Estimate maximum number of cases of death by suicide (&undetermined intent) from summaries provided by NRS up to 2017

Jan Savinc

24 June, 2019

Contents

Definition of cases

We defined cases as individuals who were born 1981 or later and died as a result of suicide (including both self-harm and events of undetermined intent) between 1st January 1991 and 31st December 2017 and were 10–34 years of age.

For the purposes of estimating the number of cases that we may be able to extract, we process the total numbers of suicides published by NRS Scotland. These are binned by age groups of 5 years, which means we need to define which age groups are to be included. This means we can estimate a lower and an upper bound of the number of cases we can expect to be able to extract from the records. The upper & lower bound can then be further adjusted by subtracting an estimated rate of missing data.

Downloading data

The data was found by searching the NRS Scotland website, specifically the Vital Events section: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/vital-events/deaths/suicides/

On the above website there is a URL link to the spreadsheet, which we'll use to download and save the data the *raw* folder.

```
url <- "https://www.nrscotland.gov.uk/files//statistics/probable-suicides/2017/prop-suicides-17-all-tabs
file <- "./raw/prop-suicides-17-all-tabs.xlsx"

download.file(
  url=url,
  destfile=file, mode = "wb"
  )
```

There are several sheets in the spreadsheet: Contents, 1 - Sex and type of cause, chart 1, 2 - Method and Usual residence, 2b - Nature of death and Method, chart 2, 3 - Age-group, 3F - Females by Age-group, 3M -

Males by Age-group, 4 - Health Board, 5 - Local Authority, figures for chart 1, figures for chart 2, and we are interested in the sheet **“3 - Age-group”**

Cleaning data

The spreadsheet contains three tables: self-harm & undetermined intent cases combined, and then each separately, all arranged vertically. We are only interested in the first, so we only read down to the end of the first table.

There are duplicate rows for 2009 to 2017 when a new coding system was in place. We only use the rows labelled **old** in this period.

The other cleaning steps involve removing rows with no information recorded on them, and reformatting the data into a long format, where each row stands for a number of cases associated to a year, and the minimum and maximum age bracket.

```r
suicides <- read_excel(file, sheet="3 - Age-group", trim_ws = TRUE,skip = 6) %>%
  slice(-1) %>%  # remove blank 1st row
  slice(1:44) %>%  # only keep rows 1 to 44
  rename(year_death = "..1") %>%
  filter(str_detect(year_death,pattern="^[0-9]{4}$|^[0-9]{4}.*old")) %>%
  mutate(year_death=parse_number(year_death)) %>%
  select(-`All3 ages`) %>%
  gather(-year_death,key="age",value="num") %>%
  separate(age, sep="-", into=c("age_lo","age_hi")) %>%
  mutate(age_lo=ifelse(age_lo=="85+","85",age_lo)) %>%
  mutate_all(as.numeric) %>%
  mutate(age_hi=ifelse(is.na(age_hi),Inf,age_hi)) %>%  # change upper bound to infinity in top category
  mutate(age_hi=ifelse(age_lo==0,4,age_hi))  # change 0-4 instead of 0-44
```

```
## New names:
## * `` -> `..1`
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 44 rows
## [749, 750, 751, 752, 753, 754, 755, 756, 757, 758, 759, 760, 761, 762, 763,
## 764, 765, 766, 767, 768, ...].
```

Now that the data is cleaned up, we can start subsetting it to sum only the cases we are interested in.

Subsetting data for upper & lower bound

we are only interested in cases born in 1981 or later, and who died aged 10 to 34 between 1991-01-01 and 2017-12-31.

For the upper bound, we’ll include the number of cases where the maximum age of cases crossed the lower threshold for each age group - i.e. the age group 10-14 will first be included in the year 1991, when the cases would be aged 10 maximum. This may overestimate the number of cases, because some of those include may have been aged e.g. 14 in 1991, and would therefore have been born before 1981.

Conversely, for the lower bound, we’ll only count age groups when the cases where at the maximum of the age bracket, i.e. the age group 10-14 will only be included from 1995, when cases would definitely be aged 14. This may underestimate the number of cases, since this way we will ignore any cases aged 10-13 who died 1991-1994.

TODO: check if there was also an upper DOB limit

```
min_age_included <- 10
max_age_included <- 34
year_start <- 1991  # an equivalent maximum year not needed - we only have data until end of 2017

suicides_cases_upper_bound <-
  suicides %>%
  filter(year_death >= year_start) %>%  # filter out all cases pre-1991
  mutate(
    cases_age_youngest = min_age_included,  # define youngest cases each year (i.e. aged 10)
    cases_age_oldest = min_age_included + year_death - year_start  # define odlest cases each year (i.e
    ) %>%
  filter(
    age_lo >= min_age_included &  # keep only rows where ages 10 or more are included
      age_hi <= max_age_included &  # keep only rows where ages 34 or less are included
      cases_age_oldest >= age_lo  # keep only rows where oldest possible cases that year cross minimum
    ) %>%
  arrange(year_death)

suicides_cases_lower_bound <-
  suicides %>%
  filter(year_death >= year_start) %>%  # filter out all cases pre-1991
  mutate(
    cases_age_youngest = min_age_included,  # define youngest cases each year (i.e. aged 10)
    cases_age_oldest = min_age_included + year_death - year_start  # define odlest cases each year (i.e
    ) %>%
  filter(
    age_lo >= min_age_included &  # keep only rows where ages 10 or more are included
      age_hi <= max_age_included &  # keep only rows where ages 34 or less are included
      cases_age_oldest >= age_hi  # keep only rows where oldest possible cases that year are at maximum
    ) %>%
  arrange(year_death)
```

```
number_of_cases_estimated_upper_limit <- sum(suicides_cases_upper_bound$num)
number_of_cases_estimated_lower_limit <- sum(suicides_cases_lower_bound$num)
```

Estimates

The estimated figures were as follows:

```
tibble(bound=c("Upper","Lower"), N=c(number_of_cases_estimated_upper_limit,number_of_cases_estimated_lo
```

bound	N
Upper	3020
Lower	2016

We can adjust these for the rate of missing data (because of missing CHI numbers). This is estimated from emails with eDRIS, who > found 28,667 deaths of persons aged between 10 and 34 years in the period 1/1/1991 – 31/12/2017 (after excluding records with missing CHI – 979)

```
num_missing <- 979
num_nonmissing <- 28667
```

```
rate_missing <- num_missing / (num_missing+num_nonmissing)

number_of_cases_estimated_upper_limit_adjusted <- round(number_of_cases_estimated_upper_limit * (1-rate_
number_of_cases_estimated_lower_limit_adjusted <- round(number_of_cases_estimated_lower_limit * (1-rate_
```

This makes for a rate of missing CHI numbers of 3.30%. The adjusted upper & lower bound are therefore as follows:

```
tibble(bound=c("Upper (adjusted)","Lower (adjusted)"), N=c(number_of_cases_estimated_upper_limit_adjuste
```

bound	N
Upper (adjusted)	2920
Lower (adjusted)	1949