2014 Documentation

DIARY SURVEY
CONSUMER EXPENDITURE
PUBLIC USE MICRODATA

September 3, 2015

U.S. Department of Labor
Bureau of Labor Statistics
Division of Consumer Expenditure Survey

Table of Contents

I. Introduction

The Consumer Expenditure Survey (CE) program provides data on the buying habits of American consumers. These data are primarily used as weights for the Consumer Price Index. However, CE also provides the data to the public for research in two formats. The first format are tabulations of average and aggregate expenditures and income in news releases, databases, and tables. The second format are individual responses to the CE Survey in the Public-Use Microdata (PUMD). For broad analysis the former format is better suited; for detailed studies the latter format may proof more useful.

This document discusses the methodology of PUMD for the Diary Survey. The primary objective of the Diary Survey is to obtain expenditures data on small, frequently purchased items, which can be difficult to recall even a few weeks later. These items include food and beverage expenditures at home and in eating places; housekeeping supplies and services; nonprescription drugs; and personal care products and services. The Diary Survey is not limited to these types of expenditures but, rather, includes all expenses that the consumer unit incurs during the survey week. Expenses incurred by family members while away from home overnight and for credit and installment plan payments are excluded. To provide novice users additional assistance, CE prepared a "Getting started with Consumer Expenditure Public-use Microdata."

This document does not list the changes from the previous year, the topcoding and other suppressions of sensitive data, and the response rates. These items can be found in the 2014 Documentation zip file.

The microdata files are in the public domain and, with appropriate credit, may be reproduced without permission. A suggested citation is: "U.S. Department of Labor, Bureau of Labor Statistics, Consumer Expenditure Survey, Diary Survey, 2014."

II. File Information

The Diary Survey microdata are provided as SAS, STATA, SPSS. or ASCII comma-delimited files. The 2014 Diary release contains two groups of files:

- **4 major data files**: FMLD, MEMD, EXPD, DTBD, and DTID
- **3 Processing files:** DSTUB, INTSTUB, and sample code

The four major data files (FMLD, MEMD, EXPD, DTBD, and DTID) are organized by the calendar quarter of the year in which the data were collected. There are four quarterly data sets for each of these files.

The FMLD files contain CU characteristics, income, and summary level expenditures; the MEMD files contain member characteristics and income data; the EXPD files contain detailed weekly expenditures at the UCC level and is structured like the Diary Survey Form (See the Diary Survey form.); the DTBD files contain the CUs' reported income values or the mean of the five imputed income values in the multiple imputation method; and the DTID files contain the five imputed income values.

The two processing files enhance computer processing and tabulation of data, and provide descriptive information on item codes. CE provides these processing files:

- **DSTUB and INTSTUB**: provide the aggregation scheme used in the published consumer expenditure survey diary tables and integrated tables. These files contain UCCs and their abbreviated titles, identifying the expenditure or demographic item represented by each UCC.
- **Sample programs** with code that approximates the tables that CE publishes. CE provides the code in SAS and R.

The processing files are further explained in Section III.F.6. Processing Files.

A. Dataset Names

The file naming convention is listed in the table below. The files are compressed and can be uncompressed with most unzip utilities.

\DIARY14\FMLD141.* (Diary FMLD file for first quarter, 2014)
\DIARY14\MEMD141.* (Diary MEMD file for first quarter, 2014)
\DIARY14\EXPD141.* (Diary EXPD file for first quarter, 2014)
\DIARY14\DTBD141.* (Diary DTBD file for first quarter, 2014)
\DIARY14\DTID141.* (Diary DTID file for first quarter, 2014)
\DIARY14\FMLD142.* (etc.)
\DIARY14\MEMD142.*
\DIARY14\EXPD142.*
\DIARY14\DTBD142.*
\DIARY14\DTID142.*
\DIARY14\FMLD143.*
\DIARY14\MEMD143.*
\DIARY14\EXPD143.*
\DIARY14\DTBD143.*
\DIARY14\DTID143.*
\DIARY14\FMLD144.*
\DIARY14\MEMD144.*
\DIARY14\EXPD144.*
\DIARY14\DTBD144.*
\DIARY14\DTID144.*

B. Record Counts

The following are number of records in each data set.

Data Set	Record Count
FMLD141	3,261
FMLD142	3,392
FMLD143	3,363
FMLD144	3,289
MEMD141	7,853
MEMD142	8,426
MEMD143	8,295
MEMD144	8,051
EXPD141	111,273
EXPD142	116,979
EXPD143	114,972
EXPD144	115,181
DTBD141	51,270

Data Set	Record Count
DTBD142	53,131
DTBD143	52,946
DTBD144	51,411
DTID141	66,520
DTID142	68,440
DTID143	68,075
DTID144	66,645

C. Data Flags

Data fields on the FMLD and MEMD files are explained by flag variables following the data field. The names of the flag variables are derived from the names of the data fields they reference.

In general the rule for naming variable flags is to add an underscore to the last position of the data field name, for example WAGEX becomes WAGEX_. However, if the data field name is eight characters in length, then the fifth position is replaced with an underscore. If this fifth position is already an underscore, then the fifth position is changed to a zero, so that PENSIONX becomes PENS_ONX, EDUC_REF becomes EDUC0REF.

Flag value	Description
A	Valid blank; a blank field where a response is not anticipated
B	Blank due to invalid nonresponse; nonresponse that is not consistent with other data reported by the CU
C	Blank due to “Don’t know,” refusal, or other nonresponse
D	Valid value, unadjusted
E	Valid value, allocated
T	Valid value, topcoded or suppressed

Allocation refers to the process of allocating an expenditure amount for unspecified items to specific items. In the Diary Survey, the variable “ALLOC” tracks allocations. Below are the codes:

Code	Description	Corresponding Flag
0	Valid value, unadjusted	D
1	Valid value, allocated	E
2	Topcoded and allocated	T
3	Topcoded, not allocated	T

D. Income Imputation

Beginning in 2004, the CE has implemented multiple imputation of income data. Imputation allows income values to be estimated when they are not reported. Many income variables and other income related variables will be imputed using a multiple imputation process. These imputed income values will be included in the FMLD, MEMD, DTBD, and DTID files. The multiple imputation process derives five imputation values and a mean imputation value per income variable. More information on the imputation process and how to appropriately use the data are found in the document "User's guide to Income Imputation in the CE."

In the public-use microdata, not all of the imputed income variables will contain the derived imputation values. For some income variables, the five derived imputations are excluded and only the mean of

those imputations is available. For these variables, there are 3 associated income variables in the FMLD and MEMD files (*INCOME*M, *INCOME*M_, and *INCOME*I). For all other imputed income variables, there are 7 associated variables in the FMLD and MEMD files:

*INCOME*1	the first imputed income value or the reported income value, if non-missing
*INCOME*2	the second imputed income value or the reported income value, if non-missing
*INCOME*3	the third imputed income value or the reported income value, if non-missing
*INCOME*4	the fourth imputed income value or the reported income value, if non-missing
*INCOME*5	the fifth imputed income value or the reported income value, if non-missing
*INCOME*M	the mean of the five imputed income values
*INCOME*M_	the flag variable for the imputed variable (see Section III.C. Data Flags)
*INCOME*I	the imputation indicator

Income variables that have imputed values as components (ex: FINCBEFM) will also have 5 imputed values and a mean based on each of the imputed components.

The imputation indicator variable is a 3 digit number that is coded as follows:

The first digit in the 3 digit code defines the imputation method. The meanings are:

1: No Imputation
2: Multiple imputation due to invalid blank only
3: Multiple imputation due to bracketing only
4: Multiple imputation due to invalid blanks and bracketing
5: Multiple imputation due to conversion of a valid blank to an invalid blank (this occurs only when initial values for all sources of income for the CU were valid blanks).

The meaning of the last two digits of the three digit code differs depending on whether you are looking at one of the components of overall income, like FWAGEXM, or you are looking at the summary level variable FINCBEFM. For the components, the last 2 digits represent the number of family members who had their data imputed for that source. For example, if a family had a value of 302 for FWAGEXI that would mean that 2 of the members in the family had their salary income imputed and that in both cases the imputation was due to bracketing only. For the summary level variable FINCBEFM which is a summation of all of the income components, the last 2 digits represent the number of income sources imputed for each member added together. For example, if a family had 3 members and 2 had salary income imputed due to invalid blank only, and 2 had self-employment income imputed due to bracketing only, and that was the only income data imputed for members of that family, then FSMPFRXI for the family would be 202, FBSNSXI would be 302, and FINCBEFI would be 404.

The DTBD file includes income UCCs mapped from the associated *INCOME*M variables and the income variables that are not imputed in the FMLD files. The DTID file includes UCCs mapped from income variables subject to income imputation, including the variable IMPNUM to indicate the imputation number 1 - 5.

E. File Notation

Every record from each data file includes the variable NEWID, the CU's unique identification number, which can be used to link records of one CU from several files. Data fields for variables on the microdata files have either numeric or character values. The format column in the diary data dictionary distinguishes whether a variable is numeric (NUM) or character (CHAR) and shows the number of field positions the variable occupies. Variables that include decimal points are formatted as NUM(t,r) where t is the total number of positions occupied, and r is the number of places to the right of the decimal.

In addition to format, the diary data dictionary gives an item description, questionnaire source and identification of codes where applicable for each variable.

An asterisk (*) is shown in front of new variables, those which have changed in format or definition, and those which have been deleted.

Some variables require special notation. The following notation is used throughout the documentation for all files:

*D(Yxxq) identifies a variable that is deleted as of the quarterly file indicated. The year and quarter are identified by the 'xx' and 'q' respectively. For example, the notation *D(Y141) indicates the variable is deleted starting with the data file of the first quarter of 2014.

*N(Yxxq) identifies a variable that is added as of the quarterly file indicated. The year and quarter are identified by the 'xx' and 'q' for new variables in the same way as for deleted variables.

*C(Yxxq) identifies a variable whose description has been changed. The year and quarter are identified by the 'xx' and 'q' for new variables in the same way as for new and deleted variables.

*L indicates that the variable can contain negative values.

F. State Identifier

The variable STATE identifies the state of residence of respondents. Since the CE survey is not designed to produce state-level estimates, summing the CU weights by state will *not* yield representative state population totals because of three reasons:

- CU's basic weight reflects its *national* probability of selection among a group of primary sampling units of similar characteristics. For example, sample units in an urban nonmetropolitan area in California may represent similar areas in Wyoming and Nevada.
- CUs are post-stratified nationally by sex-age-race. For example, the weights of CUs containing a black male, age 16-24 in Alabama, Colorado, or New York, are all adjusted equivalently.
- Some CUs are located in PSU that span over two states or are suppressed due to nondisclosure requirements by Census. For information, see 2014 Topcoding and Suppression in the 2014 Documentation zip file.

Nevertheless state-level estimates that are unbiased in a repeated sampling sense can be calculated for various statistical measures, such as means and aggregates. However, the estimates will generally be subject to large variances and may be far from the true state population.

List of state identifiers

01	Alabama	29	Missouri
02	Alaska	30	Montana
04	Arizona	31	Nebraska
05	Arkansas	32	Nevada
06	California	33	New Hampshire
08	Colorado	34	New Jersey
09	Connecticut	36	New York
10	Delaware	37	North Carolina
11	District of Columbia	39	Ohio
12	Florida	40	Oklahoma
13	Georgia	41	Oregon
15	Hawaii	42	Pennsylvania
16	Idaho	44	Rhode Island

17	Illinois	45	South Carolina
18	Indiana	46	South Dakota
20	Kansas	47	Tennessee
21	Kentucky	48	Texas
22	Louisiana	49	Utah
23	Maine	51	Virginia
24	Maryland	53	Washington
25	Massachusetts	54	West Virginia
26	Michigan	55	Wisconsin
27	Minnesota		
28	Mississippi		

G. Notes on Files

1. Consumer Unit (CU) Characteristics and Income File (FMLD)

The FMLD file, also referred to as the "Consumer Unit Characteristics and Income" file, contains CU characteristics, CU income, and characteristics and earnings of the reference person and of the spouse. The file includes weights needed to calculate population estimates and variances (see Sections V. Estimation Procedures and VI. Reliability Statement).

Summary expenditure variables in this file can be combined to derive weekly estimates for broad consumption categories. Demographic characteristics, such as family size, refer to the CU status on the date of the interview. Income variables contain annual values, covering the 12 months prior to the date of the interview. When there is a valid nonresponse, or where nonresponse occurs and there is no imputation, there will be missing values. The type of nonresponse is explained by associated data flag variables described in Section III.C. Data Flags.

Summary Expenditure Data

Some variables in the FMLD file contain summary expenditure data. They are all BLS derived. The UCCs comprising each summary expenditure variable are listed below the variable description in the data dictionary. UCCs may not be represented in all Diary quarters. When UCCs are added to or deleted from the summary variable definition, the quarter in which the addition (deletion) to the summary expenditure variable occurs is denoted by a leading character directly after the UCC code in the "Changes to the 2014 Microdata" section. For example, N141<UCC> or D141<UCC> identifies a new or deleted UCC for a given summary expenditure variable beginning in Q141.

2. Member Characteristics and Income File (MEMD)

The "MEMD" file, also referred to as the "Member Characteristics and Income" file, contains selected characteristics for each CU member, including identification of relationship to reference person. Characteristics for the reference person and spouse appear on both the MEMD file and FMLD file. Demographic characteristic data, such as age of CU member, refer to the member status at the placement of each diary. Income data are collected for all CU members over 13 years of age. Income taxes withheld and pension and retirement contributions are shown both annually and as deductions from the member's last paycheck. Income variables contain annual values for the 12 months prior to the interview month. When there is a valid nonresponse, or where nonresponse occurs and there is no imputation, there will be missing values. The type of nonresponse is explained by associated data flag variables described in Section III C. Data Flags.

3. Detailed Expenditures File (EXPD)

In the "EXPD" file, each expenditure recorded by a CU in a weekly diary is identified by UCC, gift/nongift status, and day on which the expenditure occurred. UCCs are six digit codes that identify items or groups of items. (See Dstub file in the 2014 Documentation zip file) There may be more than one record for a UCC on a single day if that is what was reported in the diary. There are no missing values in this file. If no expenditure was recorded for the item(s) represented by a UCC, then there is no record for the UCC on file.

4. Income File (DTBD)

The "DTBD" file, also referred to as the "Income" file, contains CU characteristic and income data. This file is created directly from the FMLD file and contains the same annual and point-of-placement data. It was created to facilitate computer processing when linking CU income and demographic characteristic data with EXPD expenditure data. As such, the file structure is similar to EXPD. Each characteristic and income item is identified by UCC. (See Dstub file in the 2014 Documentation zip file) There are no records with missing values in DTBD. If the corresponding FMLD file variable contained a missing value, there is no record for the UCC.

5. Imputed Income File (DTID)

As a result of the introduction of multiply imputed income data in the Consumer Expenditure Survey, the Imputed DTID file is now on the Microdata. It is very similar to the DTBD file, except that the variable "IMPNUM" will indicate the number (1-5) of the imputation variant of the income variable and it only contains UCCs from variables subject to income imputation.

6. Processing Files

Dstub File (Dstub2014.txt)

Stub files show the hierarchy or aggregation scheme used in the published consumer expenditure tables. The DStub provides the hierarchy for the Diary Survey and the IntStub to integrate both surveys. Each stub file has 7 columns. The stub files are in the sample programs folder in the documentation zip file. The files are formatted as follows:

Name	Content	Code	Format
Type	If information in this line contains aggregation data or not	**1**: Row contains content **2**: Row contains overflow space for descriptions *: Row contains title	CHAR(1)
Level	Aggregation level	Lowest number (1) corresponds to the highest level of aggregation and the highest number to the lowest level (6)	CHAR(1)
Title	Title of the line item	NA	CHAR(60)
Variable or UCC	Variable or UCC number	NA	CHAR(8)
Source	Source of the information	**I**: UCCs from Interview file **D**: UCCs from Diary file **G**: BLS derived variables **T**: Section title (no data) **S**: Administrative variables	CHAR(1)

Name	Content	Code	Format
Factor	Converts annual data into quarterly data	**1**: Leaves quarterly data as such **4**: Converts annual data into quarterly data	CHAR(1)
Group	If the item is an expenditure, income, or asset	**CUCHARS**: CU characteristics **EXPEND**: Expenditures **Food**: Food expenditures **Income**: Income data **Addenda**: Data describing variables	CHAR(7)

III. Estimation Procedure

This section provides users of the CE Diary microdata files with procedures for estimating means and variances of data associated with any U.S. subpopulation. The production of *Consumer Expenditures in 2014* used an integration methodology which incorporated information from *both* Diary and Interview Surveys. Diary data users will not be able to match published CE estimates because of this. In addition, users will not be able to match all values because of suppression of some values, due to topcoding. See the topcoding and other nondisclosure requirements in the 2014 Documentation zip file.

A. General Concepts

NEWID and CUID provide the identification number of each consumer unit (CU) across different PUMD files and interview waves. To connect data for one CU across different files use either variable. However they have a slight difference.

The CUID only identifies the CU. It consists of 7 digits. The first digit is a leading “blank” and is not visible in the data. The next 6 digits identify the CU.

The NEWID identifies the CU and the interview wave. It consists of 8 digits. The first seven digits are identical to the CUID and the last digit identifies the interview wave.

It is not possible to connect CUs in the Interview Survey to CUs in the Diary Survey, because the two surveys survey different CUs.

B. Definition of Terms

Consider the following general situation. We wish to estimate expenditures on certain food items for a special group (subpopulation) of U.S. CUs; for example, all CUs of three persons. Our specific objective is to estimate the expenditures for item *k* over a period of *q* months, where data collected over *r* months are used in the estimate. The following definitions will be helpful in formulating the above type of estimate.

Definition of Terms:

Let

S = all CUs in the subpopulation of interest
k = expenditure item(s) of interest
q = number of months for which estimate is desired
r = number of months in which expenditures were made to be used in calculating the estimate
D = number of days in each of the months in which expenditures were made
j = individual CU in subpopulation S
t = month of expenditure

Then

$X_{(j,k,t)}$ = the amount of money $CU_{(j)}$ spent on item *k* for a week during month *t*
$W_{(j,t,F21)}$ = the weight assigned to $CU_{(j)}$ during month *t*

The F21 denotes FINLWT21 which is used for population estimates.

NOTE: The CUs on the Diary Survey microdata files represent the U.S. population. Some CUs represent more of the population than others; and hence carry more weight. The weight, $W_{(j,t,F21)}$, is a complex estimate of this representation. Refer to Section X.C. Weighting for an explanation of weights. The weights have been adjusted so that the sum of all CU weights for one month approximates one third of the U.S. population. Consequently, the weights for three months (one quarter) of data approximate the total U.S. population.Using the above terminology, we may define:

$X_{(S,k)(q,r)}$ as an estimate for the expenditures of subpopulation *S* on item *k* over a period of *q* months, where data collected over *r* months are used.

and

$\overline{X}_{(S,k)(q,r)}$ as an estimate of the mean expenditures of subpopulation *S* on item *k* over a period of *q* months, where data collected over *r* months are used.

C. Estimation of Total and Mean Expenditures

As an example, let us estimate total expenditures on milk (item *k*) of subpopulation *S* over a 12-month period. Data collected over 6 months will be used to make the estimate. Users may use less than 12 months of data to perform seasonal calculations. In the notation described above, the estimate is $X_{(S,k)(12,6)}$.

$$X_{(S,k)(12,6)} = 3\left(\frac{12}{6}\right)\sum_{t=1}^{6}\left(\sum_{j=1}^{n}\left(\frac{D_{(t)}}{7}\right)W_{(j,t,F21)}X_{(j,k,t)}\right)_t \qquad (1a)$$

where the inner summation sums expenditures for all *j* in *S*, indexed from *j* = 1 through *n* and the outer summation sums over months *t* = 1 through 6. The factor "3" compensates for the fact that the weights for the CUs visited in one month have been adjusted to represent one third of the U.S. population. The factor "12" reflects our desire to estimate expenditures over a 12-month period; and the "6" is the adjustment made because data for 6 months are used. Since the data $X_{(j,k,t)}$ are in terms of weekly expenditures, the factors, (number of days in the month)/7, are used to convert weekly expenditures into their monthly equivalents.

The above formula can be generalized to estimate the total expenditures of subpopulation *S* on item *k* for *q* months, but using data collected over *r* months. The generalization is

$$X_{(S,k)(q,r)} = 3\left(\frac{q}{r}\right)\sum_{t=1}^{r}\left(\sum_{j=1}^{n}\left(\frac{D_{(t)}}{7}\right)W_{(j,t,F21)}X_{(j,k,t)}\right)_t \qquad (1b)$$

where the inner summation sums expenditures for all *j* in *S*, indexed from *j* = 1 through *n* and the outer summation sums over months *t* = 1 through *r*.

An estimate for the expenditures for two or more items may be obtained by summing those expenditures at the CU level and then proceeding as before.

The next example will give an estimate, $\overline{X}_{(S,k)(12,6)}$, of mean expenditures over twelve months (*q*), on item *k*, of CUs in subpopulation *S*, where data collected over a six month period (*r*) are used. The result is

$$\overline{X}_{(S,k)(12,6)} = \frac{3\left(12/6\right)\sum_{t=1}^{6}\left(\sum_{j=1}^{n}\left(\frac{D_{(t)}}{7}\right)W_{(j,t,F21)}X_{(j,k,t)}\right)_t}{\frac{3\sum_{t=1}^{6}\left(\sum_{j=1}^{n}W_{(j,t,F21)}\right)_t}{6}} \qquad (2a)$$

where the numerator is an estimate of aggregate expenditures as formulated in equation (1a), and where the denominator is an estimate of the population of CUs in the U.S. during the six-month period for which the expenditure data are collected. The inner summation in the denominator of (2a) sums FINLWT21 for a given month (*t*), for all *j* in *S*, indexed from *j* = 1 through *n,* and the outer summation in the denominator of (2a) sums over months *t* = 1 through 6. As in the estimate of aggregate expenditures, the factor "3" to the left of the outer summation in the denominator of equation (2a) adjusts FINLWT21 to represent the entire population for each month of data used. The proper U.S. population count is arrived at by dividing the denominator by *r*, or in this case "6", (representing the 6 month period of collected data in this example).

The above formula generalizes to $\overline{X}_{(S,k)(q,k)}$, (i.e., the estimate of the mean expenditure by subpopulation *S* on item *k* for *q* months using data collected over *r* months). In detail:

$$\overline{X}_{(S,k)(q,r)} = \frac{q\sum_{t=1}^{r}\left(\sum_{j=1}^{n}\left(\frac{D_{(t)}}{7}\right)W_{(j,t,F21)}X_{(j,k,t)}\right)_t}{\sum_{t=1}^{r}\left(\sum_{j=1}^{n}W_{(j,t,F21)}\right)_t} \qquad (2b)$$

Note: The factors "3" (adjustment of FINLWT21 to one U.S. population) and "6", (number of months, *r*, for which the data are collected), which appear both in the numerator and the denominator of (2a), cancel. These scalars are dropped from the general form of $\overline{X}_{(S,k)(q,r)}$.

The estimates for total ($X_{(S,k)(q,r)}$) and mean expenditures ($\overline{X}_{(S,k)(q,r)}$) are based on all CUs; not just the CUs with positive expenditures for item *k*. Consider the calculation for the mean expenditure of tobacco. The formula $\overline{X}_{(S,k)(q,r)}$ includes all CUs, both smoking and nonsmoking. One might be more interested in the mean expenditures on tobacco but only for those CUs that actually have expenditures. This can be accounted for by properly defining the initial subpopulation *S* so as to restrict it to CUs with positive tobacco expenditures.

D. Estimation of Mean Annual Income

Let $\overline{Z}_{(S,r)}$ be an estimate of the mean annual income of CUs in subpopulation S, where income data collected over r months is to be used.

Let $Z_{(j,t)}$ = the annual income reported by CU(j) in month *t*. Then the estimated mean annual income is

$$\overline{Z}_{(S,r)} = \frac{\sum_{t=1}^{r}\left(\sum_{j=1}^{n} W_{(j,t,F21)} Z_{(j,t)}\right)_t}{\sum_{t=1}^{r}\left(\sum_{j=1}^{n} W_{(j,t,F21)}\right)_t}$$

IV. Reliability Statement

A. Description of Sampling Error and Non-Sampling Error

Sample surveys are subject to two types of errors, sampling and non-sampling. Sampling errors occur because observations are not taken from the entire population. The standard error, which is the accepted measure for sampling error, is an estimate of the difference between the sample data and the data that would have been obtained from a complete census. The sample estimate and its estimated standard error enable one to construct confidence intervals.

Assuming the Normal Distribution applies to the means of expenditures, the following statements can be made:

(1) The chances that an estimate from a given sample would differ from a complete census figure by less than one standard error are approximately 68 out of 100.

(2) The chances that the difference would be less than 1.6 times the standard error are approximately 90 out of 100.

(3) The chances that the difference would be less than two times the standard error are approximately 95 out of 100.

Non-sampling errors can be attributed to many sources, such as definitional difficulties, differences in the interpretation of questions, inability or unwillingness of the respondent to provide correct information, mistakes in recording or coding the data obtained, and other errors of collection, response, processing, coverage, and estimation for missing data. The full extent of the non-sampling error is unknown. Estimates using a small number of observations are less reliable. A small amount of non-sampling error can cause a small difference to appear significant even when it is not. It is probable that the levels of estimated expenditure obtained in the Diary Survey are generally lower than the "true" level due to the above factors.

B. Estimating Sampling Error

1. <u>Variance Estimation</u>

Variance estimation can be done in many ways. The method illustrated below (a pseudo-replication technique) is chosen because it is accurate yet simple to understand. The basic idea is to artificially

construct several "subsamples" from the original sample data. This construction is done in a manner so that the variance information of the original data is preserved in these subsamples. These subsamples (or pseudo-replications) can then be used to obtain approximate variances for the estimates.

The Diary microdata files contain information that facilitates this form of variance estimation procedure. Specifically, 45 weights are associated with each CU. The forty-fifth weight, called FINLWT21 at BLS, (which is the weight for the total sample) is used for estimations of total or mean expenditures. The other weights (replicates 1 through 44) are used for variance estimation of the totals or means. Note that half of the weights in each replicate are zero. This reflects the fact that in this technique only half the CUs are used in each of the 44 pseudo-replicates. Recall that $X_{(S,k)(q,r)}$ is an estimate for the expenditures of subpopulation *S* on item *k* over a period of *q* months, where data collected over *r* months are used. This notation does not reveal the fact that 45 replicate weights are to be used for estimation of variance. We expand the notation to include this information. Specifically, let $X_{(S,k)(q,r),a}$ = an estimate of the same quantity as $X_{(S,k)(q,r)}$, but using the weights of the a^{th} replicate.

That is $X_{(S,k)(q,r),a}$ is an estimate of the total expenditures by CUs in subpopulation *S* on item *k* over *q* months using *r* months of collection data, and where the weights from the a^{th} replicate are used. Note that the estimate using any one of the first 44 replicate weights only uses part of the data; hence in general $X_{(S,k)(q,r),a}$ is not equal to $X_{(S,k)(q,r)}$.

An estimate for the variance of $X_{(S,k)(q,r)}$ (denoted by $V(X_{(S,k)(q,r)})$ can be calculated using the following formula:

$$V(X_{(S,k)(q,r)}) = \frac{1}{44}\sum_{a=1}^{44}\left(X_{(S,k)(q,r),a} - X_{(S,k)(q,r)}\right)^2$$

Estimates for the variances of $\overline{X}_{(S,k)(q,r)}$ and $\overline{Z}_{(S,r)}$ are similar and are given below.

$$V\left(\overline{X}_{(S,k)(q,r)}\right) = \frac{1}{44}\sum_{a=1}^{44}\left(\overline{X}_{(S,k)(q,r),a} - \overline{X}_{(S,k)(q,r)}\right)^2$$

and

$$V\left(\overline{Z}_{(S,r)}\right) = \frac{1}{44}\sum_{a=1}^{44}\left(\overline{Z}_{(S,r),a} - \overline{Z}_{(S,r)}\right)^2$$

where $\overline{X}_{(S,k)(q,r),a}$ and $\overline{Z}_{(S,r),a}$ are estimates similar to $\overline{X}_{(S,k)(q,r)}$ and $\overline{Z}_{(S,r)}$ except weights of the a^{th} replicates are used.

2. Standard Error of the Mean

The standard error of the mean, *S.E.*($\overline{x}$), is defined as the square root of the variance of the mean. *S.E.*($\overline{x}$), is used to obtain confidence intervals that evaluate how close the estimate may be to the true population mean. A 95 percent confidence interval can be constructed around an estimate, bounded by values 1.96 times the standard error less than and greater than the estimate. For example, the average weekly expenditure for food away from home for All CUs in 2013 was $47.53. The standard error for this estimate is $3.20. Hence, the 95 percent confidence interval around this estimate is from $41.26 to

$53.80. Therefore, we could conclude with 95 percent confidence that the mean weekly expenditures for food away from home for all CUs in 2013 lies within the interval $41.26 to $53.80.

3. Standard Error of the Difference between Two Means

Standard errors may also be used to perform hypothesis testing, a procedure for distinguishing between population parameters using sample estimates. The most common types of hypotheses are: 1) the population parameters are identical; versus 2) they are different.

For example, in 2013 the estimated average weekly expenditure for food away from home for CUs in the *Managers and professionals* occupation category is $62.76 and the estimate for CUs in the *Construction workers and mechanics* category is $45.44 The apparent difference between the two mean expenditures is $62.76 – $45.44 = $17.32. The standard error on the estimate of $62.75 is $1.61 and the estimated standard error for the $45.44 estimate is $3.81. The standard error (S.E.) of a difference is approximately equal to

$$S.E.(\bar{X}_1, \bar{X}_2) = \sqrt{\left(V(\bar{X}_1) + V(\bar{X}_2)\right)}$$

where

$$V(\bar{X}_i) = \left(S.E.(\bar{X}_i)\right)^2$$

This assumes that $\bar{X}_1$ and $\bar{X}_2$ are disjoint subsets of the population. Hence, the standard error of the difference in food away from home expenditures between CUs in the *Managers and professionals* occupation group and in the *Construction workers and mechanics* group is about

$$\sqrt{\left((1.61)^2 + (3.81)^2\right)} = 4.14$$

This means that the 95 percent confidence interval around the difference is from $9.04 to $25.60. Since this interval does not include zero, we can conclude with 95 percent confidence that the mean weekly food away from home expenditures for the *Managers and professionals* occupation group is more than the mean weekly food expenditures for the *Construction workers and mechanics* group.

Analyses of the difference between two estimates can also be performed on non-disjoint sets of the population, where one is a subset of the other. The formula for computing the standard error (S.E.) of the difference between two non-disjoint estimates is

$$SE\left({}_W\bar{X}_1 - {}_W\bar{X}_2\right) = \sqrt{V\left({}_W\bar{X}_1\right) + V\left({}_W\bar{X}_2\right) - 2\rho \cdot SE\left({}_W\bar{X}_1\right) \cdot SE\left({}_W\bar{X}_2\right)}$$

where

$$V(\bar{X}_i) = \left(S.E.(\bar{X}_i)\right)^2$$

and where *r* is the correlation coefficient between $\bar{X}_1$ and $\bar{X}_2$. The correlation coefficient is generally no greater than 0.2 for CE estimates.

V. Sample programs

CE provides sample code to approximate the published tables presented by income groups. The code is available in SAS and R. The variables and ranges referred to in the program are described in the diary data dictionary. The dictionary and the code are available on the PUMD home page in the Documentation zip folder below the section Most Recent Data Release.

The results of the sample code may differ from the published tables due to U.S. Census Bureau confidentiality requirements. CE provides the programs to illustrate the estimation methodology.

VI. Description of the Survey

The CE program consists of two separate components, each with its own questionnaire and independent sample:

1) A Diary or recordkeeping survey completed by the sample CUs for two consecutive 1-week periods; the sample is surveyed across a 12-month period.

2) An Interview panel survey in which each CU in the sample is interviewed once every 3 months over five consecutive quarters to obtain a year's worth of data. New panels are initiated every month of the year.

Data are collected by the Bureau of the Census under contract with BLS. All data collected in both surveys are subject to The U.S. Census Bureau confidentiality requirements, which prevent the disclosure of the CU member's identity.

The Diary survey collects expenditure data for items purchased each day over two one-week periods. This survey is designed to collect expenditure data for small, frequently purchased items such as food, beverages, food consumed away from home, gasoline, housekeeping supplies, nonprescription drugs and medical supplies, and personal care products and services. Respondents are not limited to recording expense for these items only.

A Household Characteristics Questionnaire is completed to record demographic and family characteristics data pertaining to age, sex, race, marital status, and CU relationships each CU member. Income information, such as wage, salary, unemployment compensation, child support, and alimony, as well as information on the employment of each CU member age 14 and over is collected. The expenditure collection instrument is a self-reporting, product-oriented diary on which respondents record all expenses for two consecutive one-week periods. It is divided by day of purchase and by broad classification of goods and services, a format designed to aid the respondents when recording daily purchases.

At the beginning of the two-week collection period, the interviewer uses the Household Characteristics Questionnaire to record demographic and characteristics information pertaining to CU members. Also at this time, a diary for the first week is left with the participating CU. At the completion of the first week, the interviewer picks up the diary, reviews the entries, clarifies any questions, and leaves a second diary for the following week. At the end of the second week, the diary is picked up and reviewed. At this point, the interviewer again uses the Household Characteristics Questionnaire to collect information on CU income, employment and earnings of CU members. These data, along with the other household characteristics information, permit data users to classify sample units for research purposes, and allow BLS to adjust population weights for CUs who do not cooperate in the survey.

VII. Data Collection and Processing

In addition to its data collection duties, the U.S. Census Bureau is responsible for field editing and coding, consistency checking, quality control, and data transmittal to BLS. BLS performs additional review and editing procedures in preparing the data for publication and release.

A. The US Census Bureau Activities

Data collection activities have been conducted by the U.S. Census Bureau on a continuing basis since October 1979. Due to differences in format and design, the Diary Survey and the Interview Survey data are collected and processed separately. Preliminary Diary survey data processing carried out by the U.S. Census Bureau includes programming the Computer Assisted Personal Interview (CAPI) instrument used to collect household characteristics, keying the expenditure data from the diary questionnaire, clerical data editing, and correcting for inconsistencies in the collected data.

The data collected on household characteristics using CAPI are sent directly to the Census Demographic Surveys Division (DSD). Upon completion of the written questionnaire by respondents, the diaries are sent from the regional offices to the Census National Processing Center (NPC) in Jeffersonville, IN. At the NPC, the expenditure data are keyed and codes are applied. The keyed expenditure data are sent to DSD, where they are merged with the household characteristic data. Inconsistencies and errors in the combined data are identified and corrected.

After clerical processing at the NPC, the data are transmitted to the Census Processing Center in Suitland, MD, where they pass through basic quality checks of control counts, missing values, etc. The data are then electronically transmitted to BLS in Washington, DC.

B. Bureau of Labor Statistics Activities

Upon receipt from the U.S. Census Bureau, the data undergo a series of computer edits that identify and correct irregularities and inconsistencies. Other adjustments apply appropriate sales taxes and derive CU weights based on BLS specifications. In addition, demographic and work experience items are imputed when missing or invalid. All data changes and imputations are identified with flags on the Interview data base.

Next, BLS conducts an extensive review to ensure that severe data aberrations are corrected. The review takes place in several stages: a review of counts, weighted means, and unweighted means by region; a review of family relationship coding inconsistencies; a review of selected extreme values for expenditure and income categories; and a verification of the various data transformations.

Cases of extreme data values are investigated by reviewing images of the questionnaires. Errors discovered through this procedure are corrected prior to release of the data.

Two major types of data adjustment routines--imputation and allocation--are carried out to improve and classify the estimates derived from the Diary Survey. Data imputation routines correct for missing or invalid entries among selected CU characteristic fields. Allocation routines are applied when respondents provided insufficient expenditure detail to meet tabulation requirements. For example, reports of combined expenditures for fuels and utilities are allocated among gas, electricity, and other items in this group. To analyze the effects of these adjustments, tabulations are made before and after the data adjustments.

VIII. Sampling Statement

A. Survey Sample Design

Samples for the CE are national probability samples of households designed to be representative of the total U. S. civilian population. Eligible population includes all civilian non-institutionalized persons.

The first step in sampling is the selection of primary sampling units (PSUs), which consist of counties (or parts thereof) or groups of counties. The set of sample PSUs used for the 2014 sample is composed of 91 areas. The design classifies the PSUs into four categories:

- 21 "A" certainty PSUs are Metropolitan Statistical Areas (MSA's) with a population greater than 1.5 million.
- 38 "X" PSUs, are medium-sized MSAs.
- 16 "Y" PSUs are nonmetropolitan areas that are included in the CPI.
- 16 "Z" PSUs are nonmetropolitan areas where only the urban population data will be included in the CPI.

The sampling frame (that is, the list from which housing units were chosen) for the 2013 survey is generated from the 2000 Population Census file. The sampling frame is augmented by new construction permits and by techniques used to eliminate recognized deficiencies in census coverage. All Enumeration Districts (EDs) from the Census that fail to meet the criterion for good addresses for new construction, and all EDs in non-permit-issuing areas are grouped into the area segment frame.

To the extent possible, an unclustered sample of units is selected within each PSU. This lack of clustering is desirable because the sample size of the Diary Survey is small relative to other surveys, while the intraclass correlations for expenditure characteristics are relatively large. This suggests that any clustering of the sample units could result in an unacceptable increase in the within-PSU variance and, as a result, the total variance.

Each selected sample unit is requested to keep two 1-week diaries of expenditures over consecutive weeks. The earliest possible day for placing a diary with a household is predesignated with each day of the week having an equal chance to be the first of the reference week. The diaries are evenly spaced throughout the year.

B. Weighting

Each CU included in the CE represents a given number of CUs in the U.S. population, which is considered to be the universe. The translation of sample families into the universe of families is known as weighting. However, since the unit of analysis for the CE is a CU, the weighting is performed at the CU level. Several factors are involved in determining the weight for each CU for which a diary is obtained. There are four basic steps in the weighting procedure:

1) The basic weight is assigned to an address and is the inverse of the probability of selection of the housing unit.

2) A weight control factor is applied to each diary if subsampling is performed in the field.

3) A noninterview adjustment is made for units where data could not be collected from occupied housing units. The adjustment is performed as a function of region, housing tenure, family size and race.

4) A final adjustment is performed to adjust the sample estimates to national population controls derived from the Current Population Survey. The adjustments are made based on both the CU's member composition and on the CU as a whole. The weight for the CU is adjusted for individuals within the CU to meet the controls for the 14 age/race categories, 4 regions, and 4 region/urban categories. The CU weight is also adjusted to meet the control for total number of CUs and total number of CU who own their living quarters. The weighting procedure uses an iterative process to ensure that the sample estimates will meet all the population controls.

NOTE: The weight for a consumer unit (CU) can be different for each week in which the CU participates in the survey as the CU may represent a different number of CUs with similar characteristics.

IX. Interpreting the Data

Several factors should be considered when interpreting the expenditure data. The average expenditure for an item may be considerably lower than the expenditure by those CUs that purchased the item. The less frequently an item is purchased, the greater the difference between the average for all consumer units and the average of those purchasing (see Section V.B. Estimation of Total and Mean Expenditures). Also, an individual CU may spend more or less than the average, depending on its particular characteristics. Factors such as income, age of family members, geographic location, taste and personal preference also influence expenditures. Furthermore, even within groups with similar characteristics, the distribution of expenditures varies substantially.

Expenditures reported are the direct out-of-pocket expenditures. Indirect expenditures, which may be significant, may be reflected elsewhere. For example, rental contracts often include utilities. Renters with such contracts would record no direct expense for utilities, and therefore, appear to have no utility expenses. Employers or insurance companies frequently pay other costs. CUs with members whose employers pay for all or part of their health insurance or life insurance would have lower direct expenses for these items than those who pay the entire amount themselves. These points should be considered when relating reported averages to individual circumstances.

X. Appendix 1—Glossary

Population
The civilian non-institutional population of the United States as well as that portion of the institutional population living in the following group quarters: Boarding houses, housing facilities for students and workers, staff units in hospitals and homes for the aged, infirm, or needy, permanent living quarters in hotels and motels, and mobile home parks. Urban population is defined as all persons living in a Metropolitan Statistical Area (MSA's) and in urbanized areas and urban places of 2,500 or more persons outside of MSA's. Urban, defined in this survey, includes the rural populations within MSA. The general concept of an MSA is one of a large population nucleus together with adjacent communities that have a high degree of economic and social integration with that nucleus. Rural population is defined as all persons living outside of an MSA and within an area with less than 2,500 persons.

Consumer unit (CU)
A consumer unit comprises either: (1) all members of a particular household who are related by blood, marriage, adoption, or other legal arrangements; (2) a person living alone or sharing a household with others or living as a roomer in a private home or lodging house or in permanent living quarters in a hotel or motel, but who is financially independent; or (3) two or more persons living together who use their income to make joint expenditures. Financial independence is determined by the three major expense categories: housing, food, and other living expenses. To be considered financially independent, at least two of the three major expense categories have to be provided entirely or in part by the respondent.

Reference person
The first member mentioned by the respondent when asked to "Start with the name of the person or one of the persons who owns or rents the home." It is with respect to this person that the relationship of other CU members is determined.

Income before taxes
The combined income earned by all CU members 14 years old or over during the 12 months preceding the interview. The components of income are: Wage and salary income, business income, farm income, Social Security income and Supplemental Security income, unemployment compensation, workmen's compensation, public assistance, welfare, interest, dividends, pension income, income from roomers or boarders, other rental income, income from regular contributions, other income, and food stamps.

Income after taxes
Income before taxes minus personal taxes, which includes Federal income taxes, state and local taxes, and other taxes.

Geographic regions
CUs are classified by region according to the address at which they reside during the time of participation in the survey. The regions comprise the following States:

Northeast - Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont

Midwest - Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin

South - Alabama, Arkansas, Delaware, District of Columbia, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia

West - Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, and Wyoming

XI. Appendix 3—Publications and Data Releases from the Consumer Expenditure Survey

CDs and Free Online Data
PUMD are available for 1996 forward at the PUMD website and for pre-1996 data on CD for purchase at PUMD on CD.

For information and downloading of past PUMD releases, please visit the links below. Multiple zip files can also be downloaded at one time. Please see Instructions for Downloading Consumer Expenditure Survey (CE) Microdata and Documentation for information on downloading the files.

CE microdata on CD are available from the Bureau of Labor Statistics for 1972-73, 1980-81, 1990-91, 1992-93, and for each individual year after 1993 (excluding those years which are currently available for free download online). The 1980-81 through 2013 releases contain Interview and Diary data, while the 1972-73 CD includes Interview data only. The 1980-81, and the 1990 files (of the 1990-91 CD) include selected EXPN data, while the 1991 files (from the 1990-91 CD) and the 1992-93 CD do not. In addition to the Interview and Diary data, the CDs from 1994-2004 include the complete collection of EXPN files. A 1984-94 “multi-year” CD that presents Interview FMLI file data is also available. In addition to the microdata, the CDs also contain the same integrated Diary and Interview tabulated data (1984-2009) that are found on the Consumer Expenditure Survey web site (http://www.bls.gov/cex).

XII. Inquiries, Suggestions and Comments

If you have any questions, suggestions, or comments about the survey, the microdata, or its documentation, please call (202) 691-6900 or email mailto:cexinfo@bls.gov.

Written suggestions and comments should be forwarded to:

Division of Consumer Expenditure Survey
Branch of Information and Analysis
Bureau of Labor Statistics, Room 3985
2 Massachusetts Ave. N.E.
Washington, DC. 20212-0001

The Bureau of Labor Statistics will use these responses in planning future releases of the microdata.