

THREE DIMENSIONAL MODELLING OF HOUSEHOLD SIZE DISTRIBUTIONS

Vic Jennings⁺, The University of Melbourne

Bill Lloyd-Smith, RMIT University

Duncan Ironmonger, The University of Melbourne

Consistent discrepancies between the Poisson model and observed household size distributions were noted in Jennings, Lloyd-Smith and Ironmonger (1999). This paper explores how these discrepancies may arise using household size distribution data for 35 countries stratified by the age of household reference persons and household size. A three dimensional model, the Poisson-gamma reference frame, is used to map the changes of household size distribution over the life course. A significant source of the discrepancies arises because the sum of the observed household size distributions over the life course is similar to the sum of Poisson distributions. This sum is related to the incomplete gamma function rather than an overall version of the Poisson distribution. It is shown that this effect becomes negligible for countries with large average household size but is significant for households of sizes 1 to 3 for countries with small average household size. Improved models are described.

Introduction

The aims of this paper are threefold:

To formalise the modelling of the discrete-continuous relationships of household size distributions noted in Jennings, Lloyd-Smith and Ironmonger (1999) using the Poisson-gamma reference frame. Secondly to show that for 35 countries, household size distributions classified by age of household reference person have Poisson distribution characteristics. Finally to show that the sum of such distributions, which were equivalent to the overall distribution for a country, were better modelled by the sum of Poisson distributions over a parameter range rather than a single parameter Poisson distribution. This could in the limit be derived from the difference between two incomplete gamma function values over the same range.

There is a large literature on Poisson distributions and gamma distributions. Tables which show together the incomplete gamma function and cumulative sums of the Poisson distribution have been available since at least 1965, see Abramowitz and Stegan (1965:978-983). Tables of the incomplete gamma function have been available as early as 1922, see Pearson (1922). There is also a developing literature on count data analysis and estimation, see for example Winkelmann (2000). This paper applies some of the Poisson-gamma geometric properties to mapping and analysing a large set of household distributional data.

The Poisson distribution describes the outcome of the random allocation of persons to households. The family of gamma distributions (scale parameter equals one) describe how the Poisson distribution changes as the Poisson parameter changes. These distributions are combined to give the Poisson-gamma reference frame, enabling 455 observed distributions to be compared geometrically and compactly. The frame is used

⁺ Address for correspondence Households Research Unit, Economics Department, University of Melbourne, ,Victoria 3010,Australia Email: dsi@unimelb.edu.au

both as a reference for household size distributions, and as a filter for subtracting noise (random) from the distributions leaving a signal to be interpreted. It is recognised that there are limitations to the data. There is a degree of indefiniteness in the definition of the age of a household, and the procedure for calculating the number of persons in size 6+ households varies between countries. Nevertheless the data shows consistent patterns.

The household data used here is from the United Nations and the Australian Bureau of Statistics. The data consist of a three-way stratification of household populations. Firstly into 35 countries. Secondly into households defined by the age of the household reference person, the ages being 15-19, 20-24, ..., 75+ so that there are 13 age groups in all. Finally into households according to household size which has values 1,2,3,4,5,6+. There are a total of 455 household size distributions and 2730 cells where each cell has the dimensions of country, age and size. Age is different to the other categories since it is infinitely divisible. For a description of data sources and a discussion of the use of the 'country' as the unit for investigation see Appendix 1.

Households are distinguished in this paper by their household reference person (HRP). For a more extensive definition of the HRP see United Nations (1997:107). This person is usually an adult who is a member of the household, and who is responsible for signing off on the census form. Adults in a household of a particular age are likely to be of a similar age to the HRP. Census data and general observation confirm clustering by age. The Australian Bureau of Statistics (ABS) 1986 Census one percent sample tape, ABS (1986) provides illustrations; see example in Table 1. For females of a given age the majority of male partners will have ages within a limited number of years of that age. For example, of the females aged 35 years to 39 years in partnerships with males, 64 per cent had male partners in the range 35 and 44 years. The age of the household is therefore defined in this paper to be the age of the HRP. It is not implied that all adults of a particular age group are associated with HRPs of that same age group. For example some persons aged 30-34 years may be members of households with HRPs of 25-29 years

Table 1 **The proportion of couples with similar ages given the age of the female**

Female given age	Male partner age range	Couples Per cent	Female given age	Male partner age range	Couples Per cent
15-19	15 - 24	79.5	45 - 49	45 - 54	71.4
20-24	20 - 29	87.4	50 - 54	50 - 59	77.3
25-29	25 - 34	83.2	55 - 59	55 - 64	76.9
30-34	30 - 39	73.2	60 - 64	60 - 69	76.5
35-39	35 - 44	64.0	65 - 69	65 - 74	74.0
40-44	40 - 49	62.8	70 - 74	70 - 79	84.4
			75+	75+	79.4

Source: Australian Bureau of Statistics (1986a): One per cent sample of the 1986 Australian Census and author's calculations.

Poisson-gamma reference frame

In Jennings, Lloyd-Smith and Ironmonger (1999) the distribution of persons to households was represented by a model where people (balls) were allocated to dwellings (cells) randomly. The resulting Poisson distribution was truncated at zero because of inadequate data on vacant dwellings. In this paper only occupied dwellings are considered, that is households with one or more persons. One of these persons is designated the household reference person (HRP). The distribution of persons is generated by a model where all additional persons (balls) are allocated randomly to these HRPs (cells). These cells are then classified according to the number n of additional persons allocated to them. This number is a variate with values $n = 0, 1, 2, \dots$. The resulting distribution can be shown under suitable conditions to approximate to the Poisson distribution for a large number of cells where λ is the mean number of additional persons per cell. See equation (1)

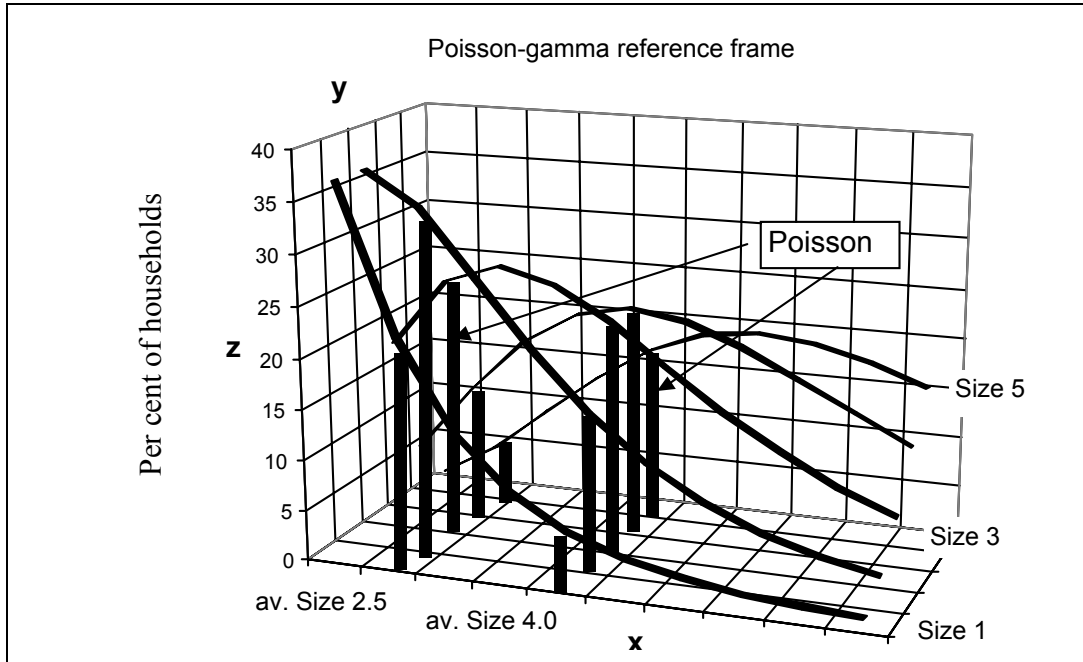
$$P(n | \lambda) = \frac{e^{-\lambda} \lambda^n}{n!} \quad (1)$$

where $P(n | \lambda)$ is the probability of a household containing n persons additional to the HRP for a given λ . The definition is extended to include $P(0 | 0) = 1$, the distribution concentrated at zero, ie all households are of size one. It also includes $P(\infty | \infty) = 1$, the distribution concentrated at infinity, ie one household (taken as equivalent to a country) contains all persons. See Kingman (1993:3-4).

The Poisson distribution described above can be directly computed, unlike the truncated Poisson which requires a look up table. For a 112 country data base similar to that used in Jennings, Lloyd-Smith and Ironmonger (1999) the Poisson distribution shown as equation (1) together with linear adjustments has a residual error about the same as the truncated Poisson distribution with linear adjustments. The model described by equation (1) is therefore used in this paper and applied to the 35-country data. For each country there are 13 age groups for HRPs. Corresponding to each age group there is a household size distribution. To encompass in a compact form the 455 distributions and differences between them the Poisson theory is extended to include the gamma distribution. See Figure 1.

Figure 1 shows the relationships between the Poisson distribution and the family of standard gamma distributions (scale parameter one). Two Poisson distributions are shown, one with an average household size of 2.5 and the other with an average size of 4.0. These two distributions have parameters at the lower and upper end of the range typically seen in countries with average household size about 3.0. As average household size changes then the ordinates of the household size distributions describe the family of standard gamma distributions shown as curves. The word 'frame' is used because the family of curves and the relationships between them are fixed, and this arrangement is analogous to a spatial frame. The Poisson distribution and the gamma distribution are related as follows.

Figure 1 Poisson-gamma reference frame



If k is the number of persons in a household then the number of persons additional to the HRP is given by $n = k - 1$. If the average household size of a population of households is μ and each household includes one HRP, then there are on average $\mu - 1$ additional persons per household ie $\lambda = \mu - 1$. Thus the distribution of these additional persons defines the distribution of the households. If the distributions are Poisson and $F(k; \mu) = \text{Prob}(\text{Household has } k \text{ persons, given that } \mu = \text{average household size})$ then

$$F(k | \mu) \equiv P(k - 1 | \mu - 1) = P(n | \lambda) \quad (2)$$

where $P(n | \lambda)$ is a discrete function of n . On the other hand, the probability density of those households as the mean λ varies, given households have n additional members is a continuous function of λ described as $g(\lambda | n + 1)$. Thus

$$P(n | \lambda) = e^{-\lambda} \left(\frac{\lambda^n}{n!} \right) = \frac{1}{n!} (\lambda^n e^{-\lambda}) = g(\lambda | (n + 1)) \quad (3)$$

Using the notation for the standard gamma distribution (scale parameter equals one) put $\lambda = x$ and $n = \alpha - 1$ then with $\Gamma(\alpha)$ as the gamma function, and taking α as a positive integer $\Gamma(\alpha) = (\alpha - 1)!$.

$$g(\lambda | n + 1) = g(x | \alpha) = \begin{cases} \frac{x^{\alpha-1} e^{-x}}{\Gamma(\alpha)} & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

and this is a standard gamma distribution. See for example Devore (1991:157).

Thus any ordinate $P(n | \lambda)$ determined by the Poisson distribution is also simultaneously the ordinate $g(x | \alpha)$ determined by the gamma distribution. Hence the family of curves generated by the standard gamma distribution with shape parameters $\alpha = 1, 2, \dots$ can be used to generate all Poisson distributions $P(n | \lambda)$ for $n = 0, 1, 2, \dots$ for $\lambda \geq 0$, and by equation (2) all $F(k | \mu)$. It is noted that since $n = k - 1$ and also $n = \alpha - 1$ that $k = \alpha$, the shape parameter used in the cumulative distribution function tables of the gamma distribution, the incomplete gamma function tables.

This limited form of the gamma distribution family was termed the functional form of the Poisson distribution in Jennings, Lloyd-Smith and Ironmonger (1999) but is referred to here as the Poisson-gamma reference frame. The family of curves are grouped together and shown in Figure 1 and Figure 4f. This family of curves is equivalent to the contours of a general surface where α , the shape parameter is continuous and greater than zero, thus the more general model is the Poisson-gamma reference surface. Sometimes in the text Poisson-gamma will be used to describe a value which can be used either as a discrete Poisson quantity or a continuous gamma quantity.

Referring to Figure 1 consider a very large number of Poisson distributions between size 2.5 and 4.0 and that they are equally spaced, and consider their ordinates for size $k = 1$, and the associated gamma distribution curve with shape parameter $\alpha = 1$. Then the Poisson mean ordinate for size one will be equal to the total area under the gamma distribution curve for size one divided by $(4.0 - 2.5)$. But the average household size over this range will be $(2.5 + 4.0) / 2 = 3.25$. The Poisson ordinate for average size 3.25 will be less than the mean ordinate calculated using the gamma since the gamma distribution curve for size one is concave. On the other hand the gamma distribution curve for size three is convex over this range so that the opposite effect is expected for size three. It is on the basis of these observations that it is expected that the overall observed household size distribution for a country is more closely matched to the sum of Poisson distributions for age groups for that country than it is to the Poisson distribution for the country as a whole. This general approach is discussed in more detail in Appendix 2.

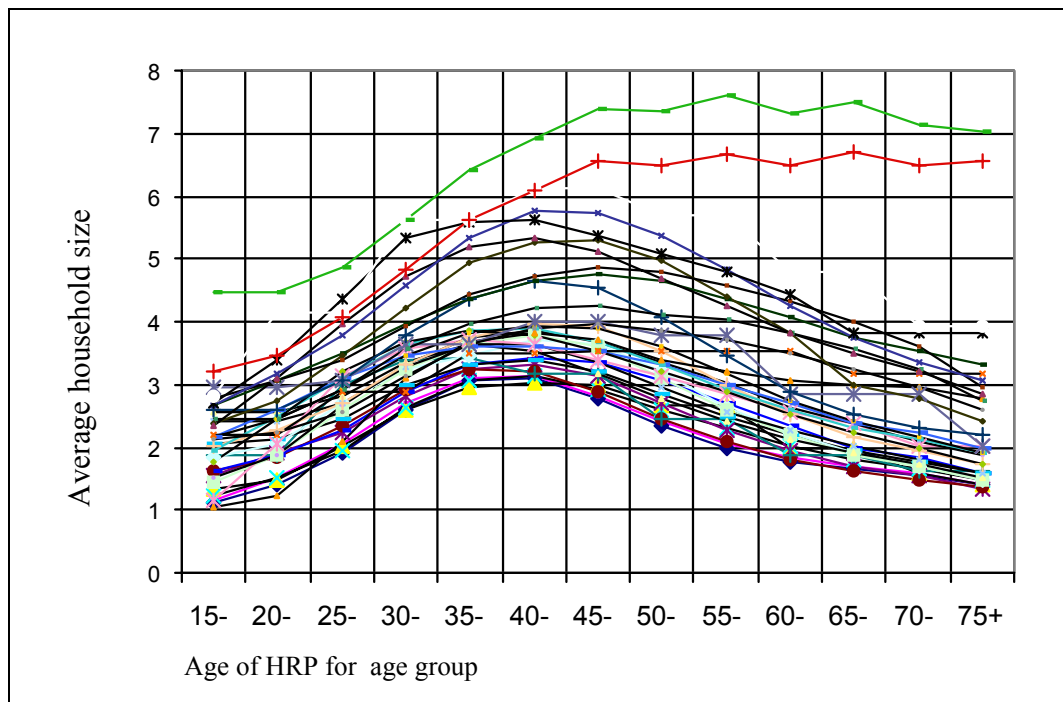
The word 'reference' is used because there is by definition only one frame satisfying equation (3) and (4). There is a unique Poisson distribution for each value of λ . The frame enables the compact comparison of large sets of distributional data which are modelled using either the Poisson distribution or the standard gamma distribution (scale parameter one) or both. For example the 455 distributions used in the 35-country data. Poisson distributions are readily calculated and there are available tables of the Incomplete gamma function for calculating cumulative values of the gamma distribution. See for example a small table in Devore (1991:675) or a more complete Table 26.7 in Abramowitz and Stegun (1965). It is also possible to simulate the fitted distribution. See Stuart and Ord (1987) in sections 9:22 to 9:24.

Some continuous characteristics of the 35-country data

Since the number of households in a country is large it is expected that changes in average household size over time or as the age of household reference person changes should be near continuous. Jennings, Lloyd-Smith and Ironmonger (1999) showed in Figure 1 that for 104 countries the percentage of households of a particular size was related to the average household size for the country. This section of the paper sets out to demonstrate further continuous characteristics of household size distributions, and the relationships of the proportions of households of a particular size to the gamma distribution.

In Figure 2 for each of the 35 countries in the 35-country data the average household size for each of the individual age groups is plotted against age of HRP. Ordinates for adjacent ages for a particular country are joined. A consistent pattern is seen.

Figure 2 Change in average household size of age groups with age^a progression for 35 countries^b around 1990.



- a Age of age groups is defined as the age of the household representative person (HRP).
- b Countries are ordered from the bottom by average households size. The order is shown in Table 2.

For most of the 35 countries average household size of age group increases until about age 40-44 and then decreases. Maximum less minimum average household size is generally about 1.9. Those countries with lowest average household size such as Sweden are represented by curves at the bottom of the stack of curves, and countries with high average household size such as Philippines are near the top of the stack. The maximum average household size for the lower curves occurs at about age 40. For those countries

near the top of the stack the maximum is around age 45. Although the curves are roughly symmetrical about a vertical axis at age 40-44 years they show some slight skewness. Bangladesh and Pakistan in 1981 show no drop in average household size with increasing age. All other countries show a drop from about age 40-44 years. The average household size for each of the countries described in Figure 2 is shown in Table 2. Each country is shown with an abbreviation for the year of the census at which the data was collected.

Over time and as discussed in Jennings, Lloyd-Smith and Ironmonger (1999) the average household size for most countries has lowered with a trend to smaller households. For example average household size for the total population of Australia has moved down from 3.55 in 1961 to 2.79 in 1991. See yearbooks ABS(1988:266) and ABS (1995:100). Thus the curves shown in Figure 2 will also move down but will probably retain their shape since the same life course pattern is likely.

Table 2 Average household (HH) size for a country in the year in which the data was collected

Country	Av. HH Size	Country	Av. HH Size	Country	Av. HH Size
Sweden 90	2.13	Australia 91	2.84	Spain 81	3.53
Denmark 91	2.23	Bulgaria 85	2.94	Macau 91	3.69
Germany 87	2.33	Greece 91	2.97	Cuba 81	4.11
Switzerland 90	2.34	Italy 81	3.01	Reunion 82	4.27
Norway 90	2.40	Slovenia 91	3.07	Mauritius 90	4.28
Finland 90	2.42	Romania 92	3.07	Bolivia 92	4.37
Hungary 90	2.60	Poland 88	3.10	Brazil 80	4.67
Luxembourg 91	2.67	Portugal 91	3.12	C African R 88	4.73
Canada 91	2.67	Japan 85	3.14	Philippines 90	5.31
Austria 81	2.70	Neths. Antilles 91	3.30	Bangladesh 81	5.75
France 82	2.70	Guadeloupe 90	3.41	Pakistan 81	6.71
New Zealand 91	2.76	China HK 91	3.43		

The change in average household size over the life course reflects varying proportions of different size households. For the 35 countries the ratio of HRPs aged 30-59 years to adults aged 30-59 in households is 0.5138 with standard error of 0.0459. Thus on average households of larger sizes than two in the age range 30-59 will usually have as additional members young persons under 30 years of age. The life course trends in household size distribution shown for Australia 1991, see Figure 3a, 3b are fairly typical of many countries. Figure 3a shows this change with the HRP being used to class households by age group. It particularly shows the small number of age 15-19 households relative to the number of households of other ages. Figure 3b shows the per cent of households given the age group chosen. The points are joined representing the near continuity in change over the life course. For groups with age 25-29 or more a similar pattern occurs in both Figure 3a and 3b. This similarity continues to age 75+(allowing for the summation of various age groups in the 75+ classification).

The general pattern shown in Figure 3b is repeated for most of the 35 countries and over the range 20-24 years to 60-64 years is roughly symmetrical about ages 40-44 years. With increasing age and the acquisition of children the proportion of larger households

increases until a maximum is reached at about age 40-44 years. Above age 40-44 years average household size declines as young adults leave home and the distribution reverts to its former shape. This peak and then a decline is consistent with the observation that on average children live at home for about twenty years while the child bearing period for women extends over twenty years. For ages 70+ mortality is a major factor.

Size two and size three are bimodal. Size two is the most popular size for young and old persons except for age 75+ where the proportions of size one exceeds size two. Size three has a local maximum at about age 30-34 when average household size is increasing and a local maximum at about age 50-54 years when average household size is decreasing. Size three can therefore be seen as representing a transitional size for many households as they grow larger or smaller. The graphs show a small amount of skewing. For countries such as the Philippines with large average household size there is a higher proportion of large households at age 40-44 than for Australia.

The curves to the left of age 40-44 roughly correspond to the curves shown in Figure 1. The curves to the right of age 40-44 years up to age 69 years are approximately the mirror image of those on the left side. If Figure 3b is folded along a vertical axis at age 40-44 and the age categories are replaced with a continuous variable, average household size, then curves of the form of those in Figure 4 are obtained. For clarity of presentation Sweden, which has a lower average household size replaces Australia, and is compared with the Philippines, which has a high average household size. Sweden is at the bottom of the stack in Figure 2 while Philippines is near the top of the stack. Sweden has rather similar patterns to those shown in Figure 3. The observed data are plotted against the appropriate Poisson-gamma reference frame curves with shape parameters $\alpha = 1, 2, 3, 4, 5$ which correspond to household sizes $k = 1, 2, 3, 4, 5$. See Figure 4. Generally the distributions for the 35 countries range between that of Sweden and the Philippines following the reference frame.

Graph points for a particular country are linked by order of age, from group aged 15-19 to group aged 75+ years. The approach adopted is rather similar to the hypothetical birth cohort of woman used in calculating the total fertility rate, see Hinde (1998:102). Thus we obtain curves representing the age progression of a hypothetical group through the 13 age strata. The Poisson-gamma reference frame shown in Figure 4f incorporates all the Poisson-gamma curves in Figures 4a to 4e.

For a particular size beginning at age 15-19, there is a progression to the right, the out path, to about age 40-44 years with increasing average household size. Then there is a return towards the starting point, the return path, as average household size reduces again. Generally the out path to age 40-44 differs from the return path to age 69 years. Thus the proportion of households for an age group depend not only on average household size but also on the age order of the HRP age group. Since adjoining age groups come from adjoining birth cohorts they are not completely independent of each other.

Figure 3a Australia 1991: Observed per cent of all household reference persons (HRP) by household size by age of HRPs.

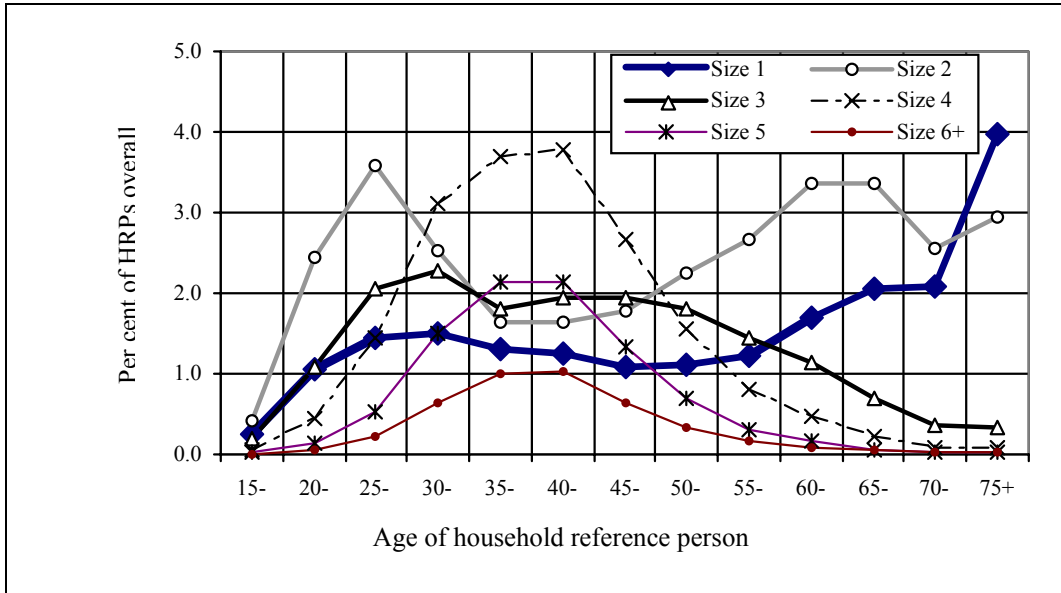
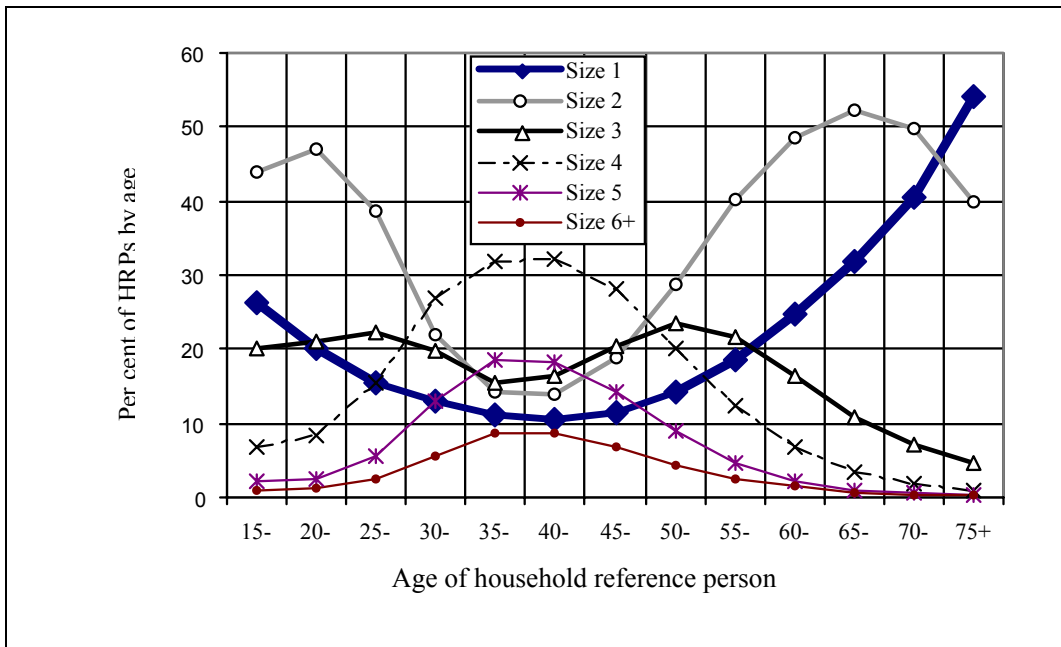


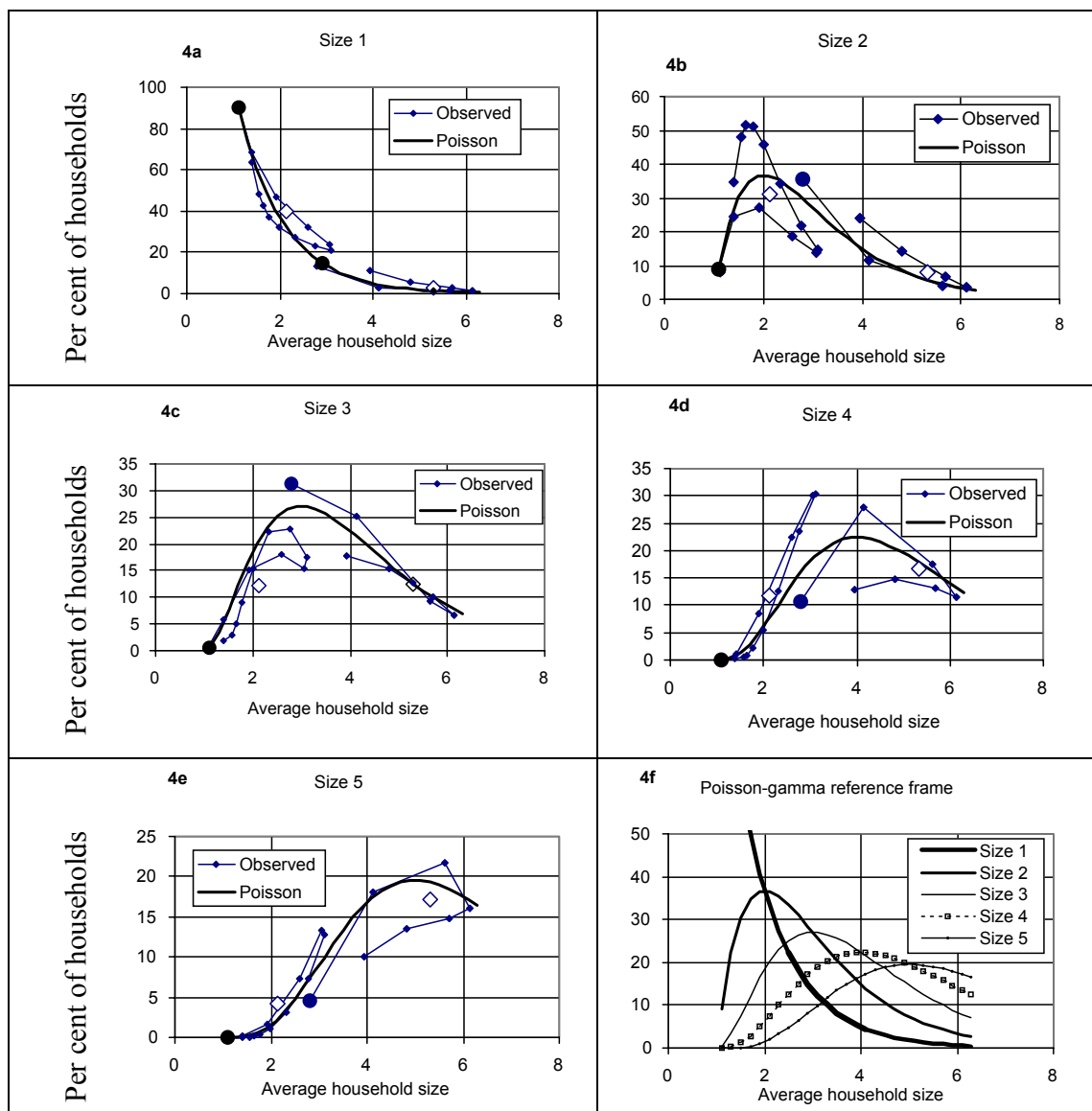
Figure 3b Australia 1991: Observed per cent of household reference persons of a given age for a given household size by age of HRPs.



Source: Australian Bureau of Statistics (2001)

The location and shape of these household circuits show similarities to the corresponding gamma distribution curves. The open diamonds indicate at the overall country level the proportion of households of a particular size. The left diamonds represent Sweden and the right the Philippines. Graphs have been plotted for all 35 countries and similar patterns are seen.

Figure 4 Life course circuits for household reference persons by average household size of age group from age 15 to 75+. Sweden 1990 is in left circuit, and Philippines 1990 is in right circuit. Start group is age 15-19 and is marked by the black dot in each case. Finishing point is age 75+ for Sweden and 70+ years for the Philippines. The open diamonds indicate the observed per cent of households size k at the country level.



Certain properties of the frame shown in Figure 4f are useful for identifying a data set as being consistent with the Poisson-gamma, or in comparing different distributions which have an underlying Poisson distributional form. These properties are as follows:

1) Adjacent curves in the family of curves in Figure 4f are related as follows. If λ is a positive integer and $\lambda = n$ then

$$P(n-1 | \lambda) = e^{-\lambda} \left(\frac{\lambda^{n-1}}{(n-1)!} \right) = e^{-\lambda} \left(\frac{\lambda^{n-1}}{(n-1)!} \right) \frac{\lambda}{n} = P(n | \lambda) \quad (5)$$

ie if a household size distribution has parameter λ equal to integer n then the proportions of households with n additional persons will be the same as the proportions of households with $n-1$ additional persons.

2) For the first differential of $P(n | \lambda)$ with respect to λ

$$\frac{d}{d\lambda} P(n | \lambda) = P(n | \lambda) \left(\frac{n-\lambda}{\lambda} \right) \quad (6)$$

so that the curves for $n \geq 1$ in Figure 4f show a maximum when $\lambda = n$ and since $n - \lambda = k - \mu$ this also corresponds to a maximum when $k = \mu$. For example the Philippines has a turning point for size four at about average household size four.

3) Consider the case of an observed distribution of the form $y_n = C_n + P(n | \lambda)$ where C_n is a constant for households with additional persons n and $\sum_n C_n = 0$. Then if values

of the observed distribution for y_n are obtained for various values of λ and hence $P(n | \lambda)$ the constants can be estimated and the observed distribution modelled using the Poisson. If only one set of distributional data is available then constants can obscure the underlying Poisson nature of a distribution. For example in Jennings, Lloyd-Smith and Ironmonger (1999) there were shown to be discrepancies between the Poisson and the observed distributions of sizes of households at a particular point in time. Graphing the data against average household size for different time periods showed consistency in the discrepancies indicating a constant as well as a variable. See Figure 3 in Jennings, Lloyd-Smith and Ironmonger (1999). In Figure 4 age groups serve the same function as time based distributions. They enable linear changes to be distinguished from Poisson-gamma like changes.

4) Each intersection of curves on the reference frame also defines a modal range for the corresponding Poisson distribution. Thus referring to Figure 4f if $n = 2$ which corresponds to size 3 the modal value is size 3 over average household size $\mu = 3$ to $\mu = 4$. Conversely if the mode is known to be size k then the average size will lie between k and $k+1$. In the case of graphs 4b to 4e the observed modes are not far from being consistent with this Poisson-gamma characteristic.

5) Using proportions instead of percentages the area under each of the gamma curves in the frame is one, while the sum of the ordinates of the Poisson distribution at any mean value λ is also one. This relationship only applies for the standard gamma distribution (scale parameter one). Thus in both the Poisson and gamma planes we can interpret ordinates as probabilities, see (Devore 1991:157). In practice the ordinates for the curve for $n = 5 +$ is calculated as 100 less the sum of the ordinates for $n = 1, 2, 3, 4$.

Summarising this section if average household size for a country changes then it will do so gradually over time. The proportions of households of various sizes also change gradually over time, as for example for Australia, as described in Jennings, Lloyd-Smith and Ironmonger (1999). However there are greater changes at the age group level. This is observed in Figure 2 where average household size for most of the 35 countries increases by about two persons per households and then decreases again with age. Figure 3 indicates that underlying this fluctuation is an even greater degree of variation in the proportions of various size households. Figure 3b shows the symmetry of the proportions of younger and older persons in some of the household sizes. Figure 4 shows clearly the turning points in the proportions of households of particular sizes with age and that the distributions roughly follow the Poisson-gamma reference frame.

Comparing household size distributions for individual countries with estimation models

The aim in this section is to establish the degree to which age group household size distributions for each of the 35 countries are Poisson like. To do this a comparison is made between estimates from three models. The distributions used for comparison are the percentage of households of size 1,2,3,4,5,6+ called the ordinates. The discrepancy $D(AG)$ is measured by the standard deviation of the six values of observed less estimated ordinates.

The measure used to summarise the discrepancies for the age groups for a country is taken to be the square root of the mean of the sum of the squares of the $D(AG)$ s for the oldest 12 age groups for a country and this is termed $D(sAG)$. Age group 15 to 19 is left out of this section of work since the number of HRP's in this age group is about 10 per cent or less of the other age groups as seen in Figure 3a. Hence to include it would bias the overall measure of variability. The $D(AG)$ s are not weighted, for example according to the number of households in an age group, since the aim here is to compare age group distributions with Poisson models rather than to estimate an overall result.

The first model used is the 'base' model where it is assumed that all ordinates have the same value, the mean ordinate, which in this case is $100/6 = 16.67$. The standard deviation of the (observed less mean ordinate) is equal to the standard deviation of the observed. To be useful the other two models should show a lower discrepancy than that of the base model. The ideal model would have a value of D equal to zero.

In the second model the Poisson distribution is used to estimate the distribution of households for an age group with parameter equal to average household size less one as in equation (1). This is termed the *Poisson* (AG12) model.

In the third model the same Poisson model is used as above together with an additional linear regression component and is termed the *Poisson* (AG12+L) model. This necessitates some calculations as follows:

Separate linear models for each of the six different sizes are fitted to the observed less Poisson values for the 12 data points. The two variables used are average household size less one for each age group, and the age order for each age group. Age order is used to distinguish between the out path and the return path as noted in Figure 4.

The Poisson estimate for the proportion of households as given in equation (2) is

$$F(k | \mu) = P(n | \lambda) = \frac{e^{-\lambda} \lambda^n}{n!} \quad n, \lambda \geq 0 \quad (7)$$

where $k = n + 1$ is the number of persons in the household and $\mu = \lambda + 1$ is the average household size for the corresponding household population. If t is the order number of an age group, starting with order number one for age group 15-19 and finishing with order number 13 for 75+ then

$$F(k | \mu_t) = P(n | \lambda_t) = \frac{e^{-\lambda_t} \lambda_t^n}{n!} \quad \text{for age group with order number } t. \quad (8)$$

For households with n additional persons the difference is taken between the observed proportions of households and the fitted Poisson value over the range λ_t for values of $t = 2, 3, \dots, 13$. This difference is regressed on λ_t and t as a linear function. As mentioned earlier the case for $t = 1$ is excluded. In symbols with suitable constants α_n, β_n and γ_n the least squares regression equations obtained are

$$\hat{O}_n - P(n | \lambda_t) = \hat{\alpha}_n + \hat{\beta}_n \lambda_t + \hat{\gamma}_n t \quad (9)$$

where

$$\sum_n \hat{\alpha}_n = \sum_n \hat{\beta}_n = \sum_n \hat{\gamma}_n = 0 \quad (10)$$

Take $\hat{\alpha}_n = a_n, \hat{\beta}_n = b_n, \hat{\gamma}_n = c_n$. Then if there are household with sizes 1 to 6+ so that n ranges from 0 to 5+ there are six linear equations per country each containing three coefficients. Any five of the linear equations are independent. The regression equations become

$$O_{nt} - P(n | \lambda_t) = \hat{O}_n - P(n | \lambda_t) + O_{nt} - \hat{O}_n = a_n + b_n \lambda_t + c_n t + \varepsilon_{nt} \quad (11)$$

where λ_t is the average household size less one for age group with order t and ε_{nt} is the error term for age order t and $\sum_t \varepsilon_{nt} = 0$ and $\sum_n \varepsilon_{nt} = 0$ since ε_{n1} is taken equal to 0.

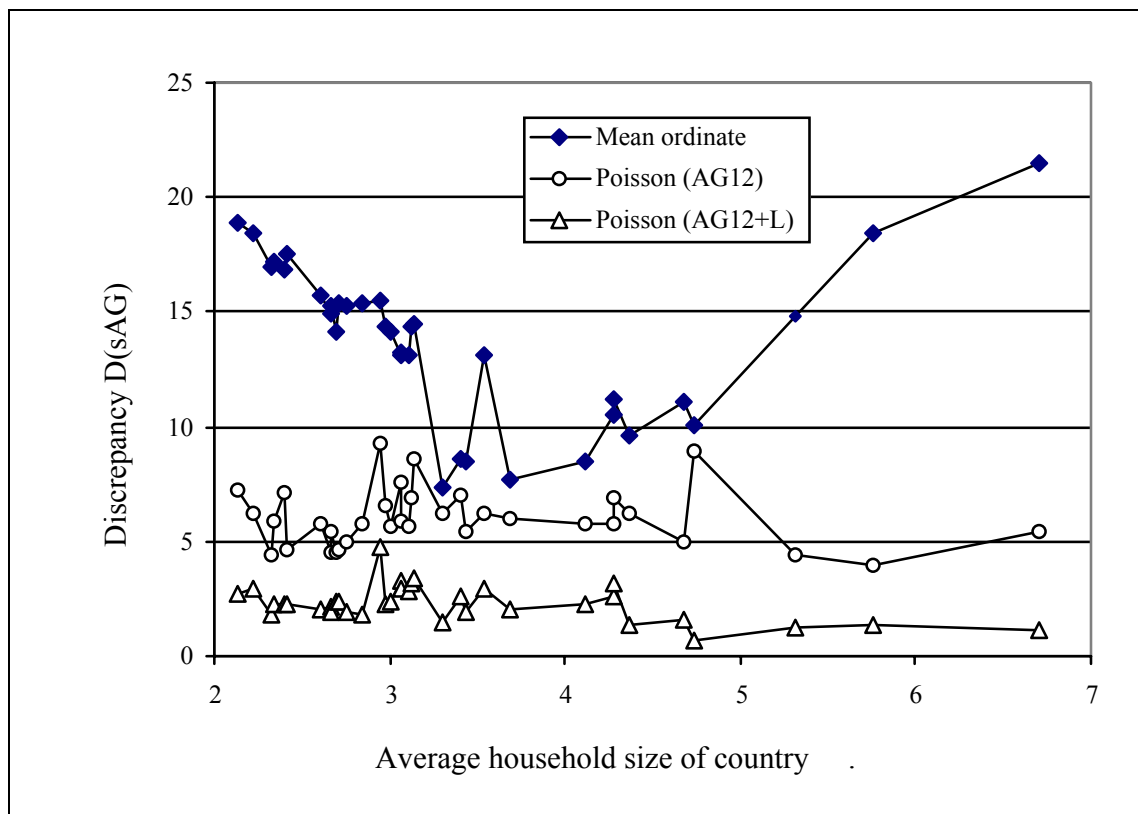
That is for these 12 age groups for a country the total sum of discrepancies over all age groups, for a particular size is zero, while the total sum of discrepancies for one age group over all sizes is also zero. The discrepancies of the estimates from the observed percentages for the 12 age group data are approximately normally distributed and thus standard statistical tests can be applied. The estimates of α, β, γ are found by ordinary least squares regression. In this way the discrepancies between observed and theoretical proportions of size 1,2,3,4,5,6+ households can be modelled.

It is necessary to ensure that the distribution of household sizes generated by the linear and a Poisson component model should have the same mean value as that of the

original observed distribution. This is achieved as follows. For any of the models the number of persons and the number of households are estimated for sizes one to five giving totals for sizes one to five. Subtract these amounts from the overall totals to give estimated values for size 6+. In this way the average household size for the model distributions will be identical to the observed. Examination of observed values of 6+ show that in some cases the same age for size 6+ households has been chosen for all age groups for a country; in some cases also the values seem too high. The variances of the distributions show little pattern.

The discrepancies arising from the three models are plotted in Figure 5.

Figure 5 Discrepancy $D(sAG)$ from observed ordinates of estimates based on the *mean ordinate*, *Poisson (AG12)* and *Poisson (AG12+L)* models



If the observed values are taken to lie on the x axis then the discrepancies can be shown in a consistent manner above the x axis. The base case shows the discrepancies arising from the *mean ordinate*. The *Poisson (AG12)* model reduce the discrepancies so that they are now more linear in shape. A further reduction occurs using the *Poisson (AG12+L)* model. Taking the square root of the mean sum of the squares of the D values for each country over all countries gives an overall discrepancy Discrepancy (all). The Discrepancy (all) for the *Poisson (AG12)* discrepancies is 43 per cent of the *mean ordinate* discrepancies and the *Poisson (AG12+L)* is 17 per cent of the *mean ordinate* discrepancies.

Thus the observed age group household size distributions for each of the 35 countries may be decomposed into the mean ordinate height, a constant 16.67, plus a Poisson component plus a linear component plus a residual error. The constant and the linear components are additive, ie their total value is independent of how they are summed. The major remaining factor, the Poisson, is non-additive. Therefore it is expected that the observed household size distribution for a country which is the sum of age group distributions would be better matched to the sum of age group Poisson estimates, rather than an overall country level Poisson estimate.

The outlier in both Figures 5 and 6 is the Central African Republic 1988 with average household size 4.73. Further data from countries close by this country could help identify the source of this distinctive difference from other countries with high average household size. The next section demonstrates how the summation of Poisson distributions can be evaluated theoretically, and compares examples similar to Sweden and the Philippines.

Age group summation properties – the gamma effect

The age group distributions show Poisson properties. In this section together with Appendix 2 the aim is to show how the sum of age group distributions relate to the gamma distribution. Appendix 2 establishes the underlying theory, and in this section this theory is applied to two cases similar to Sweden and the Philippines. Finally the observed less estimates of household size distribution for three models are compared.

Example 1

The conditions chosen in this example are chosen are designed to illustrate how the summation of Poisson distributions will influence household size distributions. Assume that for each country the 13 age strata 15-19 to age 75+ contain the same number of households. From Figure 3 we note that for individual countries the average household size of age groups increase until about age 40-44 years and then decreases. For comparison purposes take the range from low to high of average household size at 1.7 persons per household. This is about the range of average household size that occurs if we exclude the 15-19 age group. Now these age groups are obtained by a division of the age profile into 5 year age intervals. The intervals could however be one year or less so that over the range of 1.7 persons per household it is possible given a sufficiently large number of households overall to have a large number of distributions. If it is assumed that these distributions are spaced equally far apart along the x axis and they were Poisson distributions then equation (12) in Appendix 2 applies. In this case therefore following Appendix 2 the country level household size distribution is best represented by the average of many Poisson distributions and this is in turn asymptotic to the mean ordinate of the gamma distribution over the same range. As shown in Example 1 and Table 3 even a summation of six Poisson distributions over this range provides a fair approximation to the values derived from the incomplete gamma function. In practice the results will be weighted if the number of households in each of the age groups is not the same.

In this example the incomplete gamma function estimates are compared with overall Poisson estimates for two cases. The first case is comparable to Sweden 1990 where the average household size for age groups 20-24 to 40-44 range from 1.4 to 3.1 persons per household and the overall average is 2.25. Generally the range for most countries for ages 20 years or more is about 1.7 to 1.9 persons per household. The second case is comparable to the Philippines 1990 where the average household size for age groups range from 4.3 to 6.0 persons per household and the overall average is 5.15. The values of $G(k | \lambda_1, \lambda_2)$ in the first row of each country section of Table 3 may be taken as equivalent to values of a hypothetical observed distribution.

The example shows that for countries such as Sweden with low values of average household size approaching two the differences between $G(k | \lambda_1, \lambda_2)$ and *Poisson* (C) are considerably greater than for countries such as the Philippines with much higher average household size. An examination of the incomplete gamma function tables shows this pattern to be general. Where the Poisson parameter is above about four the differences between the gamma derived distributions and the overall Poisson distributions are small. However for the Poisson parameter near two the differences become quite significant. This will apply to an increasing number of countries since worldwide there is a trend to average household size reducing below four.

The two means of six Poisson distributions spaced evenly along the λ axis over the range considered are shown in the last line of each of the two cases in Table 3. These values reflect the age group range for the two cases. The difference between these levels and $G(k | \lambda_1, \lambda_2)$ are greater for the Swedish case than for the Philippines but are not sufficient to affect the main line of argument in the paper.

Table 3 Comparison between the incomplete gamma function estimate $G(k | \lambda_1, \lambda_2)$ and the overall Poisson distribution estimate Poisson (C). Two countries, one with a low average household size and the other with a high average household size. Note that $\lambda = \mu - 1$.

Household size $k = \alpha$	1 %	2 %	3 %	4 %	5 %	6+ %	Total
Like Sweden							
Area between values (1.4-1) to (3.1-1) for incomplete gamma function divided by 1.7 ie $G(k \lambda_1, \lambda_2)$	32.23	32.87	20.14	9.45	3.65	1.66	100
Poisson(C) $\lambda = 2.25 - 1$	28.65	35.81	22.38	9.33	2.91	0.91	100
$G(k \lambda_1, \lambda_2)$ less Poisson(C)	3.58	-2.94	-2.24	0.12	0.74	0.75	0.00
Standard deviation							2.33
Mean of six Poisson distributions	33.72	31.52	19.37	9.52	3.91	1.95	100
Like the Philippines							
Area between values (4.3-1) to (6-1) for incomplete gamma function divided by 1.7 $G(k \lambda_1, \lambda_2)$	1.77	6.95	13.81	18.55	18.95	39.97	100
Poisson(C) $\lambda = 5.15 - 1$	1.58	6.53	13.54	18.73	19.44	40.18	100
$G(k \lambda_1, \lambda_2)$ less Poisson(C)	0.20	0.42	0.27	-0.19	-0.49	-0.21	0.00
Standard deviation							0.39
Mean of six Poisson distributions	1.86	7.12	13.90	18.45	18.74	39.95	100

Source: Abramowitz. and Stegun 1965 Table 26.7 for incomplete gamma function values.

Comparing the two Poisson models with the base model

The previous two sections showed that the distribution of households by size changed continuously and in a Poisson-like manner with change in average household size. This indicated that the 35-country data may be modelled using the Poisson-gamma relationships. In this section the discrepancy between observed and estimated arising from a base model and two Poisson models are compared. The discrepancy D is measured by the standard deviation of the difference between observed and estimated percentage of households for sizes 1,2,3,4,5,6+at the country level.

The first model used is the ‘base model’ where it is assumed that all ordinates have the same value, this is the *mean ordinate* (C) model.

In the second model the Poisson distribution is used to estimate the percentage distribution of households at the country level using equation (1). This is called the

Poisson (C) model and uses as parameter the average household size less one at the country level.

The third model is obtained as follows. Calculate the *Poisson* distribution for each age group for a country as in equation (1) using the average household size for the age group less one as the parameter. Since the number of households are known for each of the thirteen age groups the total estimated households can be calculated for a country for each household size and hence the distribution of household sizes is obtained. This is called the *Poisson* (C.AG13) model. Results are shown in Figure 6.

The graphs show that there is a substantial reduction in the discrepancy *D* of the *Poisson* (C) model over the *mean ordinate* (C) model. Applying the *Poisson* (C.AG13) model further reduces this discrepancy. The difference between the *Poisson* (C) and the *Poisson* (C.AG13) model are greatest for average household size below 3.5 which is consistent with the theory as shown in Example 1.

Using the same data in Figure 6 as in Figure 5 it is possible to further isolate the main sources of the difference in model estimates. The data may be classified by household size, and by whether average household size for a country is less than 3.5 or 3.5 or more. Results are shown in Table 4. The method of measurement used is to take the difference between the values of percentage of households estimated using *Poisson* (C) and *Poisson* (C.AG13) for a particular size for a group of countries, and take the square root of the mean of the sum of squares of the differences. This is called a standardised difference .

The standardised difference is about two or three times as great for sizes one, two and three for countries with average household size below 3.5, as compared to the other cases. This is consistent with Example 1. For household sizes 1,2,3,5,6+ the linear regression model over 35 countries for the differences between the *Poisson* (C) and *Poisson* (C.AG13) estimates of percentage of households fitted to average household size is highly significant under the *t* test (0.001 under the null hypothesis).

Over the 35 countries for all sizes the standard deviation of discrepancies was 9.92 for the base case *Observed - mean ordinate* (C), 4.9 for *Observed -Poisson* (C) and 3.36 for *Observed - Poisson* (C.AG13). Thus it appears likely on the evidence in this paper that about a third of the observed less *Poisson* discrepancies can be removed if the *Poisson* (C) model is replaced by the *Poisson* (C.AG13) model.

Figure 6 Discrepancy D of distribution of households for the following models; *Mean ordinate (C)*, country level *Poisson (C)*, and age group level *Poisson (C.AG13)*. Estimated per cent of households for 35 countries.

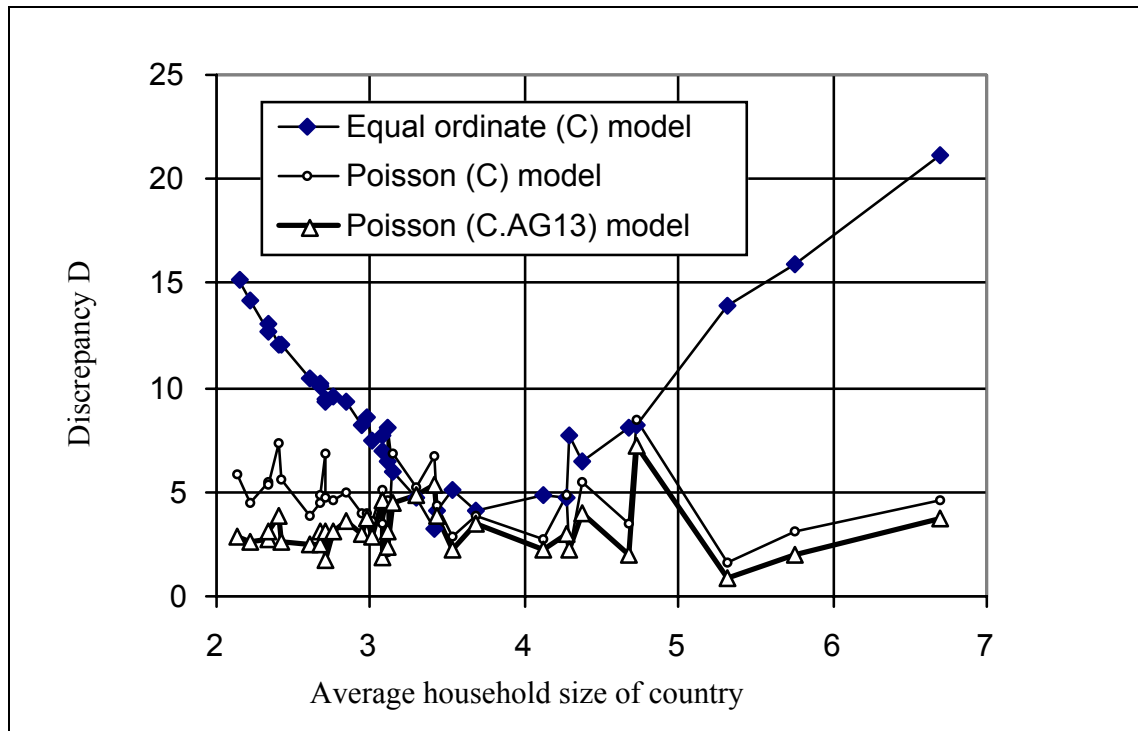


Table 4. Standardised differences between the *Poisson (C)* model and the *Poisson (C.AG13)* model, by household size and by average household size for a country

	Average household size below 3.5 No. of countries = 24	Average household size 3.5 or more No. of countries = 11	All No. of countries = 35
Household size	Standardised differences	Standardised differences	Standardised differences
1	4.48	1.48	3.81
2	2.75	1.24	2.38
3	3.37	1.05	2.85
4	1.19	1.69	1.36
5	0.70	1.45	1.00
6+	1.66	1.28	1.55

Summary

The methodology in this paper differs from that of Jennings, Lloyd-Smith and Ironmonger (1999). Instead of using the allocation of persons to houses as a basis for a model, in this paper the allocation of persons is made to household reference persons. In this instance it is implied that households are more broadly defined in terms of shared living arrangements rather than just houses. This has two advantages. Firstly the Poisson model then has as its dual the gamma distribution (scale parameter equals one), and hence the Poisson-gamma reference frame can be used to distinguish Poisson, gamma and other influences. Secondly the notion of the household is freed from that of house hence broadening the potential of the model.

The age stratified household data allowed the effects of age to be observed on average household size and on household size distribution. In particular as seen in Figure 2 the average household size for age groups for most countries has a range of about two persons per household around the overall average household size. Such a wide range indicates a significant variation in age group household size distributions. The consistency of the range provided a basis for new estimation models as described in the conclusion.

Evidence that household size distributions for the age groups had similarities to the Poisson was obtained by directly comparing discrepancies between the percentage of observed and estimated household size distributions for age groups for each country. These results were compared across 35 countries using average household size of a country as the independent variable. On average the Poisson discrepancies were 43 per cent of *mean ordinate* model discrepancies.

It was then argued that since the age groups have Poisson like properties their sum should be related to the gamma function. In Appendix 2 it is demonstrated that the sum of Poisson distributions over a range would under suitable conditions tend to the gamma function over the same range. In the text in Table 3 examples are given of how to apply this result to modelling distributions such as those of Sweden and the Philippines. The household size distributions of the 35 countries are then compared with estimates using the *mean ordinate* (C) model, the *Poisson* (C) model using average household size for a country as parameter, and age group *Poisson* (C.AG13) models. Figure 6 summarises the results and shows that there is a significant reduction in discrepancy across the models. Table 4 shows that the reduction in discrepancies of the age group Poisson models over the country level Poisson models is mainly due to reductions for households size one, two and three, for average household size of country less than 3.5.

Over the 35 countries totalling over households in all age groups it is found that the observed less Poisson distribution discrepancies are about half those of the mean ordinate model. The observed less the sum of a range of Poisson distributions is about one third of the mean ordinate model discrepancies. From Figure 5 it appears that about half the remaining discrepancy could be accounted for by a linear component and the remainder would be residual error.

Further work is indicated. The method of selection of the household reference person is not generally described in the census documents, but is left up to the members of the household. This may not matter because of the age clustering described in Table 1. However this should be explored in the future to determine whether there are any

systematic biases in the distributions resulting from this approach. For example a review by the authors (not shown) indicates that for the 35 country data there are significant differences between age group household size distributions depending upon whether a man or a woman is the household reference person.

Conclusion

Observed household size distributions at the country level are best seen as the mean of a number of Poisson distributions over a parameter range. In the limit the ordered set of Poisson distributions is the standard gamma distribution. This gamma distribution model is based upon the methods outlined in Example 1 and Table 3. If for example average household size for a country is 3.0 then it may be presumed, given Figure 2, that the average household size for age groups (20+ years) in the country range from about 2.0 to about 4.0, a range of 2.0. The equivalent range for the gamma distribution is over $\lambda = \mu - 1$ ie $\lambda_1 = 1.0$ and $\lambda_2 = 3.0$. The mean value $G(k | \lambda_1, \lambda_2)$ will then give the ordinate values for size k .

An alternative is to use a table constructed from Poisson distributions to approximate gamma distribution ordinates. This avoids the need for a look-up table. For example the mean of the ordinate values of twenty one Poisson distributions over the range (λ_1, λ_2) may be used. The lower end of this range could be $\lambda_1 = (\mu - 1) - 1$ and the upper end $\lambda_2 = (\mu - 1) + 1$. This range should give a close approximation to $G(k | \lambda_1, \lambda_2)$ in most cases. For each of the 35 countries the observed less estimates for the six ordinates are taken and the variance of the discrepancies calculated. These variances are averaged across the 35 countries and the square root of the sum taken. The ratio of this value for the Poisson summed discrepancies to the Poisson country level discrepancies is 0.76, ranging from 0.50 for Sweden 1990 to 1.08 for Mauritius 1990. Portugal 1991 at ratio of 1.0 is the only other country with a ratio of 1.0 or more. The ten countries with the lowest average household sizes, from 2.13 to 2.70, have an average ratio of 0.60.

It would be desirable to test this method against a larger number of household size distributions at the country level. Some refinements are possible particularly if age group household numbers or average household sizes are known.

References

- Abramovitz, M. and I. A. Stegun 1965 (ed.) *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards 1965 Applied Mathematical Series.55. Washington: U.S. Department of Commerce, U.S. Government Printing Office.
- Australian Bureau of Statistics (ABS) 1986a *1986 Census one percent sample tape*. Canberra.
- Australian Bureau of Statistics (ABS) 1986b *Australia in Profile Census 1986*, Catalogue No. 2502.0: 9. Canberra.
- Australian Bureau of Statistics (ABS) 1999 *Household and Family Projections, Australia 1996 to 2021*, Catalogue No. 3236.0. Canberra.
- Australian Bureau of Statistics (ABS) 2001 *1991 and 1996 Census data provided directly to authors*. Canberra.
- Australian Bureau of Statistics (ABS) 1988 *Year book Australia 1988* Catalogue No. 1301.0. Canberra.
- Australian Bureau of Statistics (ABS) 1995 *Year book Australia 1995* Catalogue No. 1301.0. Canberra.
- Devore, Jay L. 1991. *Probability and Statistics for Engineering and the Sciences*, Third Edition. California: Brooks/Cole Publishing Company.
- Hinde, A. 1998. *Demographic Methods*. London: Arnold – Hodder Headline Group.
- Jennings, V.E., Lloyd-Smith, C. W. and Ironmonger, D.S. 1999. Household size and the Poisson Distribution. *Journal of the Australian Population Association*, 16(1/2):65-82.
- Kingman, J.F.C. 1993 *Poisson Processes*. Oxford Studies in Probability. Oxford: Oxford University Press.
- Pearson, K. (Ed.) 1922 *Tables of the Incomplete Γ -Function*. London: Published for the Department of Scientific and Industrial Research by His Majesty's Stationary Office.
- Slotnik, H. 1998. International Migration 1965-96: An Overview. *Population and Development Review*. 24(3): 429-468.
- Stuart, A. and J.K. Ord. 1987. *Kendall's Advanced Theory of Statistics*. Volume 1, Fifth Edition. London: Griffin
- United Nations (UN) 1989. *1987 Demographic Year Book*. New York.
- United Nations (UN) 1997. *1995 Demographic Yearbook*. New York.
- Winkelmann, R. 2000. *Econometric Analysis of Count Data*. Third Edition, Berlin: Springer.

Appendix 1

35 country data set

This data set stratifies the population in households by the age and sex of the household reference persons and by household size. It is derived from the United Nations 1995 Demographic Yearbook, Table 30 and Table 34, United Nations (1997). There is also a similar table of 32 countries in the United Nations 1987 Demographic Yearbook United Nations (1989) Table 33 entitled 'Households by age and sex of householder and sizes of household and urban/rural residences'. Data has also been obtained on Australia from the Australian Bureau of Statistics ABS (2001).

The 35 countries used together with year of collection of data are:

Australia 1991, Austria 1981, Bangladesh 1981, Bolivia 1992, Brazil 1980, Bulgaria 1985, Canada 1991, Central African Republic 1988, China, Hong Kong SAR 1991, Cuba 1981, Denmark 1991, Finland 1990, France 1982, Germany 1987, Greece 1991, Guadeloupe, 1990, Hungary 1990, Italy 1981, Japan 1985, Luxembourg 1991, Macau 1991, Mauritius 1990, Netherlands Antilles 1992, New Zealand 1991, Norway 1990, Pakistan 1981, Philippines 1990, Poland 1988, Portugal 1991, Réunion 1982, Romania 1992, Slovenia 1991, Spain 1981, Sweden 1990, Switzerland 1990.

The country as the geographical unit

The geographical division chosen as the unit for this investigation is the country or nation. However the authors have applied the Poisson method to compare local government areas of Melbourne with minimum area population down to 30000 persons. The 'country' defines the boundary of residence for most people. Generally this boundary changes little over time. Although there is migration from country to country this is currently about 2.1 to 2.3 per cent per annum of the world population, Slotnik(1998). Most migration occurs within countries, partly because it is unrestricted whereas migration between countries is subject to government restriction with the exception perhaps of some regions of Europe. For example for Australia we may have less than one per cent per annum external immigration whilst about 18 per cent of the population will move from one house to another each year. See Australian Bureau of Statistics, ABS (1986b)

Appendix 2

Summing an infinite set of Poisson distributions spaced evenly along the parameter axis x between $x = \lambda_1$ and $x = \lambda_2$

Consider Cartesian coordinates (x, y, z) . Position a Poisson distribution $P(n | \lambda)$ in the $x = \lambda$ plane on the line with parametric equation $x = \lambda, z = 0$. For example as the Poisson distributions are positioned in Figure 1. In this case two Poisson distributions are shown with $z = P(k | \mu)$ where $k = n + 1$ and $\mu = \lambda + 1$ so that $\mu = 2.5$ and 4.0 .

Then points will be generated in this plane with positions defined by $y = n = 0, 1, 2, \dots$ and values $z = P(n | \lambda)$. Using the balls into cells model we assume that there is an equal number of cells in each of a series of Poisson distributions ordered by $x = \lambda$. Consider a set of Poisson distributions S_p equally spaced $d\lambda$ apart along the x axis from values λ_1 to λ_2 so that there are N Poisson distributions. Thus $(N - 1) \times d\lambda = \lambda_2 - \lambda_1$ where $\lambda_2 > \lambda_1$.

For households of the same size $n + 1$ over the set S_p using equation (5) the average proportion per set is given by

$$\bar{z} = \frac{1}{N} \sum_{\lambda} P(n | \lambda) \quad \text{noting also that } \sum_n P(n | \lambda) = 1 \text{ for all } \lambda.$$

Let $\lambda_2 - \lambda_1 = L$ then $(N - 1)d\lambda = L$ and hence

$$\bar{z} = \frac{1}{N} \sum_{\lambda} P(n | \lambda) = \frac{1}{L + d\lambda} \sum_{\lambda} P(n | \lambda) d\lambda \quad (12)$$

and since $P(n | \lambda)$ is continuous over λ as $d\lambda \rightarrow 0$ this is equivalent to the Stieltjes integral, see Stuart and Ord (1987:16-17). With $n = k - 1$ where k is the number of persons in a household equation (12) may be expressed in the form

$$\frac{1}{L + d\lambda} \sum_{\lambda} P(n | \lambda) d\lambda = \frac{1}{L} \int_{\lambda_1}^{\lambda_2} \frac{\lambda^n e^{-\lambda}}{n!} d\lambda = \frac{1}{L} \int_{\lambda_1}^{\lambda_2} \frac{\lambda^{k-1} e^{-\lambda}}{\Gamma(k)} d\lambda = G(k | \lambda_1, \lambda_2) \quad (13)$$

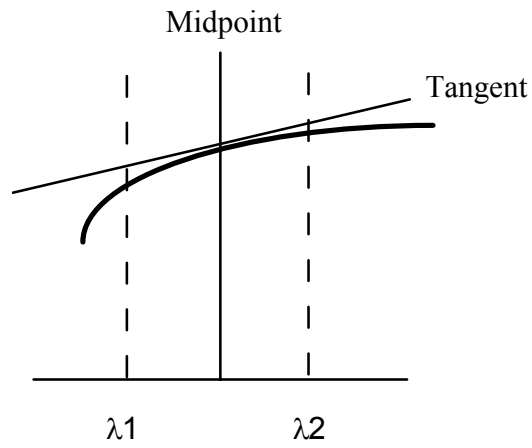
which is the area defined by the difference of two values of the incomplete gamma function divided by L noting also that the gamma function is defined by

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx.$$

Thus for household size k , over (λ_1, λ_2) the average height for N equally spaced Poisson distributions as $N \rightarrow \infty$ is equal to $G(k | \lambda_1, \lambda_2)$, the mean ordinate of the gamma distribution curve over the same range. For equally spaced distributions along the x axis over (λ_1, λ_2) the parameter for the overall Poisson distribution $P(C)$ will be $(\lambda_1 + \lambda_2)/2$.

In general over the range (λ_1, λ_2) if the gamma distribution curve is convex, that is the second differential for the curve is negative, then the mean ordinate over the range will be less than the ordinate for the overall Poisson distribution since its parameter will be at the midpoint of the range (λ_1, λ_2) , ie at $(\lambda_1 + \lambda_2)/2$ as shown in the diagram. Similarly if the gamma distribution curve is concave, that is the second differential is positive, then the mean ordinate over the range will be more than the ordinate for the overall Poisson distribution $P(C)$. As the curvature tends to zero the differences tend to zero from the negative side and the positive side and when the curves are straight the difference is zero since it coincides with the tangent. Thus for the function consisting of a

Poisson plus a linear component, the linear component makes no contribution to the differences described above. See diagram.



For this convex case the tangent defines the mean ordinate for $P(C)$ while the curve defines the mean ordinate for the gamma distribution curve, and this is under the tangent.