Appendix B.
Limitations of the Data and Methodology

Introduction

The data presented in this *County and City Data Book* came from many sources. The sources include not only the U.S. Census Bureau, but also other organizations that collect and issue statistics. 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.

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may restrict data collection to a survey universe based on some available list, which may be inaccurate or out of date.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 0 and less than or equal to 1, 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 SE is the most commonly used measure of the sampling error of an estimate. Valid estimates of the SEs of survey estimates can usually be calculated from the data collected in a probability sample.

For convenience, the SE 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 SE of 10 units has an estimated CV of 5 percent.

A sample estimate and an estimate of its SE 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 SE below the estimate to one SE above the estimate would include the average estimate derived from all possible samples; 2) approximately 90 percent of the intervals from 1.6 SEs below the estimate to 1.6 SEs above the estimate would include the average estimate derived from all possible samples; and 3) approximately 95 percent of the intervals from two SEs below the estimate to two SEs 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 SE of 10 units. An approximate 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 SEs or CVs, are available. To obtain estimates of the estimated SEs 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://fcsm.ssd.census.gov>.

Principal Databases: Beginning below are brief descriptions of some 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 Agricultural 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; agriculture products sold; total cropland; irrigated land; farm operator characteristics; livestock and poultry inventory and sales; and selected crops harvested. Market value of land and buildings, total farm production expenses, machinery and equipment, fertilizer and chemicals, and farm labor are estimated from a sample of farms.

Data Collection and Imputation Procedures: Data collection takes place by mailing questionnaires to all farmers and ranchers. Nonrespondents are contacted by telephone and correspondence follow-ups. Imputations were made for all nonresponse items/characteristics. Coverage adjustments were made to account for missed farms and ranches.

Estimates of Sampling Error: Variability in the estimates is due to the sample selection and estimation for items collected by sample and census nonresponse and coverage estimation procedures. The CVs for national and state estimates are generally very small. The response rate is approximately 81 percent.

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. Results from the evaluation program for the 2002 census indicate the net undercoverage amounted to about 18 percent of the nation's total farms.

Sources of Additional Material: U.S. Department of Agriculture, National Agricultural Statistics Service, *2002 Census of Agriculture*, Volume 1, *Subject Series C* Part 1, *Agriculture Atlas of the U.S.*; Part 2, *Coverage Evaluation*; Part 3, *Rankings of States and Counties*. See also the *2002 Census of Agriculture*, Volume 1, Chapter 1, U.S. National Level Data, Appendix C, Statistical Methodology at <http://www.nass.usda.gov/census/census02/volume1/us/us2appxc.pdf>.

U.S. BUREAU OF LABOR STATISTICS (BLS)

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 160,000 businesses and government agencies, which represent approximately 400,000 individual work sites.

Data Collection and Imputation Procedures: Each month, the state agencies cooperating with 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 SE for total nonfarm employment is 0.2 percent.

Other (nonsampling) Errors: Estimates of employment adjusted annually to reflect complete universe. The average adjustment is 0.2 percent over the last decade, with an absolute range from less than 0.05 percent to 0.5 percent.

Sources of Additional Material: U.S. Bureau of Labor Statistics, *Employment and Earnings*, monthly, Explanatory Notes and Estimates of Errors, Tables 2-A through 2-F. See also the *BLS Handbook of Methods*, Chapter 1, Labor Force Data Derived from the Current Population Survey, and Chapter 2, Employment, Hours, and Earnings from the Establishment Survey. The BLS Handbook may be found at <http://www.bls.gov/opub/hom>.

U.S. DEPARTMENT OF COMMERCE

U.S. Bureau of Economic Analysis (BEA)

Regional Economic Information System (REIS)

Universe, Frequency, and Types of Data: REIS 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 by-product of the administration of various federal and state government programs. The most important sources of these data are the state unemployment insurance programs of BLS , the social insurance programs of the Centers for Medicare and Medicaid Services, federal income tax program of the Internal Revenue Service (IRS), 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 USDA and the Census of Population and Housing conducted by the Census Bureau. Other sources may include estimates of farm proprietors' income by USDA, wages and salaries from County Business Patterns from the Census Bureau, and the Quarterly Census of Employment and Wages by the U.S. 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 datasets may affect personal income estimates.

Sources of Additional Material: U.S. Bureau of Economic Analysis, *Local Area Personal Income and Employment Methodology, 2005.* See also <http://www.bea.gov/regional/docs/lapi2005>. Methodological information on other BEA datasets, such as state personal income and gross state product, may be found at <http://www.bea.gov/bea/regional/articles.cfm?section=methods>.

U.S. CENSUS BUREAU

American Community Survey (ACS)

Universe, Frequency, and Types of Data: Nationwide survey to obtain data about demographic, social, economic, and housing characteristics of people, households, and housing units. Covers household population and excludes the population living in institutions, college dormitories, and other group quarters.

Type of Data Collection Operation: Two-stage stratified annual sample of approximately 838,000 housing units. The ACS samples housing units from the Master Address File (MAF). The first stage of sampling involves dividing the United States into primary sampling units PSUs, most of which comprise a metropolitan area, a large county, or a group of smaller counties. Every PSU falls within the boundary of a state. The PSUs are then grouped into strata on the basis of independent information, that is, information obtained from the decennial census or other sources. The strata are constructed so that they are as homogeneous as possible with respect to social and economic characteristics that are considered important by ACS data users. A pair of PSUs were selected from each stratum. The probability of selection for each PSU in the stratum is proportional to its estimated 1996 population. In the second stage of sampling, a sample of housing units within the sample PSUs is drawn. Ultimate sampling units (USUs) are housing units. The USUs sampled in the second stage consist of housing units that are systematically drawn from sorted lists of addresses of housing units from the MAF.

Data Collection and Imputation Procedures: The ACS is conducted every month on independent samples. Each housing unit in the independent monthly sample is mailed a prenotice letter announcing the selection of the address to participate, a survey questionnaire package, and a reminder postcard. These sample units receive a second (replacement) questionnaire package if the initial questionnaire is not returned by a scheduled date. In the mailout/mailback sites, sample units for which a questionnaire is not returned in the mail and for which a telephone number is available are defined as the telephone nonresponse follow-up universe. Interviewers attempt to contact and interview these mail nonresponse cases. Sample units from all sites that are still unresponsive 2 months after the mailing of the survey questionnaires and directly after the completion of the telephone follow-up operation are subsampled at a rate of 1 in 3. The selected nonresponse units are assigned to field representatives, who visit the units, verify their existence or declare them nonexistent, determine their occupancy status, and conduct interviews. After data collection is completed, any remaining incomplete or inconsistent information was imputed during the final automated edit of the collected data.

Estimates of Sampling Error: The data in the ACS products are estimates of the actual figures that would have been obtained by interviewing the entire population using the same methodology. The estimates from the chosen sample also differ from other samples of housing units and persons within those housing units.

Other (nonsampling) Errors: 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 collect and process survey data. An important goal of the ACS is to minimize the amount of nonsampling error introduced through nonresponse for sample housing units. One way of this is by following up on mail nonrespondents.

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

2002 Census of Governments

Universe, Frequency, and Types of Data: The universe of local governments enumerated for the 2002 Census of Governments includes county, municipal, town, township, and special-purpose local governments. The census of governments is taken every 5 years. The 2002 Census of Governments, similar to those taken since 1957, covers three major subject fields: government organizations, public employment, and government finances.

Type of Data Collection Operation: The census of governments includes more than 87,000 local government units and in 2002 covered 3,034 county governments, 19,429 municipal governments, 16,504 town or township governments, 35,052 special district governments, and 13,506 school district governments. The universe list used as the mail and control file for all phases of the census of governments is updated periodically to add newly established units that meet Census Bureau criteria for independent governments and to delete dissolved or inactive units.

Data Collection and Imputation Procedures: There are three phases of state and local government data collection. Phase 1 is precensus research to classify and identify all 87,525 local governments. It includes extensive legal research into government structure by state, as well as a mailout/mailback survey, and produces and updated list of all local governments and selected data. Phase 2 covers all of the state and local governments and expands the census year annual finance survey from about 14,000 to 87,575 state and local governments. It uses in-house data compilations of source documents for many of the state and largest local governments, consolidated data submissions (usually electronic files) for about 55,000 local governments, Internet data collection capabilities, and a mailout/mailback survey of the remaining governments. Phase 3 covers all of the federal civilian, state, and local governments and expands the census-year annual employment survey from about 10,000 to all 87,575 local governments. It relies on consolidated submissions from more than 30 state respondents and an Internet data collection capability, with the remainder obtained through a mailout/mailback survey.

For counties, municipalities, and townships that failed to respond, missing data were imputed by using the most recently reported prior census or annual survey data. These data were adjusted by median growth rates calculated from similar types of governments, i.e., in the same state, same type, and about the same size. The appropriate growth rate was applied to the nonreporting government's year data to get a current year imputed value. For general purpose governments that failed to report in the prior census or subsequent annual surveys, a responding county, municipality, or township from the same group as the nonrespondent was randomly selected and per capita values for the respondent were multiplied by the population of the nonrespondent to get imputed values for the nonrespondent. For special districts, because of an initial lower response to the mail canvass than for general purpose governments, the special district government portion of the 2002 Census of Governments employed secondary sources and follow-ups more extensively. These efforts included using 2001 fiscal year data if they were available; summary financial data obtained during the organization phase of the 2002 Census of Governments, the American Hospital Association, Mergent Municipal and Government Manual, Bond Buyer's database, and various state agencies and departments.

Estimates of Sampling Error: Since the census of governments is based on a survey whose coverage extends to all governments in the universe, there is no sampling error to be accounted for.

Other (nonsampling) Errors: Only 10.7 percent of counties, and 18.5 percent of the municipalities and townships, and 35.3 percent of the special districts were total nonrespondents to the census. Other nonsampling errors include response errors and processing errors, many of which are corrected through computer and analyst checks.

Sources of Additional Material: U.S. Census Bureau, 2002 Census of Governments, various reports and Web site at <http://www.census.gov/govs/www/cog2002.html> and the "Federal, State, and Local Governments, Government Finance and Employment Classification Manual" found at <http://ftp2.census.gov/govs/class/classfull.pdf>.

2002 Economic Census

Universe, Frequency, and Types of Data: Conducted every 5 years to obtain data on number of establishments, number of employees, total payroll size, total sales/receipts/revenues, and other industry-specific statistics. In 2002, the universe was all employer and nonemployer establishments primarily engaged in wholesale, retail, utilities, finance and insurance, real estate, transportation and warehousing, information, education, health care, and other service industries.

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 a 5 percent to 25 percent sample of the small employer firms. Firms with no employees were not required to file a census return.

Data Collection and Imputation Procedures: Mail questionnaires were used with both mail and telephone follow-ups for nonrespondents. Data for nonrespondents and for small employer firms not mailed a questionnaire were obtained from administrative records of other federal agencies or imputed. Nonemployer data were obtained exclusively from IRS 2002 income tax returns.

Estimates of Sampling Error: Not applicable for basic data such as sales, revenue, receipts, and payroll.

Other (nonsampling) Errors: Establishment response rates by North American Industry Classification System (NAICS) sector in 2002 ranged from 80 percent to 89 percent. Item response rates generally ranged from 50 percent to 90 percent with lower rates for the more detailed questions. Nonsampling errors may occur during the collection, reporting, and keying of data, and due to industry misclassification.

Sources of Additional Material: U.S. Census Bureau, 2002 Economic Census, Geographic Area Series Reports (by NAICS sector), Appendix C and <http://www.census.gov/econ/census02/guide/index.html>.

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 1980, 1990, and 2000, complete census for some items—age, date of birth, sex, race, 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 were included.

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 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, *1990 Census of Population and Housing*, *Content Reinterview Survey: Accuracy of Data for Selected Population and Housing Characteristics as Measured by Reinterview*, CPH-E-1; *1990 Census of Population and Housing*, *Effectiveness of Quality Assurance*, CPH-E-2; Programs to Improve Coverage in the 1990 Census, CPH-E-3. For Census 2000, 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 companies maintained and updated by the 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 persons, domestic service workers, 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 IRS and the Social Security Administration.

Estimates of Sampling Error: Not applicable.

Other (nonsampling) Error: The data are subject to nonsampling errors, such as industry classification errors, as well as errors of response, keying, and nonreporting.

Sources of Additional Materials: U.S. Census Bureau, *General Explanation of County Business Patterns*. See also "Frequently Asked County Business Patterns (CBP) Questions" at <http://www.census.gov/epcd/cbp/view/cbpfaq.html.

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: For permit-issuing places, probability sample of 850 housing units obtained from 19,000 permit-issuing places. For nonpermit places, multistage probability sample of new housing units selected in 169 PSUs. In those areas, all roads are canvassed in selected enumeration districts.

Data Collection and Imputation Procedures: Data are obtained by telephone inquiry and field visit.

Estimates of Sampling Error: Estimated CV of 3 percent to 4 percent for estimates of national totals but may be higher 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: U.S. Census Bureau, "New Residential Construction" at <http://www.census.gov/const/www/newresconstindex.html>.

Population Estimates

Universe, Frequency, and Types of Data: The 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, "State and County Total Resident Population Estimates Method: July 1, 2006," at <http://www.census.gov/popest/topics/methodology/2006_st_co_meth.html>.

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. FEDERAL BUREAU OF INVESTIGATION (FBI)

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 collected on crimes cleared by arrest; by age, sex, and race of arrestees and for victims and offenders for homicides; 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: Coverage of 94 percent of the population (95 percent in metropolitan statistical areas [MSAs], 84 percent in cities outside of metropolitan areas, and 89 percent in nonmetropolitan counties) by UCR program through a varying number of agencies reporting.

Sources of Additional Material: U.S. Federal Bureau of Investigation, *Crime in the United States*, annual. For most recent report, see Web site at <http://www.fbi.gov/ucr/ucr.htm#cius>.

For additional information, see Appendix I "Methodology" at <http://www.fbi.gov/ucr/cius_03/pdf/03sec7.pdf>.

U.S. INTERNAL REVENUE SERVICE (IRS)

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 100 percent file, also include returns from 1040NR, filed by nonresident aliens, plus certain self-employment tax returns.

Type of Data Collection Operation: Annual 2002 stratified probability sample of approximately 176,000 returns broken into sample strata based on the larger of total income or total loss amounts as well as the size of business plus farm receipts. Sampling rates for sample strata varied from 0.05 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 2002—adjusted gross income less deficit 0.12 percent; salaries and wages 0.21 percent; and tax exempt interest received 1.78 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. For recent report, see Web site at <http://www.irs.gov/taxstats/index.html>. For information about methodology and sample, see "Description of the Sample" at <http://www.irs.gov/pub/irs-soi/02insec2.pdf>.

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. Beginning 1972, data from some states received through Vital Statistics Cooperative Program

(VSCP) and complete file used; data from other states based on 50 percent sample. Beginning 1986, all reporting areas participated in the VSCP.

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 Mariana Islands.

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: Data on births and deaths believed to be at least 99 percent complete.

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

For more information about the National Vital Statistics System, see <http://www.cdc.gov/nchs/data/misc/usvss.pdf>.