

# Appendix D. 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.)
- Wrong answers (The U.S. Census Bureau does not adjust for wrong answers and does not estimate the size of the errors.)
- Sampling (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 the Census Bureau's sampling lists, the housing units in the survey do not represent all housing units in the country. The Census Bureau attempts to adjust for the deficiencies by raising the raw numbers from the survey proportionally so that the numbers published here match independent estimates of the total number of housing units. See Appendix B, Independent Total Housing Unit Adjustment.

In 2009 the Census Bureau attempted to reduce the under-coverage in two segments of the population by adding sample units selected from Census 2000 (i.e., manufactured/mobile homes built between 1980 and 2000 and special living units). The approximate housing unit under-coverage rates for the 2009 metropolitan areas range from 1.8 percent to 7.4 percent. Table D-1 lists units that have known coverage deficiencies.

**Missing data.** Some people refuse the interview or some of the questions or do not know the answers. When the entire interview is missing, other similar interviews represent the missing ones. See Appendix B, Noninterview

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 1990 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 non-permit-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	Nonresidential units at the time of the 1990 Census that converted to residential units were missed.
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.

Adjustment. For most missing answers, an answer from a similar household is copied.<sup>1</sup> The Census Bureau does not know how close the imputed values are to the actual values. For other items, "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.

Incompleteness can cause large errors since, when even 10 percent of housing units do not answer a particular question, they represent about 13 million housing units that have to be estimated on little or no basis (there are about 130 million housing units in the United States). The survey estimates adjust for them by assuming that they are like some group of housing units that did give data. This assumption is never exactly true, although it is usually better than ignoring the housing units with the missing data. Thus, 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

<sup>1</sup> Hot deck allocation is used: an answer is copied from the most recently processed similar household before the household with the missing item.

items. Again, readers should be wary of items with highly incomplete data.<sup>2</sup>

Rates of completeness were not computed for 2009. Table 2 in Appendix D of *American Housing Survey for the United States in 1995* gives the completeness rates for 1995. Due to the change in data collection methodology, the rates for 2009 may be higher or lower than in the past. However, the items that were most incomplete in 1995 are probably still the most incomplete for 2009.

**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 computerized questionnaire in 1997
- Lack of Spanish questionnaire, prior to 2009

## WRONG ANSWERS

Wrong answers happen because people misunderstand questions, cannot recall the correct answer, or do not want to give the right answer. See *American Housing Survey for the United States: 2005* for more discussion on this topic.

## 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.

**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 error from sampling. As with the other types of errors, readers should be wary of numbers with large errors from sampling.

<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 United States 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 from HUD USER at <www.huduser.org>.

Table D-3 gives a list of errors for a range of numbers for the 2009 AHS-MS metropolitan areas. For numbers not found in this table, interpolate between the numbers in the table or use the appropriate formula from Table D-4 for the 2009 AHS-MS metropolitan areas. In each formula, “A” is a number (a count of units in thousands) from the AHS. Remember in any case that the total error is larger than sampling error.

For example, suppose there are 320,000 owner-occupied housing units in the Seattle, WA MSA (that is, A = 320). The error from sampling for a 90 percent confidence interval for those 320,000 owner-occupied housing units is:

$$1.645x\sqrt{(.555x320) - (.000395x320x320)} = 19.3$$

The 90 percent confidence interval can then be formed by adding and subtracting this error to the survey estimate of 320 (that is,  $320 \pm 19.3$ ). Statements such as “the actual value is in the range 320 plus or minus 19.3 (that is, 300.7 to 339.3)” 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 320 means 320,000. The formulas are designed to use numbers directly from the publication; do not add zeros. The results are also in thousands, so 19.3 means 19,300.

**Percents.** Any subgroup can be shown as a percent of a larger group. For AHS-MS metropolitan areas, use the appropriate formula in Table D-5. In each formula, p is the percent and A is the denominator, or base of the percent in thousands.

For example, the error from sampling for a 90 percent confidence interval for 40 percent of 320 (meaning 320,000) is:

$$1.645x\sqrt{\frac{0.555x40x(100 - 40)}{320}} = 3.4$$

Statements such as “the actual percent is in the range of 36.6 percent to 43.4 percent” are right 90 percent of the time.

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>4</sup>

<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 a 95 percent confidence interval, multiply by 1.96 instead of 1.645. For a 99 percent confidence interval, multiply by 2.576 instead of 1.645.

<sup>4</sup> The error from sampling for a 90 percent confidence interval for ratio C/D is:

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

where 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.

Table D-2.  
**Errors for Incomplete Data Bias: 2009 AHS-MSAs**

(Numbers in thousands)

Size of estimate	Chicago, IL PMSA	Detroit, MI PMSA	New Orleans, LA MSA	New York, NY PMSA	Northern NJ, NJ PMSA	Philadelphia, PA PMSA	Seattle, WA MSA
0	6.7	3.9	0.9	9.9	5.3	4.2	2.8
1	9.0	5.3	1.1	13.4	7.3	5.6	3.7
5	18.1	10.7	2.1	27.6	15.1	11.5	7.3
10	29.6	17.6	3.4	45.3	25.0	18.8	11.9
15	41.0	24.5	4.7	63.1	34.8	26.1	16.5
50	121.0	72.6	13.6	187.1	103.6	77.4	48.4
100	235.4	141.2	26.4	364.4	201.9	150.6	94.1
250	578.3	347.3	48.7	896.1	496.7	370.2	231.0
500	1,150.0	690.7	(N/A)	1,782.4	988.2	736.2	459.3
750	1,721.6	1,034.1	(N/A)	2,668.6	1,479.6	1,102.2	600.6
1,000	2,293.3	1,328.7	(N/A)	3,554.9	1,971.1	1,468.2	372.3
1,250	,2864.9	985.3	(N/A)	4,441.1	2,462.5	1,277.2	144.1
1,500	3,436.5	641.9	(N/A)	5,327.4	2,328.2	911.2	(N/A)
1,750	3,771.6	298.5	(N/A)	6,213.6	1,836.8	545.2	(N/A)
2,000	3,199.9	(N/A)	(N/A)	7,099.8	1,345.3	179.2	(N/A)
2,250	2,628.3	(N/A)	(N/A)	7,986.1	853.9	(N/A)	(N/A)
2,500	2,056.6	(N/A)	(N/A)	8,872.3	362.5	(N/A)	(N/A)
2,750	1,485.0	(N/A)	(N/A)	8,037.2	(N/A)	(N/A)	(N/A)
3,000	913.4	(N/A)	(N/A)	7,151.0	(N/A)	(N/A)	(N/A)
3,250	341.7	(N/A)	(N/A)	6,264.7	(N/A)	(N/A)	(N/A)
3,500	(N/A)	(N/A)	(N/A)	5,378.5	(N/A)	(N/A)	(N/A)
3,750	(N/A)	(N/A)	(N/A)	4,492.3	(N/A)	(N/A)	(N/A)
4,000	(N/A)	(N/A)	(N/A)	3,606.0	(N/A)	(N/A)	(N/A)
4,250	(N/A)	(N/A)	(N/A)	2,719.8	(N/A)	(N/A)	(N/A)
4,500	(N/A)	(N/A)	(N/A)	1,833.5	(N/A)	(N/A)	(N/A)
4,750	(N/A)	(N/A)	(N/A)	947.3	(N/A)	(N/A)	(N/A)
5,000	(N/A)	(N/A)	(N/A)	61.0	(N/A)	(N/A)	(N/A)
5,015	(N/A)	(N/A)	(N/A)	7.9	(N/A)	(N/A)	(N/A)

(N/A) No error estimates are provided because the estimate is larger than the estimated total number of housing units in the MSA.

**Medians.** The steps in Table D-6 calculate the error for sampling for a 90 percent confidence interval for medians. This is an approximation to the error.

For small bases the error cannot be estimated reliably. To estimate a median's sampling error more accurately, find the sampling error on 50 percent as described in Table D-7 and compute the 90 percent confidence interval.

**Differences.** Two numbers from the AHS, like 34 and 40 or 40 percent and 45 percent have a "statistically significant difference" if their ranges of error from sampling for a 90 percent confidence interval do not overlap.<sup>5</sup>

<sup>5</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

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

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 number respectively.

**Formulas for error from sampling.** The letter "A" in the formulas in Tables D-4, D-5, D-6, and D-7 represents a number (a count of units in thousands) from AHS (see "Sampling Errors" section for an example of how "A" is used).

For a 90 percent confidence interval on zero for the 2009 AHS-MS refer to Table D-3 where the size of the estimate is zero. If a formula gives an error smaller than the error for zero, use the error for zero.

Table D-3.

**Errors From Sampling to Compute a 90 Percent Confidence Interval: 2009 AHS-MS**

(Numbers in thousands)

Size of estimate	Chicago, IL PMSA	Detroit, MI PMSA	New Orleans, LA MSA	New York, NY PMSA	Northern NJ, NJ PMSA	Philadelphia, PA PMSA	Seattle, WA MSA
0 .....	—	—	—	—	—	—	—
1 .....	1.9	1.5	0.6	2.4	1.8	1.6	1.2
5 .....	4.3	3.4	1.4	5.4	4.0	3.5	2.7
10 .....	6.1	4.7	2.0	7.6	5.7	4.9	3.9
15 .....	7.5	5.8	2.5	9.3	6.9	6.0	4.7
50 .....	13.6	10.5	4.3	17.0	12.6	10.8	8.5
100 .....	19.1	14.6	5.7	23.9	17.6	15.1	11.8
250 .....	29.5	22.2	6.7	37.2	27.1	23.0	17.6
500 .....	40.0	29.0	(N/A)	51.2	36.3	30.3	22.0
750 .....	46.9	32.4	(N/A)	61.0	41.8	34.2	22.9
1,000 .....	51.5	33.3	(N/A)	68.3	45.0	35.7	20.8
1,250 .....	54.5	32.1	(N/A)	74.0	46.4	35.1	14.4
1,500 .....	56.1	28.3	(N/A)	78.3	46.2	32.5	(N/A)
1,750 .....	56.5	20.8	(N/A)	81.5	44.3	27.1	(N/A)
2,000 .....	55.6	(N/A)	(N/A)	83.8	40.5	16.5	(N/A)
2,250 .....	53.5	(N/A)	(N/A)	85.1	34.1	(N/A)	(N/A)
2,500 .....	49.9	(N/A)	(N/A)	85.6	23.3	(N/A)	(N/A)
2,750 .....	44.4	(N/A)	(N/A)	85.2	(N/A)	(N/A)	(N/A)
3,000 .....	36.4	(N/A)	(N/A)	83.9	(N/A)	(N/A)	(N/A)
3,250 .....	23.1	(N/A)	(N/A)	81.8	(N/A)	(N/A)	(N/A)
3,500 .....	(N/A)	(N/A)	(N/A)	78.7	(N/A)	(N/A)	(N/A)
3,750 .....	(N/A)	(N/A)	(N/A)	74.5	(N/A)	(N/A)	(N/A)
4,000 .....	(N/A)	(N/A)	(N/A)	68.9	(N/A)	(N/A)	(N/A)
4,250 .....	(N/A)	(N/A)	(N/A)	61.8	(N/A)	(N/A)	(N/A)
4,500 .....	(N/A)	(N/A)	(N/A)	52.3	(N/A)	(N/A)	(N/A)
4,750 .....	(N/A)	(N/A)	(N/A)	38.8	(N/A)	(N/A)	(N/A)
5,000 .....	(N/A)	(N/A)	(N/A)	11.6	(N/A)	(N/A)	(N/A)
5,015 .....	(N/A)	(N/A)	(N/A)	7.0	(N/A)	(N/A)	(N/A)

— Represents or rounds to zero.

(N/A) No error estimates are provided because the estimate is larger than the estimated total number of housing units in the MSA.

Table D-4.

**Formulas for 90 Percent Confidence Intervals: 2009 AHS-MS**

MSA and estimate type <sup>1</sup>	The formula is: <sup>2</sup>
<b>Chicago, IL PMSA</b>	
All estimates .....	$1.645 \times \sqrt{(1.390 \times A) - (0.000410 \times A^2)}$
<b>Detroit, MI PMSA</b>	
All estimates .....	$1.645 \times \sqrt{(0.835 \times A) - (0.000425 \times A^2)}$
<b>New Orleans-Metairie-Kenner, LA MSA</b>	
Manufactured/mobile home estimates .....	$1.645 \times \sqrt{(0.235 \times A) - (0.013272 \times A^2)}$
All other estimates .....	$1.645 \times \sqrt{(0.155 \times A) - (0.000440 \times A^2)}$
<b>New York-Nassau-Suffolk-Orange, NY PMSAs</b>	
All estimates .....	$1.645 \times \sqrt{(2.155 \times A) - (0.000430 \times A^2)}$
<b>Northern New Jersey, NJ PMSAs</b>	
All estimates .....	$1.645 \times \sqrt{(1.195 \times A) - (0.000445 \times A^2)}$
<b>Philadelphia, PA-NJ PMSA</b>	
All estimates .....	$1.645 \times \sqrt{(0.890 \times A) - (0.000420 \times A^2)}$
<b>Seattle-Tacoma-Bellevue, WA MSA</b>	
Manufactured/mobile home estimates .....	$1.645 \times \sqrt{(0.825 \times A) - (0.155690 \times A^2)}$
All other estimates .....	$1.645 \times \sqrt{(0.555 \times A) - (0.000395 \times A^2)}$

<sup>1</sup>Some items (for example, characteristic of total housing units) may involve housing units from both the manufactured/mobile home and nonmanufactured/mobile home universe. The formulas for all other estimates should be used for these items. The formulas for manufactured/mobile home estimates should be used for items that only involve housing units from the manufactured/mobile home universe.

<sup>2</sup>The formulas in the text are based on 1.645 times the errors 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.

Table D-5.

**Formulas for 90 Percent Confidence Intervals Associated With a Percentage**

MSA and estimate type <sup>1</sup>	The formula is: <sup>2</sup>
<b>Chicago, IL PMSA</b>	
All estimates .....	$1.645 \times \sqrt{((1.390 \times p) - (100 - p)) / A}$
<b>Detroit, MI PMSA</b>	
All estimates .....	$1.645 \times \sqrt{((0.835 \times p) - (100 - p)) / A}$
<b>New Orleans-Metairie-Kenner, LA MSA</b>	
Manufactured/mobile home estimates .....	$1.645 \times \sqrt{((0.235 \times p) - (100 - p)) / A}$
All other estimates .....	$1.645 \times \sqrt{((0.155 \times p) - (100 - p)) / A}$
<b>New York-Nassau-Suffolk-Orange, NY PMSAs</b>	
All estimates .....	$1.645 \times \sqrt{((2.155 \times p) - (100 - p)) / A}$
<b>Northern New Jersey, NJ PMSAs</b>	
All estimates .....	$1.645 \times \sqrt{((1.195 \times p) - (100 - p)) / A}$
<b>Philadelphia, PA-NJ PMSA</b>	
All estimates .....	$1.645 \times \sqrt{((0.890 \times p) - (100 - p)) / A}$
<b>Seattle-Tacoma-Bellevue, WA MSA</b>	
Manufactured/mobile home estimates .....	$1.645 \times \sqrt{((0.825 \times p) - (100 - p)) / A}$
All other estimates .....	$1.645 \times \sqrt{((0.555 \times p) - (100 - p)) / A}$

<sup>1</sup>Some items (for example, characteristic of total housing units) may involve housing units from both the manufactured/mobile home and nonmanufactured/mobile home universe. The formulas for all other estimates should be used for these items. The formulas for manufactured/mobile home estimates should be used for items that only involve housing units from the manufactured/mobile home universe.

<sup>2</sup>The formulas in the text are based on 1.645 times the errors 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.

Table D-6.

**How to Compute the Error From Sampling for a 90 Percent Confidence Interval for a Median**

Steps for calculations	The formula	An example	Your data
How many total units is the median based on (in thousands, exclude "not reported" and "do not know")? .....	A	200	—
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	—
Then the error from sampling for the median is approximately: <sup>1</sup> .....	$\frac{K \times W \times \sqrt{A}}{B}$	$\frac{.747 \times 100 \times \sqrt{200}}{30.0} = \$35$	—
The 90 percent confidence interval for the median is: .....	median $\pm \frac{K \times W \times \sqrt{A}}{B}$	median $\pm$ \$35	—

<sup>1</sup>To obtain an appropriate value for K, multiply the numerator of the formula for computing the error from sampling for 50 percent by a factor of .01. Refer to Table D-5 for the appropriate formula for AHS-MS metropolitan areas. For example, for estimates consisting of only manufactured/mobile homes in the Seattle-Tacoma-Bellevue, WA MSA,  $1.645 \times \sqrt{.825 \times 50 \times (100 - 50) / A} = 74.7 / \sqrt{A}$  and for all other estimates in Seattle-Tacoma-Bellevue, WA MSA, K = .613

Table D-7.

**Calculation of the 90 Percent Confidence Interval for Medians**

(In the following example, cost data are used to calculate the 90 percent confidence interval for medians (all numbers are in thousands))

		Cumulative number of housing units
Total housing units	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	\$619	

Item	Formula	Bottom limit		Top limit	
		Example	Your data	Example	Your data
How many total units is the median based on (in thousands, exclude "not reported" and "no cash rent")? . . . . .	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 . . . . .	$74.7/\sqrt{A}$	5.3	—		
Multiply this percentage error by .01 to turn it into a fraction and by total units to give the error in housing units . . . . .	$0.75\sqrt{A}$	10.6	—		
Bottom of error range (second line minus fourth line, in thousands) . . . . .	B <sub>bottom</sub>	89.4	—		
Top of error range (second line plus fourth line, in thousands) . . . . .	B <sub>top</sub>			110.6	—
*Start adding up the housing units in the 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	—	\$600–699	—
How many housing units are in all the categories before this one (in thousands)? . . . . .	C	50	—	95	—
How many housing units are in this category (in thousands)? . . . . .	D	45	—	30	—
What is the bottom limit of this category (in dollars, rooms, or whatever the item measures)? . . . . .	E	\$500	—	\$600	—
What is the bottom limit of the next category (in dollars, rooms, etc)? . . . . .	F	\$600	—	\$700	—
Formula to calculate limits of confidence interval . . . . .	$\frac{(B-C)}{D}(F-E)+E$	$\frac{(89.4-50)}{45}(100)+500$	—	$\frac{(110.6-95)}{30}(100)+600$	—
Limits of confidence interval (in dollars, rooms, etc.) . . . . .	—	\$588	—	\$652	—

¹Starting with the starred step, this worksheet is equivalent to interpolation, for those who are familiar with this term.

¹ Statistical note: This example formula is based on the error from sampling for 50 percent using the appropriate formula  $1.645 \times \sqrt{.825 \times 50 \times (100 - 50)/A} = 74.7 / \sqrt{A}$  for medians involving estimates of only manufactured/mobile homes in Seattle-Tacoma-Bellevue, WA metropolitan area. For medians involving all other estimates in the Seattle-Tacoma-Bellevue, WA metropolitan area, use:

$$1.645 \times \sqrt{((0.555 \times p) - (100 - p)) / A}$$

Refer to Table D-5 for the appropriate formula.