nonresponse sample weights. Weight adjustment and publication estimation are described in the following subsections.

<sup>1</sup>North American Industry Classification System (NAICS) – United States, 2002. For sale by National Technical Information Service (NTIS), Springfield, VA 22161. Call NTIS at 1-800-553-6847.

### **Weight Adjustment**

For estimation purposes, each company was placed into 1 of 4 response-related categories:

1. Respondents.
2. Nonrespondents.
3. Not in business.
4. Known duplicates.

A company was considered a respondent or nonrespondent based on whether the company provided sufficient data in items 1, 2 or 3 of the survey form. Companies that went out of business prior to 2006 and duplicates were dropped from the survey. Companies that went out of business during the survey year were kept in the sample and efforts were made to collect data for the period the company was active.

**ICT-1 segment.** The following discussion assumes 658 strata (strata designation  $h = 1, 2, \dots, 658$ ) which are based on 135 industries, each normally containing five strata (including the certainty stratum), which would be a maximum of 675 strata. Where there is insufficient sample size to justify distinct strata, they were collapsed together. In 2006, 34 strata were collapsed into 17 strata.

The original stratum weights ( $W_h$ ) were adjusted to compensate for nonresponse. The adjusted weight is computed as follows:

$$W_{h(adj)} = W_h * \frac{(P_{hr} + P_{hn})}{(P_{hr})}$$

where,

$W_{h(adj)}$  is the adjusted stratum weight of the  $h^{\text{th}}$  stratum

$$W_h = \frac{N_h}{n_h}$$

$n_h$  is the original stratum weight of the  $h^{\text{th}}$  stratum

$N_h$  is the population size of the  $h^{\text{th}}$  stratum

$n_h$  is the sample size of the  $h^{\text{th}}$  stratum

$P_{hr}$  is the sum of total company payroll for respondent companies in stratum  $h$

$P_{hn}$  is the sum of total company payroll for nonrespondent companies in stratum  $h$

### Publication Estimation

Publication cell estimates were computed by obtaining a weighted sum of reported values for companies treated as respondents. For those strata undergoing nonresponse adjustment, the estimates for  $X_j$  are biased, since this method assumes that nonresponse is not a purely random event. No attempt was made to estimate the magnitude of this bias.

**ICT-1 segment.** The ICT-1 estimates were derived as follows. Each estimated cell total,  $\hat{X}_j$ , is of the form

$$\hat{X}_j = \sum_{h=1}^{658} \sum_{i \in h} (W_{h(adj)} * X_{(j),i,h})$$

where,

$W_{h(adj)}$  is the adjusted weight of the  $h^{\text{th}}$  stratum

$X_{(j),i,h}$  is the value attributed to the  $i^{\text{th}}$  company of stratum  $h$ , where  $j$  is the publication cell of interest.

Note: Although a company was assigned to and sampled in one ICT industry, it could report expenditures in multiple ICT industries. When this occurred, the reported data for all industries were inflated by the weight in the sample industry.

### RELIABILITY OF THE ESTIMATES

The data shown in this report are estimated from a sample and will differ from the data which would have been obtained from a complete census. Two types of possible errors are associated with estimates based on data from sample surveys: sampling errors and nonsampling errors. The accuracy of a survey result depends not

only on the sampling errors and nonsampling errors measured but also on the nonsampling errors not explicitly measured. For particular estimates, the total error may considerably exceed the measured errors.

### Sampling Variability

The sample used in this survey is one of many possible samples that could have been selected using the sampling methodology described earlier. Each of these possible samples would likely yield different results. The RSE is a measure of the variability among the estimates from these possible samples. The RSEs were calculated using a delete-a-group jackknife replicate variance estimator. The RSE accounts for sampling variability but does not account for nonsampling error or systematic biases in the data. Bias is the difference, averaged over all possible samples of the same design and size, between the estimate and the true value being estimated.

The RSEs presented in the tables can be used to derive the SE of the estimate. The SE can be used to derive interval estimates with prescribed levels of confidence that the interval includes the average results of all samples:

- a. intervals defined by one SE above and below the sample estimate will contain the true value about 68 percent of the time.
- b. intervals defined by 1.6 SE above and below the sample estimate will contain the true value about 90 percent of the time.
- c. intervals defined by two SEs above and below the sample estimate will contain the true value about 95 percent of the time.

The SE of the estimate can be calculated by multiplying the RSE presented in the tables by the corresponding estimate. Note, the RSE is the measure of variability presented for all estimates in this publication except for the estimates of percent changes presented in Table 2a, for which we provide the SE as the measure of variability (refer to Table 2b). Also note that RSEs in this publication are in percentage form. They must be divided by 100 before being multiplied by the corresponding estimate.

#### Examples of Calculating a Confidence Interval:

a. For a data value: using data from Table 4a and Table 4b, the SE for 2006 total nondurable manufacturing noncapitalized expenditures would be calculated as follows:

$$\hat{\sigma}(\hat{X}_j) = \left[ \frac{RSE(\hat{X}_j)}{100} \right] * \hat{X}_j = \left( \frac{1.5}{100} \right) * \$2,279$$

million = \$34.2 million

The 90-percent confidence interval can be constructed by multiplying 1.6 by the SE, adding this value to the estimate to create the upper bound, and subtracting it from the estimate to create the lower bound.

$$\hat{X}_j \pm \left[ 1.6 * \hat{\sigma}(\hat{X}_j) \right]$$

Using data from Table 4a, for 2006 total nondurable manufacturing noncapitalized equipment expenditures, a 90-percent confidence interval would be calculated as:

$$\$2,279 \text{ million} \pm 1.6 * (\$34.2 \text{ million}) = \$2,279 \pm \$54.7 \text{ million}$$

This implies 90 percent confidence that the interval \$2,224 million to \$2,334 million contains the actual total for nondurable manufacturing noncapitalized equipment expenditures, subject to further nonsampling errors.

b. For percent change: using data from Table 2a and Table 2b, the 90-percent confidence interval can be constructed by multiplying 1.6 by the SE of the percent change, adding this value to the estimated percent change to create the upper bound, and subtracting it from the estimate to create the lower bound. For example, for the noncapitalized expenditures in the Health care and social assistance sector, the estimated percent change from 2005 to 2006 is 2.3 percent (from Table 2a), and the standard error of this estimate is 4.8 percent (from Table 2b)

$$2.3 \text{ percent} \pm \left[ 1.6 * 4.8 \text{ percent} \right] = 2.3 \pm 7.7 \text{ percent}$$

This implies 90 percent confidence that the interval -5.4 percent to +10.0 percent contains the actual percent change for noncapitalized expenditures in the Health care and social assistance sector. Since this interval contains zero (0), we do not have sufficient evidence to conclude that the estimated change was statistically different from 0, i.e., the change is not statistically significant

### Examples of Calculating Absolute Differences and Percent Changes

Data for the current year along with revised data for the prior year are presented in this publication. Two numbers of interest for many data users may be the absolute difference between the prior year and the current year, and the percent change from the prior year to the current year.

The absolute difference is calculated as:

$$\hat{d}_j = \left[ \hat{x}_t - \hat{x}_{(t-1)} \right]$$

and a 90-percent confidence interval on this difference is estimated as:

$$CI(d_j) = 1.6 * \sqrt{\left[ \sigma^2(\hat{x}_t) + \sigma^2(\hat{x}_{(t-1)}) \right]}$$

As an example, for the capitalized equipment expenditures for computer and peripheral equipment in the Information sector, from Table 4c, the estimate for 2006 is \$8,144 with the RSE found in Table 4d as 2.2, and for 2005 the revised estimate from Table 4c is \$6,758 with the RSE found in Table 4d as 2.5. The above calculations would be:

$$\hat{d}_j = \left[ \$8,144 \text{ million} - \$6,758 \text{ million} \right] = \$1,386 \text{ million}$$

And the 90-percent confidence interval is estimated as:

$$\begin{aligned} CI(\hat{d}_j) &= 1.6 * \sqrt{\left[ (.022 * \$8,144)^2 + (.025 * \$6,758)^2 \right]} \\ &= 1.6 * \$246.26 \\ &= \$394.02 \approx \$394 \text{ million} \end{aligned}$$

so the 90-percent confidence interval is \$1,386 +/- \$394 million, or \$992 million to \$1,780 million.

The percent change is calculated as 100 multiplied by the ratio of the difference divided by the prior estimate.

So continuing with the example from above,

$$\begin{aligned}
 PC_j &= 100 * \left( \frac{\hat{d}_j}{\hat{x}_{(t-1)}} \right) \\
 &= 100 * \left( \frac{\$1,386}{\$6,758} \right) \\
 &= 20.51 \text{ percent}
 \end{aligned}$$

and a 90-percent confidence interval on this percent change is estimated as:

$$\begin{aligned}
 CI(PC_j) &= 1.6 * 100 * \left( \frac{\hat{x}_t}{\hat{x}_{(t-1)}} \right) * \sqrt{\left[ \left( \frac{RSE(\hat{x}_t)}{100} \right)^2 + \left( \frac{RSE(\hat{x}_{(t-1)})}{100} \right)^2 \right]} \\
 &= 1.6 * 100 * \left( \frac{\$8,144}{\$6,758} \right) * \sqrt{[(.022)^2 + (.025)^2]} \\
 &= 1.6 * 100 * 0.040131 \\
 &= 6.4210 \text{ percent} \doteq 6.42 \text{ percent}
 \end{aligned}$$

so the 90-percent confidence interval is 20.51 percent +/- 6.42 percent or 14.09 percent to 26.93 percent.

### **Nonsampling Error**

All surveys and censuses are subject to nonsampling errors. Nonsampling errors can be attributed to many sources: inability to obtain information about all companies in the sample; inability or unwillingness on the part of respondents to provide correct information; response errors; definition difficulties; differences in the interpretation of questions; mistakes in recording or coding the data; and other errors of collection, response, coverage, and estimation for nonresponse.

Explicit measures of the effects of these nonsampling errors are not available. However, to minimize nonsampling error, all reports were reviewed for reasonableness and consistency, and every effort was made to achieve accurate response from all survey participants.

Coverage errors may have a significant effect on the accuracy of estimates for this survey. The BR, which forms the basis of our survey universe frame, may not contain all businesses. Also, businesses that are contained in the BR may have their payroll misreported.