

# A TECHNIQUE FOR CORRECTING AGE DISTRIBUTIONS FOR HEAPING ON MULTIPLES OF FIVE

by Griffith Feeney

Age distributions enter into the calculation of virtually every standard demographic measure of fertility, mortality, and population growth. Where vital registration data do not exist or are deficient, moreover, age distributions are essential input data to virtually all special estimation techniques, from those based on stable population theory to the own-children type of methods. One scarcely exaggerates in saying that age distributions are an essential, structural part of the foundation on which the entire edifice of refined demographic statistics rests. If the foundation is inadequate, the structure is endangered. If we are concerned with the accuracy of demographic statistics, logic obliges us to be concerned with the accuracy of the age distributions that enter into their calculation. This note is concerned with the problem of deriving reliable information on the true age distribution of a population from recorded age distributions that exhibit substantial heaping on multiples of five.

## Redistributing excess numbers at multiples of five

The obvious way to attempt to undo the effects of heaping on ages that are multiples of five is to develop procedures for distributing the excess numbers of persons at multiples of five to the surrounding ages. Suppose we transfer numbers of persons from a given multiple of five to the eight immediately surrounding ages in such a way that (1) the adjusted numbers in the surrounding ages are proportional to the original numbers; and (2) the adjusted numbers at the lower four ages, the central age, and the upper four ages form a linear progression. For reasons that will become clear below, we consider only the numbers of persons at ages that are multiples of five and the total numbers of persons at the intermediate ages. Some elementary algebra shows that, considering any age  $x$ , a multiple of five, in isolation, the redistribution described by conditions (1) and (2) may be effected by applying the formulas

$$P'_{x-} = P_{x-} + (\Delta_x - 1)P_{x-}$$

$$P'_x = P_x - (\Delta_x - 1) [P_{x-} + P_{x+}]$$

$$P'_{x+} = P_{x+} + (\Delta_x - 1)P_{x+}$$

where  $P_x$  denotes the number of persons at age  $x$ ,  $P_{x-}$  the total number of persons at the four ages immediately below age  $x$ , and  $P_{x+}$  the number at the four ages immediately above age  $x$ ; the primes (') on the left denote preliminary adjusted values, and  $\Delta_x$  is calculated as

$$\Delta_x = \frac{8}{9} \left[ \frac{P_{x-} + P_x + P_{x+}}{P_{x-} + P_{x+}} \right]$$

These formulas apply to any multiple of five considered in isolation. When all ages are considered together, the increments to the numbers of persons intermediate between two successive multiples of five are made independently from the lower and upper multiple of five, as described by the general formulas

$$P'_x = P_x - (\Delta_x - 1) [P_{(x-5)+} + P_{x+}]$$

$$P'_{x+} = P_{x+} + (\Delta_x - 1)P_{x+} + (\Delta_{x+5} - 1)P_{x+} \\ = (\Delta_x + \Delta_{x+5} - 1)P_{x+}$$

for  $x = 0, 5, \dots, c-5$  where  $\Delta_0$  and  $\Delta_c$  are taken equal to one,  $c$  denoting the age, assumed to be a multiple of five, that begins the open-ended age group with which age distributions are generally terminated. Should this open-ended age group begin with something other than a multiple of five, the distribution should be truncated so that the open-ended age group of the truncated distribution begins with the greatest multiple of five given by the recorded distribution.

Exhibit 1 shows the application of this procedure to the Indonesian age distribution shown in Table 1.

## Iteration and convergence

The values of  $\Delta_x$  in Exhibit 1 may be taken as a rough measure of heaping on age  $x$ . The more  $\Delta_x$  exceeds one, the greater the extent of the heaping on age  $x$ . Values of  $\Delta_x$  less than one signify a deficit of persons at age  $x$ . Because numbers of persons at successive multiples of five are distributed independently to the surrounding ages, and because the surrounding ages overlap, intermediate age groups evidently get a "double dose" of redistributed persons. We would therefore expect the intermediate age groups to receive too large an increment, leaving a deficit of persons at multiples of five. This expectation can be tested by calculating  $\Delta_x$  values for the adjusted distribution. These values are shown in the last column of Exhibit 1 and bear out the expectation. What is unexpected is the magnitude of these  $\Delta_x$  values. Though the largest  $\Delta_x$  value for the recorded

Exhibit 1 Redistribution of numbers of persons at multiples of five

| Age (x) | $P_x$ | $P_{x+}$ | $\Delta_x$ | $P'_x$ | $P'_{x+}$ | $\Delta'_x$ |
|---------|-------|----------|------------|--------|-----------|-------------|
| 0       | 2,337 | 15,763   | 1.0000     | 2,337  | 15,737    | 1.0000      |
| 5       | 3,685 | 14,156   | 0.9984     | 3,734  | 14,354    | .9992       |
| 10      | 3,457 | 10,093   | 1.0156     | 3,078  | 10,398    | .9994       |
| 15      | 2,586 | 8,196    | 1.0146     | 2,319  | 9,119     | .9945       |
| 20      | 3,006 | 4,579    | 1.0980     | 1,753  | 6,039     | .9917       |
| 25      | 3,550 | 4,927    | 1.2208     | 1,451  | 7,503     | .9841       |
| 30      | 3,951 | 3,573    | 1.3021     | 1,383  | 5,965     | .9802       |
| 35      | 3,919 | 3,705    | 1.3675     | 1,244  | 6,488     | .9777       |
| 40      | 3,409 | 2,420    | 1.3836     | 1,059  | 4,301     | .9761       |
| 45      | 2,488 | 1,960    | 1.3938     | 763    | 3,675     | .9739       |
| 50      | 2,259 | 1,430    | 1.4812     | 628    | 2,570     | .9783       |
| 55      | 1,153 | 971      | 1.3157     | 395    | 1,933     | .9669       |
| 60      | 1,500 | 726      | 1.6746     | 355    | 1,515     | .9804       |
| 65      | 670   | 411      | 1.4127     | 201    | 897       | .9630       |
| 70      | 696   | 291      | 1.7702     | 155    | 515       | .9865       |

## NOTATION AND COMPUTATIONAL PROCEDURE

$P_x$  = number of persons aged  $x$  in completed years

$P_{x-}$  = number of persons aged  $x-4$  to  $x-1$  in completed years

$P_{x+}$  = number of persons aged  $x+1$  to  $x+4$  in completed years

$$\Delta_x = \frac{8}{9} \left[ \frac{P_{x-} + P_x + P_{x+}}{P_{x-} + P_{x+}} \right], x = 5, \dots, 70 (\Delta_0 \equiv \Delta_{75} \equiv 1)$$

$$P'_x = P_x - (\Delta_x - 1) [P_{(x-5)+} + P_{x+}], x = 0, \dots, 70$$

$$P'_{x+} = (\Delta_x + \Delta_{x+5} - 1)P_{x+}, x = 0, \dots, 70$$

**Table 1 Population of Indonesia, 22 provinces, by single year of age: Census of 24 September 1971**

| Age (x)    | Population at indicated age (in thousands) |       |       |       |       |       |
|------------|--|-------|-------|-------|-------|-------|
|            | x, x+5                                     | x     | x+1   | x+2   | x+3   | x+4   |
| 0          | 18,100                                     | 2,337 | 3,873 | 3,882 | 3,952 | 4,056 |
| 5          | 17,841                                     | 3,685 | 3,687 | 3,683 | 3,611 | 3,175 |
| 10         | 13,550                                     | 3,457 | 2,379 | 3,023 | 2,375 | 2,316 |
| 15         | 10,782                                     | 2,586 | 2,014 | 2,123 | 2,584 | 1,475 |
| 20         | 7,585                                      | 3,006 | 1,299 | 1,236 | 1,052 | 992   |
| 25         | 8,477                                      | 3,550 | 1,334 | 1,314 | 1,337 | 942   |
| 30         | 7,524                                      | 3,951 | 1,128 | 1,108 | 727   | 610   |
| 35         | 7,624                                      | 3,919 | 1,221 | 868   | 979   | 637   |
| 40         | 5,829                                      | 3,409 | 887   | 687   | 533   | 313   |
| 45         | 4,448                                      | 2,488 | 677   | 426   | 524   | 333   |
| 50         | 3,689                                      | 2,259 | 551   | 363   | 290   | 226   |
| 55         | 2,124                                      | 1,153 | 379   | 217   | 223   | 152   |
| 60         | 2,226                                      | 1,500 | 319   | 175   | 143   | 89    |
| 65         | 1,081                                      | 670   | 149   | 96    | 97    | 69    |
| 70         | 987  | 696   | 170   | 60    | 38    | 23    |
| 75         | 745  | 0     | 0     | 0     | 0     | 0     |
| Not stated | 15   | 0     | 0     | 0     | 0     | 0     |
| Total      | 112,627                                    | 0     | 0     | 0     | 0     | 0     |

SOURCE: *Census of Indonesia, 1971*, Series E, Numbers 1–26 (26 volumes), Table 01, pages 1–4, in each volume.

NOTE: There are 26 provinces in Indonesia; data in the table describe 22 provinces. The provinces not included here are Nusa Tenggara Barat, Nusa Tenggara Timur, Moluku, and Irian Jaya.

distribution is  $\Delta_{70} = 1.77$ , indicating considerable heaping at age 70, the lowest  $\Delta'_x$  value for the adjusted distribution is  $\Delta_{65} = 0.96$ , indicating a relatively small deficit at age 65. This striking contraction of the  $\Delta_x$  values toward one suggests that if the procedure is applied repeatedly, the result may be an adjusted distribution for which all  $\Delta_x$  values are one. This convergence does indeed occur, in various numerical cases, at any rate. We do not consider proof of convergence. The resultant values are shown in the first two columns of Exhibit 2.

#### Interpolation between corrected numbers of persons at multiples of five

It may come as a surprise that the final step in the correction procedure is not the apparently obvious one of aggregating the adjusted numbers in the first two columns of Exhibit 2 into standard five-year age groups. The convergence of the  $\Delta_x$  values to one signifies that the age distribution, as represented by the first two columns of Exhibit 2, has been forced into a series of straight line segments centered on multiples of five. In consequence, a deficit of persons in one intermediate age group may result in an excess of persons in the next intermediate age group, a deficit in the next, and so forth. This phenomenon is indicated quite clearly in Figure 1, in which numbers of persons at multiples of five are plotted by circles and numbers of persons at intermediate ages (average per year of age) by dots.

The penultimate step in the correction procedure consists of interpolating among the adjusted numbers at multiples of five (the  $P_x$  column of Exhibit 2), regarded as estimates of the age distribution density function at ages  $x + 2.5$ . Various interpolation approaches are possible. The simplest, which will often give satisfactory results, is as follows.

Interpolate linearly among adjusted numbers at multiples of five to obtain values of the age distribution density function at points 7.5, 12.5, ..., 67.5 (shown in column 3 of Exhibit 2), and multiply these values by 5 to give adjusted numbers of persons aged 5–9, 10–14, ..., 65–69 (shown in column 4). The numbers for the remaining age groups, 0–4, 70–74, and 75 and over, are taken directly from the recorded distribution.

#### Exhibit 2 The correction procedure

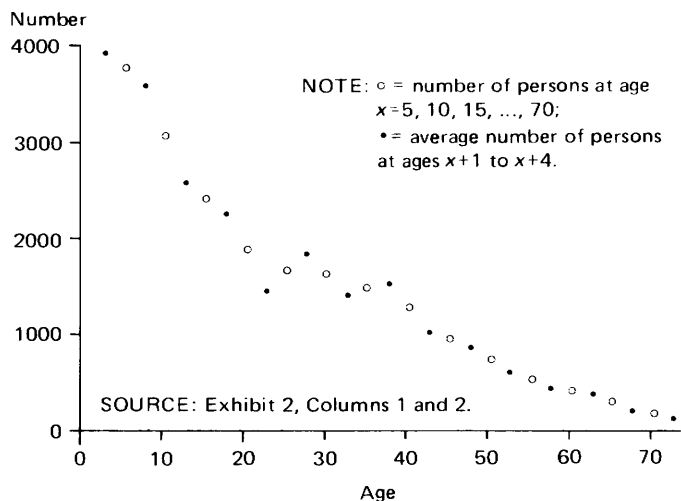
| Age (x)     | $P_x$ (1) | $P_{x+}$ (2) | $f(x+2.5)$ (3) | $5P'_x$ (4) | $5P_x$ (5) |
|-------------|-----------|--------------|----------------|-------------|------------|
| 0           | 2,337     | 15,724       | na             | 18,100      | 18,004     |
| 5           | 3,758     | 14,338       | 3,489          | 17,444      | 17,351     |
| 10          | 3,085     | 10,342       | 2,819          | 14,093      | 14,018     |
| 15          | 2,419     | 9,005        | 2,197          | 10,985      | 10,927     |
| 20          | 1,864     | 5,907        | 1,777          | 8,884       | 8,837      |
| 25          | 1,646     | 7,260        | 1,638          | 8,188       | 8,145      |
| 30          | 1,625     | 5,738        | 1,573          | 7,865       | 7,823      |
| 35          | 1,495     | 6,218        | 1,413          | 7,067       | 7,029      |
| 40          | 1,291     | 4,106        | 1,156          | 5,779       | 5,748      |
| 45          | 953       | 3,518        | 870            | 4,349       | 4,326      |
| 50          | 745       | 2,442        | 661            | 3,307       | 3,289      |
| 55          | 536       | 1,842        | 486            | 2,428       | 2,415      |
| 60          | 410       | 1,437        | 361            | 1,804       | 1,794      |
| 65          | 287       | 855          | 241            | 1,203       | 1,197      |
| 70          | 171       | 509          | na             | 987         | 982        |
| 75 and over | 745       | na           | na             | 745         | 741        |
| Not stated  | 15        | na           | na             | na          | 0          |
| Total       | 112,627   | na           | na             | 113,228     | 112,627    |

na—not applicable.

#### NOTES:

- (1) The procedure begins with repeated application of the redistribution (see Exhibit 1) until all  $\Delta_x$  values converge to one. The resulting numbers for the data in Table 1 are given in columns (1) and (2).
- (2) The adjusted numbers at age  $x$  are then regarded as values of the age distribution density function at exact age  $x + 2.5$ , for all ages  $x$  beyond 0, and linear interpolation is performed between these values to obtain values of the age distribution density function at ages 7.5, 12.5, ..., the limiting value for  $x$  in this case being 67.5. If the open-ended age group terminating the original age distribution begins at age  $c$ , a multiple of five, then the adjusted values at age  $x$ , a multiple of five, obtained by redistribution (Exhibit 1) range from 5 to  $c - 5$  in steps of five; therefore the interpolated values for ages  $x + 2.5$  range from  $x = 5$  to  $x = c - 10$ . These interpolated values are given in column (3). The interpolation weights are 0.6 and 0.4. For example, the first entry in column (3) is calculated as  $0.6 \times 3,758 + 0.4 \times 3,085$ .
- (3) The interpolated values of the age distribution density function at  $x = 7.5, 12.5, \dots$ , are multiplied by 5 to give preliminary adjusted numbers of persons in the age groups 5–9, 10–14, ... . These adjusted values are combined with the recorded values for the 0–4 age group and, in this case, the 70–74 and the 75 and over age groups (Table 1). The resulting values, and their total, are shown in column (4).
- (4) The total of the adjusted values in column (4) does not equal the original total, though the difference is very small, on the order of 0.5 percent. The last step in the procedure is to multiply the values in column (4) by a factor which makes the resulting total conform to that of the given distribution. The result of this multiplication is given in column (5).
- (5) Column (5) shows the final result of applying the correction procedure to the data given in Table 1.
- (6) Variations in the procedure are obviously possible, as for example using a higher order of interpolation than linear.

**Figure 1** Number of persons at ages that are multiples of five and at intermediate ages: Indonesia, 1971



The final step is to reconcile the total of the adjusted values with the total of the recorded age distribution, taking into account both the discrepancies introduced by the interpolation and the cases where age was not stated. The final result is shown in the last column of Exhibit 2 and is plotted in Figure 2. The corrected and the recorded distributions are compared in Table 2, which shows that the correction procedure transfers large numbers of persons between standard five-year age groups.

**Conclusion**

The term "correction" has been used in preference to "smoothing" because the procedure described here is specifically designed to undo the effects of heaping on multiples of five.

**Table 2** Comparison of recorded and corrected age distribution for 22 Indonesian provinces: Census of 24 September 1971

| Age group   | Number of persons |           | Difference | Percentage difference |
|-------------|-------------------|-----------|------------|-----------------------|
|             | Recorded          | Corrected |            |                       |
| 0-4         | 18,100            | 18,004    | +96        | +1                    |
| 5-9         | 17,841            | 17,351    | +490       | +3                    |
| 10-14       | 13,550            | 14,018    | -468       | -3                    |
| 15-19       | 10,782            | 10,927    | -145       | -1                    |
| 20-24       | 7,585             | 8,837     | -1,252     | -17                   |
| 25-29       | 8,477             | 8,145     | +332       | +4                    |
| 30-34       | 7,524             | 7,823     | -299       | -4                    |
| 35-39       | 7,624             | 7,029     | +595       | +8                    |
| 40-44       | 5,829             | 5,748     | +81        | +1                    |
| 45-49       | 4,448             | 4,326     | +122       | +3                    |
| 50-54       | 3,689             | 3,289     | +400       | +11                   |
| 55-59       | 2,124             | 2,415     | -291       | -14                   |
| 60-64       | 2,226             | 1,794     | +432       | +19                   |
| 65-69       | 1,081             | 1,197     | -116       | -11                   |
| 70-74       | 987               | 982       | +5         | +1                    |
| 75 and over | 745               | 741       | +4         | +1                    |
| Not stated  | 15                | 0         | +15        | +100                  |
| Total       | 112,627           | 112,626   | +1         | 0                     |

na—not applicable.

SOURCE: Recorded numbers from Table 1; corrected numbers from Exhibit 2, last column.

NOTE: See Exhibit 2 for description of correction procedure.

It should be understood, however, that age distributions obtained by application of the procedure will at best be better approximations to the true age distribution than the recorded age distribution. The procedure takes no account of census underenumeration and has been designed for application to age distributions that exhibit substantial heaping on multiples of five. Applied to an accurately recorded age distribution that exhibits sharp fluctuations, the "corrected" distribution may be less correct than the recorded distribution, and the most one can hope for by applying the procedure is to obtain the correct age distribution of the enumerated population.

The merits and demerits of any data analysis procedure can be argued only halfway in the abstract. The rest depends on success or failure in the actual analysis of data. Preliminary analyses on the Indonesian age data suggest that the procedure is remarkably effective, far more so than any existing procedure for smoothing or otherwise correcting age distributions. It is put forward here that others may test its efficacy in practice. □

**ACKNOWLEDGMENTS**

Much of this research was conducted during a visit to the Central Bureau of Statistics (CBS) of Indonesia during early 1978. I am grateful to Mr. M. Abdulmajid, Director General of the CBS, for the opportunity this visit provided. It is a pleasure to acknowledge many useful conversations with Dr. Sam Suharto, Director of the Data Processing Center of the CBS, and with Dr. Hananto Sigit. I am grateful also for numerous useful comments of the staff of the CBS when this material was presented at an informal seminar arranged by Dr. Sigit.

My thinking on the subject of age distribution analysis over the past year has been greatly influenced by many long conversations with Dr. Budi Utomo of the University of Indonesia School of Public Health, who spent ten months at the East-West Population Institute as a research intern. Ms. Bondan Suprptilah, of the Institute of Demography of the University of Indonesia, presently a research intern at EWPI, read the manuscript and pointed out several errors.

**REFERENCE**

Central Bureau of Statistics, Indonesia. 1974. *Census of Indonesia, 1971*, Series E, Numbers 1-26.

**Figure 2** Corrected age distribution: Indonesia, 1971

