

Weighting the CFM

By

Jay Zagorsky

December 2011

The CFM provides researchers with a number of weights. When these weights are used they adjust the raw CFM data so that it better matches results found in other national surveys and enables an easier explanation of results.

Weights were constructed to match five goals. The first goal is that CFM data match the age ranges of respondents in the Current Population Survey (CPS). This adjustment is important because raw CFM data has too many elderly people and too few young compared to the actual population in the USA. The second goal of the weights is to ensure homeownership rates in the CFM match homeownership in the actual USA. The third goal is to ensure the CFM matches the black/African American and non-black population percentages. The fourth goal is that CFM respondents have a similar income distribution as found in the March CPS. The final goal is that after weighting the total weights of all respondents in a particular time frame match the number of USA households.

Mechanics of Weighting

CFM weights are created using a mathematical concept formally called the “Generalized Raking Method.” The method is described in the paper “Generalized Raking Procedures in Survey Sampling” by Deville, Sarndal and Sautory, and published in the Journal of the American Statistical Association, Vol. 88, No. 423 (Sept. 1993), pages 1013-1020.

In simple terms the idea of raking is to come up with adjustment factors so that the percentages found in the survey match a set of known population percentages. For example, let’s assume that in the population 70% of the households own their home and 30% rent. These numbers are known with certainty. After doing a telephone survey, however, the raw telephone data show half of all respondents own a home and half do not. The raking method adjusts the telephone survey numbers and overweights respondents with state they own a home and underweights respondents who state they rent. In this example, survey respondents who own a home should have a weight of $7/5$ or 1.4, while renters should have a weight of $3/5$ or 0.6. Using these weights, results in calculations that match the national totals.

While originally CFM survey staff wrote their own computer code to do the generalized raking method, today the CFM uses computer code written by the French government. Deville, Sarndal and Sautory, the authors of the raking paper cited above, work for INSEE and provide to the public SAS raking macros, which have been thoroughly tested, to perform the computations. INSEE calls their raking method code “Calmar.” For readers familiar with the Calmar code, the CFM uses mode 2, which is the Multiplicative method. This method is used because it never returns negative weights, unlike some of the other raking methods. The computer code and documentation (only available in French) for Calmar is located at:

http://www.insee.fr/fr/methodes/default.asp?page=outils/calmar/accueil_calmar.htm

Creating the Margins

The Calmar program needs a set of “margins.” The margins are simply the percentages found in the entire USA that apply to each group. For example, the CFM stratifies respondents based on homeownership. If 64% of households in the USA own a home and 36% do not, then the margins needed by Calmar are 64% and 36%. The various percentages are the target CFM data should match after being weighted.

These percentage targets are input into Calmar using a special SAS data set called “margins.” Below is an example of the “margins” data step. Where the specific numbers come from is explained in detail after the example. In general there are a large number of tables in this document. The three rows of numbers all come from the same year. Each year after all the data are available the CFM revises the set of numbers.

Users who want to adjust the national totals need to follow very specific rules for the “margins” data set. One reason you might want to adjust the national totals is to change the numbers when new years and information becomes available. In the “margins” data set each row after the CARDS statement contains the variable name, the number of groups the variable name is tracking and then the actual percentages it should use in the calculations. The percentages on each row must total 100% and missing percentages are shown using a “.”, not a zero.

```
data margins;
input var $ n mar1 mar2 mar3 mar4 mar5 mar6 mar7 mar8 mar9 mar10 mar11 mar12
mar13;
CARDS;
INCGROUP 13 7.83 7.05 7.64 9.12 7.92 7.15 9.19 9.56 13.81 5.84 2.04 5.69 7.16
AGEGROUP 7 5.6 16.8 19.6 20.9 16.6 10.3 10.2 . . . . .
HOMEOWN 2 30 70 . . . . .
;
```

INCGROUP Data:

The CFM divides people based on four characteristics; income, race, age and homeownership. The income and race group were combined into one classification called Incgroup. The Incgroup data show what percentage of the USA population is in 13 income and race groups. Incgroup information is based on values found in the March Annual Social and Economic (ASEC) Supplement to the Current Population Survey (CPS). More information about the CPS and ASEC is found at <http://www.census.gov/cps>.

Incgroup and the associated percentages are defined in the following two tables. There are two tables because starting in 2009 the Census Bureau changed the categories they use in their public reports. The change impacts just groups 1 to 9. It does not impact groups 10 to 13.

Table 1: Income Groups Used by the CFM since 2009.

Group #	Race	Income Range	2010	2009
1	Nonblack	<\$15,000 (including negative income)	10.36%	10.01%
2	Nonblack	\$15,000 to \$19,999	5.16%	4.86%

3	Nonblack	\$20,000 to \$29,999	9.66%	9.83%
4	Nonblack	\$30,000 to \$39,999	8.67%	8.99%
5	Nonblack	\$40,000 to \$49,999	7.67%	7.83%
6	Nonblack	\$50,000 to \$59,999	6.93%	7.10%
7	Nonblack	\$60,000 to \$74,999	8.83%	9.04%
8	Nonblack	\$75,000 to \$94,999	8.77%	8.75%
9	Nonblack	\$95,000 to \$149,499	12.84%	12.78%
10	Nonblack	\$150,000 to \$249,999	5.94%	5.94%
11	Nonblack	\$250,000 and up	2.01%	1.93%
12	Black	<\$29,999 (including negative income)	6.19%	5.92%
13	Black	\$30,000 and up	6.97%	7.02%
		Total	100.0%	100.0%

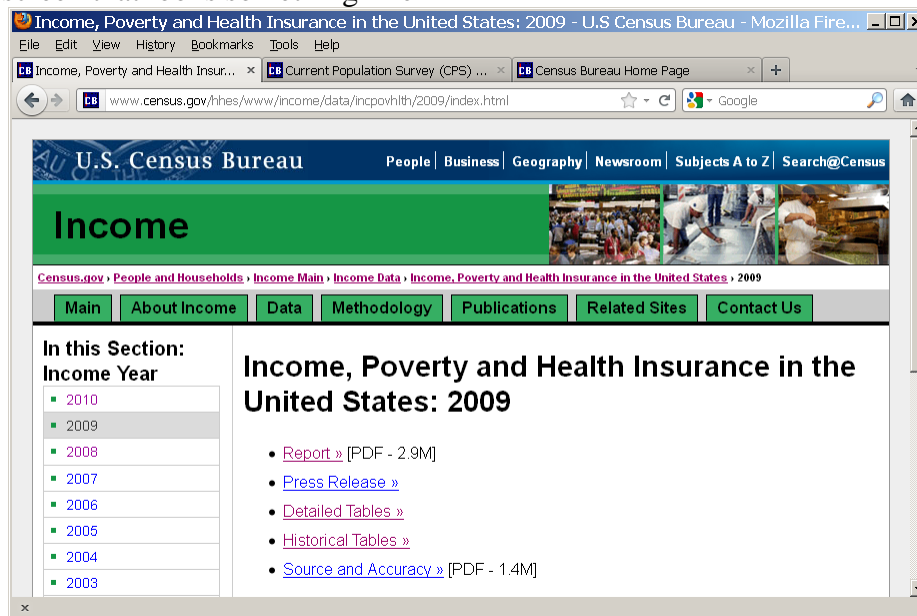
Table 2: Income Groups Used by the CFM from 2005 to 2008.

Group #	Race	Income Range	2008	2007	2006	2005
1	Nonblack	<\$12,500 (includes neg.)	7.83%	7.75%	8.08%	8.80%
2	Nonblack	\$12,500 to \$19,999	7.05%	7.26%	7.19%	7.71%
3	Nonblack	\$20,000 to \$27,499	7.64%	7.67%	7.76%	8.04%
4	Nonblack	\$27,500 to \$37,499	9.12%	8.89%	9.63%	9.61%
5	Nonblack	\$37,500 to \$47,499	7.92%	8.20%	8.44%	8.53%
6	Nonblack	\$47,500 to \$57,499	7.15%	7.41%	7.53%	7.69%
7	Nonblack	\$57,500 to \$72,499	9.19%	9.12%	9.34%	9.42%
8	Nonblack	\$72,500 to \$92,499	9.56%	9.47%	8.98%	9.14%
9	Nonblack	\$92,500 to \$149,499	13.81%	13.75%	13.18%	12.14%
10	Nonblack	\$150,000 to \$249,999	5.84%	5.80%	5.33%	4.63%
11	Nonblack	\$250,000 and up	2.04%	1.85%	1.87%	1.70%
12	Black	<\$29,999 (includes neg.)	5.69%	5.73%	5.82%	6.10%
13	Black	\$30,000 and up	7.16%	7.10%	6.86%	6.49%
		Total	100.0%	100.0%	100.0%	100.0%

Readers who want to understand exactly where the percentages in Table 1 and 2 come from or who need to update Incgroup for years beyond what are shown in the table should continue reading this section. If you do not need to understand the lowest details you can skip to the next section labeled “AGEGROUP Data.”

The particular set of percentage used to determine Incgroup are taken from the HINC-06 tables. The HINC-06 tables track household income distribution by race. There are a large number of income categories in these tables up to \$250,000 and then there is one combined category for household with more than a quarter of a million dollars of income.

The HINC-06 tables are located at census.gov. At the Census website select “People and Households” then select “Income” followed by “Income Data” and finally “Income, Poverty and Health” or jump to www.census.gov/hhes/www/income/data/incpovhlth/index.html. This should get you to a screen that looks something like



Once at this screen, select a particular year. Then select “Detailed Tables” (not historical tables). On the next screen select HOUSEHOLD. Do not select family or persons. This will bring up the following list. The file you want is on the bottom and is called HINC-06. To read this file into Excel, start Excel and pick file/open. Instead of picking a particular file from your hard drive, paste the URL for the HINC-06 table and the contents will be imported into Excel.

Household Income Table of Contents

These data come from the CPS 2010 **Annual Social and Economic Supplement**,
formerly called the March Supplement.
(2009 Income)

For information on confidentiality protection, sampling error, nonsampling error, and definitions, see www.census.gov/apsd/techdoc/cps/cpsmar10.pdf[PDF].

Table Number	Title
HINC-01	Selected Characteristics of Households by Total Money Income in 2009
HINC-02	Age of Householder--Households, by Total Money Income in 2009, Type of Household, Race and Hispanic Origin of Householder
HINC-03	People in Households--Households, by Total Money Income in 2009, Age, Race and Hispanic Origin of Householder
HINC-04	Presence of Children Under 18 Years old--Households, by Total Money Income in 2009, Type of Household, Race and Hispanic Origin of Householder
HINC-05	Percent Distribution of Households, by Selected Characteristics Within Income Quintile and Top 5 Percent in 2009
HINC-06	Income Distribution to \$250,000 or More for Households: 2009

When you open the HINC-06 file there are a variety of racial groupings. There are only two columns of information that you need is the “All Races” and “Black A.O.I.C.” A.O.I.C. stands for alone or in combination. The columns look like the below picture. To calculate the percentage for each group, divide the total number of people in an income category by the number of households in the USA. For example to get the share of low income non-blacks in the USA in 2008, which is INCGROUP 1, type “=(SUM(B13:B17)-SUM(J13:J17))/B12.” This results in 7.83%.

The screenshot shows a Microsoft Excel spreadsheet titled 'fooYear.xlsx'. The active sheet is 'Black A.O.I.C.'. The data is organized in a table with columns for 'All Races', 'White A.O.I.C.', 'White alone (1)', 'White alone, not', 'Black A.O.I.C.', and 'Black alone (2)'. Each of these columns has sub-columns for 'Number' and 'Mean Income'. The rows represent different income brackets, from 'Total' down to '\$25,000 to \$27,499'. The 'Number' column for 'Black A.O.I.C.' is highlighted in yellow.

	All Races		White A.O.I.C.		White alone (1)		White alone, not		Black A.O.I.C.		Black alone (2)	
	Number	Mean Income	Number	Mean Income	Number	Mean Income	Number	Mean Income	Number	Mean Income	Number	Mean Income
Income of Household												
Total	117,181	68,424	96,601	71,080	95,297	71,191	82,884	74,102	15,056	46,679	14,595	46,533
Under \$2,500	2,588	298	1,744	267	1,705	262	1,369	245	646	494	627	495
\$2,500 to \$4,999	971	3,792	660	3,792	642	3,781	494	3,807	253	3,787	243	3,760
\$5,000 to \$7,499	1,677	6,261	1,140	6,251	1,121	6,248	889	6,244	466	6,294	460	6,282
\$7,500 to \$9,999	3,141	8,705	2,212	8,724	2,164	8,725	1,742	8,727	814	8,648	798	8,650
\$10,000 to \$12,499	3,684	11,223	2,817	11,237	2,783	11,240	2,262	11,238	704	11,186	683	11,183
\$12,500 to \$14,999	3,163	13,687	2,494	13,705	2,465	13,704	2,060	13,695	550	13,608	546	13,609
\$15,000 to \$17,499	3,600	16,074	2,926	16,075	2,877	16,079	2,436	16,105	575	16,057	546	16,080
\$17,500 to \$19,999	3,116	18,662	2,519	18,666	2,484	18,665	2,048	18,696	489	18,653	475	18,652
\$20,000 to \$22,499	3,967	21,064	3,116	21,077	3,081	21,077	2,517	21,120	659	20,993	642	20,997
\$22,500 to \$24,999	3,117	23,698	2,538	23,699	2,498	23,699	2,128	23,708	456	23,670	447	23,670
\$25,000 to \$27,499	3,620	26,003	2,863	26,021	2,827	26,020	2,340	26,036	639	25,926	612	25,913

AGEGROUP Data:

The CFM weights also adjust individuals based on their age. There are seven different age groups used by the CFM. The age data, like the income, are based on values found in the March Supplement to the CPS. Figure 1 shows at the top of the spreadsheet the absolute number of people in each age group. The bottom of the spreadsheet shows the percentages of people who are in each age category. The bottom percentages are used in the “Margins” data set to adjust the CFM information.

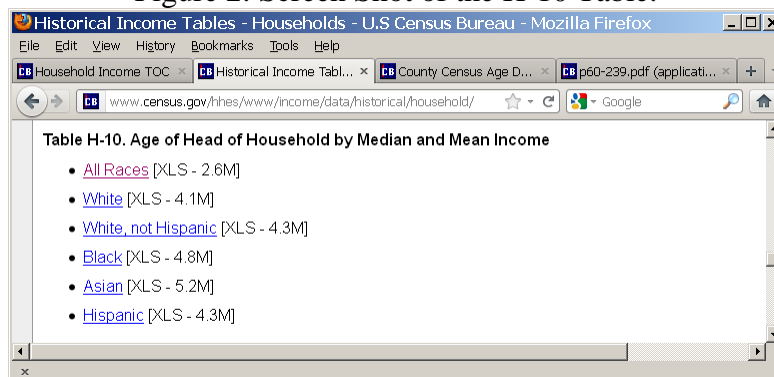
Figure 1: Seven Age Categories Used By the CFM.

	15 Years & Over	15 to 24 Years	25 to 34 Years	35 to 44 Years	45 to 54 Years	55 to 64 Years	65 to 74 Years	75 Years & Over
2010	118,682	6,140	19,572	21,250	24,530	21,828	13,348	12,015
2009	117,538	6,233	19,257	21,519	24,871	20,387	13,164	12,106
2008	117,181	6,357	19,302	22,171	24,633	19,883	12,842	11,992
2007	116,783	6,554	19,225	22,448	24,536	19,909	12,284	11,829
2006	116,011	6,662	19,435	22,779	24,140	19,266	11,926	11,803
2005	114,384	6,795	19,120	23,016	23,731	18,264	11,687	11,772
2004	113,343	6,733	19,314	23,248	23,393	17,503	11,528	11,623
2003	112,000	6,610	19,159	23,222	23,137	16,824	11,499	11,550
2002	111,278	6,611	19,055	24,069	22,623	16,260	11,360	11,299
2001	109,297	6,391	18,988	24,031	22,208	15,203	11,472	11,004
2000	108,209	6,409	19,031	24,054	21,969	14,277	11,490	10,979

	15 Years & Over	15 to 24 Years	25 to 34 Years	35 to 44 Years	45 to 54 Years	55 to 64 Years	65 to 74 Years	75 Years & Over
2010	100.0%	5.2%	16.5%	17.9%	20.7%	18.4%	11.2%	10.1%
2009	100.0%	5.3%	16.4%	18.3%	21.2%	17.3%	11.2%	10.3%
2008	100.0%	5.4%	16.5%	18.9%	21.0%	17.0%	11.0%	10.2%
2007	100.0%	5.6%	16.5%	19.2%	21.0%	17.0%	10.5%	10.1%
2006	100.0%	5.7%	16.8%	19.6%	20.8%	16.6%	10.3%	10.2%
2005	100.0%	5.9%	16.7%	20.1%	20.7%	16.0%	10.2%	10.3%
2004	100.0%	5.9%	17.0%	20.5%	20.6%	15.4%	10.2%	10.3%
2003	100.0%	5.9%	17.1%	20.7%	20.7%	15.0%	10.3%	10.3%
2002	100.0%	5.9%	17.1%	21.6%	20.3%	14.6%	10.2%	10.2%
2001	100.0%	5.8%	17.4%	22.0%	20.3%	13.9%	10.5%	10.1%
2000	100.0%	5.9%	17.6%	22.2%	20.3%	13.2%	10.6%	10.1%

The underlying data for age are found in March CPS Table H-10. The H-10 table is located at www.census.gov/hhes/www/income/data/historical/household/. The path to get to the H-10 table in case the URL changes is found by first going to census.gov. From the main menu make the following selections; “People and households,” “Income Main,” “Income Data,” Historical Tables,” and “Households.” Figure 2 shows an image of the final selection.

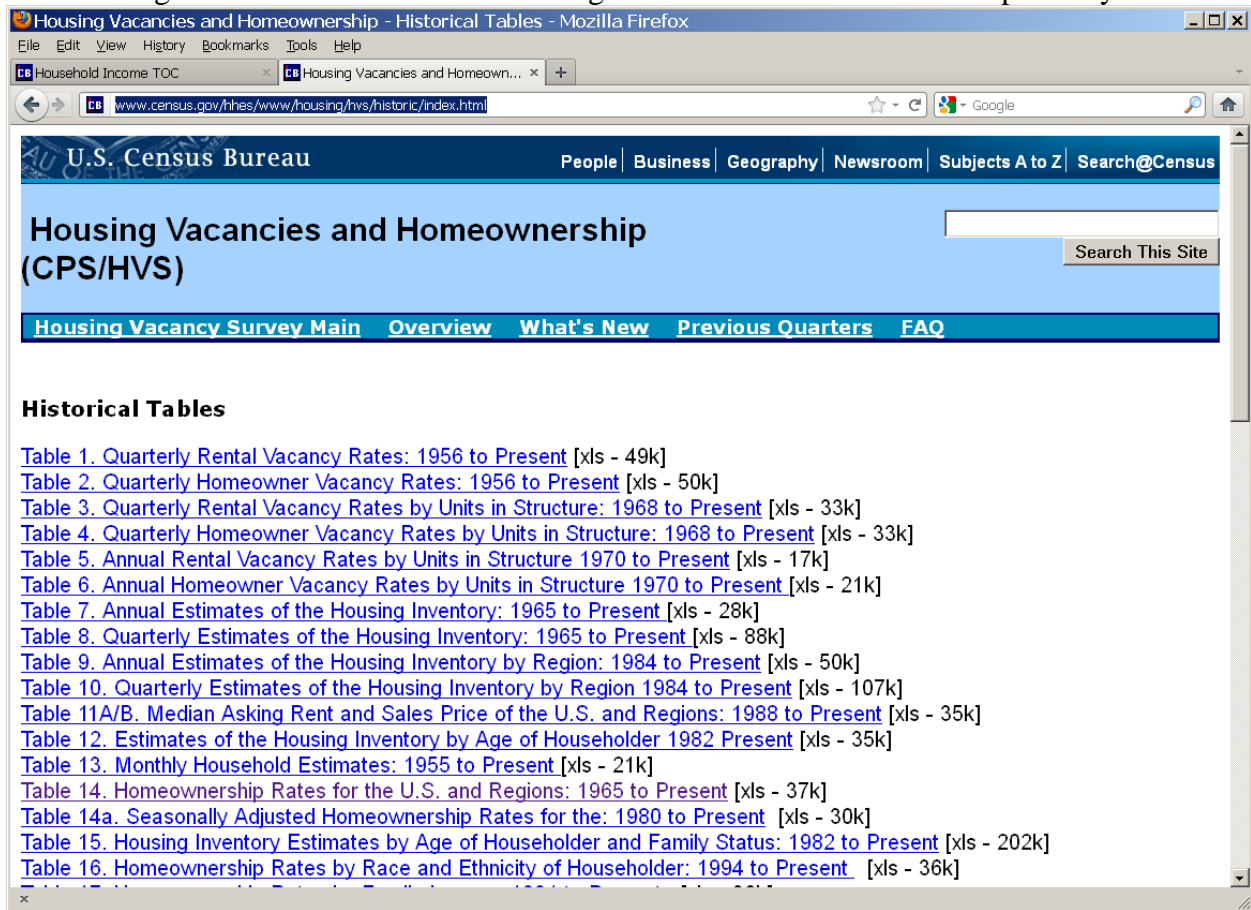
Figure 2: Screen Shot of the H-10 Table.



HOMEOWN Data:

Homeownership rates do not come from the CPS. Instead they come from the Housing Vacancies and Homeownership Survey. This survey is also run by the USA Census Bureau. The data is found by going to the below URL. Figure 3 shows a screenshot of the Housing survey (<http://www.census.gov/hhes/www/housing/hvs/hvs.html>).

Figure 3: Screen Shot of the Housing Vacancies and Homeownership Survey.



The CFM uses table 14, which is titled “Homeownership Rates for the U.S. and Regions.” This table provides quarterly homeownership rates at the national level. The CFM averages (mean) the four quarterly estimates and then uses the following figures when calculating the HOMEOWN line in the “Margins” data step.

Table 3: Home Ownership Rates Used by the CFM Since 2005.

Year	Not Own	Own
2005	31.1	68.9
2006	31.2	68.8
2007	31.9	68.2
2008	32.2	67.8
2009	32.6	67.4

2010	33.2	66.9
2011	33.8	66.2

Sum of Weights:

The average of the *original* CFM weights was “1.” This means no adjustment was made to the typical respondent. A person with a weight greater than 1 meant that they were under-represented in the CFM survey. A person with a weight less than 1 meant they were over-represented.

Starting in 2012 a revision was made to the CFM weights. Instead of the weights summing to 1, the revised weights sum to the number of households in the USA. This enables users to interpret their number as national totals. For example, with the new weighting totals it is simple to state the number of USA household who feel they have very high debt stress.

The number of households is taken each year from the HINC-06 tables, which are discussed in detail above in the INCGROUP section. Currently, the number of households (in 1,000s) used by the CFM weighting program are;

Table 4: Number of USA Households Used by the CFM Since 2005.

2010	2009	2008	2007	2006	2005
118,682	117,538	117,181	116,783	116,011	114,384

Handling Cases with Missing Information:

One problem with the generalized raking method is that to create a weight for a CFM respondent the computer program needs four pieces of information; a person’s age, their annual income, their race and if they own their home. If any of these pieces of information is missing the raking algorithm cannot produce a result, even if the respondent completed every other question in the survey. To solve this issue the CFM fills in missing information needed by the weighting program so that every respondent gets a weight.

Since the information is missing the CFM draws a random number from the uniform distribution. It then compares the random number to the cumulative distribution for the missing factor. The respondent is then assigned the value from the bin of the cumulative distribution function in which the random number falls.

For example, let’s assume a respondent did not provide a homeownership answer. In 2010 the national homeownership rates were 33.2% rented and 66.9% owned. The cumulative distribution function has two bins or groups. The first bin is from 0% to 33.2% and represents renters. The second bin is from epsilon plus 33.2% to 100% and represents homeowners. The CFM draws a random number from 0 to 1. Let’s assume the number is 0.578. The respondent is then marked as a homeowner since their random number fell into the second bin’s range.

Practical Usage:

There are three broad types of weights in the CFM; 3 month rolling weights, quarterly weights and annual weights. Which weight should you use to produce descriptive statistics? In general, you should use the weighting series that most closely approximates the time frame you are analyzing. For example, if you are looking at just a month or two of data, you should use the 3 month rolling weights. If you are looking at annual data you should use the annual series.

Because the CFM is not a multi-stage sample, users do not need to adjust for design effects. The CFM raw input for statistical purposes can be considered a simple random sample.