Framingham Case Study

The Framingham Heart Study led to the publication of more than 3,000 journal articles, the most highly cited of which is Gordon et al. *Am J Med* 1977, a seminal paper that linked high HDL cholesterol with lower risk for CHD.

We will write a literate program that uses data from the Framingham study to examine whether we can reproduce the link between HDL cholesterol and outcomes such as CHD or death from any cause.

Install the tidyverse packages, if necessary

```
if (!"tidyverse" %in% installed.packages()) install.packages("tidyverse")
```

Download the Framingham data

Download `framingham.xlsx` from the course website.

```
download.file("https://skadauke.github.io/cdar_course/framingham.xlsx",
              destfile = "framingham.xlsx")
```

Load the data

First load the `tidyverse` and `readxl` packages for data manipulation and functions and to load Excel files.

Then load the file `framingham.xlsx` into a data frame named `framingham`.

```
library(tidyverse)
library(readxl)

framingham <- read_excel("framingham.xlsx")
```

Look at the data

> Examine the structure of the data using the `str()` function.

> Use the `summary()` function on `framingham` to create a brief statistical summary of the distribution of data in each of its columns.

> What proportion of participants were smokers at the time of enrollment? What proportion had CHD? What proportion of participants developed CHD at the end of the follow up period? What proportion of participants died?

Hint: You don't need to enter any additional R code to answer these questions. Take a look at the output of the `summary()` command and remember that if the values of a vector are only `0` and `1` (or `TRUE` and `FALSE`), the mean of the vector is equal to the proportion of values that are `1` (or `TRUE`).

From the above look at the data, it appears that a number of rows are missing HDL cholesterol data. To be exact, there are a total of 236 rows with missing HDL data.

> What is the proportion of rows with missing HDL data?