2015 AHS Redesign Paper 2

Sample Sizes Determination and Decisions for the 2015 American Housing Survey and Beyond

Shawn Bucholtz
U.S. Department of Housing and Urban Development

Stephen Ash
U.S. Census Bureau

December 2015

Sample Sizes Determination and Decisions for the 2015 American Housing Survey and Beyond

Purpose

The purpose of the paper is to discuss the process that determined the national and metropolitan area sample sizes for the 2015 American Housing Survey (AHS) and beyond. The 2015 AHS sample design includes a national sample integrated with metropolitan area samples from the 15 largest metropolitan areas, by population, as of 2013. Other metropolitan area samples will be independent samples. The integration of the national and metropolitan area samples is discussed in the whitepaper "Metropolitan Area Selection Strategies and Decisions for the 2015 American Housing Survey and Beyond."

1. AHS Sample Size History

1.1 National Longitudinal Sample Size Precision Statement

The statistical precision statement for the national longitudinal sample (1985–2013), which has been included in the U.S. Department of Housing and Urban Development's (HUD's) Paperwork Reduction Act, or PRA, submissions to the Office of Management and Budget, is: "A two-year change of 10 percent in the median monthly costs for 5-percent subgroups will have a standard error of 5 percent." Although not stated explicitly, the population domain for this statement is the whole United States. This statement does not specify any particular 5-percent subgroups, however. The subgroups are meant to be general. They could be based on political geography (city), statistical geography (Census Urban Area), or demographic or socioeconomic characteristics (for example, households with a disabled person).

The precise origin of the AHS's statistical precision statement for the national longitudinal sample is unknown. It certainly makes sense from the perspective of HUD, however, that a statistical precision statement for the national longitudinal sample would focus on detecting large changes in housing costs for groups of interest. Moreover, this statement of precision assumes a repeated survey, like the AHS, although not necessarily the repeated sample that exists in the AHS.

1.2 National Longitudinal Sample Sizes, 1985–2013

The national longitudinal sample size has fluctuated somewhat since its initial use in 1985. Between 1985 and 1991, the sample size was approximately 50,000. Between 1993 and 2009, the sample size averaged about 55,000. Due to increased budget, the sample size was expanded for 2011 and 2013. To expand the sample for 2011, three additions were made. First, roughly 5,200 cases that were reduced from the 2007 sample were reinstated. Second, approximately 3,100 units that were selected during the 2000 redesign, but were never interviewed, were introduced. Finally, an oversample of approximately 5,300 HUD-subsidized housing units was introduced.

Most of the years between 1985 and 2013 included supplemental sample for areas of special interest. The neighborhood supplemental samples conducted in 1985, 1989, and 1993 were not weighted, so they did not contribute to the summary statistics. The same was true for the supplemental rural samples in 1987 and 1991. In contrast, the "Big 6" metropolitan areas samples were weighted in accordance with the rest of the national longitudinal sample (table 1).

1.3 AHS Metropolitan Area Sample Sizes, 1985–2013

Throughout the history of the AHS, metropolitan area sample sizes have varied greatly. Before the redesign of the survey in 1983, sample sizes were approximately 15,000 housing units for the largest metropolitan areas and 5,000 for the smaller metropolitan areas. After the 1983 redesign, the sample size was reduced to 8,250 housing units for the largest metropolitan areas and 4,250 for smaller metropolitan areas. A further reduction to 3,300 was implemented in the mid-1980s.

In 1995, a standard of 4,500 was adopted for all metropolitan areas, although this standard was not implemented consistently between 1995 and 2009. For example, facing budget restrictions, HUD reduced the sample sizes for the independent metropolitan areas to approximately 2,600 for 2007 and 2009. The sample

Table 1. Historical Sample Sizes of AHS

Year	National Longitudinal Sample Size	Supplemental National Sample Size	Name of Supplements
1985	49,000	12,000	Neighborhood sample
1987	47,400	6,100	Rural sample
1989	47,500	9,000	Neighborhood sample
1991	51,400	6,100	Rural sample
1993	53,000	6,000	Neighborhood sample
1995	53,500	6,000	Big 6 metropolitan areas*
1997	53,100		
1999	54,130	6,570	Big 6 metropolitan areas*
2001	55,700		
2003	56,570	6,430	Big 6 metropolitan areas*
2005	59,450		
2007	55,000		
2009	55,700	6,300	Big 5 metropolitan areas**
2011***	69,700	2,700	Los Angeles
2013	68,850	15,550	Big 5 metropolitan areas**

AHS = American Housing Survey.

* Includes Chicago, Detroit, Los Angeles, New York, Northern New Jersey, and Philadelphia.

** Includes Chicago, Detroit, New York, Northern New Jersey, and Philadelphia.

*** For 2011, 28 additional metropolitan samples totaling 114,000 cases were integrated into the national sample. They are left out of this table to facilitate comparison with other years, however.

sizes for the Big 6 metropolitan areas, which were integrated into the national sample starting in 1995, averaged around 2,500. Between 2011 and 2013, the AHS included 44 metropolitan area surveys, each with a sample of approximately 4,500 housing units.

Unfortunately, it is unclear how the statistical precision standard of 4,500 housing units was determined. This sample size is certainly not large enough to achieve the statistical precision standard established for the national sample.

2. Considerations for AHS Sample Size, 2015 and Beyond

2.1 Considerations for the National Sample Size

In a discussion of AHS sample size for 2015 and beyond, at least six key points need to be recognized:

1. The AHS is a repeated survey because HUD's goal is to measure changes in the housing stock over time. A statistical precision statement should reflect HUD's goal of measuring changes in the housing stock.
2. The AHS is a longitudinal survey. If a key indicator requires responses in two or more time periods (for example, tenure switching), the precision of the estimates of that indicator will be impacted by longitudinal nonresponse or attrition.
3. The AHS is a demographic survey, as opposed to a demographic experiment. As such, the goal is to pick a sample size sufficient to estimate a characteristic (indicator) with a given level of precision, as opposed to testing an effect with a given power. For instance, the goal may be to estimate median housing cost with a margin of error of 5 percent.
4. The AHS covers several topics related to housing. For the purposes of determining sample size, one particular indicator or subgroup may not necessarily take precedence over another.
5. Unlike the American Community Survey (ACS), the AHS will never have a sample size large enough to produce reliable estimates for small areas such as census tracts. Furthermore, the AHS does not attempt to have sample in every county or city.
6. Historically, the domain for the national AHS sample has been the U.S. housing stock. Establishing additional domains, such as census divisions, requires multiplying the desired sample size by the number of domains of interest.

2.2 Considerations for the Metropolitan Area Sample Size

Each of the six points for consideration for the national sample size is also true of metropolitan area sample size. In a discussion of the AHS metropolitan area sample sizes for 2015 and beyond, a few additional points need to be considered:

1. The mathematics of probability are such that, for a given desired precision, the necessary sample size to achieve that precision will be the same if your population is 320,000 or 320,000,000. For the AHS, it is not desirable to apply the same desired precision standard to the national sample and the metropolitan area samples. Doing so would require metropolitan area sample sizes well beyond any conceivable budget for the AHS.
2. Given the history of the AHS's fluctuating budgets, it may not be prudent to develop a statistical precision statement that assumes a repeated sample every 4 or 6 years.

3. Options for Determining an Alternative AHS Sample Size Precision Statement

With these key points in mind, a few different methods can develop the sample size precision statement.

1. The sample size is based on detecting a change in an estimate of a single key indicator, for a desired level of precision.

   Example: A sample size sufficient to detect a change of 10 percent in median housing costs with a margin of error of 5 percent.

2. The sample size is based on detecting a change in an estimate of a single key indicator for specific subgroups, for a desired level of precision.

   Example: A sample size sufficient to detect a change of 10 percent in median housing costs for low-income renters, with a margin of error of 5 percent.

3. The sample size is based on detecting a change in an estimate of a single key indicator for general subgroups, for a desired level of precision.

   Example: A sample size sufficient to detect a change of 10 percent in median housing costs for 5-percent subgroups, with a margin of error of 5 percent.

4. The sample size is based on detecting a change in an estimate of proportions of general subgroups, for a desired level of precision.

   Example: A sample size sufficient to detect a change of 1 percent in 5-percent subgroups, with a margin of error of 0.5 percent.

5. The sample size is based on detecting a change in an estimate of a single key indicator where the indicator is a longitudinal measure, for a desired level of precision.

   Example: A sample size sufficient to detect a change of 1 percent in the number of housing units switching tenure, with a margin of error of 0.5 percent.

4. Decision on 2015 Sample Sizes

Any decision about survey sample size must be made within the context of the available budget. Although budgets have fluctuated during the past 30 years, HUD has generally secured enough funding to conduct the national sample and has conducted metropolitan area samples based on the remaining budget. The expectation is that this pattern will persist. As such, the subset of options for developing national and metropolitan area precision statements is constrained by expected survey budget. HUD evaluated options that were considered feasible, given the expected survey budgets, and made the following decisions.

4.1 National Sample Size Precision Statement

After careful consideration of numerous options for the national sample size precision statement, HUD decided to retain the current national sample precision statement:

> *For the US as a whole, a two-year change of ten percent in median monthly housing costs for 5-percent subgroups will have a standard error of 5 percent.*

This precision statement, when applied at the 95-percent confidence level, will permit the detection of a 10-percent change in median monthly housing costs for 5-percent subgroups. This precision statement is not designed for any particular 5-percent subgroup. Rather, the statement is intended to be applied to generic 5-percent subgroups, whether they are based on geography, demographics, housing characteristics, or other indicators of interest.

4.2 Base National Sample Size Calculation

The calculation of the national sample size necessary to achieve the precision goal is detailed in appendix A. In short, the sample size was determined by developing a simulation model based on data from 2009 and 2011. The simulation model first identified 5-percent subgroups that had a median monthly housing cost change of 10 percent or greater, then subsampled from the groups at various sample sizes to determine the minimum sample size necessary to achieve a 5-percent standard error.

The results of the simulation showed that a 5-percent subgroup size of 2,000 housing units was large enough to achieve a standard error of 5 percent. With a 5-percent subgroup size of 2,000 housing units, the total sample size would be 40,000.

One adjustment was made to determine the final national sample size. It is common with surveys to assume a nonresponse rate and to increase the sample size accordingly. The single-survey nonresponse rates for the AHS during the past several survey cycles have been approximately 14 percent. As such, to achieve a national sample size of 40,000 complete interviews, we would need to interview approximately 46,500 housing units.

A single-survey nonresponse rate is not necessarily the appropriate adjustment for the AHS, however. The AHS is longitudinal, and some 5-percent subgroups of interest are based on a longitudinal measure. Housing units that experienced a tenure switch (renter to owner or owner to renter) constitute a common example. To calculate a longitudinal measure, the AHS housing unit must have complete data on *both* years of the survey. A review of two-survey longitudinal nonresponse rates during the past few survey cycles shows the two-survey nonresponse rate to be approximately 18 percent.

To ensure a national sample size of 40,000 housing units with complete data in two adjacent survey years, the national sample size must be at least 48,780 housing units. After further consideration of sample design issues and consultations with the U.S. Census Bureau, the national sample size was slightly increased to approximately 50,011.

4.3 Metropolitan Sample Size Precision Statement

As previously mentioned, the AHS metropolitan area sample sizes have fluctuated greatly since the early 1980s. They have ranged from 2,500 to 8,500. Moreover, a precision statement does not appear to have been developed for metropolitan areas.

After careful consideration of numerous options, HUD decided to adopt a precision statement similar to the national precision statement:

> *For the metropolitan area as a whole, a two-year change of 10 percent in median monthly housing costs will have a standard error of 5 percent.*

The notable difference between the national and metropolitan area precision statements is that the metropolitan area statement does not include the requirement to detect a 10-percent change in median monthly housing costs for 5-percent subgroups. As mentioned in section 2.2, the metropolitan area sample sizes necessary to achieve the same statistical precision as the national sample are not feasible given budget constraints.

It must be noted that HUD considered developing a metropolitan area precision statement based on a single-year cross-sectional estimate, as opposed to two-survey-year change estimate. HUD considered this option because of the history of AHS budget fluctuations. Because HUD made the decision to include the 15 largest metropolitan areas in every survey cycle, however, they ultimately decided to base the precision statement on a two-survey change.

4.4 Metropolitan Sample Size Calculation

Two methods were used to calculate the metropolitan area sample size necessary to achieve the statistical precision goal: simulation and direct calculation. Appendix B details the simulation method, and appendix C details the direct calculation method.

Metropolitan areas are different in their housing unit profile and their housing cost distribution. As was expected, both methods revealed that the sample size necessary to achieve the desired precision statement varies among the metropolitan areas, ranging from 500 housing units to more than 3,000 housing units. The simulation method, which used ACS data from the 40 largest metropolitan areas, revealed that a sample size of 2,750 housing units was large enough to achieve a standard error of 5 percent in 35 of the 40 largest metropolitan areas. The direct calculation method revealed that a sample size of 2,750 housing units was large enough to achieve a standard error of 5 percent in 13 of the 15 largest metropolitan areas.

A sample size of 2,750 was determined sufficient. A final decision was made to increase the sample size from 2,750 to 3,000 to account for two-survey longitudinal nonresponse rate.[1]

4.5 Integrating Top 15 Metropolitan Areas Into the National Sample

HUD determined that the 2015 AHS national sample will include an oversample of housing units from each of the 15 largest metropolitan areas,[2] which is referred to as the *integrated national sample*.

HUD has determined that, to achieve their desired precision standards, the AHS must have a national sample size of 50,000 and metropolitan area sample sizes of 3,000 for each of the top 15 metropolitan areas. A simple way to achieve this goal is to add 45,000 top-15 metropolitan area housing units (15 x 3,000) to the national sample. The national sample already has a significant number of housing units from the top 15 metropolitan areas, however, by virtue of it being a nationally representative sample in a nation where one-third of the housing units are in the top 15 metropolitan areas.

As such, to achieve HUD's desired precision size for the national and metropolitan area samples, the national sample needs to be augmented with just enough housing units such that each of the top 15 metropolitan areas has 3,000 housing units within the integrated national sample. The number of cases necessary to achieve this goal is approximately 30,124.

4.6 Integrating HUD-Assisted Oversample Into the National Sample

For the 2011 and 2013 AHSs, HUD included an oversample of HUD-assisted units from public housing and various multifamily programs. The sample size was approximately 5,250 cases. For 2015, HUD will continue this practice by including 5,258 cases.

4.7 Final Integrated National Sample Size

As described in the previous sections, the AHS will be an integrated national sample. This sample includes a nationally representative sample of about 50,011 housing units, supplemented with approximately 30,124 cases from the top 15 metropolitan areas and 5,258 cases from HUD-assisted housing units. The final integrated national sample size will be 85,393 housing units.

[1] Although 3,000 was the adopted standard, some metropolitan area sample sizes will be slightly more or slightly less based on additional design considerations.

[2] See Bucholtz, S., "Metropolitan Area Selection Strategies and Decisions for the 2015 American Housing Survey and Beyond" (2015), for further explanation.

Appendix A. Technical Details for the Simulation Method for Determining the National Sample Size

This appendix describes the technical details of estimating the sample size for the American Housing Survey (AHS) national sample using the simulation method.

Challenges of Determining the Sample Size for AHS

The statistical precision goal for the AHS national sample is "[a] two-year change of 10 percent in the median monthly costs for 5-percent subgroups will have a standard error of 5 percent." This requirement for the sample size provides several challenges, including:

- **The goal is not defined in terms of a specific variable or specific subgroup.** We interpreted the goal to mean that the expected value (or mean) of the standard error across "many variables" and "many 5-percent subgroups" would be 5 percent.
- **In terms of statistics, medians are more complicated than totals or means.** Medians do not have simple expressions for their variance because they are defined by the cumulative distribution function.
- **The AHS has a complex sample design and a sophisticated weighting methodology.** There are many formulas for the sample size of different sample designs; however, they make many simplifying assumptions and they do not address the complexity of the AHS sample design and weighting methodology.
- **We only have past samples; we do not have a complete known universe.** We have only one sample from 2009 and 2011.

Simulation Method

The simulation used the 2009 and 2011 AHS samples to generate multiple samples of varying size. With each sample, we generated estimates of the percent difference for multiple variables and multiple 5-percent subgroups. Because we were simulating from the 2009 and 2011 samples, we did not identify 10-percent differences from the full 2009 and 2011 samples and treat them as known 10-percent differences. We estimated the standard error as the mean over all observed 7.5- to 12.5-percent differences in our simulation. This estimation admits two types of classification error: (1) actual 10-percent differences that were not observed between 7.5 and 12.5 percent, and (2) differences that were not 10 percent and were observed between 7.5 and 12.5 percent. Because the errors offset each other to some degree and because we averaged over the interval 7.5 to 12.5 percent, the estimate of the standard error of the percent difference is still reasonable.

Assumptions of the Simulation Method

- The 2015 first-stage sample design would be similar to the 1985–2013 first-stage sample design. We would have certainty and noncertainty Primary Sampling Units (PSUs), and their sample size would also be comparable.
- We removed expansion cases because we are interested in the steady-state AHS and not the impact of an expansion.

Simulation Steps

For varying national sample sizes, the following steps were completed:

1. Allocated the national sample size to each first-stage stratum proportional to size. Because one PSU was selected from each first-stage stratum, the stratum sample size is also the PSU sample size. This methodology ensured that the sample size was well distributed and therefore did not have a random sample size within the PSU.
2. Identified 5-percent subgroups. Two sets of 5-percent subgroups were generated for each simulated sample. First, we identified several combinations of variables that defined 5-percent subgroups in the past. Examples include "owners and renters built between 1970 to 1974" and "HUs [housing units] in the Central City with 1 bedroom." Second, we generated additional 5-percent subgroups by applying different sort variables to the original sample and then identified the 20 possible 5-percent subgroups from the sorted order. For example, we sorted by year built and then identified the first 5 percent of the list, the next 5 percent, and so on.
3. Selected bootstrap samples from the 2009 and 2011 AHS samples.

4. Calculated simple base weights and a nonresponse adjustment with a reduced set of cells as compared with regular weighting. In addition, we applied a ratio for the HU totals for a reduced set of cells. The weighting was completed independently for the 160 replicates.

5. Estimated the weighted medians for 2009 and 2011 and used them to estimate the percent difference from 2009 to 2011.

6. Estimated variances for the percent difference using the replicate estimates.

The simulation produced multiple estimates of the percent difference for totals of different housing characteristics and their associated standard errors. In our analysis, we kept all the percent differences that were "close" to 10 percent, that is, we kept all percent differences between 7.5 and 12.5 percent. The final step was to average the standard errors.

Simulation Results

Table A-1 summarizes the results of the simulation. We see that a standard error of 5 percent is obtained between the sample sizes of 1,000 and 2,000. For simplicity, we decided on the sample size of 2,000.

Table A-1. National Sample Sizes from the Simulation

Sample Size	Standard Error Percent
1,000	5.5
2,000	4.3
3,000	4.0
4,000	3.8
5,000	3.7
6,000	3.6
7,000	3.5
8,000	3.2

Appendix B. Technical Details for the Simulation Method for Determining the Metropolitan Sample Size

The metropolitan area sample size was determined by developing a simulation model based on American Community Survey (ACS) data from 2005, 2007, and 2009. The ACS was used because it had consistent data in each year for metropolitan areas, as well as sample sizes large enough from which to subsample, for the 100 largest metropolitan areas.

The simulation model first identified metropolitan areas that had a median monthly housing cost change of 10 percent or greater between 2005 and 2007 or 2007 and 2009. These metropolitan areas were then subsampled at various sample sizes (500 through 4,500, by 250) to determine the minimum sample size necessary to achieve a 5-percent standard error for the change in median monthly housing costs.

Simulation Steps

The following steps were completed:

1. Extract the ACS single-year microdata for 2005, 2007, and 2009, including the median housing cost variable and metropolitan area,[3] for the top 40 metropolitan areas.[4]
2. For each metropolitan area and for each year, generate 25 subsamples of ACS housing units, for each subsample size ranging from 500 to 4,500, by 250. This step created 425 subsamples of various sizes (25 subsamples x 17 subsample sizes) for each metropolitan area.
3. For each of the 425 subsamples for a metropolitan area, calculate the median housing cost and the lower and upper confidence intervals (95 percent) of the median. Repeat this step 25 times for each subsample, using ACS replicate weights 1 through 25.[5] This process yields 10,625 estimates of the median housing cost (and lower and upper confidence limits), for each year and metropolitan area.
4. For 2005 and 2007, merge the median estimates by metropolitan area using the iteration number (1 through 10,625), which produces 10,625 pairs of median estimates. Repeat this same process for 2007 and 2009. This end result of this step is 21,250 pairs of median estimates for each metropolitan area.
5. For each of the 21,250 pairs of median estimates, determine if the pair of median estimates produced a change of between 9 and 11 percent. If not, the pair of median estimates is discarded.[6]
6. For each pair of median estimates that produced a change of between 9 and 11 percent, evaluate the change to determine if the 95-percent confidence intervals overlapped. If the confidence intervals did not overlap, that indicates that the standard error of the change is less than 5 percent. The pair of median estimates is deemed "acceptable." If the confidence intervals did overlap, this indicates that the standard error of the change is 5 percent or greater. The pair of median estimates is then deemed "not acceptable."
7. For each metropolitan area and sample size, count the percentage of pairs of median estimates that were deemed acceptable.
8. For each metropolitan area, determine the minimum sample size necessary to ensure that 95 percent of pairs of median estimates are acceptable.

Results

Table B-1 shows the minimum metropolitan area sample size necessary to achieve the desired statistical precision. The results show that a sample size of 2,750 is sufficient for 35 of the 40 largest metropolitan areas.

[3] Metropolitan area is not included in the ACS Public Use Microdata Sample (PUMS). The metropolitan area was inferred by map of Public Use Microdata Areas (PUMAs) to metropolitan areas. If a PUMA was more than 75 percent contained within a metropolitan area, it was assigned to be within that metropolitan areas.

[4] Only housing units in the top 40 metropolitan areas were used because metropolitan areas outside of the top 40 do not large enough ACS samples to conduct the simulation.

[5] The lower and upper confidence intervals were calculated using SAS Proc Surveymeans, BRR variance method. ACS replicate weights 26 through 80 were used in the variance calculation.

[6] Although the precision standard is a 10-percent change in median monthly housing costs, 10 percent is too exact for conducting this analysis, so a small range around 10 percent is used.

Table B-1. Simulation Method Results for 40 Largest Metropolitan Areas

Metropolitan Statistical Area Name	Minimum Sample Size Necessary To Achieve Precision Goal
Indianapolis-Carmel-Anderson, IN	1,250
Nashville-Davidson--Murfreesboro--Franklin, TN	1,500
Virginia Beach-Norfolk-Newport News, VA-NC	1,500
Atlanta-Sandy Springs-Roswell, GA	1,750
Austin-Round Rock, TX	1,750
Baltimore-Columbia-Towson, MD	1,750
Cincinnati, OH-KY-IN	1,750
Jacksonville, FL	1,750
Kansas City, MO-KS	1,750
San Antonio-New Braunfels, TX	1,750
Boston-Cambridge-Newton, MA-NH	2,000
Minneapolis-St. Paul-Bloomington, MN-WI	2,000
Orlando-Kissimmee-Sanford, FL	2,000
Phoenix-Mesa-Scottsdale, AZ	2,000
Charlotte-Concord-Gastonia, NC-SC	2,250
Cleveland-Elyria, OH	2,250
Dallas-Fort Worth-Arlington, TX	2,250
Detroit-Warren-Dearborn, MI	2,250
Los Angeles-Long Beach-Anaheim, CA	2,250
Milwaukee-Waukesha-West Allis, WI	2,250
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	2,250
Portland-Vancouver-Hillsboro, OR-WA	2,250
St. Louis, MO-IL	2,250
Houston-The Woodlands-Sugar Land, TX	2,500
Las Vegas-Henderson-Paradise, NV	2,500
Miami-Fort Lauderdale-West Palm Beach, FL	2,500
New York-Newark-Jersey City, NY-NJ-PA	2,500
Providence-Warwick, RI-MA	2,500
San Diego-Carlsbad, CA	2,500
San Francisco-Oakland-Hayward, CA	2,500
Tampa-St. Petersburg-Clearwater, FL	2,500
Chicago-Naperville-Elgin, IL-IN-WI	2,750
Columbus, OH	2,750
Denver-Aurora-Lakewood, CO	2,750
Pittsburgh, PA	2,750
Riverside-San Bernardino-Ontario, CA	3,000
Sacramento--Roseville--Arden-Arcade, CA	3,000
San Jose-Sunnyvale-Santa Clara, CA	3,000
Seattle-Tacoma-Bellevue, WA	3,000
Washington-Arlington-Alexandria, DC-VA-MD-WV	3,250

Appendix C. Technical Details for the Direct Method for Determining the Metropolitan Sample Size

This appendix discusses our general method for calculating the sample sizes. In part 1, we simplify the original problem of the difference between two medians so that we have an expression in terms of one median and not two. In part 2, we use the simplification of part 1 with "Woodruffing" in reverse to express the variance in terms of the sample size.

Part 1: Simplify the Percent Difference

We want the standard error for the percent difference of medians m_1 and m_2 at two times to be 5 percent; that is,

$se\left(\frac{m_2 - m_1}{m_1}\right) = 0.05$, when the percent difference is 10 percent; that is,

$$\frac{m_2 - m_1}{m_1} = 0.10. \qquad \text{(C.1)}$$

We know from Yates (1948; equation 7.5.k) that

$$v\left(\frac{m_2 - m_1}{m_1}\right) = \left(\frac{m_2 - m_1}{m_1}\right)^2 \left[\frac{v(m_2 - m_1)}{(m_2 - m_1)^2} + \frac{v(m_1)}{m_1^2}\right].$$

Next, we apply (C.1) and we get

$$\begin{aligned} v\left(\frac{m_2 - m_1}{m_1}\right) &\le (0.10)^2 \left[\frac{2v(m_1)}{(0.10)^2 m_1^2} + \frac{v(m_1)}{m_1^2}\right] \\ &= (0.10)^2 \left[\frac{2v(m_1)}{(0.10)^2 m_1^2} + \frac{v(m)}{m_1^2}\frac{(0.10)^2}{(0.10)^2}\right] \\ &= \frac{(0.10)^2}{(0.10)^2}\left[\left(2 + (0.10)^2\right)\frac{v(m_1)}{m_1^2}\right] \\ &= \left(2 + (0.10)^2\right)\frac{v(m_1)}{m_1^2} \end{aligned}$$

where $v(m_1 - m_2) = v(m_1) + v(m_2) - 2cov(m_1, m_2) \le v(m_1) + v(m_2) = 2v(m_1)$ and where we assume that the variances of the medians m_1 and m_2 are equivalent; that is, $v(m_1) = v(m_2)$.

So when $se\left(\frac{m_2 - m_1}{m_1}\right) = 0.05$,

$$v\left(\frac{m_2 - m_1}{m_1}\right) = (0.05)^2 = \left(2 + (0.10)^2\right)\frac{v(m_1)}{m_1^2},$$

and this implies that

$$\begin{aligned} \left(2 + (0.10)^2\right)\frac{v(m_1)}{m_1^2} = (0.05)^2 &\Rightarrow \frac{v(m_1)}{m_1^2} = \frac{(0.05)^2}{2 + (0.10)^2} \\ &\Rightarrow \frac{v(m_1)}{m_1^2} = \frac{(0.05)^2}{2 + (0.10)^2} \\ &\Rightarrow se(m_1) = (0.0352)m_1 \quad \text{or} \quad cv(m_1) = 3.52\%. \end{aligned}$$

So we need a sample size that provides a standard error of the median of m_1 that is 3.52 percent of the size of m_1.

Part 2: Reverse "Woodruffing" To Find the Sample Size

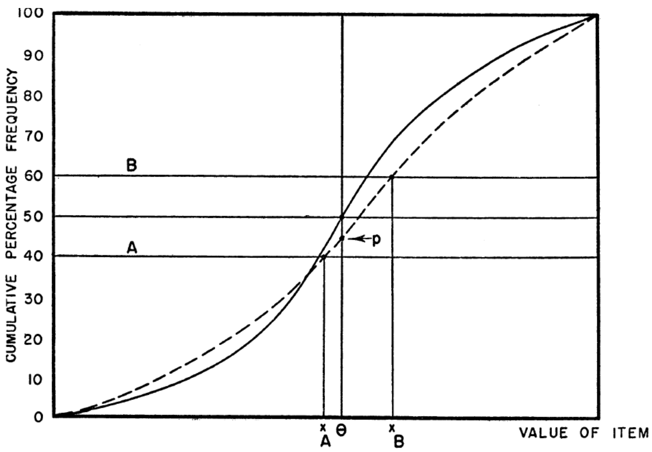
At this point, we want an expression of the desired variance in terms of the sample size. With this type of expression, we can solve for the sample size in terms of the desired variance.

For the median, the Woodruff (1952) method is the usual way of forming confidence intervals. Instead of forming confidence intervals, we want to do the reverse—we know the confidence interval we want for a median and we want the sample size n that will produce it. The key to calculating confidence intervals and our approach for sample sizes is the following from Woodruff (p. 638):

> If the sampling is done with replacement or from a very large population the variance of the sample number is less than θ [the median] is nPQ where $P = Q = 0.5$. The one standard derivation limits are $(n/2) \pm \sqrt{n}/2$. To secure the corresponding one standard derivation limits it is necessary only to find the value of the $\left(n + \sqrt{n}\right)/2$ and the $\left(n - \sqrt{n}\right)/2$ items of the sample (arrayed in order of size).

Figure C-1 shows how the sample size on the vertical axis is related to the values of x_k on the horizontal axis. When estimating confidence intervals, we start with the vertical axis and the sample size n and then find the corresponding upper and lower points on the horizontal axis. The approach we suggest is to calculate sample sizes doing the reverse: finding the desired confidence interval on the horizontal axis, which is then related to the n on the vertical axis.

Figure C-1. Reproduction of Woodruff (1952) Cumulative Percentage Frequency Graph



We now describe how we relate the sample size to the confidence interval of the median. We first define the cumulative distribution function of a variable x_k as $\Phi_a = \sum_{x_k \le a} w_k$. Note that k is indexing the sample housing units. Then the weighted median of a variable x_k is defined as the value of m such that $\Phi_m \le \frac{\hat{N}}{2}$ and the sample estimate of the size of the universe is $\hat{N} = \sum_{k \in s} w_k$. We also define the cumulative point that is one standard error to the left of the median as Φ_L and one standard error to the right of the median Φ_U; that is, $L = m - se(m)$ and $U = m + se(m)$. Now, we can relate the sample size to the cumulative distribution as

$$\frac{se(n/2)}{n/2} = \frac{\left|\hat{N}/2 - \Phi_L\right|}{\hat{N}/2}, \qquad \text{(C.2)}$$

and

$$\frac{se(n/2)}{n/2} = \frac{\left|\hat{N}/2 - \Phi_U\right|}{\hat{N}/2}. \qquad \text{(C.3)}$$

Because of the result that follows, we can estimate the sample sizes as $n_L = \left(\frac{\hat{N}}{\left|\hat{N}/2 - \Phi_L\right|}\right)^2 \frac{1}{4}\sqrt{deff}$ or $n_U = \left(\frac{\hat{N}}{\left|\hat{N}/2 - \Phi_U\right|}\right)^2 \frac{1}{4}\sqrt{deff}$.
If the original distribution is symmetric, $\left|\hat{N}/2 - \Phi_L\right| = \left|\hat{N}/2 - \Phi_U\right|$, so $n_L = n_U$. In practice, we hardly ever have symmetry, so in our calculations, we choose the more conservative $n = \max(n_L, n_U)$ instead of averaging n_L and n_U.

As a result, $n_L = \left(\frac{\hat{N}}{\left|\hat{N}/2 - \Phi_L\right|}\right)^2 \frac{1}{4}\sqrt{deff}$.

From (C.2), we know

$$\frac{se(n/2)}{n/2} = \frac{\left|\hat{N}/2 - \Phi_L\right|}{\hat{N}/2} \Rightarrow \frac{\sqrt{\frac{n}{4}deff}}{n} = \frac{\left|\hat{N}/2 - \Phi_L\right|}{\hat{N}}$$

$$\Rightarrow \frac{\frac{1}{2}\sqrt{deff}}{\sqrt{n}} = \frac{\left|\hat{N}/2 - \Phi_L\right|}{\hat{N}}$$

$$\Rightarrow n = \left(\frac{\hat{N}}{\left|\hat{N}/2 - \Phi_L\right|}\right)^2 \frac{1}{4}\sqrt{deff},$$

where

$$v(np)\big|_{p=\frac{1}{2}} = n^2 \frac{p(1-p)}{n} deff\Big|_{p=\frac{1}{2}}$$

$$= n\frac{1}{2}\left(1-\tfrac{1}{2}\right)deff$$

$$= \frac{n}{4}deff.$$

Similarly, $n_U = \left(\frac{\hat{N}}{\left|\hat{N}/2 - \Phi_U\right|}\right)^2 \frac{1}{4}\sqrt{deff}$.

Results

For the metropolitan area sample sizes, we estimated the sample sizes so that 10-percent differences would have a 5-percent standard error. In addition, the sample sizes were calculated in terms of a single variable: total housing costs.

The far right column of table C-1 provides specific estimated sample sizes.

Table C-1. Metropolitan Sample Sizes From the Direct Calculation

AHS Metropolitan Area Name	$\hat{N}$	$\lvert \hat{N} - \Phi_L \rvert$	$\lvert \hat{N} - \Phi_U \rvert$	Design Effect (*deff*)	NR factor	Sample size *n*
New York	3,889,007	97,718	86,669	1.17	2.38	1,406
Los Angeles	1,206,897	45,229	48,960	1.05	2.74	512
Chicago	1,238,037	45,091	32,642	1.14	1.82	745
Dallas	1,310,991	47,526	66,549	1.07	2.10	427
Philadelphia	1,380,894	33,190	28,315	1.20	1.94	1,385
Houston	1,169,223	50,105	16,955	1.10	2.37	3,079
Washington	1,238,409	17,179	31,251	1.06	2.06	2,835
Miami	812,384	12,952	12,039	1.07	2.18	2,644
Atlanta	1,148,859	29,585	41,999	1.07	1.91	769
Boston	111,299	2,569	3,318	1.13	3.70	1,961
San Francisco	1,032,858	24,390	38,623	1.05	2.26	1,067
Detroit	1,068,419	29,839	29,779	1.05	1.90	643
Riverside	871,234	19,030	37,750	1.09	2.14	1,221
Phoenix	906,192	30,589	20,584	1.32	2.27	1,448
Seattle	832,478	28,370	13,943	1.07	2.21	2,112

NR = nonresponse.