

## Appendix D. Nonsampling and Sampling Errors

All numbers from the American Housing Survey (AHS), except for sample size, are estimates. As in other surveys, errors come primarily from the following sources:

- **Incomplete data.** Incomplete data are adjusted by assuming that the respondents are similar to those not answering and the size of these errors is estimated, to include coverage errors and missing data.
- **Wrong answers.** The U.S. Census Bureau does not adjust for wrong answers and does not estimate the size of the errors.
- **Sampling Variability.** Sampling errors are not adjusted and the size of the error is estimated.

Incomplete data and wrong answers are usually the largest source of errors, larger than sampling errors. For example, in the *American Housing Survey-National* (AHS-N), the changes in weighting in 1991 and 2003 corrected some of the error due to incomplete data; that one correction averaged 2.5 percent in 1991 and 1.0 percent in 2003. Worse errors from incomplete data and from wrong answers apply to some items, discussed below. Additional information on the quality of AHS data can be obtained from the U.S. Census Bureau, *American Housing Survey: A Quality Profile*, Series H121/95-1.

### INCOMPLETE DATA

**Coverage errors.** Because of deficiencies with our sampling lists, the housing units in the AHS do not represent all housing units in the country. The Census Bureau attempts to compensate for the deficiencies by adjusting the raw numbers from the survey proportionally so that the published numbers match independent estimates of the total number of housing units. See Appendix B, “New Construction Adjustment”.

In 2005, the Census Bureau attempted to reduce the undercoverage in two segments of the population by adding sample units selected from the 2000 census (i.e., manufactured/mobile homes built between 1980 and 2000 and special living units). The approximate housing unit undercoverage for the 2013 AHS national sample ranges from less than 1 percent to 22.8 percent. Table D-1 lists units that have known coverage deficiencies.

**Missing data.** Some people refuse to answer some or all of the questions and some people do not know the answers to all of the questions. When entire interviews are missing, survey estimates are adjusted for them by assuming that they are similar to some group of housing units that did report data. With this assumption, housing units that responded are adjusted

accordingly to account for non-response. This assumption is never exactly true, although it is usually better than ignoring the housing units that were not interviewed (see Appendix B, “Noninterview Adjustment”).

Incompleteness or “missing answers”, also known as item nonresponse, can cause large errors. For example, if 10 percent of housing units did not respond to a particular question, they represent about 13.2 million housing units (there are about 132 million housing units in the U.S.). For each case of item nonresponse, other similar interview answers are used to represent those missing interview answers; this is, for most missing answers, an answer from a similar household is copied<sup>1</sup>. For other items not recorded in the interview, "Not reported" is used as an answer category. The items with the most missing data are primarily those that people forget or consider personal: mortgages, other housing costs, and income.

It is not surprising that large biases, as shown in Table D-2, are possible when the survey has data for only 50 to 90 percent of housing units for particular items. Readers should be wary of items with highly incomplete data.<sup>2</sup>

Rates of completeness were not computed for 2011. Table D-2 in Appendix D of *American Housing Survey for the U.S. in 1995* gives the completeness rates for 1995. Due to the change in data collection methodology, the rates for 2011 may be higher or lower than in the past. However, the items that were most incomplete in 1995 are probably still most incomplete for 2011.

**Change over time.** Several aspects of the AHS make estimates of change from previous data unreliable. These changes may elicit different answers from the past, even if nothing changed in the housing unit. Some examples of changes that may have affected answers include:

- Question wording
- Order of questions
- Switch from paper to computer questionnaire in 1997
- Lack of Spanish questionnaire, prior to 2009

---

<sup>1</sup>Hot deck imputation is used: missing answers are copied from the most recently processed similar household for the household with the missing items.

<sup>2</sup>Statistical note: The November 1990 paper, *How Response Error, Missing Data and Undercoverage Bias Survey Data*, estimates that 90 percent of errors from incomplete data are less than:  $1.645 \times (.0012 \times U + .0363 \times (\text{lesser of } A \text{ or } U-A))$  where A is any count from the AHS and U is the total number of housing units in the U.S. or metropolitan area (both in thousands, result also in thousands). Weights are adjusted to reduce these errors, but it is not known how much error remains. *How Response Error, Missing data and Undercoverage Bias Survey Data*, order number HUD-6458, is available by e-mailing [helpdesk@huduser.org](mailto:helpdesk@huduser.org) or calling 800-245-2691.

**Table D-1. Poorly Covered Units**

Type of Unit	Type of Deficiency
Manufactured/mobile homes, boats, and recreational vehicles (RVs)	No coverage of new manufactured/mobile home parks, new marinas, and new RV parks since April 1980 for AHS-N in areas where addresses are complete and permits are required for new construction.
Conventional new construction in permit issuing areas	No coverage of permits issued fewer than 8 months before interviewing or housing units built without permits where permits are required. In addition, eligible units could be missed and ineligible units included because of incorrect answers to questions used to screen out ineligible units.
New construction in special places (for example, college campuses, prisons)	Not covered in either permit-issuing or nonpermit-issuing areas.
Group quarters and houses moved in	Eligible units could be missed because of incorrect answers to questions used to screen out ineligible units.
Conversions from nonresidential units	Minimal coverage of nonresidential units in buildings with no living quarters at the time of the 1980 census that converted to housing units by 1991 (and no coverage since 1991) in areas where addresses are complete and permits are required for new construction.
Within-structure additions	Some extra apartments created illegally or occupied by fugitives are probably missed because people do not report them for fear of penalties.
Whole structure additions not covered by permit sampling	These units are chosen with the aid of screening questions. Eligible units could be missed and ineligible units included because of incorrect answers to the screening questions.

## WRONG ANSWERS

Wrong answers happen because people misunderstand questions, cannot recall the correct answer, or do not want to give the right answer. Table 1 in Appendix D of the *American Housing Survey for the United States in 1995* shows inconsistency rates for items (pages D-4 to D-6 located here:

[https://www.census.gov/content/dam/Census/library/publications/1997/demo/h150\\_95rv.pdf](https://www.census.gov/content/dam/Census/library/publications/1997/demo/h150_95rv.pdf)). These results are from 1995, but the inconsistency rates for 2011 are probably similar.

## SAMPLING ERRORS

**Sampling errors definition.** Error from sampling reflects how estimates from a sample vary from the actual value. (Note: "Actual value" means the value derived if all housing units had been interviewed under the same conditions, rather than only a sample). A confidence interval is a range that contains the actual value with a specified probability. The range of nonsampling error is usually larger than this confidence interval. AHS utilizes two approaches to providing sampling errors:

- Providing Replicate Weights in the Public Use File (PUF) to allow users to calculate their own sampling errors they see fit; and
- By providing Generalized Variance Formulas (GVFs), so that a user can calculate sampling errors for any estimate. Note that using GVFs may be an overestimate, or a “conservative” estimate of sampling errors (Table D-4).

**Counts.** Most numbers from the AHS are counts of housing units (for example, units with basements or units with an elderly person). These counts have variation due to the sample selection methodology that we describe as the sampling errors. As with the other types of errors, readers should be wary of numbers with large errors from sampling.

Table D-3 gives a convenient list of associated sample estimates and standard errors for the 2011 AHS-N sample. For numbers not in Table D-3, use the appropriate formula from Table D-4. In each formula, “A” is the estimate (a count of housing units in thousands) from the AHS. Remember that in any case that the total error is larger than sampling error.

The error from sampling is approximated using the following formula for constructing a 90-percent confidence interval:

$$1.645 \times \sqrt{6.70 \times A - 0.000050 \times A^2}$$

where A is a number (a count of units in thousands) from the AHS. This formula is an overestimate for most items because it applies to any characteristic of AHS. For more accurate estimates, use the formulas in Table D-4.

For example, suppose a specific domain of interest has an estimated value of 200,000, (that is, A = 200). The error from sampling for a 90 percent confidence interval for that value of 200,000 units is:

$$1.645 \times \sqrt{6.70 \times 200 - .000050 \times 200^2} = 60$$

The 90-percent confidence interval can be formed by adding and subtracting this error to the survey estimate of 200 (that is, 200 plus or minus 60). Statements such as "the actual value is in the range 200 plus or minus 60 (140 to 260)," are right 90 percent of the time and wrong 10 percent of the time.<sup>3</sup>

Numbers in the publication are printed in thousands, so 200 means 200,000. The formulas are designed to use numbers directly from the publication; do not include zeros. The result is also in thousands, so 60 means 60,000.

**Percents.** Any subgroup can be shown as a percent of a larger group. For AHS-N, the error from sampling for a 90-percent confidence interval for this percent is:

$$1.645 \times \sqrt{\frac{6.70 \times p \times (100 - p)}{A}}$$

where p is the percent A is the denominator, or base of the percent, in thousands.

For example, suppose that out of the estimate of 200,000 housing units, 40% were 3-bedroom units. The error from sampling for a 90-percent confidence interval for 40 percent of 200 (meaning 200,000) is:

$$1.645 \times \sqrt{\frac{6.70 \times 40 \times 60}{200}} = 14.8$$

Statements such as "the actual percent is in the range 25.2 percent to 54.8 percent" are right 90 percent of the time.

---

<sup>3</sup>The formula in the text is based on 1.645 times the standard error from sampling. This formula gives "90-percent confidence interval errors." For 95-percent confidence interval errors, multiply by 1.96 instead of 1.645; for 99-percent confidence, multiply by 2.576 instead of 1.645.

This formula is an overestimate for most items. To get a more accurate estimate for AHS-N, replace the first number under the square root sign above with the first number under the square root sign of the appropriate formula from Table D-4.<sup>4</sup>

Note that when a ratio C/D is computed where C is not a subgroup of D (for example, the number of Hispanics as a ratio of the number of Blacks), the error from sampling is different.<sup>5</sup>

**Medians.** The steps in Table D-5 calculate the error from sampling for a 90 percent confidence interval for medians. This is an approximation of the error.

For small bases, the confidence interval on medians cannot be estimated reliably. To estimate a median’s sampling error more accurately, use Table D-6 to find the sampling error on 50 percent and apply it to compute the 90 percent confidence interval for the median.

**Differences.** Two numbers from the AHS, like 210 and 324 or 34 percent and 55 percent have a “statistically significant” difference if their ranges of error from sampling for a 90 percent confidence interval do not overlap.<sup>6</sup>

**Formulas for error from sampling.** The letter “A” in the formulas in Tables D-4, D-5, and D-6 represents a number (the estimated count of housing units in thousands) from AHS (see “Sampling Errors” text for an example of how “A” is used). For AHS-N, the minimum error from sampling is ±15 (meaning ±15,000).<sup>7</sup> If a formula gives an error smaller than 15, use 15.

---

<sup>4</sup>This formula is actually  $1.645 \times \sqrt{p(100 - p)/n}$ , since  $6.70/A$  adjusts the data to the effective sample size.

<sup>5</sup>The error from sampling for a 90 percent confidence interval for a ratio C/D is:

$$C/D \sqrt{(\text{error for } C/C)^2 + (\text{error for } D/D)^2}$$

when the error for C should be interpreted as the error for a 90 percent confidence interval for C. Likewise, the error for D should be interpreted as the error for a 90 percent confidence interval for D.

<sup>6</sup> When ranges of error from sampling for a 90 percent confidence interval do overlap, numbers are still statistically different if the result of subtracting one from the other is more than

$$\sqrt{(\text{error for first number})^2 + (\text{error for second number})^2}$$

The error for the first and second numbers should be interpreted as the error for a 90 percent confidence interval for the first and second numbers respectively.

<sup>7</sup> This minimum error formula is based on the following binomial 90 percent confidence interval on zero  $U \times (1 - .1^{6.70/U}) \approx 15$  (where U is the total number of housing units from the AHS). For a 95 percent confidence interval, substitute .05 for .1 in the above formula. For a 99 percent confidence interval, substitute .01 for .1. More discussion and other approximations are in the paper “Sampling Errors for Small Groups”, abstract available at <http://www.huduser.org/portal/searchbiblio/Bibliography?id=8509>.

For AHS-N, if an item falls into two different categories in Table D-4, use the formula that gives the largest error.

The formulas in this appendix give the errors for a 90 percent confidence interval. For a 95 percent confidence interval, multiply by 1.960 instead of 1.645. For a 99 percent confidence interval, multiply by 2.576 instead of 1.645.

Historically, separate formulas were provided for several individual characteristics such as fuels and neighborhood characteristics. In 2011, variance formulas were calculated separately for Total Housing Units, Total Occupied Housing Units, and for the two tenures (Owner & Renter). These formulas adequately approximate the errors for the previously separated characteristics.

**Replicate Weights.** Each year starting in 2009, a file of replicate weights for that year's data is provided at <http://www.census.gov/programs-surveys/ahs/data.html>; additionally, this website provides a detailed explanation on how to use the replicate weight file. This file is merged with the Public Use File to calculate the exact errors used to calculate confidence intervals. These replicate weights simulate the drawing of multiple samples from the population; these multiple simulations are used to estimate the variability observed in repeated sampling. Note that one year's replicate weight file is specific to that year, and the replicate weights are not used to calculate population estimates.

## REFERENCES

Schenker, Nathaniel and Gentleman, Jane F. (2001). On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals. *The American Statistician*, 55(3), 182-186

## REVISION HISTORY

The 2011 AHS-N was issued in September 2013. In 2016, the Census Bureau updated the 2011 weighting methodology by applying the same methodology used in 2013; however, the methodology for estimating variances did not change. The values in the tables of this Appendix have changed to reflect the revised sample and weighting methodology.

**Table D-2. Errors for Incomplete Data Bias<sup>2</sup>: 2011 AHS-N, using 1995 model  
(Numbers in thousands)**

When the AHS gives one of the following numbers	The chances are 90 percent that the complete value is inside the range of plus or minus
0	261
10	262
100	267
1,000	321
2,500	411
5,000	560
10,000	859
25,000	1,754
50,000	3,247
75,000	3,691
100,000	2,198
110,000	1,601
120,000	1,004
125,000	705
132,000	287

**Table D-3. Errors from Sampling: 2011 AHS-N  
(Numbers in thousands)**

When the AHS gives one of the following numbers	The chances are 90 percent that the actual value is inside the range of plus or minus
0	15*
25	21
100	43
1,000	134
2,500	211
5,000	295
10,000	410
25,000	607
50,000	754
75,000	774
100,000	678
110,000	598
120,000	477
125,000	390
132,000	189

Source: These errors were computed based on a formula in Table D-4 with high error.

\* This value was obtained by using the minimum error formula (see footnote 7).

**Table D-4. Formulas for 90 Percent Confidence Intervals:<sup>8</sup> 2011 AHS-N**

**Housing Unit Characteristics**

Characteristic	Total & Total Occupied	Owner Occupied	Renter Occupied
Total Units, Regions, Divisions, Units In Structure	$1.645 \times \sqrt{6.70 \times A - .000050 \times A^2}$	$1.645 \times \sqrt{6.93 \times A - .000075 \times A^2}$	$1.645 \times \sqrt{3.48 \times A - .000069 \times A^2}$
New Construction	$1.645 \times \sqrt{3.12 \times A - .000397 \times A^2}$	$1.645 \times \sqrt{3.14 \times A - .000652 \times A^2}$	$1.645 \times \sqrt{3.21 \times A + .000330 \times A^2}$
Mobile Homes	$1.645 \times \sqrt{5.95 \times A - .000458 \times A^2}$	$1.645 \times \sqrt{5.68 \times A - .000625 \times A^2}$	$1.645 \times \sqrt{4.34 \times A + .000046 \times A^2}$

**Household Characteristics**

Characteristic	Total & Total Occupied	Owner Occupied	Renter Occupied
Black	$1.645 \times \sqrt{2.89 \times A - .000151 \times A^2}$	$1.645 \times \sqrt{3.01 \times A - .000395 \times A^2}$	$1.645 \times \sqrt{2.59 \times A - .000268 \times A^2}$
Hispanic	$1.645 \times \sqrt{2.90 \times A - .000086 \times A^2}$	$1.645 \times \sqrt{2.90 \times A - .000263 \times A^2}$	$1.645 \times \sqrt{2.38 \times A - .000229 \times A^2}$
Elderly	$1.645 \times \sqrt{2.95 \times A - .000094 \times A^2}$	$1.645 \times \sqrt{3.13 \times A - .000131 \times A^2}$	$1.645 \times \sqrt{1.90 \times A - .000312 \times A^2}$
Below Poverty	$1.645 \times \sqrt{3.25 \times A - .000051 \times A^2}$	$1.645 \times \sqrt{3.53 \times A - .000156 \times A^2}$	$1.645 \times \sqrt{2.77 \times A - .000058 \times A^2}$

**Inside MSA Characteristics**

Characteristic	Total & Total Occupied	Owner Occupied	Renter Occupied
Central City	$1.645 \times \sqrt{4.96 \times A + .000184 \times A^2}$	$1.645 \times \sqrt{4.67 \times A + .000250 \times A^2}$	$1.645 \times \sqrt{3.10 \times A + .000098 \times A^2}$
Suburbs	$1.645 \times \sqrt{5.38 \times A + .000151 \times A^2}$	$1.645 \times \sqrt{5.02 \times A + .000158 \times A^2}$	$1.645 \times \sqrt{3.20 \times A + .000096 \times A^2}$
Outside MSA	$1.645 \times \sqrt{18.8 \times A + .00158 \times A^2}$	$1.645 \times \sqrt{15.3 \times A + .00139 \times A^2}$	$1.645 \times \sqrt{7.87 \times A + .00211 \times A^2}$

<sup>8</sup> The formula is based on 1.645 times the standard error from sampling. This formula gives “90 percent confidence interval” bounds. For 95 percent confidence intervals, multiply by 1.96 instead of 1.645; for 99% confidence, multiply by 2.576 instead of 1.645.

### Other Characteristics

Characteristic	Total & Total Occupied	Owner Occupied	Renter Occupied
Rural	$\frac{1.645}{\times \sqrt{14.6 \times A - .000128 \times A^2}}$	$\frac{1.645}{\times \sqrt{10.4 \times A - .000162 \times A^2}}$	$\frac{1.645}{\times \sqrt{5.17 \times A - .000053 \times A^2}}$
Urban	$\frac{1.645}{\times \sqrt{5.13 \times A - .000014 \times A^2}}$	$\frac{1.645}{\times \sqrt{4.31 \times A - .0000013 \times A^2}}$	$\frac{1.645}{\times \sqrt{3.00 \times A - .000031 \times A^2}}$

### Seasonal and Vacant Housing Units

Characteristic	Formula
Seasonal	$\frac{1.645}{\times \sqrt{11.3 \times A + .000437 \times A^2}}$
Total Year-round Vacant	$\frac{1.645}{\times \sqrt{5.33 \times A + .000335 \times A^2}}$
For Rent	$\frac{1.645}{\times \sqrt{2.27 \times A + .000093 \times A^2}}$
For Sale Only	$\frac{1.645}{\times \sqrt{2.07 \times A - .000391 \times A^2}}$
Rented or Sold	$\frac{1.645}{\times \sqrt{3.15 \times A + .000767 \times A^2}}$
Occasional Use / URE	$\frac{1.645}{\times \sqrt{6.71 \times A + .000735 \times A^2}}$
Other Vacant	$\frac{1.645}{\times \sqrt{3.29 \times A + .000052 \times A^2}}$

**Table D-5. Public Use File definitions to identify the various domains:**

<b>Domain</b>	<b>Definition</b>
Occupied	STATUS = '1'
Owner	TENURE = '1'
Renter	TENURE in ('2', '3')
Northeast	REGION = '1'
New England	DIV = '1'
Middle Atlantic	DIV = '2'
Midwest	REGION = '2'
East North Central	DIV = '3'
West North Central	DIV = '4'
South	REGION = '3'
South Atlantic	DIV = '5'
East South Central	DIV = '6'
West South Central	DIV = '7'
West	REGION = '4'
Mountain	DIV = '8'
Pacific	DIV = '9'
New Construction	NEWC = '1'
One unit, detached	NUNIT2 = '1'
One unit, attached	NUNIT2 = '2'
Multiunit	NUNIT2 = '3'
Number of units	NUNITS = 2, 3,...
Manufactured / Mobile Homes	NUNIT2 = '4'
Black alone	HHRACE = '02'
Hispanic	HHSPAN = '1'
Elderly	HHAGE >= 65
Below Poverty	POOR < 1000
Central City	METRO3='1'
Suburbs	METRO3 in ('2','3')
Outside MSA	METRO3 in ('4','5')
Rural	METRO3 in ('3','5')
Urban	METRO3 in ('1','2','4')
Seasonal	8 <= VACANCY <= 11
Total Year-round Vacant	1 <= VACANCY <= 7
For Rent	STATUS = '3' and 1 <= VACANCY <= 2
For Sale Only	STATUS = '3' and VACANCY = 3
Rented or Sold	STATUS = '3' and 4 <=VACANCY<= 5
Occasional Use / URE	(STATUS = '3' and VACANCY = 6) or (STATUS = '2' and 1 <=VACANCY<= 7)
Other Vacant	STATUS ='3' and VACANCY=7

In the following two examples in Tables D-7 and D-8, cost data from Table D-6 are used to calculate the 90 percent confidence interval for medians for a hypothetical entry in the publication tables (all numbers are in thousands):

**Table D-6. Hypothetical Publication Table Example**

Publication table example		Cumulative number of housing units
<b>Total Housing Units</b>	209	
Less than \$500	50	50
\$500 to \$599	45	95
\$600 to \$699	30	125
\$700 to \$799	20	145
\$800 or more	55	200
Not reported	9	
Median	\$627	

**Table D-7. How to Compute the Error From Sampling for a 90 Percent Confidence Interval for a Median**

Steps for Calculations	The formula	An example
How many total units is the median based on (in thousands, exclude “not reported” and “don’t know”)?	A	200
What is the estimated standard error of a 50% characteristic with a base equaling the total units?	$\sigma = \sqrt{\frac{6.70(0.5)(1 - 0.5)}{A}}$	$\sqrt{\frac{6.70(0.5)(1 - 0.5)}{200}} = 0.092$
What are the end points of the category the median is in?	X – Y	\$600 – 699
What is the width of this category (in dollars, rooms, or whatever the item measures)?	W	\$100
How many housing units are in this median category (in thousands)?	B	30
What is the estimated proportion of the total units falling in the category containing the sample median?	$P = \frac{B}{A}$	$\frac{30}{200} = 0.15$
Then the standard error from sampling for the median is approximately:	$se_{median} = \frac{\sigma \times W}{P}$	$\frac{0.092 \times \$100}{0.15} = \$61$
The 90 percent confidence interval for the median is:	$Median \pm 1.645 \times se_{median}$	$Median \pm \$100$

**Table D-8. Calculation of the 90 Percent Confidence Interval for Medians**

Item	Formula	Bottom limit example	Top limit example
How many total units is the median based on (in thousands, exclude “not reported”)?	A	200	
Half the total, for the median (in thousands)	A/2	100	
Error from sampling for 50 percent of the base of this median (first line)	$1.645 \times \sqrt{\frac{6.70(0.5)(1 - 0.5)}{A}}$ $= \frac{2.13}{\sqrt{A}}$	$\frac{2.13}{\sqrt{200}} = 0.150$	
Multiply this percentage by total units to give the error in housing units.	$\frac{2.13}{\sqrt{A}} \times A = 2.13\sqrt{A}$	$0.150 \times 200 = 30$	
Bottom of error range (second line minus fourth line, in thousands)	B <sub>bottom</sub>	70**	
Top of error range (second line plus fourth line, in thousands)	B <sub>top</sub>		130**
*Start adding up the housing units in this table, category by category, cumulatively from the beginning of the table, until you exceed the starred number above. What interval does the starred number fall in?		\$500 – 599	\$700 – 799
How many housing units are in all the categories before this one (in thousands)?	C	50	125
How many housing units are in this category (in thousands)?	D	45	20
What is the bottom limit of this category (in dollars, rooms, or whatever the item measures)?	E	\$500	\$700
What is the bottom limit of the next category (in dollars, rooms, etc.)?	F	\$600	\$800
Formula to calculate limits of confidence interval	$\frac{B - C}{D}(F - E) + E$	$\frac{70 - 50}{45}(100) + 500$	$\frac{130 - 125}{20}(100) + 700$
Limits of confidence interval (in dollars, rooms, etc.)		\$544	\$725

\*\*Starting with the starred step, this worksheet is equivalent to interpolation, for those who are familiar with this term.