Applied Statistics II, Lab 1

J. Arturo Esquivel

11/01/23

Table of contents

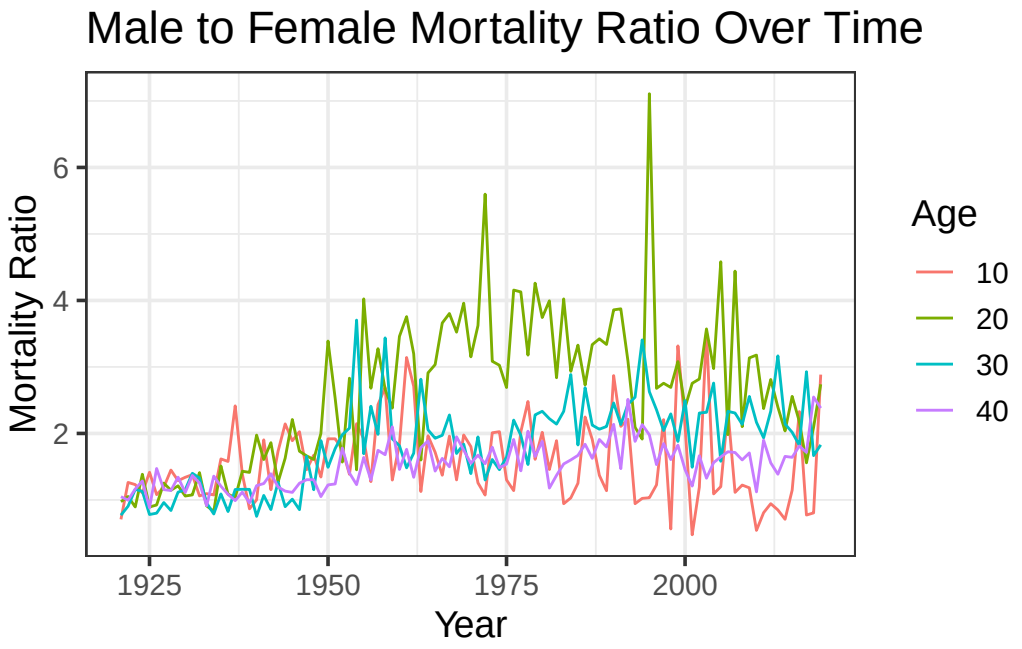

```
#install.packages("tidyverse")
library(tidyverse)
```

```
dm <- read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt", skip = 2, col_t
```

Lab Exercises

1. Plot the ratio of male to female mortality rates over time for ages 10,20,30 and 40 (different color for each age) and change the theme

```
# Contruct the ration and keep only the desired ages
  dm_ratio <- dm |>
  mutate(mf_ratio = Male/Female) |>
  filter(Age==10 | Age==20 | Age==30 | Age==40)

# Plot graph
p <- ggplot(data = dm_ratio, aes(x = Year, y = mf_ratio))
p +
  geom_line(aes(x = Year, y = mf_ratio, color = Age)) +
  labs(title = "Male to Female Mortality Ratio Over Time",
       y = "Mortality Ratio") +
  theme_bw(base_size = 14)
```



2. Find the age that has the highest female mortality rate each year

```
summary_max <- dm |>
  group_by(Year) |>
  # Assign to each year the age with maximum mortality rate in the group
  mutate(Max_Mort_Age = Age[which.max(Female)]) |>
  # Keep one entry per year
  summarize(Max_Mort_Age = max(Max_Mort_Age, na.rm = TRUE))

head(summary_max)
```

```
# A tibble: 6 x 2
   Year Max_Mort_Age
  <dbl> <chr>
1  1921 106
2  1922 98
3  1923 104
4  1924 107
5  1925 98
6  1926 106
```

3. Use the `summarize(across())` syntax to calculate the standard deviation of mortality rates by age for the Male, Female and Total populations.

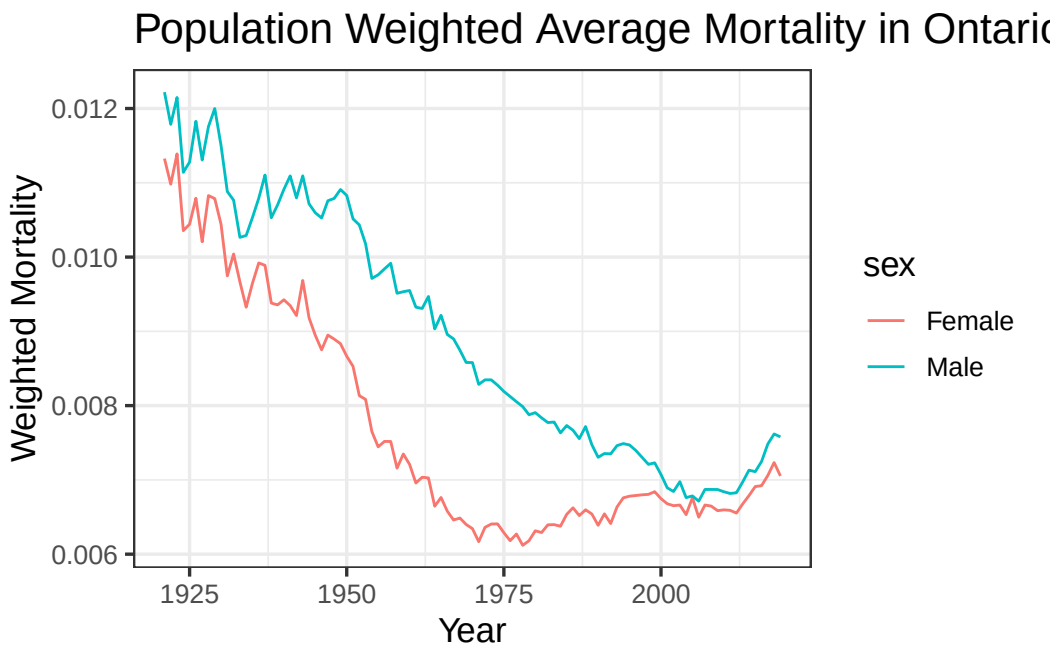
```r
sd_age <- dm |>
  # Turn age column into integers so that the dataframe is correctly sorted when shown
  mutate_at(2, as.integer) |>
  group_by(Age) |>
  summarize(across(Female:Total, sd, na.rm = TRUE))

head(sd_age)
```

```
# A tibble: 6 x 4
    Age   Female     Male    Total
  <int>    <dbl>    <dbl>    <dbl>
1     0 0.0256   0.0330   0.0294
2     1 0.00352  0.00396  0.00374
3     2 0.00154  0.00175  0.00164
4     3 0.00113  0.00127  0.00120
5     4 0.000925 0.000987 0.000947
6     5 0.000748 0.000820 0.000776
```

4. The Canadian HMD also provides population sizes over time (https://www.prdh.umontreal.ca/BDLC/data
   Use these to calculate the population weighted average mortality rate separately for males and females, for every year. Make a nice line plot showing the result (with meaningful labels/titles) and briefly comment on what you see (1 sentence). Hint: `left_join` will probably be useful here.

```r
data <- read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Population.txt", skip = 1)

data <- data |>
  pivot_longer(Female:Total, names_to = "sex", values_to = "Size")

# Get the total population size per year and sex
totals <- data |>
  group_by(Year, sex) |>
  summarize(Total_size = sum(Size, na.rm = TRUE))

# Compute the weight for each each per year
data_totals <- left_join(data, totals, by = c("Year", "sex")) |>
  mutate(weight = Size/Total_size)

dm_long <- dm |>
  pivot_longer(Female:Total, names_to = "sex", values_to = "mortality")
```

```r
# Compute mortality per year by adding the weighted mortality of all ages
plot_data <- left_join(data_totals, dm_long, by = c("Year", "sex", "Age")) |>
  # weighted mortality
  mutate(w_mortality = weight*mortality) |>
  group_by(Year, sex) |>
  # Sum all weighted moralities per year
  summarize(Total_mortality = sum(w_mortality, na.rm = TRUE)) |>
  # Remove 2020 due to missing data
  filter(sex=="Female" | sex=="Male", Year<2020)

# Plot graph
p <- ggplot(plot_data, aes(x = Year, y = Total_mortality))
p +
  geom_line(aes(x = Year, y = Total_mortality, color = sex)) +
  labs(title = "Population Weighted Average Mortality in Ontario Through Time",
       y = "Weighted Mortality") +
  theme_bw(base_size = 13)
```



Female mortality is consistently lower than male's. In both cases mortality has had a decreasing trend, stabilizing after 1990 for females, and around 2000 for males.