Fluctuations in the death statistics for datazones

Vital Events statistics are produced from complete counts of all the events which were registered, and so are not subject to some of the kinds of errors that may affect the results of sample surveys. However, the figures for a datazone may be subject to large percentage fluctuations from (say) year to year. A separate section 'Datazone Codes' explains what datazones are and can be found in the Geographical basis of Vital Events statistics PDF document. It follows that the figure for a datazone for any given year may provide an unreliable indicator of the usual annual number of events. This note illustrates this point using the numbers of deaths for individual datazones for 2001 to 2007 (the years for which the National Records of Scotland (NRS), formerly the General Register Office for Scotland (GROS), conducted the ad-hoc analysis whose results are described below). A workbook containing the tables and chart is available via a link at the foot of the page.

Deaths for each datazone for each year from 2001 to 2007

55,986 deaths were registered in Scotland in 2007. As there are 6,505 datazones, this represents an average of 8.6 deaths per datazone in that year. However, the number of deaths varies greatly between datazones. Areas with establishments like old people's homes and hostels for people with particular types of problem will tend to have higher numbers of deaths than areas whose residents are mainly families with young children. Table 1 shows how many datazones had each number of deaths (0, 1, 2, 3, …) in 2007. The results may be summarised as follows:
- 26% of datazones had 0 to 4 deaths each;
- 27% had 5 to 7 deaths each;
- 24% had 8 to 11 deaths each; and
- 23% had 12 or more deaths each. 20 datazones had 40 or more deaths each, including one datazone which had 70 deaths.

When the data for 2001 to 2007 were used to calculate the (rounded) average number of deaths per year for each datazone, a broadly-similar pattern was found:
- 21% of datazones had an average of 0 to 4 deaths per year;
- 30% had an average of 5 to 7 deaths per year;
- 27% had an average of 8 to 11 deaths per year; and
- 22% had an average of 13 or more deaths per year. 12 datazones had averages of 40 or more deaths per year, including one datazone which had an average of 70 deaths per year.
Table 2 shows the numbers of datazones which had each (rounded) average number of deaths.

However, while there might not be much difference between years in the overall distribution of the number of deaths per datazone, there are some large percentage year-to-year fluctuations in the figures for individual datazones. Table 3 shows the numbers of deaths in each year for the 20 datazones which have the largest total number of deaths over the seven years, together with the total and the annual average. The datazone with most deaths had 491 over the seven years, or an average of 70 per year, and its figures for individual years ranged from 56 (in 2005) to 80 (in 2003) - so it had a 30% fall in deaths between 2003 and 2005. The datazone with the second largest total number of deaths had an average of 58 per

year, and its numbers in individual years were between 49 (in 2001) and 68 (in 2002) - so it had a 39% increase in deaths between 2001 and 2002. Other large year-to-year fluctuations can be seen in the figures for the other datazones. The chart illustrates them. The general pattern is one of 'random' rises and falls in the number of deaths in each datazone, although there may be an occasional datazone for which there could be a trend. For example, the number of deaths in a datazone could decline over a number of years if its population were falling due to, for example, the demolition of blocks of flats and the subsequent use of the land for, say, lower density housing; the number of deaths could rise if (e.g.) formerly vacant land were used to build accommodation for old people. Of the 20 datazones shown, the most likely to be subject to such trends are the eighth (its figures are 83, 49, 48, 45, 31, 36 and 19) and, to much less marked extent, the fifth (36, 41, 48, 48, 70, 46, 56).

The likely range of random fluctuation in the number of deaths for a datazone

Because the number of deaths in a typical datazone fluctuates 'randomly' from year to year, one can calculate 95% confidence intervals which indicate the likely range of random variation in its figures: based on statistical theory, one would expect a figure outwith that range in only about one year in twenty. The range is calculated on the assumption that the number of deaths can be represented as the outcome of a 'Poisson process' (a separate page provides more information about this). Table 3 shows the likely lower and upper 'limits' for each of the twenty datazones with the most deaths, and has columns which indicate, for each datazone for each year, whether the number of deaths was below the lower 'limit' or above the upper 'limit'. With 20 datazones, one would expect that (on average) one datazone per year would have a figure which was outwith its '95% confidence interval', and therefore seven cases of figures outwith the 'limits' in the seven years. In fact, there were 12 such cases. The most likely reasons for more cases than predicted being outwith the 'limits' are:

- trends in the figures for some datazones (e.g. due to redevelopment) meaning that the underlying rate changes during the period - so the overall average annual rate does not represent the underlying rate in, say, the years towards each end of the period, and the 'limits' for those years are not based on their true underlying rate. Three of the 12 'outwith limit' cases were for the eighth datazone (whose numbers fell from 83 in 2001 to 19 in 2007), and two more were for the fifth datazone (whose numbers rose from 36 to 56);
- factors, such as the opening (or closing) of an establishment like an old people's home, which could change markedly the underlying rate part-way through the period - again the overall average annual rate would not represent the underlying rate in, say, the years towards each end of the period, and the 'limits' for those years would not be based on their true underlying rate;
- the fact that deaths may not occur 'randomly' - for example, several people who live in the same area may die as the result of a single event (e.g. a fire in a care home, or a bad car crash), so the scale of the fluctuations may be greater than predicted by statistical theory.

Table 4 shows 20 of the datazones with the smallest non-zero numbers of deaths in total over the seven years: two datazones with 1 death, two with 2 deaths, ten with 3 deaths and only six of the datazones with 4 deaths (because Table 4 shows only 20

datazones in total). Some of these datazones had two deaths in a single year, even though they had at most 4 deaths in seven years.

Table 5 illustrates how the figures for individual datazones can fluctuate from year to year, by showing a 'randomly-selected' example datazone for each (rounded) value of the annual average number of deaths, up to an average of 36 deaths per year (because Table 3's list of the 20 datazones with the most deaths shows ones with averages of 37 and more). For example, in the cases of:

- a datazone with an average of 2.6 deaths per year, the numbers varied between 0 (in 2001 and 2004) and 8 (in 2005);
- a datazone with an average of 5.6 deaths per year, the numbers varied between 2 (in 2007) and 9 (in 2005);
- a datazone with an average of 9.6 deaths per year, the numbers varied between 6 (in 2005) and 14 (in 2002).

Table 5 shows likely lower and upper '95% confidence intervals' (based on the unrounded overall annual average) for each of the example datazones, and whether each year's figure was outwith those 'limits'. One would expect about 1-in-20 of the 259 numbers (one for each of the 7 years for each of the 37 example datazones) to be outwith the '95% confidence intervals' - and that is so (although, with 12 cases below the lower 'limit' and only one above the upper 'limit', the outcome is less 'balanced' than expected).

The workbook containing the tables and chart is available through the following link: Fluctuations in the death statistics for datazones (77 Kb Excel file)