

Appendix B.

Limitations of the Data and Methodology

INTRODUCTION

The data presented in this *State and Metropolitan Area Data Book* came from many sources. The sources include not only federal statistical bureaus and other organizations that collect and issue statistics as their principal activity, but also governmental administrative and regulatory agencies, private research bodies, trade associations, insurance companies, health associations, and private organizations such as the National Education Association and philanthropic foundations. Consequently, the data vary considerably as to reference periods, definitions of terms and, for ongoing series, the number and frequency of time periods for which data are available.

The statistics presented were obtained and tabulated by various means. Some statistics are based on complete enumerations or censuses while others are based on samples. Some information is extracted from records kept for administrative or regulatory purposes (school enrollment, hospital records, securities registration, financial accounts, social security records, income tax returns, etc.), while other information is obtained explicitly for statistical purposes through interviews or by mail. The estimation procedures used vary from highly sophisticated scientific techniques to crude “informed guesses.”

Each set of data relates to a group of individuals or units of interest referred to as the *target universe* or *target population*, or simply as the *universe* or *population*. Prior to data collection the target universe should be clearly defined. For example, if data are to be collected for the universe of households in the United States, it is necessary to define a “household.” The target universe may not be completely tractable. Cost and other considerations may restrict data collection to a survey universe based on some available list, such list may be inaccurate or out of date. This list is called a *survey frame* or *sampling frame*.

The data in many tables are based on data obtained for all population units, a census, or on data obtained for only a portion, or sample, of the population units. When the data presented are based on a sample, the sample is usually a scientifically selected probability sample. This is a sample selected from a list or sampling frame in such a way that every possible sample has a known chance of selection and usually each unit selected can be assigned a number, greater than zero and less than or equal to one, representing its likelihood or probability of selection.

For large-scale sample surveys, the probability sample of units is often selected as a multistage sample. The first stage of a multistage sample is the selection of a probability sample of large groups of population members, referred to as primary sampling units (PSUs). For example, in a national multistage household sample, PSUs are often counties or groups of counties. The second stage of a multistage sample is the selection, within each PSU selected at the first stage, of smaller groups of population units, referred to as secondary sampling units. In subsequent stages of selection, smaller and smaller nested groups are chosen until the ultimate sample of population units is obtained. To qualify a multistage sample as a probability sample, all stages of sampling must be carried out using probability-sampling methods.

Prior to selection at each stage of a multistage (or a single-stage) sample, a list of the sampling units or sampling frame for that stage must be obtained. For example, for the first stage of selection of a national household sample, a list of the counties and county groups that form the PSUs must be obtained. For the final stage of selection, lists of households, and sometimes persons within the households, have to be compiled in the field. For surveys of economic entities and for the economic censuses the Census Bureau generally uses a frame constructed from the Census Bureau’s Business Register. The Business Register contains all establishments with payroll in the United States including small single establishment firms as well as large multiestablishment firms.

Wherever the quantities in a table refer to an entire universe, but are constructed from data collected in a sample survey, the table quantities are referred to as *sample estimates*. In constructing a sample estimate, an attempt is made to come as close as is feasible to the corresponding universe quantity that would be obtained from a complete census of the universe. Estimates based on a sample will, however, generally differ from the hypothetical census figures. Two classifications of errors are associated with estimates based on sample surveys: (1) *sampling error*—the error arising from the use of a sample, rather than a census, to estimate population quantities and (2) *nonsampling error*—those errors arising from nonsampling sources. As discussed below, the magnitude of the sampling error for an estimate can usually be estimated from the sample data. However, the magnitude of the nonsampling error for an estimate can rarely be estimated. Consequently, actual error in an estimate exceeds the error that can be estimated.

The particular sample used in a survey is only one of a large number of possible samples of the same size, which could have been selected using the same sampling procedure. Estimates derived from the different samples would, in general, differ from each other. The *standard error* (SE) is a measure of the variation among the estimates derived from all possible samples. The standard error is the most commonly used measure of the sampling error of an estimate. Valid estimates of the standard errors of survey estimates can usually be calculated from the data collected in a probability sample. For convenience, the standard error is sometimes expressed as a percent of the estimate standard error is sometimes expressed as a percent of the estimate and is called the relative standard error or *coefficient of variation* (CV). For example, an estimate of 200 units with an estimated standard error of 10 units has an estimated CV of 5 percent.

A sample estimate and an estimate of its standard error or CV can be used to construct interval estimates that have a prescribed confidence that the interval includes the average of the estimates derived from all possible samples with a known probability. To illustrate, if all possible samples were selected under essentially the same general conditions, and using the same sample design, and if an estimate and its estimated standard error were calculated from each sample, then: 1) approximately 68 percent of the intervals from one standard error below the estimate to one standard error above the estimate would include the average estimate derived from all possible samples; 2) approximately 90 percent of the intervals from 1.6 standard errors below the estimate to 1.6 standard errors above the estimate would include the average estimate derived from all possible samples; and 3) approximately 95 percent of the intervals from two standard errors below the estimate to two standard errors above the estimate would include the average estimate derived from all possible samples.

Thus, for a particular sample, one can say with the appropriate level of confidence (e.g., 90 percent or 95 percent) that the average of all possible samples is included in the constructed interval. Example of a confidence interval: an estimate is 200 units with a standard error of 10 units. An approximately 90 percent confidence interval (plus or minus 1.6 standard errors) is from 184 to 216.

All surveys and censuses are subject to nonsampling errors. Nonsampling errors are of two kinds *random and nonrandom*. Random nonsampling errors arise because of the varying interpretation of questions (by respondents or interviewers) and varying actions of coders, keyers, and other processors. Some randomness is also introduced when respondents must estimate. Nonrandom nonsampling errors result from total nonresponse (no usable data obtained for a sampled unit), partial or item nonresponse (only a portion of a response may be usable), inability or

unwillingness on the part of respondents to provide correct information, difficulty interpreting questions, mistakes in recording or keying data, errors of collection or processing, and coverage problems (overcoverage and undercoverage of the target universe). Random nonresponse errors usually, but not always, result in an understatement of sampling errors and thus an overstatement of the precision of survey estimates. Estimating the magnitude of nonsampling errors would require special experiments or access to independent data and, consequently, the magnitudes are seldom available.

Nearly all types of nonsampling errors that affect surveys also occur in complete censuses. Since surveys can be conducted on a smaller scale than censuses, nonsampling errors can presumably be controlled more tightly. Relatively more funds and effort can perhaps be expended toward eliciting responses, detecting and correcting response error, and reducing processing errors. As a result, survey results can sometimes be more accurate than census results.

To compensate for suspected nonrandom errors, adjustments of the sample estimates are often made. For example, adjustments are frequently made for nonresponse, both total and partial. Adjustments made for either type of nonresponse are often referred to as *imputations*. Imputation for total nonresponse is usually made by substituting for the questionnaire responses of the nonrespondents the “average” questionnaire responses of the respondents. These imputations usually are made separately within various groups of sample members, formed by attempting to place respondents and nonrespondents together that have “similar” design or ancillary characteristics. Imputation for item nonresponse is usually made by substituting for a missing item the response to that item of a respondent having characteristics that are “similar” to those of the nonrespondent.

For an estimate calculated from a sample survey, the *total error* in the estimate is composed of the sampling error, which can usually be estimated from the sample, and the nonsampling error, which usually cannot be estimated from the sample. The total error present in a population quantity obtained from a complete census is composed of only nonsampling errors. Ideally, estimates of the total error associated with data given in these tables should be given. However, due to the unavailability of estimates of nonsampling errors, only estimates of the levels of sampling errors, in terms of estimated standard errors or coefficients of variation, are available. To obtain estimates of the estimated standard errors from the sample of interest, obtain a copy of the referenced report, which appears at the end of each table.

Source of Additional Material: The Federal Committee on Statistical Methodology (FCSM) is an interagency

committee dedicated to improving the quality of federal statistics <<http://www.fcsm.gov/>>.

Principal databases: Beginning below are brief descriptions of 18 of the sample surveys, censuses, and administrative collections that provide a substantial portion of the data contained in this publication.

U.S. DEPARTMENT OF AGRICULTURE

National Agriculture Statistics Service (NASS)

Census of Agriculture

Universe, Frequency, and Types of Data: Complete count of U.S. farms and ranches conducted once every 5 years with data at the national, state, and county level. Data published on farm numbers and related items/ characteristics.

Type of Data Collection Operation: Complete census for number of farms; land in farms; farm income; agriculture products sold; farms by type of organization; total cropland; irrigated land; farm operator characteristics; livestock and poultry inventory and sales; and selected crops harvested. Market value of land, buildings, and products sold, total farm production expenses, machinery and equipment, and fertilizer and chemicals.

Data Collection and Imputation Procedures: Data collection is by mailing questionnaires to all farmers and ranchers. Producers can return their forms by mail or online. Nonrespondents are contacted by telephone and correspondence follow-ups. Imputations were made for all nonresponse item/characteristics and coverage adjustments were made to account for missed farms and ranches. The response rate for the 2007 census was 85.2 percent.

Estimates of Sampling Error: Weight adjustments were made to account for the undercoverage and whole-unit nonresponse of farms on the Census Mail List (CML). These were treated as sampling errors.

Other (Nonsampling) Errors: Nonsampling errors are due to incompleteness of the census mailing list, duplications on the list, respondent reporting errors, errors in editing reported data, and in imputation for missing data. Evaluation studies are conducted to measure certain nonsampling errors such as list coverage and classification error. It is a reasonable assumption that the net effect of nonmeasurable errors is zero (the positive errors cancel the negative errors).

Sources of Additional Material: U.S. Department of Agriculture, National Agricultural Statistics Service (NASS), 2007 Census of Agriculture, Appendix A-1 Census of Agriculture Methodology, Appendix B-1 General Explanation and Census of Agriculture Report Form.

Multiple Frame Surveys

Universe, Frequency, and Types of Data: Surveys of U.S. farm operators to obtain data on major livestock inventories, selected crop acreage and production, grain stocks, and farm labor characteristics, farm economic data, and chemical use data. Estimates are made quarterly, semi-annually, or annually depending on the data series.

Type of Data Collection Operation: Primary frame is obtained from general or special purpose lists, supplemented by a probability sample of land areas used to estimate for list incompleteness.

Data Collection and Imputation Procedures: Mail, telephone, or personal interviews used for initial data collection. Mail nonrespondent follow-up by phone and personal interviews. Imputation based on average of respondents.

Estimates of Sampling Error: Estimated CVs range from 1 percent to 2 percent at the U.S. level for crop and livestock data series and 3 to 5 percent for economic data. Regional CVs range from 3 to 6 percent, while state estimate CVs run 5 to 10 percent.

Other (Nonsampling) Errors: In addition to above, replicated sampling procedures used to monitor effects of changes in survey procedures.

Sources of Additional Material: U.S. Department of Agriculture, National Agricultural Statistics Service (NASS), USDA's National Agricultural Statistics Service: The Fact Finders of Agriculture, March 2007.

U.S. BUREAU OF LABOR STATISTICS

Current Employment Statistics (CES) Program

Universe, Frequency, and Types of Data: Monthly survey drawn from a sampling frame of over 8 million unemployment insurance tax accounts in order to obtain data by industry on employment, hours, and earnings.

Type of Data Collection Operation: In 2006, the CES sample included about 150,000 businesses and government agencies, which represent approximately 390,000 individual worksites.

Data Collection and Imputation Procedures: Each month, the state agencies cooperating with Bureau of Labor Statistics (BLS), as well as BLS Data Collection Centers, collect data through various automated collection modes and mail. BLS Washington staff prepares national estimates of employment, hours, and earnings while states use the data to develop state and area estimates.

Estimates of Sampling Errors: The relative standard error for total nonfarm employment is 0.1 percent. From April 2002 to March 2003, the cumulative net birth/death model added 469,000.

Other (nonsampling) Errors: Estimates of employment adjusted annually to reflect complete universe. Average adjustment is 0.2 percent over the last decade, with an absolute range from less than 0.1 percent to 0.6 percent.

Sources of Additional Material: U.S. Bureau of Labor Statistics, Employment & Earnings Online. See <<http://www.bls.gov/opub/ee/home.htm>>.

U.S. DEPARTMENT OF COMMERCE U.S. BUREAU OF ECONOMIC ANALYSIS (BEA)

Regional Economic Information System (REIS)

Universe, Frequency, and Types of Data: The Regional Economic Information System contains estimates of personal income and its components and employment for local areas such as states, counties, metropolitan areas, and micropolitan areas.

Type of Data Collection Operation: The estimates of personal income are primarily based on administrative-records data, census data, and survey data.

Data Collection and Imputation Procedures: The data are collected from administrative records, which may come from the recipients of the income or from the sources of the income. These data are a byproduct of the administration of various Federal and state government programs. The most important sources of these data are—the state unemployment insurance programs of the Bureau of Labor Statistics (BLS), the social insurance programs of the Centers for Medicare and Medicaid Services, federal income tax program of the Internal Revenue Service, veterans benefit programs of the U.S. Department of Veterans Affairs, and military payroll systems of the U.S. Department of Defense.

The data from censuses are mainly collected from the recipients of income. The most important sources for these data are the Census of Agriculture at the U.S. Department of Agriculture (USDA) and the Census of Population and Housing conducted by the U.S. Census Bureau. Other sources may include estimates of farm proprietors' income by the USDA, wages and salaries from County Business Patterns from the Census Bureau, and the Quarterly Census of Employment and Wages by the Department of Labor.

Estimates of Sampling Error: Not applicable, except component variables may be subject to error.

Other (Nonsampling) Errors: Nonsampling errors in the administrative data sets may affect personal income estimates.

Sources of Additional Material: Methodological information on other Bureau of Economic Analysis (BEA) datasets such as "State Personal Income" and "Gross State Product" may be found at <<http://www.bea.gov/regional/methods.cfm>>.

U.S. CENSUS BUREAU

American Community Survey (ACS)

Universe, Frequency, and Types of Data: Nationwide survey to obtain annual data about demographic, social, economic, and housing characteristics of housing units and the people residing in them. It covers the household population and, beginning in 2006, also includes the group quarter population living in prisons, nursing homes and college dormitories, and other group quarters.

Type of Data Collection Operation: Housing unit address sampling is performed twice a year in both August and January. First-phase of sampling defines the universe for the second stage of sampling through two steps. First, all addresses that were eligible for the second-phase sampling within the past 4 years are excluded from eligibility. This ensures that no address is in sample more than once in any 5-year period. The second step is to select a 20 percent systematic sample of "new" units, i.e., those units that have never appeared on a previous Master Address File (MAF) extract. All new addresses are systematically assigned to either the current year or to one of four back-samples. This procedure maintains five equal partitions of the universe. The second-phase sampling is done on the current year's partition and results in approximately 3,000,000 housing unit addresses in the United States and 36,000 in Puerto Rico. Group quarter sampling is performed separately from the housing unit sampling. The sampling begins with separating the small (15 persons or fewer) and the large (more than 15 persons) group quarters. The target sampling rate for both groups is a 2.5 percent sample of the group quarters population. It results in approximately 200,000 group quarter residents being selected in the United States, and an additional 1,000 in Puerto Rico.

Data Collection and Imputation Procedures: The American Community Survey is conducted every month on independent samples. Each housing unit in the independent monthly samples is mailed a prenotice letter announcing the selection of the address to participate, a survey questionnaire package, and a reminder postcard. These sample units addresses receive a second (replacement) questionnaire package if the initial questionnaire has not been returned by mid-month. Sample addresses for which a questionnaire is not returned in the mail and a telephone number is not available is forwarded to telephone centers for follow-up. Interviewers attempt to contact and interview these mail nonresponse cases by telephone. Sample addresses that are still unresponsive after 2 months of attempts are forwarded for a possible personal visit. Unresponsive addresses are subsampled at rates between 1 in 3 and 2 in 3. Those addresses selected through this process are assigned to Field Representatives (FRs), who visit the addresses, verify their existence, determine their

occupancy status, and conduct interviews. Collection of group quarters data is conducted by FRs only. Their methods include completing the questionnaire while speaking to the resident in person or over the telephone, or leaving paper questionnaires for residents to complete for themselves and then pick them up later. This last option is used for data collection in federal prisons. If needed, a personal interview can be conducted with a proxy, such as a relative or guardian. After data collection is completed, any remaining incomplete or inconsistent information on the questionnaire are imputed during the final automated edit of the collected data.

Estimates of Sampling Error: The data in the ACS products are estimates and can vary from the actual values that would have been obtained by conducting a census of the entire population. The estimates from the chosen sample addresses can also vary from those that would have been obtained from a different set of addresses. This variation causes uncertainty, which can be measured using statistics such as standard error, margin of error, and confidence interval. All ACS estimates are accompanied by margin of errors to assist users.

Other (Nonsampling) Errors: Nonsampling Error—In addition to sampling error, data users should realize that other types of errors may be introduced during any of the various complex operations used to select, collect, and process survey data. An important goal of the ACS is to minimize the amount of nonsampling error introduced through coverage issues in the sample list, nonresponse from sample housing units, and transcribing or editing data. One way of accomplishing this is by finding additional sources of addresses, following up on nonrespondents, and maintaining quality control systems.

Sources of Additional Material: U.S. Census Bureau, American Community Survey Web site available on Internet, <<http://www.census.gov/acs>>, U.S. Census Bureau, American Community Survey Accuracy of the Data documents available on the Internet, <<http://www.census.gov/acs/www/UseData/Accuracy/Accuracy1.htm>>.

Annual Survey of Manufactures (ASM)

Universe, Frequency, and Types of Data: The Annual Survey of Manufactures is conducted annually, except for years ending in “2” and “7” for all manufacturing establishments having one or more paid employees. The purpose of the ASM is to provide key intercensal measures of manufacturing activity, products, and location for the public and private sectors. The ASM provides statistics on employment, payroll, worker hours, payroll supplements, cost of materials, value added by manufacturing, capital expenditures, inventories, and energy consumption. It also provides estimates of value of shipments for 1,800 classes of manufactured products.

Type of Data Collection Operation: The ASM includes approximately 50,000 establishments selected from the census universe of 346,000 manufacturing establishments. Approximately 24,000 large establishments are selected with certainty, and the remaining 26,000 other establishments are selected with probability proportional to a composite measure of establishment size. The survey is updated from two sources: Internal Revenue Service (IRS) administrative records are used to include new single-unit manufacturers and the Company Organization Survey identifies new establishments of multiunit forms.

Data Collection and Imputation Procedures: Survey is conducted by mail with phone and mail follow-ups of nonrespondents. Imputation (for all nonresponse items) is based on previous year reports, or for new establishments in survey, on industry averages.

Estimates of Sampling Error: Estimated relative standard errors for number of employees, new expenditures, and for value added totals are given in annual publications. For U.S.-level industry statistics, most estimated relative standard errors are 2 percent or less, but vary considerably for detailed characteristics.

Other (Nonsampling) Errors: The unit response rate is about 85 percent. Nonsampling errors include those due to collection, reporting, and transcription errors, many of which are corrected through computer and clerical checks.

Sources of Additional Material: U.S. Census Bureau, *Annual Survey of Manufactures*, and Technical Paper 24.

State Government Tax Collections (STC)

Universe, Frequency, and Types of Data: The universe for the State Tax Collections Survey covers the 50 state governments only. No local governments are included in the universe for each state. The data have been collected annually since 1939. Statistics on the State Government Tax Collections Survey include measurement of tax by category: Property Tax, Sales and Gross Receipts Taxes, License Taxes, Income Taxes, and Other Taxes. Each tax category is broken down into subcategories (e.g., motor fuel sales, alcoholic beverage sales, motor vehicle licenses, alcoholic beverage licenses). There are currently 25 different tax codes that state tax revenue may fall into.

Type of Data Collection Operation: Most of the data in this report were gathered by a mail canvass of appropriate state government offices that are directly involved with state-administered taxes. There are approximately one hundred offices that are canvassed to collect data from all fifty states. Follow-up procedures include the use of mail, telephone, and e-mail until data are received.

Data Editing and Imputation Procedures: Data are processed from several collection methods including direct response to survey forms from state government officials,

as well as from the compilation of administrative records and supplemental sources. Regardless of the collection method, these data are edited using ratio edits of the current year's value to the prior year's value. The fifty state governments provide the Census Bureau with administrative records from their central accounting system. These administrative records are unique to each state as each state is legally organized differently from every other state and, as such, each state has a unique organizational and accounting structure. It is the responsibility of the Census Bureau to classify the different accounting and organizational structures into uniform tax categories so that entities with different methods of government accounting can be presented on a comparable basis. The records represent the core, or central, state government and are limited to tax revenue. Data on state government tax revenues are compiled from state administrative records by Census Bureau employees, according to the Census Bureau's classification methodology. When state records do not include full tax revenue detail or reporting units do not respond, supplemental data sources from external financial reports or the Census Bureau's *Annual Survey of State Government Finances and Quarterly Summary of State and Local Government Tax Revenue* are required to complete the data sets. This procedure is called imputation. Supplemental records are merged with data from the state governments. Although every effort is made to obtain financial information from all state government entities, financial statements may not be available at the time the Census Bureau closes the processing, or governmental entities may not respond to our requests. Every year the data are subject to revisions as new data become available.

Estimates of Sampling Error: These data are not subject to sampling error because this is a complete enumeration of all 50 state governments.

Other (Nonsampling) Errors: Despite efforts made in all phases of collection, processing, and tabulation to minimize errors, the survey is subject to nonsampling errors such as the inability to obtain data for every variable for all units, inaccuracies in classification, mistakes in keying and coding, and coverage errors.

Sources of Additional Material: For further information, see the *Government Finance and Employment Classification Manual* and the *2007 Census of Governments*.

Annual Survey of Public Employment and Payroll (ASPEP)

Universe, Frequency, and Types of Data: The population of interest for this survey includes the civilian employees of all federal government agencies (except the Central Intelligence Agency, the National Security Agency, and the Defense Intelligence Agency), all agencies of the 50 state governments, and 89,476 local governments (i.e., counties,

municipalities, townships, special districts, and school districts) including the District of Columbia.

Data have been collected annually since 1957. A census is conducted every 5 years (years ending in "2" and "7"). A sample of state and local governments is used to collect data in the intervening years. A new sample is selected every 5 years (in years ending in "4" and "9"). The survey provides data on full-time and part-time employment, part-time hours worked, full-time equivalent employment, and payroll statistics by governmental function (i.e., elementary and secondary education, higher education).

Type of Data Collection Operation: Data collected for the Annual Survey of Government Employment are public record and are not confidential, as authorized by Title 13, U.S. Code, Section 9. Census Bureau staff compiled federal government data from records of the U.S. Office of Personnel Management (OPM). These data are based on the Monthly Report of Federal Civilian Employment. Census Bureau staff collected some state government data through special arrangements, referred to as central collection agreements, wherein data for multiple state agencies or school districts are reported by a central respondent generally in an electronic file. Forty-five of the state governments provided data from central payroll records for all or most of their agencies/institutions. Data for agencies and institutions for the remaining state governments were obtained by mail canvass questionnaires. Local governments were also canvassed using a mail questionnaire. All respondents receiving the mail questionnaire had the option of responding electronically using the Web site developed for reporting data.

Data Editing and Imputation Procedures: Editing is a process that ensures survey data are accurate, complete, and consistent. Efforts are made at all phases of collection, processing, and tabulation to minimize errors. Although some edits are built into the Internet data collection instrument and the data entry programs, the majority of the edits are performed post collection. Edits consist primarily of two types: (1) *consistency edits* and (2) *historical ratio edits* of the current year's reported value to the prior year's value. The *consistency edits* check the logical relationships of data items reported on the form. For each function where employees are reported, the *historical ratio edits* compare data from two different time periods.

Not all respondents answer every item on the questionnaire. There are also questionnaires that are not returned despite efforts to gain a response. Imputation is the process of filling in missing or invalid data with reasonable values in order to have a complete data set for analytical purposes. For nonresponding governments, the imputations were based on recently reported historical data from either a prior year annual survey or the most recent Census of Governments. These data were adjusted by a growth

rate that was determined by the growth of responding units that were similar (in size, geography, and type of government) to the nonrespondent. If there was no recent historical data available, the imputations were based on the data from a randomly selected responding donor that was similar to the nonrespondent. In cases where good secondary data sources exist, the data from those sources were used.

Estimates of Sampling Error: The intercensal data come from a sample rather than a census of all possible units. The particular sample that was selected is one of a larger number of possible samples of the same size and sample design that could have been selected. Each sample would have yielded different estimates. The estimated coefficients of variation, which are provided for each estimate on <www.census.gov/govs>, are an estimate of this sampling variability.

Other (Nonsampling) Errors: Although every effort is made in all phases of collection, processing, and tabulation to minimize errors, the data are subject to nonsampling errors such as inability to obtain data for every variable from all units in the population of interest, inaccuracies in classification, response errors, misinterpretation of questions, mistakes in keying and coding, and coverage errors. The data processing section describes our efforts to mitigate errors due to nonresponse, keying, reporting errors, etc.

Sources of Additional Material: For further information, see the *Government Finance and Employment Classification Manual* and the *2007 Census of Governments*.

Annual Finance Survey (AFS)

Universe, Frequency, and Types of Data: The population of interest for this survey contains the 50 state governments and 89,476 local governments (counties, municipalities, townships, special districts, and school districts) including the District of Columbia. In years ending in "2" and "7" the entire universe is canvassed. In intervening years, a sample of the population of interest is surveyed. The survey coverage includes all state and local governments in the United States.

The survey collects financial data. Revenue data include taxes (i.e., property, sales, tobacco, motor vehicle, licensing and permit), charges, interest, and other earnings. Expenditure data include total by function (i.e., education, highways, airports, water and sewerage, health, hospitals, corrections, fire and police protection), and by accounting category (i.e., current operations and capital outlays). Debt data include issuance, retirement, and amounts outstanding. Financial assets data include securities and other holdings, by type.

Type of Data Collection Operation: The data collection for the state and local finance survey (both census and sample survey) is made up of three modes to obtain data: mail canvass, Internet collection, and central collection from state sources. Collection methods vary by state and type of government. Administrative data are compiled for most state government agencies and the 48 largest and most complex county and municipal governments. The survey melds several government finance surveys, including the Survey of Local Government Finances, Survey of Public-Employee Retirement Systems, Integrated Post-secondary Educational Data System (IPEDS) from the National Center for Education Statistics (NCES), State Government Finances Survey, and the Survey of Public Elementary-Secondary Education Finances.

Data Editing and Imputation Procedures: Not all respondents answer every item on the questionnaire. There are also questionnaires that are not returned despite efforts to gain a response. Imputation is the process of filling in missing or invalid data with reasonable values in order to have a complete data set for analytical purposes. For nonresponding governments, imputations for missing units are based on recently reported historical data from either a prior year annual survey or the most recent census, adjusted by a growth rate. If no historical data are available, data from a randomly selected similar unit are used as the impute.

Editing is a process that ensures data are accurate, complete, and consistent. Efforts are made at all phases of collection, processing, and tabulation to minimize errors. Although some edits are built into the Internet data collection instrument and the data entry programs, the majority of the edits are performed post collection. Data are checked for internal consistency within the questionnaire and for historical accuracy.

Estimates of Sampling Error: In census years, all of the units in the population are surveyed, and there is no sampling error. In the intercensal years, the population is sampled, and the estimates are subject to sampling error. The coefficient of variation is a measure of sampling variability expressed as a percentage of the estimated total. Generally, the estimated coefficients of variation for state and local government revenues, expenditures, debt, or assets are under 3 percent in each state. Coefficients of variation for the estimates are given in the tables on the Web site, <<http://www.census.gov/govs/estimate/index.html>>.

Other (Nonsampling) Errors: Although every effort is made in all phases of collection, processing, and tabulation to minimize errors, the data are subject to nonsampling

errors such as inability to obtain data for every variable from all units in the population of interest, inaccuracies in classification, response errors, misinterpretation of questions, mistakes in keying and coding, and coverage errors.

Sources of Additional Material: For more information on the survey, see <<http://www.census.gov/govs/estimate/index.html>>. On that site, see the *Survey Methodology and Government Finance and Employment Classification Manual*.

Federal Programs:

Consolidated Federal Funds Report (CFFR) Federal Aid to States (FAS) Federal Assistance Award Data System (FAADS)

Universe, Frequency, and Types of Data: The federal statistics included in these tables come from three sources. The Consolidated Federal Funds Report (CFFR) covers all states, the District of Columbia, and U.S. Outlying Areas. CFFR data were obtained from federal government expenditures or obligations to government agencies. Thirty-three departments and agencies of the executive branch of the federal government with grant making authority are generally reporting quarterly to FAADS.

The Federal Assistance Award Data System (FAADS) is authorized by Title 31, Section 6102(a), U.S. Code. Reporting covers approximately 600 federal assistance programs. While primarily concerned with assistance to state and local governments, all major programs providing transfer payments to individuals, discretionary project grants, loans, or insurance are also covered. The CFFR reports have been prepared annually by the Census Bureau since 1983 as authorized by Titles 13 and 31, U.S. Code and a 1982 designation by the Office of Management and Budget. Data are obtained on the amount of virtually all federal expenditures, including grants, loans, direct payments, insurance, procurement, salaries and wages, and other awards (such as price supports and research awards). Data collected for CFFR come from the following sources: U.S. Department of Defense, Federal Assistance Awards Data System (FAADS), Federal Procurement Data System, Office of Personnel Management, and the U.S. Postal Service. Selected data from the information on payments to state or local governments reported by federal agencies come from the Federal Aid to States Report.

Type of Data Collection Operation: FAADS is a repository of data on federal financial assistance award transactions; the administrative data are compiled quarterly. The CFFR aggregates this quarterly data and supplements the data with additional records from the federal agencies to get the complete array of variables reported by CFFR. FAS uses imported electronic data files that have been completed by the various federal agencies.

Data Editing and Imputation Procedures: When respondents submit the data, sometimes there are errors due to a misinterpretation of the request, a keying error, an inadvertent misclassification, etc. To mitigate these types of errors, the Census Bureau edits the data by verifying the data totals and geographic coding.

Estimates of Sampling Error: These data are not subject to sampling error because this is a complete enumeration of all governments in the universe.

Other (Nonsampling) Errors: Coverage errors occur when there is a failure to cover the entire population of interest for a survey. Since we may not have a complete list of agencies, we may have some coverage error. When data are missing due to nonresponse, the Census Bureau imputes for the missing data items in order to have a complete data set for analytical purposes. Errors may also arise in our geographic coding due to agency data entry input errors. These errors are mitigated by following editing procedures. Routine edits applied to FAADS data are primarily intended to identify and correct keying or calculation errors made by respondents.

Sources of Additional Material: For more information on the federal data, see <<http://www.census.gov/govs/cffr/index.html>> for information on the Consolidated Federal Funds Report, <<http://www.census.gov/govs/www/faads.html>> for information on the Federal Assistance Award Data System, or <<http://www.census.gov/prod/2009pubs/fas-08.pdf>> for information on the most recent Federal Aid to States.

2007 Economic Census

(Industry Series, Geographic Area Series, and Subject Series Reports) (for NAICS sectors 21 to 81).

Universe, Frequency, and Types of Data. Conducted every 5 years to obtain data on number of establishments, number of employees, payroll, total sales/receipts/revenue, and other industry-specific statistics. The universe is all establishments with paid employees excluding agriculture, forestry, fishing, and hunting, and government. (Nonemployer Statistics, discussed separately, covers those establishments without paid employees.)

Type of Data Collection Operation: All large employer firms were surveyed (i.e., all employer firms above payroll-size cutoffs established to separate large from small employers) plus, in most sectors, a sample of the small employer firms.

Data Collection and Imputation Procedures: Mail questionnaires were used with both mail and telephone follow-ups for nonrespondents. Businesses also had the option to respond electronically. Data for nonrespondents and for small employer firms not mailed a questionnaire

were obtained from administrative records of other federal agencies or imputed.

Estimates of Sampling Error: Not applicable for basic data such as sales, revenue, receipts, payroll, etc., for sectors other than Construction (NAICS 23). Estimates of sampling error for construction industries are included with the data as published on the Census Bureau Web site.

Other (Nonsampling) Errors: Establishment response rates by NAICS sector in 2002 ranged from 80 percent to 89 percent. Nonsampling errors may occur during the collection, reporting, keying, and classification of the data.

Sources of Additional Material U.S. Census Bureau, see <<http://www.census.gov/econ/census07/www/methodology/>>.

Census of Population

Universe, Frequency, and Types of Data: Complete count of U.S. population conducted every 10 years since 1790. Data obtained on number and characteristics of people in the United States.

Type of Data Collection Operation: In the 1990 and 2000 censuses, the 100 percent items included: age, date of birth, sex, race, Hispanic origin, and relationship to householder. In 1980, approximately 19 percent of the housing units were included in the sample; in 1990 and 2000, approximately 17 percent.

Data Collection and Imputation Procedures: In 1980, 1990, and 2000, mail questionnaires were used extensively with personal interviews in the remainder. Extensive telephone and personal follow-up for nonrespondents was done in the censuses. Imputations were made for missing characteristics.

Estimates of Sampling Error: Sampling errors for data are estimated for all items collected by sample and vary by characteristic and geographic area. The coefficients of variation (CVs) for national and state estimates are generally very small.

Other (Nonsampling) Errors: Since 1950, evaluation programs have been conducted to provide information on the magnitude of some sources of nonsampling errors such as response bias and undercoverage in each census. Results from the evaluation program for the 1990 census indicated that the estimated net undercoverage amounted to about 1.5 percent of the total resident population. For Census 2000, the evaluation program indicated a net overcount of 0.5 percent of the resident population.

Sources of Additional Material: U.S. Census Bureau, *The Coverage of Population in the 1980 Census*, PHC80-E4; *Content Reinterview Study: Accuracy of Data for Selected Population and Housing Characteristics as Measured by Reinterview*, PHC80-E2; *1980 Census of Population*, Vol.

1, (PC80-1), Appendixes B, C, and D. *Content Reinterview Survey: Accuracy of Data for Selected Population and Housing Characteristics as Measured by Reinterview*, 1990, CPH-E-1; *Effectiveness of Quality Assurance*, CPH-E-2; *Programs to Improve Coverage in the 1990 Census*, 1990, CPH-E-3. For Census 2000 evaluations, see <<http://www.census.gov/pred/www/>>.

County Business Patterns

Universe, Frequency, and Types of Data: County Business Patterns is an annual tabulation of basic data items extracted from the Business Register, a file of all known single- and multilocation employer companies maintained and updated by the U.S. Census Bureau. Data include number of establishments, number of employees, first quarter and annual payrolls, and number of establishments by employment size class. Data are excluded for self-employed individuals, private households, railroad employees, agricultural production workers, and most government employees.

Type of Data Collection Operation: The annual Company Organization Survey provides individual establishment data for multilocation companies. Data for single establishment companies are obtained from various Census Bureau programs, such as the Annual Survey of Manufactures and Current Business Surveys, as well as from administrative records of the Internal Revenue Service, the Social Security Administration, and the Bureau of Labor Statistics.

Estimates of Sampling Error: Not applicable.

Other (Nonsampling) Error: The data are subject to nonsampling errors, such as inability to identify all cases in the universe; definition and classification difficulties; differences in interpretation of questions; errors in recording or coding the data obtained; and estimation of employers who reported too late to be included in the tabulations and for records with missing or misreported data.

Sources of Additional Materials: U.S. Census Bureau, County Business Patterns, <<http://www.census.gov/econ/cbp/index.html>>.

Current Population Survey (CPS)

Universe, Frequency, and Types of Data: Nationwide monthly sample designed primarily to produce national and state estimates of labor force characteristics of the civilian noninstitutionalized population 16 years of age and older.

Type of Data Collection Operation: Multistage probability sample that currently includes 72,000 households from 824 sample areas. Sample size increased in some states to improve data reliability for those areas on an annual average basis. A continual sample rotation system

is used. Households are in sample 4 months, out for 8 months, and in for 4 more. Month-to-month overlap is 75 percent; year-to-year overlap is 50 percent.

Data Collection and Imputation Procedures: For first and fifth months that a household is in sample, personal interviews; other months, approximately 85 percent of the data collected by phone. Imputation is done for item nonresponse. Adjustment for total nonresponse is done by a predefined cluster of units, by state, metropolitan status and CBSA size; for item nonresponse imputation varies by subject matter.

Estimates of Sampling Error: The national total estimates of the civilian labor force and of employment have monthly CVs of about .2 percent and annual average CVs of about .1 percent. Unemployment is a much smaller characteristic and consequently has substantially larger CVs than the civilian labor force or employment. The national unemployment rate, the most important CPS statistic, has a monthly CV of about 2 percent and an annual average CV of about 1 percent. Assuming a 6 percent unemployment rate, states have annual average CVs of about 8 percent. The estimated CVs for family income and poverty rate for all persons in 2005 are .4 percent and 1.2 percent, respectively. CVs for subnational areas, such as states, tend to be larger and vary by area.

Other (Nonsampling) Errors: Estimates of response bias on unemployment are available. Estimates of unemployment rate from reinterviews range from -2.4 percent to 1.0 percent of the basic CPS unemployment rate (over a 30-month span from January 2004 through June 2006). Eligible CPS households are approximately 82 percent of the assigned households, with a corresponding response rate of 92 percent.

Sources of Additional Material: U.S. Census Bureau and Bureau of Labor Statistics, Current Population Survey: Design and Methodology, (Technical Paper 66), available on the Internet <<http://www.census.gov/prod/2006pubs/tp-66.pdf>> and the Bureau of Labor Statistics, <<http://www.bls.gov/cps/>> and the *BLS Handbook of Methods*, Chapter 1, available on the Internet at <http://www.bls.gov/pub/hom/homch1_a.htm>.

Monthly Survey of Construction

Universe, Frequency, and Types of Data: Survey conducted monthly of newly constructed housing units (excluding mobile homes). Data are collected on the start, completion, and sale of housing. (Annual figures are aggregates of monthly estimates.)

Type of Data Collection Operation: A multistage probability sample of approximately 900 of the 20,000 permit-issuing jurisdictions in the United States was selected. Each month in each of these permit offices, field representatives list and select a sample of permits for which to collect

data. To obtain data in areas where building permits are not required, a multistage probability sample of 80 land areas (census tracts or subsections of census tracts) was selected. All roads in these areas are canvassed and data are collected on all new residential construction found. Sampled buildings are followed up until they are completed (and sold, if for sale).

Data Collection and Imputation Procedures: Data are obtained by telephone inquiry and/or field visit. Nonresponse/undercoverage adjustment factors are used to account for late reported data.

Estimates of Sampling Error: Estimated CV of 5 percent to 6 percent for estimates of national totals of units started, but may be higher than 20 percent for estimated totals of more detailed characteristics, such as housing units in multiunit structures.

Other (Nonsampling) Errors: Response rate is over 90 percent for most items. Nonsampling errors are attributed to definitional problems, differences in interpretation of questions, incorrect reporting, inability to obtain information about all cases in the sample, and processing errors.

Sources of Additional Material All data are available on the Internet at <<http://www.census.gov/const/www/newresconstindex.html>>.

Further documentation of the survey is also available at that site.

Nonemployer Statistics

Universe, Frequency, and Types of Data: Nonemployer statistics are an annual tabulation of economic data by industry for active businesses without paid employees that are subject to federal income tax. Data showing the number of firms and receipts by industry are available for the United States, states, counties, and metropolitan areas. Most types of businesses covered by the Census Bureau's economic statistics programs are included in the nonemployer statistics. Tax-exempt and agricultural-production businesses are excluded from nonemployer statistics.

Type of Data Collection Operation: The universe of nonemployer firms is created annually as a byproduct of the Census Bureau's Business Register processing for employer establishments. If a business is active but without paid employees, then it becomes part of the potential nonemployer universe. Industry classification and receipts are available for each potential nonemployer business. These data are obtained primarily from the annual business income tax returns of the Internal Revenue Service (IRS). The potential nonemployer universe undergoes a series of complex processing, editing, and analytical review procedures at the Census Bureau to distinguish nonemployers from employers and to correct and complete data items used in creating the data tables.

Estimates of Sampling Error: Not applicable.

Other (Nonsampling) Errors: The data are subject to nonsampling errors, such as industry misclassification as well as errors of response, keying, nonreporting, and coverage.

Sources of Additional Material: U.S. Census Bureau, Nonemployer Statistics <<http://www.census.gov/econ/nonemployer/index.html>>.

Population Estimates

Universe, Frequency, and Types of Data: The U.S. Census Bureau annually produces estimates of total resident population for each state and county. County population estimates are produced with a component of population change method, while the state population estimates are solely the sum of the county populations.

Type of Data Collection Operation: The Census Bureau develops county population estimates with a demographic procedure called an “administrative records component of population change” method. A major assumption underlying this approach is that the components of population change are closely approximated by administrative data in a demographic change model. In order to apply the model, Census Bureau demographers estimate each component of population change separately. For the population residing in households the components of population change are births, deaths, and net migration, including net international migration. For the nonhousehold population, change is represented by the net change in the population living in group quarters facilities.

Estimates of Sampling Error: Not applicable.

Other (Nonsampling) Errors: Not available.

Sources of Additional Material: U.S. Census Bureau, “Estimates and Projections Area Documentation, State and County Total Population Estimates,” at <<http://www.census.gov/popest/topics/methodology/2008-st-co-meth.pdf>>. Also see <<http://www.census.gov/popest/topics/methodology/>>.

For methodological information on other population estimates datasets, such as “Housing Unit Estimates” and “State Population Estimates by Age, Sex, Race, and Hispanic Origin,” see <<http://www.census.gov/popest/topics/methodology/>>.

U.S. DEPARTMENT OF EDUCATION

National Center for Education Statistics Integrated Postsecondary Education Data Survey (IPEDS), Completions

Universe, Frequency, and Types of Data: Annual survey of all Title IV (federal financial aid) eligible postsecondary institutions to obtain data on earned degrees and other

formal awards, conferred by field of study, level of degree, sex, and by racial/ethnic characteristics (every other year prior to 1989, then annually).

Type of Data Collection Operation: Complete census.

Data Collection and Imputation Procedures: Data are collected through a Web-based survey in the fall of every year. Missing data are imputed by using data of similar institutions.

Estimates of Sampling Error: Not applicable.

Other (Nonsampling) Errors: For 2005–06, the response rate for degree-granting institutions was 100.0 percent.

Sources of Additional Material: U.S. Department of Education, National Center for Education Statistics (NCES), *Postsecondary Institutions in the United States: Fall 2007 and Degrees and Other Awards Conferred: 2006–07 and 12-month enrollment, 2006–07*. See <<http://www.nces.ed.gov/ipeds/>>.

U.S. FEDERAL BUREAU OF INVESTIGATION

Uniform Crime Reporting (UCR) Program

Universe, Frequency, and Types of Data: Monthly reports on the number of criminal offenses that become known to law enforcement agencies. Data are also collected on crimes cleared by arrest or exceptional means; age, sex, and race of arrestees and for victims and offenders for homicides, number of law enforcement employees, on fatal and nonfatal assaults against law enforcement officers, and on hate crimes reported.

Type of Data Collection Operation: Crime statistics are based on reports of crime data submitted either directly to the FBI by contributing law enforcement agencies or through cooperating state UCR Programs.

Data Collection and Imputation Procedures: States with UCR programs collect data directly from individual law enforcement agencies and forward reports, prepared in accordance with UCR standards, to the FBI. Accuracy and consistency edits are performed by the FBI.

Estimates of Sampling Error: Not applicable.

Other (Nonsampling) Errors: During 2007, law enforcement agencies active in the UCR Program represented 94.6 percent of the total population. The coverage amounted to 95.7 percent of the U.S. population in metropolitan statistical areas, 88.0 percent of the population in cities outside metropolitan areas, and 90.0 percent in nonmetropolitan counties.

Sources of Additional Material: U.S. Department of Justice, Federal Bureau of Investigation, *Crime in the United States*, annual, *Hate Crime Statistics*, annual, *Law Enforcement Officers Killed and Assaulted*, annual, <<http://www.fbi.gov/ucr/ucr.htm>>.

U.S. INTERNAL REVENUE SERVICE

Individual Income Tax Returns

Universe, Frequency, and Types of Data: Annual study of unaudited individual income tax returns, Forms 1040, 1040A, and 1040EZ, filed by U.S. citizens and residents. Data provided on various financial characteristics by size of adjusted gross income, marital status, and by taxable and nontaxable returns. Data by state, based on the population of returns filed, also include returns from 1040NR, filed by nonresident aliens plus certain self-employment tax returns.

Type of Data Collection Operation: Stratified probability sample of 321,006 returns for tax year 2006. The sample is classified into sample strata based on the larger of total income or total loss amounts, the size of business plus farm receipts, and other criteria such as the potential usefulness of the return for tax policy modeling. Sampling rates for sample strata varied from 0.01 percent to 100 percent.

Data Collection and Imputation Procedures: Computer selection of sample of tax return records. Data adjusted during editing for incorrect, missing, or inconsistent entries to ensure consistency with other entries on return.

Estimates of Sampling Error: Estimated CVs for tax year 2006: adjusted gross income less deficit 0.09 percent; salaries and wages 0.16 percent; and tax-exempt interest received 1.17 percent. (State data not subject to sampling error.)

Other (Nonsampling) Errors: Processing errors and errors arising from the use of tolerance checks for the data.

Sources of Additional Material: U.S. Internal Revenue Service, *Statistics of Income, Individual Income Tax Returns*, annual, (Publication 1304).

NATIONAL CENTER FOR HEALTH STATISTICS (NCHS)

National Vital Statistics System

Universe, Frequency, and Types of Data: Annual data on births and deaths in the United States.

Type of Data Collection Operation: Mortality data based on complete file of death records, except 1972, based on 50 percent sample. Natality statistics 1951–1971, based on 50 percent sample of birth certificates, except a 20 percent to 50 percent sample in 1967, received by NCHS.

Data Collection and Imputation Procedures: Reports based on records from registration offices of all states, District of Columbia, New York City, Puerto Rico, Virgin Islands, Guam, American Samoa, and Northern Marianas.

Estimates of Sampling Error: For recent years, there is no sampling for these files; the files are based on 100 percent of events registered.

Other (Nonsampling) Errors: It is believed that more than 99 percent of the births and deaths occurring in this country are registered.

Sources of Additional Material U.S. National Center for Health Statistics, *Vital Statistics of the United States*, Vol. I and Vol. II, annual, and the *National Vital Statistics Reports*. See the NCHS Web site at <<http://www.cdc.gov/nchs/nvss.htm>>.

NATIONAL HIGHWAY TRAFFIC SAFETY ADMINISTRATION (NHTSA)

Fatality Analysis Reporting System (FARS)

Universe, Frequency, and Types of Data: FARS is a census of all fatal motor vehicle traffic crashes that occur throughout the United States including the District of Columbia and Puerto Rico on roadways customarily open to the public. The crash must be reported to the state /jurisdiction and at least one directly related fatality must occur within 30 days of the crash.

Type of Data Collection Operation: One or more analysts, in each state, extract data from the official documents and enter the data into a standardized electronic database.

Data Collection and Imputation Procedures: Detailed data describing the characteristics of the fatal crash, the vehicles and persons involved are obtained from police crash reports, driver and vehicle registration records, autopsy reports, highway department, etc. Computerized edit checks monitor the accuracy and completeness of the data. The FARS incorporates a sophisticated mathematical multiple imputation procedure to develop a probability distribution of missing blood alcohol concentration (BAC) levels in the database for drivers, pedestrians, and cyclists.

Estimates of Sampling Error: Since this is census data, there are no sampling errors.

Other (Nonsampling) Errors: FARS represents a census of all police-reported crashes and captures all data reported at the state level. FARS data undergo a rigorous quality control process to prevent inaccurate reporting. However, these data are highly dependent on the accuracy of the police accident reports. Errors or omissions within police accident reports may not be detected.

Sources of Additional Material: The FARS Coding and Validation Manual, ANSI D16.1 Manual on Classification of Motor Vehicle Traffic Accidents (Sixth Edition).