Census Evaluation and Household Size Distributions

Griffith Feeney

Demography - Statistics - Information Technology, Letter No. 5, 15 March 2014

Figure 1 shows a distribution of households by number of persons. The example is hypothetical, but based on real data examples.

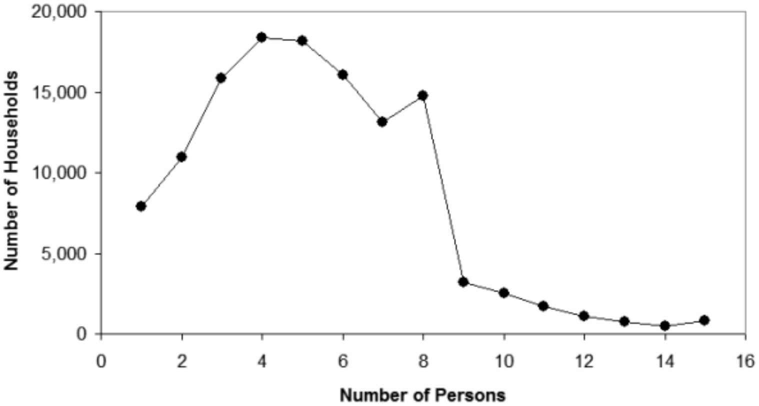

Figure 1 Illustrative Census Household Size Distribution

The 8-person household spike is striking and anomalous. The true household size distribution almost certainly does not look like this. How to explain the spike?

If the data were collected on household forms providing for eight persons per schedule, there are two possible explanations.

First, data processing may have failed to match some continuation forms for households of more than 8 persons to the forms they continue. This results in some 9-person, 10-persons ... households being identified as two households, an 8-person household and a 1-person, 2-person, ... household.

Second, enumerators may in some cases may have neglected to complete a continuation form for a household of more than 8 persons. This results in a 9-person, 10-person, ... household being identified as an 8-person household.

It may be possible to correct data processing errors by reprocessing. Correcting field work errors usually requires returning to the field, which is not generally done.

Let us assume that any data processing errors have been corrected, so that the spike is the result of fieldwork.

The challenge is then to devise a way to (a) decide how many of the households identified as having 8 persons are in fact 9 or higher person households and (b) how to distribute these larger households to 9-person, 10-persons, ... households.

1. Modelling the error

Let p denote the proportion of households incorrectly reported as having 8 persons.

The probability that an enumerators does not complete a continuation questionnaire may be plausibly assumed to decline with the number of persons who will be omitted as a result. A continuation form is more likely to be omitted for a 9-person household, for example, than for a 15-person household.

To capture this decline with a single parameter, suppose that omission of continuation forms for larger households follows a geometric series. Specifically, if N denotes the number of 9-person households wrongly identified as 8-person households, then, the number of 10-person households wrongly identified as 8-person households will be Nr, the number of 11-person households wrongly identified 8-person households Nr^2, and so on, where r is a number between zero and one.

The next step is to estimate the parameters p and r from the observed distribution and use them to calculate an adjusted distribution.

2. Minimum roughness

We want parameter values p and r that give an adjusted household size distribution with no spike. That is to say we want a *smoother* distribution, which means we want a *less rough* distribution. Our approach will therefore be to (i) define a measure of "roughness" and then (ii) choose parameter values to minimize this measure.

For each point in Figure 1 except the first and the last, calculate the difference between the point and the average of the adjacent points, square the difference, and sum these squared differences. This defines a measure of roughness $R(p, r)$.

Finding values of p and r that minimize this measure is easily accomplished by numerical minimization. We may for example construct an Excel spreadsheet with cells for the observed distribution, the parameters, the adjusted distribution, and the measure of roughness, and then use SOLVER to find parameter values that minimize roughness.

For the observed distribution in Figure 1 this gives $p = 0.347$ and $r = 0.389$. The adjusted distribution is calculated by (A) removing p times the reported number of 8-person households out of this group and (B) distributing these households to 9-person, 10-person, ... households in such a way that the number allocated to (i+1)-person households is r times the number of i-person households, i = 9, 10, ..., 15.

The adjusted household size distribution is shown in Figure 2 together with the original distribution.

The rightmost plotted point represents households of 15 or more persons, which is why it is slightly higher than the point to the left. The model could be extended to incorporate omission of continuation forms for 17-24 person (and larger) households, but the value added will usually be low.

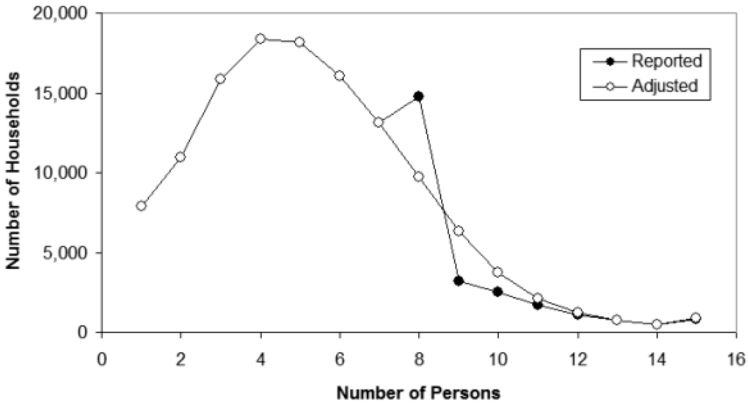

Figure 2 Observed and Adjusted Household Size Distribution

3. Implied census omission

A household size distribution implies a number of persons, namely 1 times the number of 1-person households plus 2 times the number of 2-persons households plus 3 times the number of 3-person households times 3 and so on.

In this example the reported household size distribution gives 661,663 persons and the adjusted distribution gives 669,667 persons, implying an omission of 8,334 persons, or 1.24 percent of the adjusted number of persons. This is not a very large omission, but neither is it negligible.

4. Conclusion

Evaluation of population census results should include scrutiny of household size distributions. A spike at the maximum number of persons accommodated by a household schedule signals a likely error in data processing and/or field work.

Errors due to data processing should be corrected, if possible, by reprocessing. If returning to the field to complete omitted continuation forms is not possible, the model presented above may be used to adjust the household size distribution and estimate the census omission of persons resulting from the omission.

5. Resources

The example in this letter is contained in this spreadsheet. The Excel SOLVER Add-In may not installed by default, but it should be installable using Tools > Add-Ins.

Issue No. 5 of the The DEMOGRAPHY, STATISTICS, and INFORMATION TECHNOLOGY letter, 15 March 2014. To subscribe send email to feeney@gfeeney.com with "subscribe" in the subject field.