# CREDIBLE INTERVAL ESTIMATES FOR OFFICIAL STATISTICS WITH SURVEY NONRESPONSE

Charles F. Manski
Department of Economics and Institute for Policy Research
Northwestern University

First Public Draft: February 2013
This Draft: April 2013

This research was supported in part by National Science Foundation grant SES-1129475. I am grateful to Aanchal Jain for excellent research assistance and to Bruce Spencer for helpful discussions. I have benefitted from the opportunity to present this work in a seminar at the Institute for Social and Economic Research, University of Essex.

# Abstract

Government agencies commonly report official statistics based on survey data as point estimates, without accompanying measures of error. In the absence of agency guidance, users of the statistics can only conjecture the error magnitudes. Agencies could mitigate misinterpretation of official statistics if they were to measure potential errors and report them. Agencies could report sampling error using established statistical principles. It is more challenging to report nonsampling errors because there are many sources of such errors and there has been no consensus about how to measure them. To advance discourse on practical ways to report nonsampling error, this paper considers error due to survey nonresponse. I summarize research deriving interval estimates that make no assumptions about the values of missing data. In the absence of assumptions, one can obtain computable bounds on the population parameters that official statistics intend to measure. I also explore the middle ground between interval estimation making no assumptions and traditional point estimation using weights and imputations to implement assumptions that nonresponse is conditionally random.

## Introduction

Government agencies commonly report official statistics based on survey data as point estimates, without accompanying measures of error. Agency publications documenting the data and methods used to produce official statistics may acknowledge that the estimates are subject to sampling and nonsampling error, but the publications do not quantify error magnitudes. News releases that communicate official statistics to the public present the estimates with little if any mention of the possibility of error. Some prominent American examples include the reporting of employment and income statistics by the Bureau of Labor Statistics (BLS) and the Census Bureau.

Reporting official statistics as point estimates without error measures encourages the public to believe that errors are small and inconsequential. In the absence of agency guidance, persons who understand that official statistics are subject to error must fend for themselves and conjecture the error magnitudes. Thus, users of official statistics may misinterpret the information that the statistics provide.

Government agencies could mitigate misinterpretation of official statistics if they were to measure potential errors and report them in their news releases and other publications. Using established statistical principles, agencies such as BLS and Census could report sampling error for levels and temporal changes in important statistics based on survey data. The BLS could report confidence intervals for key employment statistics, such as the level and month-to-month changes in the unemployment rate. The Census Bureau could do likewise for key income statistics, such as the poverty rate and median household income.

It is more challenging for agencies to report nonsampling errors for official statistics. There are many sources of such errors and there has been no consensus about how to measure them. Yet these facts do not justify ignoring nonsampling error. Having agency analytical staffs make good-faith efforts to measure nonsampling error would be more informative to the public than having agencies report official statistics as if they are truths.

To advance discourse on practical ways to report nonsampling error, this paper considers error in official statistics due to survey nonresponse. I study nonresponse error for three reasons. First, nonresponse is common in the surveys used to compute important official statistics. Unit and item nonresponse regularly make key data missing for substantial fractions of the persons sampled. For example, the official statistics on household income reported by the Census Bureau are based on the Annual Social and Economic (ASEC) Supplement to the Current Population Survey (CPS). During the period 2002-2012, 7 to 9 percent of the sampled households yielded no income data due to unit nonresponse and 41 to 47 percent of the interviewed households yielded incomplete income data due to item nonresponse.

Second, statistical agencies have used traditional but untenable assumptions—namely that nonresponse is random conditional on specified observed covariates—to form point estimates with nonresponse. These assumptions have been implemented as weights for unit nonresponse and imputations for item nonresponse, despite the fact that agencies do not know the consequences. A Census Bureau document describing the American Housing Survey is revealing. The document states (U. S. Census Bureau, 2011):

> "Some people refuse the interview or do not know the answers. When the entire interview is missing, other similar interviews represent the missing ones . . . . For most missing answers, an answer from a similar household is copied. The Census Bureau does not know how close the imputed values are to the actual values."

Indeed, lack of knowledge of the closeness of imputed values to actual ones is common.

Third, methodological research has shown how to form interval estimates that make no assumptions about nonresponse; see Manski (1989, 1994, 2003) and Horowitz and Manski (1998, 2000) inter alia. This research has also shown how to form confidence intervals that jointly measure sampling error and potential nonresponse error (e.g., Horowitz and Manski, 2000; Imbens and

Manski, 2004). Thus, we know how to report credible interval estimates for official statistics with survey nonresponse.

Section 2 gives several illustrations of official statistics reported as point estimates without error measures. I discuss the reporting of employment statistics by the BLS, income statistics by the Census, and GDP growth by the Bureau of Economic Analysis.

Section 3 summarizes basic results and specific findings on interval estimation that make no assumptions about nonresponse. The central problem is identification rather than statistical inference from finite samples. That is, in the absence of assumptions about nonresponse, one can only obtain bounds on the population parameters that official statistics intend to measure. These bounds provide the basis for formation of interval estimates that may be computed using available survey data.

Applying these findings, Section 4 uses data from the ASEC and the monthly CPS to form interval estimates for median household income, the family poverty rate, and the unemployment rate. I provide one set of estimates that take into account item nonresponse alone and another that recognizes unit response as well. The estimates show vividly that item nonresponse poses a huge problem for inference on the American income distribution, and that unit nonresponse exacerbates the problem. While item nonresponse is a relatively minor source of error for the unemployment rate, unit nonresponse is highly consequential.

Section 5 explores the middle ground between interval estimation making no assumptions on nonresponse and traditional point estimation assuming that nonresponse is conditionally random. Interval estimation making no assumptions has maximal credibility, but it may be excessively conservative if agency analysts have some understanding of the nature of nonresponse. Point estimation assuming random nonresponse has maximal precision but may have little credibility. The middle ground derives interval estimates based on assumptions that may include random nonresponse as one among various possibilities. I do not recommend adoption of a particular middle-ground assumption for reporting of official statistics. I only pose some alternatives that statistical agencies may want to consider.

Section 6 makes concluding comments on measurement of nonresponse error in official statistics. I urge statistical agencies to report statistics that quantify potential error openly.

## 2. Reporting Official Statistics as Point Estimates: Some Illustrations

To illustrate current practice, I describe the reporting of widely publicized national economic statistics by three agencies of the federal government. Missing data may be a significant concern in each case.

### 2.1. BLS Reporting of Employment Statistics

On the first Friday of each month, the BLS issues *The Employment Situation*, a monthly news release reporting official employment statistics for the previous month. For example, the BLS reported this on October 5, 2012 (U. S. Bureau of Labor Statistics, 2012): "The unemployment rate decreased to 7.8 percent in September, and total nonfarm payroll employment rose by 114,000." The unemployment-rate statistic is based on data on households sampled in the CPS. The one on nonfarm employment is based on data collected from employer establishments sampled in the Current Employment Statistics survey (CES).

The BLS monthly news release reports employment statistics as point estimates, without measures of potential error. A Technical Note issued with the news release contains a section on *Reliability of the estimates* that acknowledges the possible presence of errors, beginning with the statement "Statistics based on the household and establishment surveys are subject to both sampling and nonsampling error." The section describes the conventional use of standard errors and confidence intervals to measure sampling error, providing a few numerical illustrations.

The Technical Note then turns to nonsampling errors, stating that they "can occur for many reasons, including the failure to sample a segment of the population, inability to obtain information for all respondents in the sample, inability or unwillingness of respondents to provide correct information on a timely basis, mistakes made by respondents, and errors made in the collection or processing of the data." The Note does not indicate the magnitudes of the nonsampling errors that may be present in the employment statistics.

### 2.2. Census Reporting of Income Statistics

Each year the Census Bureau reports statistics on the income distribution based on data collected in the ASEC supplement to the CPS. In a news release issued September 12, 2012, the Census Bureau declared (U. S. Census Bureau, 2012A): "The nation’s official poverty rate in 2011 was 15.0 percent, with 46.2 million people in poverty. After three consecutive years of increases, neither the poverty rate nor the number of people in poverty were statistically different from the 2010 estimates." Thus, the Census release provided point estimates, acknowledged but did not quantify sampling error, and did not mention nonsampling error.

A Census Bureau publication gives this explanation for the decision of the Bureau not to report standard errors for income statistics (U. S. Census Bureau, 2012B, p. 7):

> "While it is possible to compute and present an estimate of the standard error based on the survey data for each estimate in a report, there are a number of reasons why this is not done. A presentation of the individual standard errors would be of limited use, since one could not possibly predict all of the combinations of results that may be of interest to data users. Additionally, data users have access to CPS microdata files, and it is impossible to compute in advance the standard error for every estimate one might obtain from those data sets."

This reasoning explains why the Bureau cannot measure sampling error for every logically possible application of the CPS data. It does not explain why the Bureau chooses not to report sampling error for the income statistics that it highlights in news releases.

### 2.3. Bureau of Economic Analysis Reporting of GDP Growth

The Bureau of Economic Analysis (BEA) of the U.S. Department of Commerce reports quarterly estimates of growth in gross domestic product (GDP). The BEA initially reports an "advance" estimate based on incomplete data and then reports revisions one and two months later as further data become available. For example, a November 29, 2012 news release stated (Bureau of Economic Analysis, 2012):

> "Real gross domestic product . . . . increased at an annual rate of 2.7 percent in the third quarter of 2012. . . . . The GDP estimate released today is based on more complete source data than were available for the 'advance' estimate issued last month. In the advance estimate, the increase in real GDP was 2.0 percent."

A journal article describing the measurement of GDP explains the reasons for such revisions as follows (Landefeld, Seskin, and Fraumeni, 2008, p. 194):

> "For the initial monthly estimates of quarterly GDP, data on about 25 percent of GDP—especially in the service sector—are not available, and so these sectors of the economy are estimated based on past trends and whatever related data are available. . . . . The initial monthly estimates of quarterly GDP based on these extrapolations are revised as more complete data become available. . . . . The successive revisions can be significant, but the initial estimates provide a snapshot of economic activity much like the first few seconds of a Polaroid photograph in which an image is fuzzy, but as the developing process continues, the details become clearer."

Although this passage recognizes that initial quarterly estimates of GDP growth are subject to error, BEA practice has been to report these estimates without providing quantitative measures of potential error.

## 3. Interval Estimation Without Assumptions on Nonresponse

Many official statistics aim to measure parameters of a probability distribution characterizing a population. For example, the poverty rate is intended to be the fraction of persons or families with income below specified levels. The unemployment rate is defined to be the fraction of unemployed persons among those who are in the labor force. In each case the statistic is a parameter of a probability distribution $P(y|x \in X^*)$, where y is a specified outcome taking values in a space Y, x is a specified covariate taking values in a space X, and $X^* \subset X$ is a specified subset of covariate values. The parameter of interest, say $\theta[P(y|x \in X^*)]$, may be the mean or a quantile of y within the sub-population with covariate values in $X^*$.

Surveys such as the CPS draw stratified random samples of population units, ask sample members to report their values of (y, x), and use the responses to estimate official statistics. Some sample members provide complete information and some choose not to respond to a subset of the questions posed (item nonresponse). Some provide no information at all (unit nonresponse), either because they choose not to be interviewed or because survey staff are unable to contact them. Thus, data on either or both of y and x may be missing for some sample members.

Estimation of population parameters with no assumptions about nonresponse is basically a matter of contemplating all the values that the missing data might take. Section 3.1 explains the inferential problem in abstraction. Section 3.2 summarizes findings on outcome nonresponse, which is relatively simple to study.

### 3.1. Basic Ideas

To study inference with missing data, it is productive to first consider identification and then statistical inference. In identification analysis, one supposes that all population units are sample members. Hence, one knows the distributions of y and x for units who report them. These distributions are $P(y|z_y = 1)$, $P(x|z_x = 1)$, and $P(y, x|z_y = z_x = 1)$, where $z_y = 1$(or 0) if a population unit would (or not) report y, and $z_x = 1$(or 0) if the unit would (or not) report x. One also knows the distribution $P(z_y, z_x)$ of response.

Given knowledge of these observable distributions, one may determine what this implies about the parameter $\theta[P(y|x \in X^*)]$. The generic finding is that $\theta[P(y|x \in X^*)]$ lies in a set of values, say $H\{\theta[P(y|x \in X^*)]\}$, called its *identification region* or *identified set*. The parameter is *point-identified* if the identification region contains just one point. It is *partially identified* if the region contains multiple values but is a proper subset of the space of all logically possible values of the parameter (Manski, 2003).

With identification determined, one may study statistical inference when a sampling process draws a finite number of population units. Suppose for simplicity that one draws a random sample of N units. A natural way to estimate the parameter is to use the empirical distributions $P_N(y|z_y = 1)$, $P_N(x|z_x = 1)$, $P_N(y, x|z_y = z_x = 1)$, and $P_N(z_y, z_x)$ to estimate their population counterparts. This yields an estimate $H_N\{\theta[P(y|x \in X^*)]\}$ of the identification region of the parameter. Given ordinary regularity conditions, the Strong Law of Large Numbers implies that this set-valued estimate is consistent; that is, it converges almost surely to the set $H\{\theta[P(y|x \in X^*)]\}$ as $N \rightarrow \infty$.

One can also form confidence sets to measure the uncertainty created by sampling variation. Recall the standard definition of a confidence set for a real parameter $\theta$. One first specifies a *coverage probability* $\alpha$, where $0 < \alpha < 1$. One next considers alternative ways to use the sample data to form sets on the real line. Let $C(\cdot)$ be a set-valued function that maps the data into a set on the real line; thus, for each possible value $\psi$ of the sample data, $C(\psi)$ is the set that results when the data are $\psi$. Then $C(\cdot)$ gives an $\alpha$-confidence set for $\theta$ if $\text{Prob}[\psi: \theta \in C(\psi)] = \alpha$. In words, an $\alpha$-confidence set contains the true value of $\theta$ with probability $\alpha$ as the sampling process is engaged repeatedly to draw independent data samples.

It typically is not possible to determine the exact coverage probability of a confidence set. Hence, statisticians seek asymptotically valid confidence sets, whose coverage probabilities can be shown to converge to $\alpha$ as the sample size grows. A common practice begins with a consistent estimate of $\theta$, say $\theta_N(\psi)$, and constructs an interval of the form $[\theta_N(\psi) - \delta_{0N}(\psi), \theta_N(\psi) + \delta_{1N}(\psi)]$, where $\delta_{0N}(\psi) > 0$ and $\delta_{1N}(\psi) > 0$ are chosen so that $\text{Prob}\{\psi: \theta \in [\theta_N(\psi) - \delta_{0N}(\psi), \theta_N(\psi) + \delta_{1N}(\psi)]\}$ converges to $\alpha$ as N increases.

Although the statistics literature has focused on parameters that are point-identified, the standard definition of a confidence set also applies to parameters that are partially identified. In addition, one can define confidence sets for identification regions. Let $H(\theta)$ be the identification region for $\theta$. Then $C(\cdot)$ gives an $\alpha$-confidence set for $H(\theta)$ if $\text{Prob}[\psi: H(\theta) \subset C(\psi)] = \alpha$. An $\alpha$-confidence set for $H(\theta)$ necessarily covers $\theta$ with probability at least $\alpha$. This holds because the true value of $\theta$ lies in $H(\theta)$; hence, $\text{Prob}[\psi: \theta \in C(\psi)] \geq \text{Prob}[\psi: H(\theta) \subset C(\psi)]$. See Imbens and Manski (2004).

### 3.2. Outcome Nonresponse

In practice, surveys have complex patterns of nonresponse. Some sample members may not respond to questions about outcomes, others may not provide covariate data, and others may have jointly missing outcomes and covariates. Derivation of identification regions is particularly simple when only the outcome variable y is subject to nonresponse, the conditioning covariates x always being observed. I summarize findings for this case here, drawing on Manski (1989, 1994, 2003).

See Horowitz and Manski (1998, 2000) for analysis of identification and estimation with covariate nonresponse.

For notational simplicity, I condition throughout on the event that the covariate x takes a particular value rather than on the event that x lies in a specified set $X^*$. Analogous findings hold for any specification of $X^*$.

With nonresponse on outcomes alone, the structure of the inferential problem is displayed by the Law of Total Probability

$$(1) \qquad P(y|x) \;=\; P(y|x, z = 1)P(z = 1|x) + P(y|x, z = 0)P(z = 0|x),$$

where $z \equiv z_y$. Supposing that all population units are sample members, the empirical evidence reveals $P(z = 1|x)$ and $P(z = 0|x)$, the probabilities that an outcome is observed or missing. It also reveals the distribution $P(y|x, z = 1)$ of observable outcomes when $P(z = 1|x) > 0$. The evidence is uninformative regarding the distribution $P(y|x, z = 0)$ of missing outcomes, which may be any probability distribution on Y. Hence, the evidence reveals that $P(y|x)$ lies in the identification region

$$(2) \qquad H[P(y|x)] \;\equiv\; [P(y|x, z = 1)P(z = 1|x) + \gamma P(z = 0|x), \gamma \in \Gamma_Y],$$

where $\Gamma_Y$ denotes the set of all probability distributions on the set Y. This region is a proper subset of $\Gamma_Y$ whenever the probability $P(z = 0|x)$ of missing data is less than one, and is the single distribution $P(y|x, z = 1)$ when $P(z = 0|x) = 0$. Hence, $P(y|x)$ is partially identified when $0 < P(z = 0|x) < 1$ and is point-identified when $P(z = 0|x) = 0$.

The above concerns identification of the entire outcome distribution. Now consider identification of a parameter $\theta[P(y|x)]$. The identification region is the set of all the values it can take when $P(y|x)$ ranges over all of its feasible values. Thus, $H\{\theta[P(y|x)]\} = \{\theta(\eta), \eta \in H[P(y|x)]\}$.

### 3.2.1. Event Probabilities

The above derivation is straightforward but abstract. To begin to show its practical implications, I now suppose that one wants to learn the probability $P(y \in B|x)$ that y falls in a specified set B.

By the Law of Total Probability,

$$(3) \qquad P(y \in B|x) \ = \ P(y \in B|x, z = 1)P(z = 1|x) + P(y \in B|x, z = 0)P(z = 0|x).$$

The empirical evidence reveals $P(y \in B|x, z = 1)$, $P(z = 1|x)$, and $P(z = 0|x)$, but provides no information on $P(y \in B|x, z = 0)$. The last quantity necessarily lies between zero and one. This yields the following sharp bound on $P(y \in B|x)$, developed in Manski (1989):

$$(4) \quad P(y \in B|x, z = 1)P(z = 1|x) \ \le \ P(y \in B|x) \ \le \ P(y \in B|x, z = 1)P(z = 1|x) + P(z = 0|x).$$

Bound (4) is so easy to understand and compute that one might think its use would be standard practice in analysis of survey data. However, I have not been able to find any application in the reporting of official statistics. An early application in academic survey research was performed by Cochran, Mosteller, and Tukey (1954) in their study of statistical problems in the Kinsey report on sexual behavior. On pages 274-282, the authors used bounds of form (4) to measure the possible effects of nonresponse to the Kinsey survey. However, the subsequent statistical literature on analysis of survey data did not follow up on the idea. Indeed, the textbook of Cochran (1977) dismissed bounding the effects of survey nonresponse. Using the symbol $W_2$ to denote the probability of missing data, Cochrane stated (p. 362): "The limits are distressingly wide unless $W_2$ is very small."

Although bound (4) and analogous bounds on other population parameters sometimes are "distressingly wide," I feel that this fact should not dissuade government agencies and researchers from reporting them. Even when wide, the bounds are valuable for two reasons. First, the bounds are maximally credible because they impose no assumptions on the values of missing data. Second, the bounds make explicit the fundamental role that assumptions play in inferential methods that yield tighter findings. Wide bounds reflect real uncertainties that cannot be washed away by assumptions lacking credibility.

### 3.2.2. The Distribution Function and Quantiles

The bound on event probabilities has many applications. An immediate one is the bound that it implies on the distribution function of a real-valued outcome. Suppose that y is real-valued. The distribution function for $P(y|x)$ is $P(y \le t|x)$, $t \in R$. Application of (4) to $P(y \le t|x)$ gives

$$(5) \quad P(y \le t|x, z = 1)P(z = 1|x) \le P(y \le t|x) \le P(y \le t|x, z = 1)P(z = 1|x) + P(z = 0|x).$$

The feasible distribution functions are all increasing functions of t that take values between the lower and upper bounds in (5) for all values of t.

Bound (5) can be inverted to obtain sharp bounds on quantiles of $P(y|x)$. Consider the α-quantile $Q_\alpha(y|x)$. Let $y_0$ and $y_1$ denote the smallest and largest logically possible values of y; thus, $y_0 \equiv \min_{y \in Y}$ and $y_1 \equiv \max_{y \in Y}$. Manski (1994) shows that the sharp lower and upper bounds on $Q_\alpha(y|x)$ are $r(\alpha, x)$ and $s(\alpha, x)$, where

$$(6) \quad r(\alpha, x) \equiv [\alpha - P(z = 0|x)]/P(z = 1|x) \text{ quantile of } P(y|x, z = 1) \text{ if } P(z = 0|x) < \alpha,$$
$$\equiv y_0 \text{ otherwise.}$$

$$(7) \quad s(\alpha, x) \equiv \alpha/P(z = 1|x) \text{ quantile of } P(y|x, z = 1) \text{ if } P(z = 0|x) \le 1 - \alpha,$$
$$\equiv y_1 \text{ otherwise.}$$

### 3.2.3. Means of Functions of the Outcome

Now let g(·) be a specified function of y and consider inference on the conditional mean E[g(y)|x]. Suppose first that g(·) has bounded range, with sharp lower and upper bounds $g_0$ and $g_1$.

The Law of Iterated Expectations gives

$$(8)\qquad E[g(y)|x] \;=\; E[g(y)|x, z = 1]P(z = 1|x) \;+\; E[g(y)|x, z = 0]P(z = 0|x).$$

The empirical evidence reveals E[g(y)|x, z = 1] and P(z|x). However, the evidence reveals nothing about E[g(y)|x, z = 0], which can take any value in the interval $[g_0, g_1]$. Hence, the identification region for E[g(y)|x] is the interval

$$(9)\quad H\{E[g(y)|x]\}$$

$$= \big[E[g(y)|x, z = 1]P(z = 1|x) + g_0P(z = 0|x),\; E[g(y)|x, z = 1]P(z = 1|x) + g_1P(z = 0|x)\big].$$

The width of the interval is $(g_1 - g_0)P(z = 0|x)$. Thus, the severity of the identification problem varies directly with the probability of missing data.

The situation changes if g(·) is unbounded from below or above; that is, if $g_0 = -\infty$ or $g_1 = \infty$. In such cases, result (9) still holds but has different implications whenever $P(z = 0|x) > 0$. Inspection of (9) shows that the lower bound on E[g(y)|x] is $-\infty$ if $g_0 = -\infty$ and is $\infty$ if $g_1 = \infty$. The identification region has infinite width but remains informative if g(·) is bounded from at least one side. However, it is the entire real line if g(·) is unbounded from both below and above. Thus, the presence of missing data makes assumptions a prerequisite for inference on the mean of an unbounded random variable.

### 3.2.4. Estimation of the Bounds

The sharp bounds on quantiles and the mean outcome may be obtained by the same simple argument. Wherever outcome data are missing, insert the values of y that yield the smallest and largest values of the parameter of interest to obtain the lower and upper bounds.

Given this, estimation of these bounds with sample data is easy. To estimate the lower bound, one supposes that $y_i = y_0$ for every sample member i with missing data. One then computes the usual point estimate of the parameter. To estimate the upper bound, one likewise supposes that $y_i = y_1$ whenever observation i is missing. Thus, estimation of the bound simply requires two imputations of each case of missing data.

It is important to recognize that the present imputations differ from *hot-deck* imputations applied to the CPS and other surveys. The BLS and Census describe the hot deck this way (U. S. Census Bureau, 2006, p. 9-2):

> "This method assigns a missing value from a record with similar characteristics, which is the hot deck. Hot decks are defined by variables such as age, race, and sex. Other characteristics used in hot decks vary depending on the nature of the unanswered question. For instance, most labor force questions use age, race, sex, and occasionally another correlated labor force item such as full- or part-time status."

Thus the agency staff select some sub-vector of covariates (say $x_0$) for which response is complete and determines the empirical distribution $P_N(y|x_0, z_y = 1)$ among sample members who have this value of $x_0$ and who report their outcomes. An outcome is imputed to a sample member i with missing data by drawing a realization at random from $P_N(y|x_0 = x_{0i}, z_y = 1)$. The CPS document offers no evidence that hot-deck imputation yields an outcome distribution for missing data that is close to the actual distribution of such outcomes.

## 4. Interval Estimates for Median Household Income, Family Poverty, and Unemployment

Official statistics describing income and employment in the United States are based on data collected in the Current Population Survey. The basic CPS sampling unit is a household, which may include one or more families or unrelated individuals who reside together. Data on annual income of household members in the preceding calendar year are collected in the ASEC Supplement, which is administered in the period February through April. Data on the current employment status of civilian adult household members are collected in the monthly administration of the CPS.

Applying the findings of Section 3, I use ASEC data collected in 2002-2012 to form interval estimates of median household income and the fraction of families with income below the official poverty threshold in the years 2001-2011. I use monthly CPS data to form interval estimates of the unemployment rate in March of 2012-2012.

To keep the application simple, I do not seasonally adjust the unemployment rate, as does the BLS in its official monthly estimate. Nor do I report sampling confidence intervals for the interval estimates. The ASEC and regular CPS sample sizes are so large that sampling uncertainty is a trivial consideration relative to the identification problem stemming from lack of knowledge of missing data.

### 4.1. Estimates Recognizing Item Nonresponse

The interval estimates in Table 1 show the potential implications of item nonresponse alone, ignoring unit nonresponse. The bound on the unemployment rate is easy to compute. The CPS questionnaire asks whether each adult member of the household is employed, looking for work, or out of the labor force. Data are missing when a response to this question is not obtained. The lower bound on the unemployment rate is the value that the rate would take if all persons with missing data were employed. The upper bound is the value that the rate would take if all persons with missing data were looking for work. See Horowitz and Manski (1998) for the proof.

The bounds on median household income and the family poverty rate would be similarly simple if the ASEC questionnaire were to pose one question asking for total household income and another asking for the total income of each family sub-unit of the household. Then the findings of Section 3.2 would apply immediately. In the absence of assumptions, the lower (upper) bound on median household income would be the value that the median would take if all households with missing data were to have zero (infinite) income. The lower (upper) bound on the family poverty rate would be the value that the rate would take if all families with missing data were to have infinite (zero) income.

In actuality, the ASEC questionnaire inquires about eighteen separate income components, ranging from earnings and pensions to dividends and public assistance. To determine total household income, the Census Bureau sums the responses obtained for these income components across the members of the household, imputing the values of missing data. The ASEC data shows a wide range of item nonresponse patterns, as households differ in the data they provide about the various income components.

For example, nonresponse to the question asking for earnings on the primary job ranges from 17 to 19 percent over the ten-year period. Nonresponse to the question on interest income ranges from 23 to 27 percent over the period. Nonresponse on dividend income ranges from 9 to 12 percent.

Whenever there is missing data on an income component, I take the lower bound on this component to be zero. Hence, the lower bound on total household or family income is the sum of the relevant reported income components. There is no similarly obvious finite upper bound on the value of a missing income component. One could set the upper bound as infinity, but this seems excessively cautious. The interval estimates in Table 1 take the upper bound on each missing income component to be the highest value reported by any member of the ASEC sample. Hence, the upper bound on total household or family income is the sum of the relevant income components, using reported values when available and the maximum value occurring in the sample when data are missing.

Examination of the table shows bounds that vary in their informativeness, which depends both on the fraction of cases with missing data and on the statistic under consideration. First consider median household income and the family poverty rate. The fraction of interviewed households with data missing on some income components is huge. In every year at least 0.41 of households and 0.38 of families had some missing income data. The bounds on the statistics are consequently very wide. For example, the bound on median household income in 2001 is [$32000, $102000] and the bound on the family poverty rate is [0.11, 0.32]. The bounds for 2011 are [$38000, $100000] and [0.14, 0.34] respectively. Thus, in the absence of assumptions on the missing data, the CPS data reveal little about median income or the poverty rate.

Now consider the unemployment rate. Among all civilian adults who are interviewed, the fraction with missing employment data is relatively small but increases over time, ranging from 0.0028 in 2002 to 0.0076 in 2012. The bounds on the unemployment rate correspondingly widen over time, ranging from [0.057, 0.062] in March 2002 to [0.078, 0.090] in March 2012.

Table 1 also shows point estimates that use Census Bureau imputations of missing data. The point estimates always lie within the bounds. This is necessarily the case given that the imputations are logically possible values of the missing data.

[TABLE 1 HERE]

## 4.2. Estimates Recognizing Item and Unit Nonresponse

The above bounds, as wide as they are, only recognize item nonresponse. Consideration of unit nonresponse widens the bounds further. During the period 2002-2012, the fraction of sampled

ASEC households that were not interviewed ranged from 0.069 to 0.089. These unit nonresponse rates are small relative to other major household surveys, yet they still are consequential in the absence of assumptions on the missing data. Unit nonresponse turns out to be particularly consequential for inference on the unemployment rate.

Table 2 reports interval estimates that recognize both item and unit nonresponse. Whereas Table 1A took sample size to be the number of interviewed households, Table 2A takes sample size to be the larger number of households at which an interview was attempted. Sample size excludes housing units that Census interviewers found to be vacant or demolished.

Table 2 does not include a column with point estimates. The Census Bureau does not impute data missing from unit nonresponse. Instead, the Bureau weights the interviewed households with weights derived from an assumption that unit nonresponse is random.

Interval estimation of median household income is straightforward. Each case of unit nonresponse yields exactly one missing household whose income is entirely unknown. Interval estimation of the family poverty rate and the unemployment rate runs into the conceptual difficulty that we do not know the composition of non-interviewed households. To simplify the application, I assume that every non-interviewed household contains exactly one family and one civilian adult in the labor force. Then each case of unit nonresponse yields exactly one missing family and one potentially unemployed person. To the extent that non-interviewed households contain multiple families or multiple individuals in the labor force, the intervals reported in Table 2 would widen further.

[TABLE 2 HERE]

To summarize the table, consider the years 2001 and 2011. Taking unit nonresponse into account, the bound on median household income in 2001 is [$28200, $123643] and the bound on the family poverty rate is [0.10, 0.36]. The bounds for 2011 are [$32132, $100000] and [0.13, 0.39] respectively. The bounds on the unemployment rate range from [0.053, 0.123] in March 2002 to [0.071, 0.162] in March 2012.

## 5. Interval Estimation with Assumptions on Nonresponse

Interval estimates of official statistics that place no assumptions on the values of missing data have maximal credibility. Yet they may be excessively conservative if agency analysts have some understanding of the nature of nonresponse. Traditional point estimates assuming random nonresponse have maximal precision. However, they suppose more understanding of nonresponse than agencies typically possess.

There is much middle ground between interval estimation with no assumptions and point estimation assuming random nonresponse. The middle ground obtains interval estimates based on assumptions that may include random nonresponse as one among various possibilities. This section considers identification under such assumptions. For simplicity, I restrict attention to outcome nonresponse and suppose that the outcome takes finitely many values.

It is unlikely that any one middle-ground assumption about the nature of nonresponse will be appropriate in all settings. Hence, I do not propose adoption of any particular assumption for reporting of official statistics. I only pose some alternatives that statistical agencies may want to consider.

Section 5.1 discusses assumptions that directly constrain the distribution of missing data. Section 5.2 considers ones that constrain the response propensities of persons with different outcomes. Section 5.3 discusses ones that relate unknowns at different points in time.

### 5.1. Assumptions on the Distribution of Missing Outcome Data

Assumptions that constrain the distribution of missing outcome data place $P(y|x, z = 0)$ within some set of outcome distributions. Abstractly, one might assume that $P(y|x, z = 0) \in \Gamma_{cY}$, where $\Gamma_{cY}$ is a specified constrained set of outcome distributions. Recall that, using the empirical evidence alone, the Law of Total Probability (1) yields identification region (2) for $P(y|x)$. The evidence combined with the assumption that $P(y|x, z = 0) \in \Gamma_{cY}$ analogously yields the identification region

$$(10) \qquad H_c[P(y|x)] \equiv [P(y|x\ z = 1)P(z = 1|x) + \gamma P(z = 0|x), \gamma \in \Gamma_{cY}].$$

Whereas the assumption of random nonresponse supposes that $P(y|x, z = 0) = P(y|x, z = 1)$, a middle-ground assumption might assert that $P(y|x, z = 0)$ lies in a neighborhood of $P(y|x, z = 1)$. The specific assumption depend on how one defines neighborhoods. I discuss two possibilities, embodying different perspectives on how missing data may differ from observed outcomes.

First, one might think it credible to assume that some fraction of nonresponse is random and that the remaining fraction arises from an unknown mechanism. This assumption asserts that

$$(11) \qquad P(y|x, z = 0) = (1 - \delta)P(y|x, z = 1) + \delta\gamma,$$

where $\gamma$ is an unknown outcome distribution and $\delta$ is the fraction of nonresponse drawn from $\gamma$. Then $\Gamma_{cY} = [(1 - \delta)P(y|x, z = 1) + \delta\gamma, \gamma \in \Gamma_Y]$. It follows that

$$(12) \quad H_c[P(y|x)] = \{P(y|x, z = 1)[P(z = 1|x) + (1 - \delta)P(z = 0|x)] + \gamma[\delta P(z = 0|x)], \gamma \in \Gamma_Y\}.$$

Identification region (12) has the same form as the region (2) that would be obtained without assumptions, if the fraction of missing data were $\delta P(z = 0|x)$ rather than $P(z = 0|x)$.

Alternatively, one might think it credible to assume that the probability with which each missing outcome value occurs is not too different from the corresponding probability for observed outcomes. Let $0 \leq \lambda_k \leq 1$ be the assumed maximum deviation between the probability that $y = k$ conditional on the outcome being missing and observed. The assumption asserts that

$$(13) \quad |P(y = k|x, z = 0) - P(y = k|x, z = 1)| \leq \lambda_k, \quad \text{all } k \in Y.$$

Hence,

$$(14) \qquad \Gamma_{cY} = [\gamma \in \Gamma_Y\text{: } |\gamma(k) - P(y = k|x, z = 1)| \leq \lambda_k, \quad \text{all } k \in Y].$$

5.2. Assumptions on Response Propensities Conditional on Outcomes

When outcome data are missing, the evidence reveals the response propensities $P(z = 1|x)$ that condition on covariates but only partially reveals the propensities $P(z = 1|x, y)$ that condition on covariates and outcomes. Assumptions that constrain the latter response propensities may have identifying power for $P(y|x)$.

The tools used to exploit such assumptions are Bayes Theorem and the fact that probabilities of mutually exclusive and exhaustive events sum to one. Bayes Theorem gives

$$(15) \qquad P(y = k|x) = P(y = k|x, z = 1)P(z = 1|x)/P(z = 1|x, y = k), \qquad k \in Y.$$

For each k, the evidence reveals P(y = k|x, z = 1) and P(z = 1|x). Hence, constraints on P(z = 1|x, y = k) imply restrictions on P(y = k|x). The maximum (minimum) feasible value of P(z = 1|x, y = k) gives the minimum (maximum) feasible value of P(y = k|x).

To begin, P(z = 1|x, y = k) is a probability, so $0 \le P(z = 1|x, y = k) \le 1$. Summing P(y = k|x) across $k \in Y$ gives a constraint on the vector [P(z = 1|x, y = k), $k \in Y$] that uses the empirical evidence alone, namely

$$(16) \qquad 1 \;=\; \sum_{k \in Y} P(y = k|x) \;=\; \sum_{k \in Y} P(y = k|x, z = 1)P(z = 1|x)/P(z = 1|x, y = k).$$

Further constraints may be posed as assumptions asserting that [P(z = 1|x, y = k), $k \in Y$] lies in a specified set of |Y|-dimensional vectors, say F. Combining the fact that [P(z = 1|x, y = k), $k \in Y$] is a vector of probabilities that solves (16) and the assumption that the vector lies in F yields restrictions on [P(y = k|x), $k \in Y$] via (15).

For example, one might think it credible to assume that each component of [P(z = 1|x, y = k), $k \in Y$] lies in some neighborhood of P(z = 1|x). That is, one might assume that

$$(17) \qquad \alpha_k P(z = 1|x) \;\le\; P(z = 1|x, y = k) \;\le\; \beta_k P(z = 1|x), \qquad k \in K$$

for specified constants $0 \le \alpha_k \le 1 \le \beta_k$, $k \in Y$. Then $F = \times_{k \in Y} [\alpha_k P(z = 1|x), \beta_k P(z = 1|x)]$. The assumption of random nonresponse is the polar case where $\alpha_k = 1 = \beta_k$, $k \in Y$. In this case, application of (15) yields the familiar result P(y = k|x) = P(y = k|x, z = 1), $k \in Y$.

### 5.3. Assumptions Restricting Temporal Variation in Unknowns

I have thus far considered identification of official statistics characterizing outcomes in a single time period (perhaps a month or year), using only the data collected in that period. For this purpose, it has not been necessary to use notation explicitly marking the period of interest.

Now suppose that data have been collected in periods t = 1, . . . , T and index all probability distributions by the period to which they pertain. Suppose that one wants to use the data to learn the time-series of official statistics $\theta[P_t(y|x)]$, t = 1, . . . , T. Then identification analysis depends on whether one maintains assumptions relating unknowns across time periods.

#### 5.3.1. Identification without Assumptions Restricting Temporal Variation

Consider first identification of the time-series vector of official statistics without assumptions relating unknowns across time periods. Then the period-specific findings reported earlier extend immediately to the time series. With no assumptions relating unknowns across time periods, the joint identification region for the time-series vector of statistics is the Cartesian product of the period-specific regions. That is,

$$(18)\quad H\{\theta[P_t(y|x)], t = 1, \ldots, T\} \;=\; \underset{t = 1, \ldots, T}{\times} H_t\{\theta[P_t(y|x)]\},$$

where $H_t\{\theta[P_t(y|x)]\}$ is the identification region for $\theta[P_t(y|x)]$ obtained using only the data collected in period t and the maintained period-specific assumptions.

Result (18) provides the basis for determination of the identification region for any function of the time-series vector of statistics. For example, a common concern of public discourse is to learn the change $\theta[P_t(y|x)] - \theta[P_{t-1}(y|x)]$ in an official statistic between two adjacent periods t − 1 and t. The Cartesian product form of the joint identification region for $\{\theta[P_{t-1}(y|x)], \theta[P_t(y|x)]\}$ implies that the sharp lower (upper) bound on the temporal change in the statistic is the lower (upper) bound of $H_t\{\theta[P_t(y|x)]\}$ minus the upper (lower) bound of $H_{t-1}\{\theta[P_{t-1}(y|x)]\}$.

### 5.3.2. Identification with Assumptions Restricting Temporal Variation

Given assumptions relating unknowns across time periods, the joint identification region for the time-series vector of statistics may be a proper subset of the region (18) obtained without these assumptions. The joint region generally has no simple explicit form, but it can be determined numerically.

Many temporal assumptions may be reasonable to conjecture. Among them, ones supposing that unknown quantities do not change too rapidly with time may be particularly credible when considering statistics on employment, income, and other quantities that are thought to vary relatively smoothly with time. There are multiple ways to formalize the idea. One might restrict the time-series variation of the period-specific distributions of missing outcome data $P_t(y|x, z = 0)$, $t = 1, \ldots, T$. One might restrict the time-series variation of the period-specific response propensities $P_t(z = 1|x, y)$, $t = 1, \ldots, T$. Or one might restrict the time-series variation of the period-specific population outcome distributions $P_t(y|x)$, $t = 1, \ldots, T$.

Each type of assumption yields constraints on the time-series vector of official statistics, the specifics depending on how one formalizes the notion of restricting time-series variation. Moreover, one may combine temporal-variation assumptions with period-specific assumptions of the forms studied in Section 5.1 and 5.2 to achieve additional identifying power.

## 6. Conclusion

The present norm in the reporting of official statistics is to acknowledge nonresponse errors verbally but not quantitatively. The news releases and technical documentation published by government statistical agencies caution readers that point estimates of official statistics are subject to nonresponse and other nonsampling errors. However, the estimates themselves contain no information on the potential magnitude of the errors. The documents published by statistical agencies neither specify nor attempt to justify the assumptions of conditionally random nonresponse used to generate imputations and weights. Hence, users of official statistics must fend for themselves and conjecture error magnitudes.

Statistical agencies could better inform the public if they were to measure potential errors and report them. This paper has shown how to form interval estimates that face up to nonresponse. Section 3 presented maximally credible estimates that make no assumptions about the nature of nonresponse. The empirical applications of Section 4 made plain that nonresponse creates a potentially severe problem for interpretation of American official statistics on household income and civilian unemployment.

The most appealing way to mitigate the identification problem created by nonresponse is to improve the response rates of our surveys. Short of this, the only way to tighten the interval estimates shown in Section 4 is to make assumptions that either directly or indirectly constrain the distribution of missing data. Section 5 suggested various assumptions that may yield narrower intervals. I did not opine on whether any of these assumptions is sufficiently credible that agencies should feel comfortable using them when reporting particular statistics. I recommend that analysts at statistical agencies should consider carefully the types of assumptions they deem credible enough to maintain, determine their identifying power, and report interval estimates accordingly.

References

Bureau of Economic Analysis, U. S. Department of Commerce (2012), *News Release: Gross Domestic Product*, November 29, http://www.bea.gov/newsreleases/national/gdp/gdpnewsrelease.htm, accessed November 28, 2012.

Cochran, W. (1977), *Sampling Techniques*, Third Edition, New York: Wiley.

Cochran, W., F. Mosteller, and J. Tukey (1954), *Statistical Problems of the Kinsey Report on Sexual Behavior in the Human Male*, Washington, DC: American Statistical Association.

Horowitz, J. and C. Manski (1998), "Censoring of Outcomes and Regressors due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations," *Journal of Econometrics*, 84, 37–58.

Horowitz, J. and C. Manski (2000), "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95, 77–84.

Imbens, G. and C. Manski (2004), "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857.

Landefeld, J., E. Seskin, and B. Fraumeni (2008), "Taking the Pulse of the Economy: Measuring GDP," *Journal of Economic Perspectives*, 22, 193-216.

Manski, C. (1989), "Anatomy of the Selection Problem," *Journal of Human Resources*, 24, 343–360.

Manski, C. (1994), "The Selection Problem. "in C. Sims ed. *Advances in Econometrics, Sixth World Congress*, Cambridge: Cambridge University Press.

Manski, C. (2003), *Partial Identification of Probability Distributions*, New York: Springer-Verlag.

U. S. Bureau of Labor Statistics (2012), *Employment Situation News Release*, October 5, www.bls.gov/news.release/archives/empsit_10052012.htm, accessed November 28, 2012.

U. S. Census Bureau (2006), *Current Population Survey Design and Methodology*, Technical Paper 66, Washington, DC: U. S. Census Bureau.

U.S. Census Bureau (2011), *Current Housing Reports, Series H150/09, American Housing Survey for the United States: 2009*, Washington, DC: U.S. Government Printing Office.

U. S. Census Bureau (2012A), *Income, Poverty and Health Insurance Coverage in the United States: 2011*, September 12, www.census.gov/newsroom/releases/archives/income_wealth/cb12-172.html, accessed November 28, 2012.

U. S. Census Bureau (2012B), *Source and Accuracy of Estimates for Income, Poverty, and Health Insurance Coverage in the United States: 2011*, Publication P60_243sa, www.census.gov/hhes/www/p60_243sa.pdf, accessed November 28, 2012.

TABLE 1: Interval Estimates Recognizing Item Nonresponse

| A. Annual Median Household Income | | | | |
|---|---|---|---|---|
| Year | Number of Interviewed Households | Fraction with Missing Data | Interval Estimate | Point Estimate with Imputations |
| 2001 | 78265 | 0.461 | [32000, 102000] | 44200 |
| 2002 | 78310 | 0.466 | [31225, 104612] | 44231 |
| 2003 | 77149 | 0.464 | [32040, 106000] | 45100 |
| 2004 | 76447 | 0.463 | [33000, 106847] | 46324 |
| 2005 | 75939 | 0.440 | [35600, 101000] | 48078 |
| 2006 | 75477 | 0.423 | [36462, 99999] | 50002 |
| 2007 | 75872 | 0.428 | [38000, 101747] | 52000 |
| 2008 | 76185 | 0.417 | [39157, 100100] | 52004 |
| 2009 | 76260 | 0.416 | [37481, 99999] | 51157 |
| 2010 | 75188 | 0.414 | [36909, 100000] | 51016 |
| 2011 | 74838 | 0.408 | [38000, 100000] | 52000 |

B. Annual Family Poverty Rate

| Year | Number of Interviewed Families | Fraction with Missing Data | Interval Estimate | Point Estimate with Imputations |
|---|---|---|---|---|
| 2001 | 89063 | 0.436 | [0.110, 0.315] | 0.146 |
| 2002 | 89098 | 0.440 | [0.112, 0.328] | 0.152 |
| 2003 | 87948 | 0.438 | [0.117, 0.331] | 0.158 |
| 2004 | 87149 | 0.437 | [0.118, 0.331] | 0.160 |
| 2005 | 86882 | 0.415 | [0.121, 0.319] | 0.161 |
| 2006 | 86222 | 0.400 | [0.120, 0.323] | 0.154 |
| 2007 | 86955 | 0.403 | [0.120, 0.324] | 0.154 |
| 2008 | 87562 | 0.392 | [0.126, 0.320] | 0.162 |
| 2009 | 88957 | 0.388 | [0.138, 0.338] | 0.176 |
| 2010 | 87076 | 0.390 | [0.138, 0.344] | 0.178 |
| 2011 | 86038 | 0.384 | [0.139, 0.339] | 0.176 |

### C. March Unemployment Rate

| Year | Number of Interviewed Civilian Adults | Fraction with Missing Data | Interval Estimate | Point Estimate with Imputations |
|---|---|---|---|---|
| 2002 | 108901 | 0.0028 | [0.057, 0.062] | 0.058 |
| 2003 | 110375 | 0.0032 | [0.060, 0.065] | 0.061 |
| 2004 | 108221 | 0.0030 | [0.057, 0.062] | 0.057 |
| 2005 | 106913 | 0.0035 | [0.052, 0.057] | 0.052 |
| 2006 | 106234 | 0.0030 | [0.047, 0.051] | 0.047 |
| 2007 | 105392 | 0.0039 | [0.044, 0.050] | 0.044 |
| 2008 | 105643 | 0.0047 | [0.049, 0.057] | 0.050 |
| 2009 | 106923 | 0.0052 | [0.085, 0.093] | 0.086 |
| 2010 | 107582 | 0.0053 | [0.095, 0.103] | 0.096 |
| 2011 | 105774 | 0.0072 | [0.085, 0.096] | 0.086 |
| 2012 | 105314 | 0.0076 | [0.078, 0.090] | 0.079 |

TABLE 2: Interval Estimates Recognizing Item and Unit Nonresponse

| A. Annual Median Household Income | | | | |
|---|---|---|---|---|
| Year | Number of Households in Sample | Fraction with Unit Nonresponse | Fraction with Item Nonresponse | Interval Estimate |
| 2001 | 84831 | 0.077 | 0.425 | [28200, 123643] |
| 2002 | 85092 | 0.080 | 0.429 | [27141, 104611] |
| 2003 | 84116 | 0.083 | 0.425 | [28000, 106000] |
| 2004 | 83932 | 0.089 | 0.422 | [28240, 106487] |
| 2005 | 83009 | 0.085 | 0.403 | [30652, 101000] |
| 2006 | 82554 | 0.086 | 0.387 | [31309, 99999] |
| 2007 | 82235 | 0.077 | 0.395 | [33200, 95008] |
| 2008 | 81904 | 0.070 | 0.388 | [35000, 100100] |
| 2009 | 81938 | 0.069 | 0.387 | [33004, 99999] |
| 2010 | 81737 | 0.080 | 0.381 | [31800, 100000] |
| 2011 | 81573 | 0.088 | 0.372 | [32132, 100000] |

B. Annual Family Poverty Rate

| Year | Number of Families in Sample | Fraction with Unit Nonresponse | Fraction with Item Nonresponse | Interval Estimate |
|---|---|---|---|---|
| 2001 | 95629 | 0.069 | 0.406 | [0.102, 0.362] |
| 2002 | 95880 | 0.071 | 0.409 | [0.104, 0.376] |
| 2003 | 94915 | 0.073 | 0.406 | [0.108, 0.380] |
| 2004 | 94634 | 0.079 | 0.403 | [0.109, 0.384] |
| 2005 | 93952 | 0.075 | 0.384 | [0.112, 0.370] |
| 2006 | 93299 | 0.076 | 0.370 | [0.111, 0.374] |
| 2007 | 93318 | 0.068 | 0.376 | [0.112, 0.370] |
| 2008 | 93281 | 0.061 | 0.368 | [0.118, 0.362] |
| 2009 | 94635 | 0.060 | 0.365 | [0.130, 0.378] |
| 2010 | 93625 | 0.070 | 0.363 | [0.128, 0.390] |
| 2011 | 93228 | 0.077 | 0.354 | [0.128, 0.390] |

C. March Unemployment Rate

| Year | Number of Civilian Adults in Sample | Fraction with Unit Nonresponse | Fraction with Item Nonresponse | Interval Estimate |
|---|---|---|---|---|
| 2002 | 113963 | 0.044 | 0.0027 | [0.053, 0.123] |
| 2003 | 115116 | 0.041 | 0.0031 | [0.056, 0.123] |
| 2004 | 113461 | 0.046 | 0.0028 | [0.054, 0.126] |
| 2005 | 112648 | 0.051 | 0.0033 | [0.048, 0.129] |
| 2006 | 111620 | 0.048 | 0.0029 | [0.043, 0.120] |
| 2007 | 111028 | 0.051 | 0.0037 | [0.040, 0.122] |
| 2008 | 110761 | 0.046 | 0.0045 | [0.046, 0.122] |
| 2009 | 111541 | 0.041 | 0.0050 | [0.073, 0.149] |
| 2010 | 112190 | 0.041 | 0.0051 | [0.089, 0.159] |
| 2011 | 111061 | 0.048 | 0.0069 | [0.078, 0.162] |
| 2012 | 111121 | 0.052 | 0.0072 | [0.071, 0.162] |