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:

- 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 Sample (AHS-N), the changes in weighting in 1981 and 1991 (see Appendix C) corrected some of the error due to incomplete data. That one correction averaged 2.5 percent in 1991. Worse errors from incomplete data and from wrong answers apply to some items, as discussed below.

Additional information on the quality of AHS data can be obtained from the U.S. Census Bureau's *American Housing Survey: A Quality Profile,* Series H121/95-1.

INCOMPLETE DATA

Coverage errors. Because of deficiencies with the 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, "New construction adjustment" and "Demographic adjustment." The independent estimates changed around 2.5 percent in both 1981 and 1991 (after the 1980 and 1990 censuses, respectively), which implies that some error existed in the independent estimation procedures in the years just before the censuses. By comparison, the independent estimates changed by 0.8 percent in 2003 (after the 2000 census).

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). Overall, housing unit undercoverage is about 3.2 percent for the 2007 AHS-N.

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 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	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.

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 adjustment"). For most missing answers, an answer from a similar household is copied.[1] 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

[1]Hot deck allocation is used: an answer is copied from the most recently processed similar household before the household with the missing item.

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 are missed by 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 U. S.). The survey estimates them by assuming that they are like some group of housing units that did give data, an assumption that 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 items. Again, readers should be wary of items with highly incomplete data.[2]

Rates of completeness were not computed for 2007. Table 2 of the *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 2007 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 2007.

Effect on income. The nonsampling errors interact particularly badly for income. Income questions are inconsistently answered, incompletely answered, and the totals fall short of totals known from the National Income Accounts, especially for the elderly.[3]

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
- Lack of Spanish questionnaire

WRONG ANSWERS

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

SAMPLING ERRORS

Sampling errors definition. Errors from sampling reflect 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.

Table D-3 gives a convenient list of errors for a range of numbers for the 2007 AHS-N. The error from sampling cannot be known exactly. For numbers not in Table D-3, the error from sampling is approximated using the following formula for constructing a 90-percent confidence interval:

$$1.645 \times \sqrt{(4.23 \times A) - (.000033 \times A^2)}$$

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

For example if A is 200:

$$1.645 \times \sqrt{(4.23 \times 200) - (.000033 \times 200^2)} = 48$$

The 90-percent confidence interval can then be formed by adding and subtracting this error to the survey estimate of 200 (that is, 200 plus or minus 48). Statements such as "the actual value is in the range 200 plus or minus 48 (152 to 248)," are right 90 percent of the time and wrong 10 percent of the time.[4]

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 add zeros. The result is also in thousands, so 48 means 48,000.

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:

$$1.645 \times \sqrt{(4.23 \times p \times (100 - p))/A}$$

[2]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 x (.0012 x U + .0363 x (lesser of A 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 upon request from HUD USER by e-mailing at <helpdesk@huduser.org> or calling 1-800-245-2691.

[3]Data are in the *Codebook for the American Housing Survey Volume 1,* available from <www.huduser.org/datasets/ahs/ahsprev.html>.

[4]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 1.960 instead of 1.645; for 99-percent confidence, multiply by 2.576 instead of 1.645.

where p is the percent; 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 200 (meaning 200,000) is:

$$1.645 \times \sqrt{(4.23 \times 40 \times 60)/200} = 11.7$$

Statements such as "the actual percent is in the range 28.3 percent to 51.7 percent" are right 90 percent of the time.

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.[5]

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.[6]

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, find the sampling error on 50 percent as described in Table D-6 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.[7]

Formulas for error from sampling. The letter "A" in the formulas in Tables D-4, D-5, and D-6 represents a number (a count of 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 ±9 (meaning ±9 thousand).[8] If a formula gives an error smaller than 9, use 9.

For AHS-N, if an item falls into two different categories in Table D-4, use the formula that gives the largest error. For example, for Hispanics' income in the South, use the formulas for the South (since there is no specific formula for income and errors for the South will be bigger than those for Hispanics). For the following neighborhood characteristics, use the neighborhood formulas:

- Opinion of neighborhood
- Street noise or traffic
- Neighborhood crime
- Odors
- Other bothersome neighborhood conditions
- Elementary school
- Academic comparison to other area elementary schools
- Public transportation
- Neighborhood shopping
- Police protection
- Parking lots
- Description of area (except open space, park, farm, or ranch) within 300 feet
- Age of other residential buildings within 300 feet
- Other buildings vandalized or with interior exposed within 300 feet
- Bars on windows of buildings within 300 feet
- Conditions of streets within 300 feet
- Trash, litter, or junk on streets or any properties within 300 feet
- Manufactured/mobile homes in group

For the following special characteristics, use the special characteristics formulas. The following items are defined as special characteristics:

- Cooperatives or condominiums
- No complete bathroom
- Less than 1,500 square feet of detached one-family or manufactured/mobile homes
- Well serving 1 to 5 units

[5]This formula is actually $1.645 \times \sqrt{(p(100-p)/n)}$, since 4.23/A adjusts the data to the effective sample size.

[6]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.

[7]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.

[8]This minimum error formula is based on the binomial 90-percent confidence interval on zero $U \times (1-.1^{4.23/U}) = 9$ (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. "Sampling Errors for Small Groups," order number HUD-8509, is available upon request from HUD USER by e-mailing <helpdesk@huduser.org> or calling 1-800-245-2691.

- Manufactured/mobile homes in a group
- Area within 300 feet includes open space, park, farm, or ranch
- Septic tank, cesspool, chemical toilet
- Five or more acres in lot size
- No bedroom
- Lacking complete kitchen facilities
- Lacking some plumbing facilities
- No flush toilet
- Major street repairs needed

Table D-2. **Errors for Incomplete Data Bias: 2007 AHS-N**

[Numbers in thousands]

When the AHS gives one of the following numbers–	The chances are 90 percent that the complete value[1] is inside the range of plus or minus
0	246
10	246
100	252
1,000	305
2,500	395
5,000	544
10,000	843
25,000	1,738
50,000	3,231
75,000	3,195
100,000	1,703
110,000	1,105
120,000	508
125,000	210
128,000	31

[1]"Complete value" means the value derived if there were no missing data.

Table D-3. **Errors From Sampling: 2007 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	9
10	11
100	34
1,000	107
2,500	168
5,000	235
10,000	325
25,000	480
50,000	558
75,000	597
100,000	502
110,000	423
120,000	296
125,000	188
128,000	46

Source: These errors were computed based on a formula with high sampling error in Table D-6. This table represents a conservative example.

Table D-4. **Formulas for 90-Percent Confidence Intervals:[1] 2007 AHS-N**

Characteristics	General formulas—	Other formulas	
	All characteristics except those listed under other formulas	Fuels, heating/cooling equipment, and neighborhood characteristics	Special characteristics
Total units, Midwest, West, Elderly, Black, new construction, manufactured/mobile homes, vacants	$1.645 \times \sqrt{3.47 \times A - 0.000027 \times A^2}$	$1.645 \times \sqrt{3.47 \times A - 0.000027 \times A^2}$	$1.645 \times \sqrt{4.23 \times A + 0.000255 \times A^2}$
Northeast, central city, Hispanic, urban, suburbs	$1.645 \times \sqrt{2.76 \times A - 0.000022 \times A^2}$	$1.645 \times \sqrt{2.76 \times A - 0.000022 \times A^2}$	$1.645 \times \sqrt{4.23 \times A + 0.000255 \times A^2}$
Rural, South, outside (P)MSAs	$1.645 \times \sqrt{3.32 \times A - 0.000026 \times A^2}$	$1.645 \times \sqrt{4.23 \times A - 0.000033 \times A^2}$	$1.645 \times \sqrt{4.23 \times A + 0.000255 \times A^2}$
Special living sample units	$1.645 \times \sqrt{1.58 \times A - 0.000012 \times A^2}$	$1.645 \times \sqrt{1.58 \times A - 0.000012 \times A^2}$	$1.645 \times \sqrt{3.85 \times A + 0.000255 \times A^2}$

[1]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.

Table D-5. **How to Compute the Error From Sampling for a 90-Percent Confidence Interval for a Median[1]**

Steps for calculations	The formula	An example	Your data
How many total units is the median based on (in thousands, exclude "not reported" and "don't know")?	A	200	___
What are the endpoints 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:[2]	$\frac{K \times W \times \sqrt{A}}{B}$	$\frac{1.69 \times 100 \times \sqrt{200}}{30.0} = \$80$	___
The 90-percent confidence interval for the median is:	median $\pm \frac{K \times W \times \sqrt{A}}{B}$	median ± $80	___

[1]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.

[2]Note: 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.

Table D-6. **Calculation of the 90-Percent Confidence Interval for Medians**[1]

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	$627	

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)[2]	$1.69/\sqrt{A}$	12			
Multiply this percentage error by .01 to turn it into a fraction and by total units to give the error in housing units	$1.69\sqrt{A}$	23.9			
Bottom of error range (second line minus fourth line, in thousands)	B_{bottom}	*76.1			
Top of error range (second line plus fourth line, in thousands)	B_{top}			*123.9	
* 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	$\frac{(B-C)}{D}(F-E)+E$	$\frac{(76.1-50)}{45}(100)+500$		$\frac{(123.9-95)}{30}(100)+600$	
Limits of confidence interval (in dollars, rooms, etc.)		$558		$696	

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

[1]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.

[2]Statistical note: This formula is based on the error from sampling for 50 percent (using the appropriate formula, $1.645 \times \sqrt{(4.23 \times 50) \times (100-50)/A} = 169/\sqrt{A}$).