Estimating Households by Household Size Using the Poisson Distribution

ABSTRACT
Infrastructure planners often require detail about the number of households by household size at very small levels of geography (census tract or smaller) to calibrate their models. In addition, these data must also be projected into the future in order to support planning efforts.

This article documents a statistical technique for estimating the distribution of households by household size using a modified application of the Poisson distribution. This technique is valuable to demographers as it provides a simple and reliable tool for estimating the distribution of household sizes at nearly any level of geography for a given point in time.

There are a wide variety of applications of the Poisson distribution in biology and engineering. However, there are only few documented applications in demographics. This article puts forth two key advancements over prior published work:
(1) an entirely new, and greatly simplified method for applying the distribution,
(2) evidence of the reliability of the technique for estimating household size distributions in small geographic areas (e.g. counties and census tracts).

Tests on U.S. Census data (1990-2010) suggest that the model is suitable for use in estimating the distribution of households by household size at the state, county, and census tract level.

PURPOSE

Infrastructure planners, such as transportation engineers and water and wastewater planners, often require detail about the number of households by household size at very small levels of geography (e.g. census tract or smaller). Because the size of a household may affect behaviors such as carpooling and household water usage, these data are crucial in calibrating infrastructure needs models. In addition to requiring accurate current-year estimates, these models often require reliable future-year projections in order to support planning efforts.

The calculation of average household size is relatively straightforward:

Household Size = Household Population / Occupied Households

However, while the total number of household and the average household size are commonly estimated and projected, the distribution of households by size is rarely modeled.

This article develops and documents a statistical technique for estimating and forecasting households by household size using a modified application of the Poisson distribution.

The U.S. Census Bureau American Community Survey (ACS) now provides rolling 5-year averages for households by size at the census tract level, but the Poisson technique illustrated herein remains useful for four key reasons:

(1) there is still a need for point-in-time estimates,
(2) the technique fulfills demand for estimates below the census tract level,
(3) there are documented discrepancies between household size distributions and household population estimates in the ACS, and
(4) the technique fulfills the demand for forecasted data on household size distribution.

BACKGROUND

Households have a very distinct definition. According to the U.S. Census Bureau:

> *A household includes all of the people who occupy a housing unit. (People not living in households are classified as living in group quarters.) A housing unit is a house, an apartment, a mobile home, a group of rooms, or a single room occupied (or if vacant, intended for occupancy) as separate living quarters. Separate living quarters are those in which the occupants live separately from any other people in the building and that have direct access from the outside of the building or through a common hall. The occupants may be a single family, one person living alone, two or more families living together, or any other group of related or unrelated people who share living quarters.*
>
> *In 100-percent tabulations, the count of households or householders always equals the count of occupied housing units.*
>
> U.S. Census Bureau, Public Use Microdata Sample 2000: Technical Documentation

Prior to implementation of the American Community Survey (ACS), data on household size distribution was available in the United States only once every decade. ACS now offers household size distribution data, at the census tract level in the 5-year data products.

Unfortunately, due to the independent controlling techniques used for population and household variables, and the treatment of group quarters controlling, there can be wide discrepancies between the reported household population and the household population implied by household size distributions. For example, if a census tract contains only 100 households of size 2, the household population is expected to be 200. Unfortunately the reported household population for the tract may vary widely from the expected total based on the distribution of households by size. (Jarosz 2010)

Thus, it becomes useful for demographers to have a technique available for developing household size estimates. This resolves the size discrepancy issue, allows point-in-time estimation, and allows estimation for geographic units smaller than the census tract. The Poisson distribution has been found to produce a reasonable approximation of households by household size.

In the 1990s several researchers published work on using the Poisson distribution to estimate household size at the nation- and state-level in Australia and New Zealand (Jennings, et. al. 1992 and 1999). The only other known use of the Poisson distribution in demography was published in 1955 (Aitchison).

This research puts forth two key advancements over prior published work:

(1) substantially simplified method for applying the distribution,
(2) evidence of the reliability of the technique for estimating household size distributions in small geographic areas (e.g. counties and census tracts).

The Poisson Distribution:

One of the most appealing features of the Poisson distribution for the purpose of modeling household size is that it requires only one parameter, the mean, which is generally known (average household size = household population / households), as described above.

> *The Poisson distribution describes the possible number of objects you'll find in a certain volume, or the number of events you'll observe in a particular time span.*
>
> Motulsky, Intuitive Biostatistics (2010), p53

There are a wide variety of applications of the Poisson distribution in biology and engineering (Lloyd Johnson, et. al. 2005). The Poisson distribution is used in cases where:

- Events can be counted in positive integers, and there is no upper limit on the number of events that may occur.
- Occurrences are independent, random.
- Each event is counted only once, and it is irrelevant how many cases have not occurred.
- Average frequency (or population mean) is known.

As applied to the distribution of household by household size, this means:

- Households are counted in positive integers, and there is no (official) upper limit to household size, as described in more detail below.
- The size of any given household is independent and random. For example, if households A and B are neighbors, the number of people in household A is completely independent of the number of people in neighboring household B.
- Each household is counted only once. It is irrelevant how many households of, for example size 8, have not occurred. It is only relevant how many have occurred.
- As described above, it is possible to estimate and project average household size.

As the criteria above suggest, one of the keys assumptions of the Poisson distribution is that there is no upper limit to household size. While the probability of exceedingly large households is quite low, there is no actual limit to the number of people who might choose to group together into a household. Indeed, U.S. Census data show nearly 2.3 million households of size 7 or larger in the United States in 2010 (1.9 percent of all households)(U.S. Census Bureau, Census 2010). In reporting the data, the U.S. Census Bureau allows from 0 to 97 persons to be counted, with 97 representing the top-coded 97+ category. Excerpting from the documentation:

> *D PERSONS 2 106 107*
> *T Number of person records following this housing record*
> *V 00 . Vacant unit*
> *V 01 . Householder living alone or any person in group quarters*
> *R 02..**97 . Number of persons in household***

U.S. Census Bureau, Public Use Microdata Sample 2000:Technical Documentation (emphasis added)

Coding aside, there are observed examples of very large household sizes. For example in San Diego County the 2000 Census 5 percent PUMS file shows at least one record (case weighted to 19 households) with 18 persons per household. Similarly, in the 2005-09 American Community Survey PUMS file for California, there are seven records (case weighted to 95 households) with a top-code of 20+ persons in the household. Thus, for the purposes expressed in this paper, we will presume that there is no upper limit to the size of household that might exist. In reality and the model-derived results these cases will be rare, indeed.

Given that each of four key criteria can be met, the Poisson distribution is a suitable candidate for estimating the distribution of households by size.

The Poisson Function can be written as:

$$f(n; M) = \frac{M^n e^{-M}}{n!}$$

Where

n = the number of occurrences (where n = 0 -> infinity)
M = the population mean
f(n; M) = probability of case n, when average = M

Applied to average household size, this means that in any geographic region a probability distribution can be estimated for the percent of households of size 0, 1, 2, 3, etc… given the mean household size for that region.

However, households of size zero present a practical problem. In essence, a 0-person household is a vacant unit, and is not considered a "household." The term household only applies to occupied units. (See definition above.) To solve this problem, Jennings, Lloyd-Smith, and Ironmonger (1992) went to extensive lengths to estimate "habitable" vs. "non-habitable" housing structures, and then modified the Poisson model to address size zero households. Their modified version requires regression fitting to find the M parameter.

While this technique yielded close-fitting results (Jennings, et. al. 1992), the method is needlessly complicated. Moreover, modeling 0-person households is logically inconsistent with the very definition of households, which are equivalent to *occupied* housing units. (See definition above.) In practice, vacancy rates (and thus unoccupied households, or households of size 0) are often estimated independently and need not be re-estimated as part of the household distribution (SANDAG, 2011).

A much simpler and equally sound solution is to shift the distribution categories such that households of size 1 become the "0" term, size 2 becomes the "1" term, etc... Thus, to estimate household by size using the Poisson distribution, the user must subtract one from each household size category (n) and also from the mean (M). An example of the modified distribution is illustrated in Table 1, below.

Table 1: Example of Modified Distribution

Avg HH Size	Modified Mean (M = Avg HH Size - 1)	Size 1 → n = 0	Size 2 → n = 1	Size 3 → n= 2	etc…
2.0	1.0	f(0,1) = 37%	f(1,1) = 37%	f(2,1)=18%	…
3.5	2.5	f(0,2.5) = 8%	f(1,2.5) = 21%	...	…
4.0	3.0	f(0,3.0) = 5%	...	...	…

DATA AND METHOD

To test the model, this analysis relies upon the following data sets to compare reported totals against model-derived estimates: state-level data for all 50 states, plus Washington D.C. for the year 1990, and state-level data including Puerto Rico for years 2000 and 2010; California counties from the 2010 Census; census tracts, within San Diego County from the 2010 Census. A summary of the data used is show in Table 2, below.

Table 2: Summary of Data Used in Analysis

Year	Geography	Minimum	Maximum	Range	Cases
1990	states	2.26	3.15	0.89	51 (incl. DC)
2000	states	2.16	3.13	0.10	52 (incl. DC, PR)
2010	states	2.11	3.10	0.99	52 (incl. DC, PR)
2010	counties, California	2.16	3.36	1.20	58 counties
2010	census tracts, San Diego County, CA	1.34	4.95	3.61	622 (excl. military group quarters tracts)

For state-level and county-level analysis, all cases were included in the comparison between model-derived and observed results. For census-tracts five cases were dropped. Within San

Diego County, there are several large military bases. These bases result in a handful of census tracts with large group quarters populations, but fewer than 20 households. For the purposes of this analysis, such small household counts skew the analysis of model error. More importantly, because they are comprised of military households, these tracts can be assumed to violate the principle of random distribution. There are rules regulating eligibility for on-base housing, which reduce or eliminate randomness. In practice, the demographic characteristics for these “special population” census tracts are modeled separately from non-military tracts (SANDAG, 2011).

For each year and geography, predicted results from the modified Poisson technique are compared with the U.S. Census Bureau decennial census counts. The mean algebraic error and the mean absolute error are reported for each year and geography type.

RESULTS

Comparing the model-derived estimates with observed census counts across the five geography/year datasets yields the following results:

Table 3: Mean Algebraic Error

Mean Algebraic Error
modeled percent – observed percent

Year	Geography	Size 1	Size 2	Size 3	Size 4	Size 5	Size 6	Size 7+
1990	states	-4.6%	-0.2%	8.6%	-1.2%	-1.0%	-0.6%	-0.9%
2000	states	-4.8%	-0.5%	8.9%	-0.8%	-1.1%	-0.7%	-0.9%
2010	states	-5.3%	-0.6%	9.2%	-0.2%	-1.2%	-0.9%	-1.1%
2010	counties, California	-5.3%	-2.6%	10.1%	1.3%	-0.6%	-0.9%	-1.9%
2010	census tracts, San Diego County, CA	-3.5%	-2.4%	7.3%	0.3%	-0.1%	-0.2%	-1.4%

Across states within the United States and within counties and census tracts in California, households of size 1 are generally under-estimated using the Poisson distribution estimation technique while households of size 3 are generally over-estimated. This is demonstrated by the fact that the magnitude of the mean algebraic error (Table 3) is the same as the mean absolute error (Table 4).

Table 4: Mean Absolute Error

Mean Absolute Error
abs(modeled percent – observed percent)

Year	Geography	Size 1	Size 2	Size 3	Size 4	Size 5	Size 6	Size 7+
1990	states	4.6%	1.2%	8.6%	1.6%	1.1%	0.6%	0.9%
2000	states	4.8%	1.4%	8.9%	1.2%	1.2%	0.7%	0.9%
2010	states	5.3%	1.5%	9.2%	0.8%	1.2%	0.9%	1.1%
2010	counties, California	5.4%	2.9%	10.1%	1.5%	0.8%	0.9%	1.9%
2010	census tracts, San Diego County, CA	4.4%	3.8%	7.3%	2.5%	1.0%	0.7%	1.6%

All other household sizes show relatively small estimation error, less than 2 percentage points, in either direction for state-level estimates. Error increases as the size of the population decreases, with county-level and tract-level estimates showing errors generally higher than state-level estimates.

CONCLUSIONS AND APPLICATIONS

Tests on U.S. Census data (1990-2010) suggest that the modified Poisson distribution model is suitable for use in estimating the distribution of households by household size at the state, county, and census tract level. The model tends to under-estimate 1-person households, and over-estimate 3-person households. However, these known errors can be reduced through implementation of an error-term adjustment process. The model tends to fit well for households of size 2 and for sizes 4 and larger.

In applying this method to household size estimates, the San Diego Association of Governments modifies the process as follows: an initial distribution for the number of households, across each household size bin, is estimated using the modified Poisson method; the initial household values in each size bin are multiplied by an error adjustment factor, calibrated against the most current census year of data available; an iterative proportional fitting procedure ensures that both the sum of households across size terms matches the known local household total, and that the sum across geographies (e.g. tracts) matches the known regional household total for each size bin. These adjustment factors help to reduce error in the final estimates.

REFERENCES

Aitchison, John. 1955. "On the Distribution of A Positive Random Variable Having a Discrete Probability Mass at the Origin." Journal of the American Statistical Association. Vol. 50. No. 271. (Sept 1955) http://www.jstor.org/pss/2281175

Jarosz, Beth. 2010. "Household Size in the American Community Survey." Presentation for the MPO/COG Mini-Conference on Socio-Economic Modeling. San Diego, CA.

Jennings, Vic and Bill Lloyd-Smith. 1992. "Household Change, Distribution of Household Size, and Fertility Rates." Presented at the Australian Population Association Conference. Sydney.

Jennings, Vic, Bill Lloyd-Smith, and Duncan Ironmonger. 1999. "Household Size and the Poisson Distribution." Journal of the Australian Population Association. Vol 16. Nos 1/2.

Lloyd Johnson, Norman, Adrienne W. Kemp, Samuel Kotz. Univariate Discrete Distributions: Third Edition. Wiley-Interscience. Hoboken, NJ. 2005.

Motulsky, Harvey. Intuitive Biostatistics. (Second Edition) Oxford University Press. New York, NY. 2010.

SANDAG. CONCEP 2010 (estimates model documentation). April 2011. www.sandag.org

U.S. Bureau of the Census. "Public Use Microdata Sample 2000: Technical Documentation." October 2008. http://www.census.gov/prod/cen2000/doc/pums.pdf Accessed on Sept. 15, 2011.

DATA SOURCES

Census 2010, tables H12 and H13 www.census.gov (downloaded August 16, 2011)
Census 2000, tables H12 and H13 www.census.gov (downloaded August 16, 2011)
Census 1990, tables H017 and H017A www.census.gov (downloaded August 16, 2011)