ESTIMATING THE HOMELESS POPULATION: UNDUPLICATED ENUMERATION IN THE PRESENCE OF MASSIVELY MISSING DATA FROM INSTITUTIONAL SURVEYS

Mack C. Shelley, II, and Paula W. Dail, Iowa State University
and Scott T. Fitzgerald, University of Iowa
Mack C. Shelley, II, ISU, Department of Statistics, 323 Snedecor Hall, Ames, IA 50011-1210

Key Words: Estimation, social statistics, survey research

The research problem

Homelessness is a fluid social problem. Most people who are homeless move into and out of the condition of homelessness more or less at random, as part of a lifestyle of chronic poverty and/or family abuse. This gives rise to major difficulties of recording incidents of homelessness. The number of reporting agencies changes appreciably over time, and agency reporting techniques and record-keeping abilities of agencies and shelters are subject to change. In addition, there is good reason to believe that the incidence of homelessness is underreported by some agencies, owing perhaps to inadequate record-keeping, and overreported by other agencies with reasons to prefer that larger numbers be reported.

Accurately estimating the number of homeless depends upon two critical issues: (a) defining the problem and (b) determining the best methodology for attempting the count, given the constraints imposed by available research dollars and access to relevant data sources. Many problems are associated with enumerating the homeless population accurately.

A central difficulty is to provide an operational definition of homelessness. Section 103 of the McKinney Homeless Assistance Act (1987), codified as Title 42-The Public Health and Welfare, Chapter 119, Homeless Assistance, Subchapter I (General Provisions 11302) provides a general definition of a homeless individual. This definition states that, for purposes of the Act, the term "homeless" or "homeless individual" includes: (1) an individual who lacks a fixed, regular, and adequate nighttime residence; and (2) an individual who has a primary nighttime residence that is: (A) a supervised publicly or privately operated shelter designed to provide temporary living accommodations (including welfare hotels, congregate shelters, and transitional housing for the mentally ill); (B) an institution that provides temporary residence for individuals intending to be institutionalized; or (C) a public or private place not designed for, or ordinarily used as a regular sleeping accommodation for human beings. Excluded is any individual imprisoned or otherwise detained pursuant to an Act of Congress or a State law (PL 100-77; July 22, 1987). Agencies that administer homeless assistance programs sometimes broaden this definition to include individuals who are residing in transitional or supportive housing.

This definition is supplemented by U. S. Department of Education (1989) guidelines suggesting that counts of homeless children should include children who are living in shelters for runaways, on the streets, in abandoned buildings, or in other facilities unfit for human habitation; children who do not have an adequate home base that serves as a permanent home; children living in camping areas (or trailer parks) because they lack adequate accommodations; children in transitional emergency shelters; sick or abandoned children living in state institutions because of no other suitable alternative; runaway/throwaway children living together as a group in a suitable shelter; and children living with friends or relatives. The guidelines suggest that children living in foster homes and in trailer parks with adequate, long-term accommodations; children incarcerated for violations of the law; and children of migrant workers who are living doubled-up should not be included in a count of the homeless.

Other complications arising in the process of operationally defining the incidence of homelessness include whether to include those who are "doubled-up," especially as a response to poverty and/or domestic violence; previously unsuccessful efforts to assess the problem, such as the S-night national homeless census attempted in 1990. An additional definitional problem is presented by the wildly disparate estimates provided by homeless advocates, government officials, and social scientists, who often find themselves in adversarial relationships with one another.

Another difficulty is to define and to generate data from an appropriate sampling frame, which is essential for the accuracy of the numbers resulting from any such counting effort. There is disagreement over whether the optimal counting methodology is at one or more fixed points in time or annual and/or continuous, and over the relative virtues of capture-recapture (i.e., count-recount) methods (Cowan, 1991), sampling in space and time (Cowan, 1991), key informants (key person surveys), or random sample surveys. In a national random sample telephone survey of households, Link, Phelan, Bresnahan, Stueve, Moore, & Susser (1995) found that 14% of respondents considered themselves to have been homeless at one time during their lives. More generally, methods used to estimated the incidence of homelessness span partial counts, extrapolations from partial counts, windshield street surveys, and area probability designs.

We have included in our estimates of homelessness those who are living on the streets or in abandoned buildings; in a public or a private homeless shelter; doubled-up with family/friends; in transitional housing for the mentally ill; in a single room occupancy facility; in a transitional housing project; in a home or apartment; in a youth group home; or "other" (those living in campgrounds, temporary trailers, or other makeshift arrangements not specified in other categories. This definition does not include the "near homeless," that is, adults and children who also are referred to as being "imminently homeless".

Survey Methodology

Two spread-sheet type questionnaires were distributed, one to schools across all districts of Iowa, and the other to social service agencies. Following appropriate pretesting, the survey instruments were mailed to all public schools in Iowa and to all known shelters, Community Action Program agencies, County General Relief Offices, transitional housing programs, and county Department of Human Services offices in the state, as well as to miscellaneous programs such as medical outreach services serving the homeless population. A stamped, preaddressed return envelope was mailed out with the survey form, instructions, and a cover letter signed by the Director of the Iowa State Department of Education.

Respondents from the schools were asked to identify all homeless children known to them during the current academic year to date and the service needs of this population of children. The agencies were asked to provide information about each homeless person they had served during a one-month period (March 15-April 15, 1997), and what they perceived to be the most acute service needs of the homeless in their service delivery area. Follow-ups conducted with nonrespondents included, for agencies, two reminder postcards, followed by telephone calls and e-mail messages, and for schools, two written requests. Response rates were 53.8% overall; 55.2% for schools, and 49.2% for all agencies combined. Further details regarding response rates are provided in Table 1.

Elimination of Reporting Duplications

Controlling for duplication in the reported data took place in three stages: within the schools data, within the agencies data, and between the schools data and the agencies data. A "unique identifier" was created from the first four letters of the last name and the last four digits of the Social Security number, when these were known; this was used to locate and remove multiple data lines representing a single individual.

For the schools data, when a unique identifier appeared more than once the first data line was coded "0" (unduplicated data line) and the other(s) was (were) coded "99" (duplicate data line). An algorithm was created to facilitate assessment of probable duplication status for the data lines that were missing one or both components of the "unique identifier." Examples of a likely duplication and a probable nonduplication are given in Figure 1. This assessment is based on awarding 5 points for name and Social Security number, 3 points for age, and 1 point each for gender, race, county, district, and school building. This resulted in a range of scores from 5 to 18. When both name and Social Security number were missing, the result was coded as unknown ("88").

Assessing likelihood of duplication for the agencies data followed essentially the same process as for the schools, except that the school district and building variables were replaced by the agency name (which was awarded 1 point). The resulting range of scores varied from 5 to 17. When the schools and agencies data were merged together and checked in tandem for further possible duplication, the same process of scoring was followed, except that district, building, and agency were dropped. The range of scores was from 5 to 16.

Low (M3), midrange (M4), and high (M5) unduplicated estimates were based on assumptions regarding the probability of duplication. The low estimate (Merge3) is the most conservative. It assumes that all weighted coded items are duplicates. One-half of all such paired entries were recoded "0" (nonduplicate) and the remaining one-half were recoded "99" (duplicate). All items coded "99"' and "88" (unknown) were deleted. For the midrange estimate (Merge4), items coded "5"-"10" were assumed to be nonduplicative and were recoded "0" (nonduplicate). Items coded "11"-"18" were assumed to be duplicates; one-half of all such pairs were recoded "0" (nonduplicate) and the remaining one-half were recoded "99". All items coded "99" (duplicate) and "88" (unknown) were deleted. The high estimate (Merge5) is the least conservative. It assumes that all items coded "88" (unknown) and "5"-"18" were nonduplicates and therefore were retained in the data set. Items coded "99" (duplicate) were deleted.

Of the 1,881 cases of homelessness identified by the schools, 53 were duplicates, leaving 1,828 unduplicated cases in the school data. Of 3,665 cases of homelessness identified by agencies and shelters, 479 duplicates/unknowns were removed, leaving 3,186 unduplicated cases. When the data sets were merged, 31 additional duplicates were eliminated, leaving a total of 4,983 unduplicated cases altogether. Approximately 10% of the total reported number of people homeless were apparent duplications. These estimates, by type of responding institution, are given in Table 2.

Approximate 95% Confidence Interval for M4 (Midrange) Unduplicated Reported Numbers of Homeless

Following Kish (1965, pp. 134-136), the variance of a stratified total is found by:

$$\Sigma \left[(1 - (n_h/N_h)) (N^2_h) (s^2_h) / n_h\right]$$

where n_h is the sample size of the h'th stratum, N_h is the stratum population size, and s^2_h is the stratum variance.· The variance of the M4 (midrange) unduplicated number of reported incidents, V_{M4}, assuming independence of the three strata, is

$$\begin{aligned} V_{M4} &= V_{\text{shelters}} + V_{\text{other agencies}} + V_{\text{schools}} \\ &= (1 - (47/82)) (82)^2 (1{,}859.9) / 47 \\ &+ (1 - (176/371)) (371)^2 (610.4550649) / 371 \\ &+ (1 - (861/1{,}560)) (1{,}560)^2 (396.0458499) / 1{,}560 \\ &= 113{,}572.6169 + 119{,}038.7377 + 276{,}836.049 \\ &= 509{,}447.4036 \end{aligned}$$

The standard deviation of the M4 (midrange) unduplicated number of reported incidents, S_{M4}, is 713.7558432. An approximate 95% confidence interval is, then, $M4 +/- Z_{.025}S_{M4} = 4{,}983 +/- 1.96(713.7558432)$, or approximately 4,983+/-1,399, for an interval of (3,584 to 6,382). These results should be compared against the values reported in Table 2 of 4,828 for M3(Low) and 5,291 for M5(High).

Inflating For Nonreporting

The low response rate (54% overall) made it necessary to adjust for nonreporting. Response rate adjustments were calculated separately using the respective reciprocals of the response rates from schools, shelters, other agencies, General Relief agencies, Department of Human Services agencies, Community Action Programs, transitional housing providers, miscellaneous agencies, and shelters. For shelters, these adjustments were refined further by using newly available shelter-bed capacity information, measured as the number of available beds per shelter. Shelter-bed capacity rate (SBCR) was calculated for the responding shelters. This was defined as the ratio of the number of reported clients for one month to the number of available beds on any given night. For the low estimate this information was not used, based on the assumption that nonreporting shelters had zero homeless to report. For the midrange estimate, the ratio was 1,481/1,236, so SBCR = 1.201. For the high estimate, the ratio was 1,672/1,413, and SBCR = 1.185.

Shelters

The shelter data were adjusted first for duplications. Estimates then were derived based on different assumptions about nonresponse and shelter bed capacity. Three different data sets were generated initially, under different assumptions regarding duplication. SheltM3 was defined to include the total number of unduplicated data lines reported by shelters in the Merge3 data set, and assumes that all weighted coded cases are duplicates. SheltM4 was defined to include the total number of unduplicated data lines reported by shelters in the Merge4 data set, and assumes that all cases coded 11-18 are duplicates. SheltM5 was defined to include the total number of unduplicated data lines reported by shelters in the Merge5 data set, and assumes no duplicates.

Estimates then were derived from different assumptions about nonreporting shelters and about shelter bed capacity. The low estimate was calculated as 2a(low) = SheltM3 + 0, which assumes that the nonreporting shelters had zero homeless to report. Consequently, the number reported was not adjusted for the low estimate. For the midrange estimate, 2a(mid) = SheltM4 + [SBCR*(shelter bed capacity for nonreporting shelters/2)], which assumes that, on average, one-half of nonreporting shelters maintained the same shelter bed capacity as did the reporting shelters during the reporting period. The midrange estimate also assumes that one-half of the nonreporting shelters had zero homeless to report. For the high estimate, 2a(high) = SheltM5 + (SBCR*shelter bed capacity for nonreporting shelters). The high estimate assumes that all nonreporting agencies maintained the same shelter bed capacity as did the reporting shelters.

Other Agencies (General Relief, Department of Human Services, Community Action Programs, Transitional Housing Providers, Miscellaneous)

Three different data sets were generated initially, using different assumptions regarding duplication. AgencM3 was defined as the total number of unduplicated data lines reported by nonshelter agencies in the Merge3 data set, and assumes that all weighted cases are duplicates. AgencM4 was defined as the total number of unduplicated data lines reported by the nonshelter agencies in the Merge4 data set, with cases coded 11-18 assumed to be duplicates. AgencM5 was defined to be the total number of unduplicated data lines reported by nonshelter agencies in the Merge5 data set, assuming that no cases are duplicates.

Estimates then were derived from different assumptions about nonreporting. The low estimate was calculated as 2b(low) = AgencM3 + 0. This assumes that the nonreporting agencies had zero homeless to report. The raw number reported was not adjusted. The midrange estimate was calculated as 2(mid) = 0.5[AgencM4{(1/response rate)+1}]. It assumes that one-half of the nonreporting agencies had, on average, the same number of homeless as the reporting agencies during the reporting period, while the other one-half of the nonreporting agencies had zero homeless to report. The high estimate was calculated as 2b(high) = AgencM5*(1/response rate). This assumes that nonreporting agencies, on average, had the same average number of homeless as were reported by the reporting agencies.

Schools

Again, three different data sets were generated initially, using different assumptions regarding duplication. SchoolM3 was defined as the total number of unduplicated data lines reported by schools in the Merge3 data set, assuming that all weighted coded cases are duplicates. SchoolM4 was defined as the total number of unduplicated data lines reported by schools in the Merge4 data set, with cases coded 11-18 assumed to be duplicates. SchoolM5 was defined to be the total number of unduplicated data lines reported by schools in the Merge5 data set, and assumes no duplicates.

Estimates then were derived from different assumptions about nonreporting. The low estimate, calculated as 2c(low) = SchoolM3 + 0, assumes that the nonreporting schools had zero homeless to report. The midrange estimate, calculated as 2c(mid) = 0.5[SchoolM4{(1/response rate)+1}], assumes that one-half of the nonreporting schools had, on average, the same number of homeless as the reporting schools during the reporting period, while the other one-half of the nonreporting schools had zero homeless to report. The high estimate, calculated as 2c(high) = SchoolM5(1/response rate), assumes that, on average, nonreporting schools had the same average number of reported homeless provided by the reporting schools.

Inflating For Time

To estimate the number of incidents of homelessness, it was necessary to inflate for time for the nonshelter agency data. An incident of homelessness refers to one episode, of indeterminate length between 1 and 30 days in our data set, of homelessness for one individual. Each incident, by definition, is mutually exclusive of all other incidents of homelessness for the individual in question. For example, if an individual were homeless for an entire year, for our data analysis this would be interpreted as 12 monthly incidents of homelessness.

Inflating for time is complicated because the reporting intervals were not uniform for the different sources of data. The data provided by shelters and other agencies covered a one-month period (from mid-March to mid-April 1997), while the data provided by schools were for the 1996/1997 school year. To produce an annualized estimate of the number of incidents of homelessness, an inflation factor of 12 was applied to the shelter data and to the data from the other agencies. This value assumes that the reporting period represents an average number of homeless in one given month of a 12-month period.

Nine different inflated totals thus were possible. For shelters: 3a(low) = 2a(low)*12; 3a(mid) = 2a(mid)*12; and 3a(high) = 2a(high)*12. For other agencies: 3b(low) = 2b(low)*12; 3b(mid) = 2b(mid)*12; 3b(high) = 2b(high)*12. For schools: 3c(low) = 2c(low); 3c(mid) = 2c(mid); and 3c(high) = 2c(high). Note that the schools estimates (2c) are not multiplied by 12, because the data from schools were reported on an annual basis.

To produce statewide estimates of the number of incidents of homelessness, the estimates for shelters, agencies, and schools were summed across the respective low, midrange, and high range categories. Thus, the total state low estimate of incidents of homelessness across all categories was calculated as 3a(low) + 3b(low) + 3c(low). The total state midrange estimate of incidents of homelessness across all categories was calculated as 3a(mid) + 3b(mid) + 3c(mid). The total state high estimate of incidents of homelessness across all categories was calculated as 3a(high) + 3b(high) + 3c(high).

Following these calculations, the resulting time-adjusted estimates of the number of statewide incidents of homelessness in 1997 were: low = 38,950, midrange = 59,558, and high = 83,502.

Approximate 95% Confidence Interval for M4 (Midrange) Adjusted Estimate of the Number of Incidents of Homelessness

Again using Kish's (1965) framework, we can find the estimated variance of the M4 (midrange) adjusted estimate of the number of incidents of homelessness, assuming independence, and adding squared constants representing the various adjustment ratios:

$\Sigma\ [(1 - (n_h/N_h))\ (N^2_h)\ (s^2_h)\ (N^2_h/n^2_h)\ (t)^2\ (SBCR)^2\ /\ n_h]$

where n_h, N_h, and s^2_h are as defined previously, t is the time inflation factor (12 for shelters and other agencies, 1 for schools), and SBCR is the shelter bed capacity rate as defined above (relevant only to the shelter data). The variance of the M4 (midrange) unduplicated number of reported incidents, V_{M4}, assuming independence of the three strata, is

$V_{M4adjusted} = V_{shelters,\ adjusted} + V_{other\ agencies,\ adjusted} + V_{schools,\ adjusted}$

$$\begin{aligned} &= (1 - (47/82))\,(82)^2\,(1{,}859.9)\,(82/47)^2\,(12)^2 \\ &\quad (1.201)^2 \,/\, 47 \\ &+ (1 - (176/371))\,(371)^2\,(610.4550649)\,(371/176)^2 \\ &\quad (12)^2 \,/\, 371 \\ &+ (1 - (861/1{,}560))\,(1{,}560)^2\,(396.0458499) \\ &\quad (1{,}560/861)^2\,(12)^2 \,/\, 1{,}560 \\ &= 71{,}804{,}907.72 + 76{,}168{,}129.17 + \\ &\quad 908{,}794.1781 = 148{,}881{,}831.1 \end{aligned}$$

The standard deviation of the M4 (midrange) adjusted estimate of the number of incidents of homelessness, $S_{M4adjusted}$, is 12,201.7. An approximate 95% confidence interval is, then, $M4_{adjusted}$ +/- $Z_{.025}S_{M4adjusted}$ = 59,558 +/- 1.96 (12,201.7), or approximately 59,558 +/- 23,915, for an interval of (35,643 to 83,473). These results should be compared against the sample-based time-adjusted low estimate of 38,950 and the time-adjusted high estimate of 83,502.

County Totals

Using the aggregated results, it was desired also to estimate the incidents of homelessness for each county in the state. To accomplish this, the 99 counties in the state were divided into three categories (Bruner, 1993), stratified into 3 levels: 8 large metro counties (42% of statewide population), with their largest population center in excess of 50,000; 45 small metro counties (40% of statewide population), with their largest population center between 5,000 and 49,999; and 46 rural counties (18% of statewide population), with their largest population center less than 5,000.

Each individual county's population was reexpressed as a proportion of the relevant total county-type population, to provide county-level estimates of incidents of homelessness. For example, the total county-type population for the large metro counties is 1,183,275. The population of each of the 8 large metro counties was divided by 1,183,275.

The proportion of total state population represented by each of the three county-types was multiplied by the state total estimated number of incidents of homelessness to produce an estimated number of incidents for each of the three county-types. The estimated total number of incidents of homelessness for each county-type then was multiplied by the proportion that each county's population represented of its county-type population, to obtain an estimated number of incidents of homelessness for each county. These calculations were conducted separately for two different definitional categories of homelessness: sheltered and unsheltered homeless, and doubled-up and transitional housing/other. Detailed results of county estimates are presented in Dail, Shelley, Fitzgerald, and Baker (1998).

References

Bruner, C. (1993). *Reinventing common sense: Indicators of well-being for Iowa's children.* Des Moines, IA: Child and Family Policy Center.

Cowan, C. D. (1991). Estimating census and survey undercounts through multiple service contacts. *Housing and Policy Debate, 2*(3), 869-882.

Dail, P. W., Shelley, M. C., Fitzgerald, S., & Baker, J. (1998). *Homeless in Iowa: Findings from the 1997 statewide study.* Des Moines, IA: Iowa Department of Education.

Kish, L. (1965). *Survey sampling.* New York, NY: John Wiley & Sons,Inc.

Link, B., Phelan, J., Bresnahan, M., Stueve, A., Moore, R., & Susser, E/ (1995). Lifetime and five year prevalence of homelessness in the United States: New evidence on an old debate. *American Journal of Orthopsychiatry, 65*(3), 347-354.

United States Department of Education. (1989). *Memo concerning guidelines for defining and counting homeless children.* Washington, DC: United States Department of Education.

The authors are deeply indebted to Kathy Shelley of the Statistical Laboratory at Iowa State University for the valuable technical assistance she provided in preparing this paper and related documents. In addition, important insights into many aspects of this project were provided by the Data Committee of the Interagency Task Force on Homelessness, and the Iowa Coalition for Housing and the Homeless.

This project was supported through grant #91781 awarded to Professor Dail by the Iowa Department of Education. For further information contact Dr. Paula Dail, Iowa State University, Ames, IA 50011-1120, or Dr. Ray Morley, of the Iowa Department of Education.

Table 1
Response Rates

Data Source	Sent	Returned	Response Rate (%)
Schools	1,560	861	55.2
Homeless Shelters	82	47	57.3
General Relief	101	35	34.7
County Department of Human Services Offices	104	73	70.2
Community Action Agencies	119	52	43.7
Transitional Housing Programs	32	6	18.8
Miscellaneous	15	10	66.7
Total	2,013	1,084	53.8*

* This is calculated as the number of total returns divided by the number of questionnaires sent.

Figure 1
Likelihood of Duplication (School Data Samples)

Example Case #1 (a likely duplication)

Entry#	Name	SS#	Age	Gender	Race	County	District	Building		
100	aaaa		16	2	1	57	1111	109		
101	aaaa		16	2	1	57	1111	109		
Score =	5	0	3	1	1	1	1	1	=	13

Example Case#2 (not a likely duplication)

Entry#	Name	SS#	Age	Gender	Race	County	District	Building		
200	bbbb		11	1	2	57	2222	209		
201	bbbb		16	2	1	57	1111	109		
Score =	5	0	0	0	0	1	0	0	=	6

Table 2
Unduplicated Reported Number of Homeless in All Categories

	M3 (Low)	M4 (Midrange)	M5 (High)
Shelters	1,435	1,481	1,672
Agencies	1,667	1,697	1,774
Schools	1,726	1,805	1,845
TOTAL	4,828	4,983	5,291