

## Income Imputation

### Imputing a dollar value for each Individual Income range

The 2001 Census was required to produce data on income for persons, families and households. Data was also required for counts of families and households within the various income ranges. The data was collected in ranges, not actual dollars, as this has proven to be the most reliable way to collect income data. The 2001 Census collected the gross weekly income for each person (INCP) using the following income ranges:

<i>Range Identifier</i>	<i>Individual Income (weekly)</i>
1	Negative income
2	Nil income
3	1-\$39
4	\$40-\$79
5	\$80-\$119
6	\$120-\$159
7	\$160-\$199
8	\$200-\$299
9	\$300-\$399
10	\$400-\$499
11	\$500-\$599
12	\$600-\$699
13	\$700-\$799
14	\$800-\$999
15	\$1,000-\$1,499
16	\$1,500 or more

These ranges, which are used on the Census form, were chosen after analysing data from the Survey of Income and Housing (SIHC), in which income was collected in actual dollars rather than ranges.

### Household and family income

Household and family incomes (HIND and FINF) were not collected in the Census but were derived from person level income data. It is not possible to aggregate person income ranges to derive household and family incomes. To overcome this, data from the 1999/2000 SIHC were used to impute an income value for each person. The imputed values for each person were then aggregated to create imputed household and family level incomes.

### The imputation process

The process involved analysis of the SIHC data to determine the imputation values to be used. Each of the 12,000 SIHC person records had the appropriate Census income range identifier (as defined above) allocated to it. For each range, the weighted mean, median, and midpoint of the range (with an arbitrarily assigned value used as the midpoint of the \$1,500 or more range) were calculated. Each of these measures were then aggregated to derive imputed household and family level incomes. These imputes were then compared with the actual household and family incomes reported in SIHC to determine which would be used to impute Individual Incomes for Census records:

- Initial analysis involved comparing the proportion of households and families assigned to their correct Census income range using the different methods of imputing personal income. From this analysis, the median imputation method gave the best results.
- Other analysis involved comparing the weighted relative frequencies for the different imputed household and family income ranges, with the actual income range weighted relative frequency distributions from SIHC. This analysis was done for Australia, state, and metropolitan/ex metropolitan regions. Again, the median imputation method gave the best results.

However, differences between person and household income ranges caused some problems. The ranges used for household and family level income are slightly finer than person level income:

<i>Range Identifier</i>	<i>Weekly Household or Family Income</i>
1	Negative income
2	Nil income
3	1-\$39
4	\$40-\$79
5	\$80-\$119
6	\$120-\$159
7	\$160-\$199
8	\$200-\$299
9	\$300-\$399
10	\$400-\$499
11	\$500-\$599
12	\$600-\$699
13	\$700-\$799
14	\$800-\$999
15	\$1,000-\$1,199
16	\$1,200-\$1,499
17	\$1,500-\$1,999
18	\$2,000 or more

At the higher end of the scale some one-income households and families were assigned to incorrect income ranges due to the fixed person level imputes used. To overcome this problem, the use of randomly assigned income values was investigated. Randomly assigned person level imputes were generated using assorted relative frequency distributions obtained from weighted SIHC data. These were used to generate household and family level incomes as described above. The resulting imputed household and family income distributions, compared to the actual income distribution from SIHC, were marginally better than imputed distribution from using median imputes. However, the randomly assigned imputes resulted in significantly more households and families being assigned to incorrect income ranges compared to the median imputes. Thus, the conclusion reached was that the median imputes were the most appropriate to use for the 2001 Census household and family income imputation.

## The imputed values

The imputed values (Estimated income value) for each person income range, calculated using SIHC median values, were:

<i>Range Identifier</i>	<i>Individual Income (weekly)</i>	<i>Estimated income value</i>
1	Negative income	0
2	Nil income	0
3	1-\$39	15
4	\$40-\$79	60
5	\$80-\$119	100
6	\$120-\$159	150
7	\$160-\$199	180
8	\$200-\$299	246
9	\$300-\$399	349
10	\$400-\$499	449
11	\$500-\$599	548
12	\$600-\$699	654
13	\$700-\$799	750
14	\$800-\$999	887
15	\$1,000-\$1,499	1,154
16	\$1,500 or more	1,831

These are the values used when summing records to create household and family incomes, and in calculating median values. NOTE: Individual Income is only published in ranges, so these estimated values will not apply.

### Median Values for open-ended ranges

To calculate a median value for Individual, Family or Household Income, it is necessary to identify the range in which the median lies and then to estimate where within the range the median would be. When the median lies within a range which does not have two specified finite end points (ie: \$2,000 or more in the case of HIND and FINF) the default value (ie: \$2,000 in the case of HIND and FINF) is retained. This is generally indicated as a table note, that when this value appears in the table the true median income is some value in the range \$2000 or more.

### Introduction of Sampling Error

The median income values from the SIHC used as the impute values are subject to sampling error, since SIHC is a sample survey. It would be appropriate to indicate that the household and family incomes, and therefore any derivations based on these values such as median income values for these units, are subject to both sampling and non-sampling error.