Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

ENSURING THE QUALITY OF THE LABOR FORCE ESTIMATES FROM THE CURRENT POPULATION SURVEY: A TOTAL SURVEY PERSPECTIVE

Lawrence Cahoon, Patrick Flanagan, and Karen Deaver[1]

ABSTRACT

The Current Population Survey is the primary source of labor force data for the United States. It is critical that the quality of the data be insured throughout the survey process. This paper discusses how quality is addressed at alls t e the development of the sample frame, the sampling operations, sample control, data collection, editing, imputation, estimation, questionnaire development, and quality evaluations built into the survey process. The paper concludes with a discussion of current research and possible future improvements to the survey.

KEY WORDS: Data Quality, Sampling Error, Nonsampling Error, Quality Control

1. INTRODUCTION

1.1 The Current Population Survey

In the United States, the Current Population Survey (CPS) is the Nation's primary source of monthly labor force status statistics and demographic characteristics of workers. National social and economic programs base policy changes and the distribution of Federal funds in part on the statistics produced by the CPS. Also, studies by researchers from all over the world use CPS data to test all types of hypotheses. With such enormous responsibility associated with accuracy of data and statistical estimates, all aspects of data quality take on increased importance.

1.2 CPS History

The CPS has a lineage that goes back to 1940. The original Sample Survey of Unemployment was started by the Work Projects Administration in March 1940. It was transferred to the Bureau of the Census in August 1942 and renamed in 1943 to the "Monthly Report on the Labor Force." In 1947 it became the Current Population Survey. Over the years the sample design and survey methods have been continuously improved, taking into account the latest methodology and the results of countless studies of CPS data quality carried out by researchers. Bureau of the Census (2000) provides a step by step evolution of the CPS design from the early years through to the present, including the decennial redesign of the CPS sample.

[1]Lawrence Cahoon, Patrick Flanagan, and Karen Deaver, U. S. Bureau of the Census, Washington, DC, 20233, USA

1.3 The CPS Decennial Redesign

Every decade following the United States Decennial Census, the CPS along with other major demographic surveys uses the updated information on the country to update the design and select new samples for the upcoming decade.

2. CPS DATA QUALITY CONTROL

2.1 Survey Design

The CPS survey sample design is primarily focused on the estimate of employment status of civilian (i.e., not military), noninstitutionalized population, age 16 and older. More specifically, it is focused on the unemployment rate. The goals of the design are:

- To achieve a coefficient of variation (CV)[2] of 1.9 percent on the monthly level of unemployed, assuming an unemployment rate of 6 percent. This is based on a requirement to detect a 0.2 percent change in the unemployment rate in consecutive months.
- To achieve a maximum CV of 8 percent on the estimate of the annual level of unemployed for each state, assuming an unemployment rate of 6 percent.

The above goals would be easy to meet with an unlimited budget, but that, of course, is not the case. The current CPS design makes use of cluster sampling to gain efficiency and is able to meet the goals with a sample size of 50,000 housing units per month containing approximately 112,000 people.

The sample selection is a two-stage design. The first stage is the selection of geographical Primary Sampling Units (PSUs). The second stage is the selection of sample housing units.

Each selected household is in sample for four consecutive months, then is out of sample for eight months, and is back in sample for four final consecutive months. These months are referred to the household months-in-sample (MIS) 1 through 8.

2.2 Sample Selection

2.2.1 Sample Selection - First Stage

2.2.1.1 Overview

The United States are broken up into 2,007 PSUs, where each consists of one or more counties. Many of the largest PSUs in population are selected with certainty. They are called the Self-Representing (SR) PSUs. In fact, those PSUs that are part of the 150 largest Metropolitan Statistical Areas are automatically made SR. The remaining PSUs, which are called Non-Self-Representing (NSR) PSUs are formed into strata within their respective states and one PSU is selected from each stratum.

One additional algorithm is also employed in the selection of NSR PSUs. This algorithm increases the probability that a selected PSU in the 1990 design will be selected again in the 2000 redesign while holding the overall probabilities of selection fixed in accordance with a probability-proportional to size. The data

[2] The coefficient of variation (CV), also called the relative standard deviation, is the standard deviation of the estimate divided by its expected value.

quality impact from the use of this algorithm is in the variance of the estimates, as it creates a dependency across strata within a state. No studies have been published regarding the extent of this impact.

The primary sampling unit (PSU) selection is made up of four processes. The input to PSU selection comes from official Decennial Census files which have their own verification process.

2.2.1.2 PSU Definition File Development

A county level file is constructed indicating the PSU number beside each county. PSUs are defined in accordance with specific geographic criteria listed in Bureau of the Census (2000) related to interviewing workloads and constraints. The PSU definitions are created starting with the PSUs from the previous decade. After making county definition changes to the present, we solicit input from our Regional Office personnel regarding problems encountered with the prior definitions. The new definitions also take into account projections made regarding new Metropolitan Statistical Areas that need to be represented as SR.

Quality Control: Once the file is created, it is split among the members of a working group by state. These verifiers examine each PSU in their states for compliance with the PSU definition requirements and agreed upon implementation of "field" recommendations. The verifiers also use maps to compare new PSU definitions to old ones and to ensure all counties are included.

2.2.1.3 Master Stratification Input File (MSIF) Development

Given the PSU definitions file and the input file from the Decennial Census, the next step is to construct the MSIF. The basic MSIF is a county level file with county definition, a PSU number, 16+ population size for the county, the number of housing units, and the selected stratification variables. The stratification variables are variables that are highly correlated with unemployment. From that initial file a composite file at the PSU level is formed for the stratification process. While constructing the file, some variables are used in the same form as provided in the input file from the Census. Other variables are computed from input variables (e.g., county summary data).

Quality Control: The software specifications and the compilation software go through an extensive systems test before production to verify the algorithms and ensure all variables are present. The input files are also verified using verification programs. A three-state initial production set is heavily verified to ensure each variable required is present, that the ranges of values are within requirements, and compared to prior redesign files for outlier detection. The remaining states are verified by summary calculations.

2.2.1.4 PSU Stratification

The stratification process uses a program called the Stratification Search Program (SSP), which uses a variance estimation based on the Friedman-Rubin algorithm. Within each state it forms random strata based on size restriction input and homogeneity of stratification variables. It then calculates variance estimates for key variables for each state and for the nation. It continues through an iterative process of swapping PSUs between the strata in each state until an optimal stratification is achieved. The SSP also takes into account workload balance based on goals provided as input.

Quality Control: This software evolved from the Friedman-Rubin based software used for the 1990 stratification. The most important quality control (QC) has been in the confirmation of the algorithms used and the software walk-throughs confirming that the software was executing the algorithms as specified. A workgroup of Mathematical Statisticians performed most of this QC. They documented their results which were further reviewed by our Research and Development Division for technical accuracy. Finally, the program was extensively tested with simulated actual data based on 1990 research data files. The verification during the production phase will include comparisons of results between the last decade and the current redesign in

each state, checking the strata definitions, the number of strata, and the estimates of between PSU variances.

2.2.1.5 PSU Selection

The PSU selection is really two processes. The first process cycles through each NSR stratum separately and computes the probabilities of selection that will be used for the second process, the selections themselves. The first process calculates the "overall" probability of selection for each PSU using a probability-proportional-to-size (PPS), where size is the 16+ population for the PSU. Given that calculation, this first process needs one more input: the selection information from the 1990 design. With that information, the conditional probabilities of selection for each PSU are computed using an algorithm that treats the 1990 and 2000 selections as one experiment. It holds the overall probabilities of selection for each PSU constant and then computes conditional probabilities of selection for the 2000 redesign PSUs given which PSUs were selected in the 1990 design. This overlap maximization dramatically reduces the number of PSUs that change from one design to another. That reduction avoids the damage to data quality that would occur if the Census Bureau had to hire new interviewers in the new PSUs and release interviewers in the old PSUs. More detail on the algorithm can be found in Bureau of the Census (2000).

Quality Control: The initial quality control for this process is the verification that algorithms to be employed are statistically correct. The methodology was defined in a planning document which was reviewed by the designing statisticians in the developmental working group and further reviewed by our Research and Development Division for statistical accuracy. From there, a specification was written spelling out how the algorithms are to be implemented in the software. That specification was cross-referenced to the methodology plan for compliance. The resulting software was then extensively tested with a combination of actual data and simulated data comparing results of all calculations against independent results. The final production run is also verified to ensure a wide range of summary calculations meet specified totals and ranges.

2.2.2 Sample Selection - Second Stage

Once the PSU selection is complete, the selection of the housing units from within each selected PSU has to be carried out. The primary source of housing unit information comes from the housing unit list from the Census. In the 2000 Census, that list is called the Master Address File (MAF). But, that list has flaws for the purpose of the CPS:

2.2.2.1 Problems Using the Decennial MAF as the Sole Frame

The Census was primarily mail based. In some areas the addresses reflect delivery boxes and not necessarily housing unit locations. There are also a lot of housing units on the list with just physical descriptions. In addition, the MAF list will need to be updated with housing changes throughout the decade. To cover these problems, we use four frames for sample selection.

2.2.2.2 Four Frame Coverage

In those block areas where the MAF has 96 percent or more "good" addresses, we use the MAF as the frame. For housing units, it is called the **unit frame**. For "group quarters" like college dormitories, it is called the **group quarters frame**. For other block areas, we create an **area frame**, where we select housing units and group quarters for sample using blocks and block counts and then list the block to identify the actual housing units in sample. To handle new construction, we have a **permit frame** in those areas of the country that issue building permits. In areas where permits are not issued, we count on the area frame to detect new construction.

2.2.2.3 Coverage Weaknesses in the Frame

The most obvious weakness with the frame coverage is the errors in the unit frame from the Census. Houses

missing from the Census list and within the unit frame will not be covered. As a coverage factor this is a very small factor due to the Block Canvassing effort in the Census and the fact that the blocks contained in the unit frame are usually the areas where locating addresses is easier. A second coverage factor, referred to as the permit lag is a group of permits issued before the Census Day, but are built after Census Day. The CPS permit frame includes permits back to January before Census Day to strike a balance between the permit lag problem and a possible overcoverage problem with having those new buildings selected twice. A third problem is coverage of mobile homes. In areas within the unit frame, no system is currently set up to update lists in the unit frame to include new mobile homes moving into the area.

Once the PSU selection is complete, the selection of the housing units in each frame has to be carried out. The frame development and housing unit selection are made up of three major processes, each with computer program development system testing and detailed production verification to ensure it is carried out properly. The input to this stage is a Master Address File extract from the Decennial Census, which has gone through its own verification process, and permit office listings from the Survey of Construction files.

2.2.2.4 Screening for Frame Determination

In the selected PSUs, we have to screen each block to determine which blocks have sufficient "good" addresses and which do not. The blocks with the "good" addresses will be in the Unit and Group Quarters frames (see para. 2.2.2.2), while the remainder will be in the Area frame. The screening is done automatically using a computer program.

The second screening will be the determination of which blocks within the selected PSUs are covered by permit offices and which are not. Those blocks with "good" addresses that are not covered by a permit office are moved from the Unit frame into the Area frame.

The separation of addresses into the Group Quarters frame is relatively simple in that the Decennial Census files flag group quarters allowing us to separate these addresses from the Unit frame and update our existing group quarters universe list.

Quality Control: During the development of the screening software, the criteria for defining "good" addresses are specified and reviewed by all affected groups. After the screening software is completed, a sample set of actual addresses is screened by the program and by independent software to confirm that the process is functioning properly.

2.2.2.5 Frame Creation

This process is primarily an exercise in setting up the files for selection. The unit and group quarters frames are primarily set up from the Decennial Census files, although some aggregate data from the Census Long Form is added for the sample selection sorts. For the area frame, the files are set up by blocks. Though the addresses are not useful in carrying out selection, the Census counts by block are used to set up a structure for selection that will be filled with specific addresses in the listing operation. Finally, the permit frame builds a "skeleton" frame based on anticipated growth at the block level.

Quality Control: Similar to the other quality control, the processing is verified using test files of actual Census data with measures block level measures. The steps in this process are checked by a combination of independent software comparisons and summary calculations.

2.2.2.6 Sample Selection

For each frame and within each PSU, the housing units are grouped into Ultimate Sampling Units (USU) of four units per USU to take advantage of the efficiency of cluster sampling. The USUs are then sorted by a

series of variables including urbanization, race, tenure and gender. Finally a systematic sample is drawn selecting a "hit string" of 21 USUs at each selection point. The result will be the 21samples to be used for the decade.

Quality Control: A sample set of six states worth of data is run through the process. Each step is verified either by hand or by independent software comparisons. Summary statistics are verified at this stage and during production.

2.2.2.7 Area Frame Listing QC

The quality control of the area listing operation is a combination of initial training in the listing process combined with a listing check operation to verify that the individual listers are listing to an acceptable level of quality. Listers are checked annually by selecting a sample of their listing and a senior field representative lists the block or segments a second time. If the lister has an unacceptable error rate, they are either retrained or assigned to activities other than listing.

2.3 Data Collection

2.3.1 The Questionnaire

The CPS questionnaire was most recently revised in the conversion to automation in 1994. The wording of the questions, terminology, and sequence of questions has been evolving over time benefitting from extensive studies on the accuracy of the resulting data, cognitive tests, and response error analysis. The conversion to computer-aided personal interviewing (CAPI) and computer-aided telephone interviewing (CATI) has further reduced problems with following "skip patterns" in the survey.

2.3.2 Mode of Interview

The first month-in-sample (MIS) is done primarily (80 percent) in person to establish the identity of the people in the household and to collect household information. Thereafter, most interviewing (85 percent) is done by telephone. The exception is MIS 5, where about 65 percent are interviewed in person.

2.3.3 Interviewing Process

Interviewing in CPS is done primarily using CAPI assignments to Field Representatives (FRs) through the Regional Offices. However, some of the cases (about 10%) are assigned to CATI at one of three telephone centers. The amounts assigned to each are determined through a workload balancing process. Despite the name, CAPI assignments may be done by personal visit or telephone. We use the CAPI term for cases assigned to FRs where they use a laptop computer for data entry on scene. Whether done by CAPI or CATI, the questionnaire automatically walks through the survey allowing the interviewer to enter data where it is required.

2.3.4 Proxy Respondents

The CPS asks questions about the household in general and then asks questions about each member within the household. If each of the members is present, they can respond for themselves. If they are not, the CPS allows one household member (15 years old or older) to respond for other household members. Proxy reporting tends to cover about 50 percent of all sample people.

2.3.5 Interviewer Impact

The differences in data quality from one interviewer to the next is a well-documented phenomenon (Lyberg

and Kasprzyk, 1991). This goes beyond the level of response achieved by the interviewer. In the positive sense, some interviewers are better at establishing a rapport with respondents. Since CPS is a panel survey, this is more important than with cross-sectional surveys. Bias can be introduced via the interviewer by differences in how they answer respondent questions and in how the respondent reacts to the interviewer. For example, if a respondent feels embarrassed about losing a job and they feel a level of respect for the interviewer, he or she may not report it accurately.

Training of interviewers for CPS generally is a two-stage process (assuming the interviewer has already been trained in the basics of interviewing). The interviewer is given initial training consisting of home study, classroom training, and two days of on-the-job observation. The second month, the new interviewer is given home study and one day of observation. Finally, the initial training is completed with a home study and a final review test prior to the third month. After that, each month the interviewer completes a home study course, and twice a year the interviewer attends refresher training.

2.3.6 Roster Errors

An element of undercoverage occurs if the roster compiled during the initial interview does not include everyone in the household. A study for the decennial Census has indicated that about 32 percent of all omitted persons were in households where some, but not all of the people were listed on the household roster (Killion, 1993).

2.3.7 QC Reinterview

Each month a sample of the Field Representatives (FRs) assigned to CPS is selected and a sample of the cases assigned to them is selected. These cases are contacted a second time to confirm that the FR conducted the interview in accordance with procedures. Significant errors detected or suspected falsifying of an interview are turned over to the CPS supervisor for followup action.

2.3.8 CATI Monitoring

Similar to the reinterview, the CATI facilities have a monitoring program. Under this program a "coach" selects random cases being performed by interviewers and listens to the interview in progress. They note whether all procedures were followed and debrief the interviewer on their performance after the fact.

2.3.9 Case Tracking

As discussed in paragraph 2.6.1, cases are tracked in Census Bureau Headquarters, the Regional Offices, and the CATI facilities to ensure no cases are lost in the file transfers.

2.4 Data Processing

2.4.1 Data Transmission

After an FR completes a case, it is transmitted electronically via modem back to the Regional Office database. If the case is a completed interview, it is then grouped with other completed interviews and sent back to Census Headquarters. Cases that an FR cannot complete may be reassigned by the CPS Supervisor, usually to a senior field representative. Accountability for all cases is maintained at both the Regional Offices and Headquarters to ensure that every case is completed and returned.

2.4.2 Coding

The CPS processing codes text entries for each sample person's industry and occupation. Coders specially trained for this operation work in the Census Bureau's National Processing Center.

The industry and occupation coding quality is monitored by a recoding process. A sample of each qualified coder's work is selected and coded by a second coder. The second coder is an experienced coder called a referralist[3]. If a coder has too many errors in a given month, he or she will be sent through retraining. Due to the time constraints on the CPS processing, this system is not run as an acceptance testing of the data set. The system has proven to be stable, and this information system is used to monitor the coding system to ensure it remains stable.

2.4.3 Editing & Imputation

During CPS processing a wide range of edits take place. Entries out of range, "Don't Knows," and "Refuses" may be replaced by values inferred from other entries or similar question responses in previous months. Item nonresponse to questions is imputed using the "hot deck" procedure. This procedure uses classing and sorting to pinpoint a donor who is like the sample person according to variables like age, race, and gender. The answer to the same question from the donor is imputed as a good estimate. Due to the low percentage of imputed values for any given question (often < 0.5 percent), the impact on primary data from imputation is regarded as negligible.

Each time the editing and coding systems are changed, they go through an extensive system testing. Beyond that, after the editing and separately after the imputation, the before and after distributions for each of the data elements are examined for any unusual changes.

2.5 Estimates

2.5.1 Weighting - Basic

Since the CPS is not a simple random sample, weights are used in an attempt to achieve unbiased estimates. The basic design attempts to create a self-weighting situation within each state, so that each housing unit will have the same probability of selection. The probability of selection for a given state, or more accurately its inverse the overall state sampling interval, is chosen to achieve the goal for the state CV.

Each time the weighting system is changed, it goes through an extensive system testing. Beyond that, after the weights are calculated, the weight distributions are compared to prior months to see if any unusual weights are created.

Secondly, each month the person-level weights produced by the production processing are compared to an independent weighting program. The cumulative absolute difference in these weights must be below a specified tolerance for the weighting to be accepted.

[3] A referralist is actually an experienced coder who routinely handles coding that others cannot code. The referralist does coding QC as an additional task.

2.5.2 Variance Estimation

Variances are estimated for the CPS using the "successive differences" method discussed in Fay and Train (1995), employing 160 replicates. Each time the variance estimation design or programming is changed, it goes through an extensive system testing.

3. CPS ERROR MITIGATION

3.1 Weighting Adjustments

Given that certain forms of nonsampling error, especially nonresponse, are unavoidable, some adjustments are made to the case weights in an attempt to reduce the biasing and variance effects they have on the estimates.

3.1.1 Nonresponse Adjustment

In CPS we do a simple nonresponse adjustment within cells that assumes within those cells that nonresponse is randomly distributed in relation to unemployment. The cells are determined by the Metropolitan Statistical Area (MSA) status and MSA size of each PSU. The weight from the nonresponding households (type A only) is redistributed to the responding cases in the same cell.

3.1.2 Ratio Adjustment Using "Controls"

CPS does a ratio adjustment to weights in two stages. The first stage adjusts the weights for all cases in each selected NSR PSU for possible imbalance of racial representation caused by PSU selection. The second adjustment is carried out for all case weights. It is a three-way raking using controls for the state 16+ population, Hispanic/gender/age groups, and a race/gender/age groups iteratively for six iterations. The "controls" used for these adjustments are independent estimates of the 16+ civilian noninstitutionalized population derived from the Decennial Census adjusted first for the undercount and second using births, deaths, and net migration data from a combination of other survey sources.

3.2 Seasonal Adjustment

Many of the estimates produced from CPS data are seasonably adjusted using an X-11 ARIMA program to separate seasonal changes from actual changes in the labor market.

4. CPS QUALITY ASSESSMENT

4.1 Response Error Assessment

In the current CPS, estimates of response error are limited to response variance. Data is collected for response variance in a monthly response error reinterview. A random selection of cases is selected and the questionnaire is administered in the same manner as the original interview, except that personal interviews are usually reinterviewed by telephone. The cumulative response variance for one year is then analyzed to assess whether any changes have occurred from past estimates for each question or composite variable, like major labor recode. The data will also be used to identify the questions that need improvement when the next questionnaire redesign occurs.

4.2 Industry & Occupation Coding Quality

Beyond the coding quality control of coder qualification discussed in paragraph 3.4.1, the coding quality data is also analyzed to identify problem codes. These results are used to update the computer assisted coding that the coders use.

4.3 Coverage Ratios

One measure of coverage is to estimate the population of a group before the weights are adjusted using the population controls and then divide that estimate by the control for that group. Using this measure, CPS tends to cover about 93 percent of the 16+ civilian noninstitutionalized population. However, among the various subgroups within the population broken down by age, gender, race, and ethnicity, the coverage varies from 66 (Black males, 20 - 29 years old) to 102 percent (Other race male, 70+ years old).

4.4 Nonresponse Rates

The nonresponse rate is an indicator of the possible bias introduced by nonresponse. As with all regular surveys, CPS has noticed a gradual increase in refusals over time, which has been balanced out some by better contacting methods. The CPS is currently experiencing a 7 percent nonresponse rate.

5. FUTURE PLANS

5.1 CPS to 2000 Census Matching Study

Currently underway is an analysis of CPS data quality using data from the 2000 Census. It will help us assess the accuracy of every aspect of our data quality. The CPS/Census match will attempt to measure housing unit and within-housing unit coverage for the CPS, get Census information for CPS nonrespondents to improve our nonresponse adjustment, and compare responses from CPS and Census to similar questions.

5.2 Latent Class Model Analysis

A tool for data quality analysis and adjustment that is showing a great deal of promise is latent class analysis (Biemer and Bushery, 2000). The Markov Latent Class model is able to use the panel design of the CPS to estimate response error broken down into misclassification error probabilities. We are also looking into the use of Latent Class Analysis to estimate the effect of rotation group bias that has been documented since the 1970s (Bailar, 1975), and possibly adjust the data to eliminate its effect. Though the theory supports its effectiveness as a possible estimator of these factors, we still need to go through proof of concepts studies, such as simulations and administrative record matching such as the matching to the Census discussed above.

6. CONCLUSION

The quality of data produced from the CPS is a complex combination of many different factors. Beyond the survey design focused on data quality goals, it is crucial to ensure that each step is carried out successfully to achieve those goals. For a project as complex as CPS, the quality control steps are not trivial and must be planned out carefully to ensure a successful survey. However, given each level of quality achieved, the quality control system should ensure continuous improvement to more lofty goals.

REFERENCES

Bailar, Barbara A. (1975), "The Effects of Rotation Group Bias on Estimates from Panel Surveys, " *Journal of the American Statistical Association*, 70, 23 - 30.

Biemer, Paul P. and Bushery, John M. (2000), "On the Validity of Markov Latent Class Analysis for Estimating Classification Error in Labor Force Data," *Survey Methodology*, Vol. 26, No. 2, pp. 139 - 152.

Fay, Robert and Train, George (1995), "Aspects of survey and model-based postcensal estimation of income and poverty for states and counties," *American Statistical Association Proceedings of the Section on Government Statistics*, pp. 154 - 159.

Killion, Ruth A. (1993), "The Impact of Housing Unit Coverage," Housing Unit Coverage Study Results Memorandum Number 2, Census Bureau Memorandum from Killion to Walsh dated June 24, 1993.

Lyberg, Lars and Kasprzyk, Daniel (1991), "Data Collection Methods and Measurement Error: An Overview," in P. P. Biemer et al. (eds.) *Measurement Errors in Surveys*, New York: Wiley, pp. 237 - 257.

U. S. Bureau of the Census, *Technical Paper 63, Current Population Survey, Design and Methodology*, March 2000