STA2201 Week 1 Lab

Yeonjoon Choi

2023-01-11

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
2. Find the age that has the highest female mortality rate each year
3. Use the `summarize(across())` syntax to calculate the standard deviation of mortality rates by age for the Male, Female and Total populations.
4. The Canadian HMD also provides population sizes over time (https://www.prdh.umontreal.ca/BDLC/data/ont/Population.txt). 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.

Part 1.

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.4.0      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.2
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
## Warning: package 'tibble' was built under R version 4.2.2
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

```
## Warning: package 'forcats' was built under R version 4.2.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

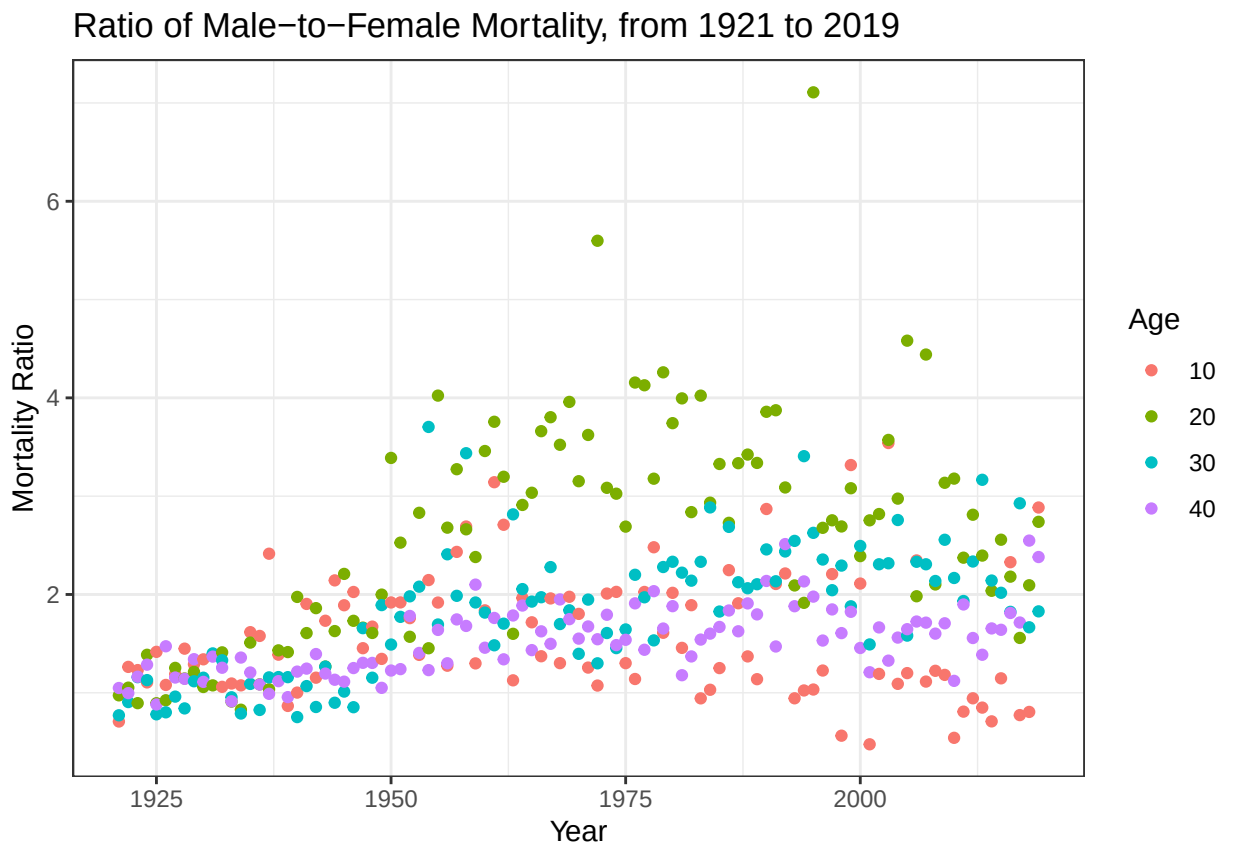
```
dm = read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt",
                skip = 2, col_types = "dcddd")
```

```
## Warning: 494 parsing failures.
## row    col               expected actual                                                file
## 108 Female no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt'
## 109 Female no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt'
## 110 Female no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt'
## 110 Male   no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt'
## 110 Total  no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt'
## ... ...... ..................... ...... ..................................................
## See problems(...) for more details.
```

```
#We get the mortality ratio for each age
data =  dm |>
    filter(Age == 10|Age == 20| Age == 30| Age ==40)|>
    select(Year:Male)|>
    mutate(ratio = Male/Female)

data
```

```
## # A tibble: 396 x 5
##     Year Age    Female    Male ratio
##    <dbl> <chr>   <dbl>   <dbl> <dbl>
##  1  1921 10    0.00239 0.00169 0.708
##  2  1921 20    0.00298 0.00290 0.974
##  3  1921 30    0.00486 0.00375 0.771
##  4  1921 40    0.00618 0.00648 1.05
##  5  1922 10    0.00159 0.00201 1.26
##  6  1922 20    0.00280 0.00294 1.05
##  7  1922 30    0.00510 0.00462 0.906
##  8  1922 40    0.00642 0.00640 0.996
##  9  1923 10    0.00140 0.00172 1.23
## 10  1923 20    0.00301 0.00270 0.894
## # ... with 386 more rows
```

```
#We plot the data
data |>
  ggplot(aes(Year, ratio))+
  geom_point(aes(colour = Age))+
  theme_bw()+
  labs(title = "Ratio of Male-to-Female Mortality, from 1921 to 2019")+
  ylab ("Mortality Ratio")
```



Part 2.

```
#We group by Year first
grouped_year = dm |>
  group_by(Year)

#and find the age that has
#the max mortality each year
mortality_age = summarize(grouped_year,
          Max_Age=Age[which.max(Female)])


#We will see the first 10 years and the last 10 years
head(mortality_age, n =10)
```

```
## # A tibble: 10 x 2
##     Year Max_Age
##    <dbl> <chr>
##  1  1921 106
##  2  1922 98
##  3  1923 104
##  4  1924 107
##  5  1925 98
##  6  1926 106
##  7  1927 106
##  8  1928 104
```

```
##  9  1929 104
## 10  1930 105
```

```
tail(mortality_age, n =10)
```

```
## # A tibble: 10 x 2
##     Year Max_Age
##    <dbl> <chr>
##  1  2010 108
##  2  2011 110+
##  3  2012 109
##  4  2013 110+
##  5  2014 110+
##  6  2015 110+
##  7  2016 110+
##  8  2017 110+
##  9  2018 110+
## 10  2019 110+
```

Part 3

```
#We group by age
grouped_age = dm |>
  group_by(Age)

#and calculate the standard deviation
summarize(grouped_age, across(Female:Total, sd))
```

```
## # A tibble: 111 x 4
##    Age      Female      Male     Total
##    <chr>     <dbl>     <dbl>     <dbl>
##  1 0      0.0256    0.0330    0.0294
##  2 1      0.00352   0.00396   0.00374
##  3 10     0.000474  0.000561  0.000509
##  4 100    0.0928    0.138     0.0729
##  5 101    0.125     0.158     0.0995
##  6 102    0.143     0.214     0.114
##  7 103    0.252     0.371     0.208
##  8 104    0.449    NA         0.363
##  9 105   NA        NA        NA
## 10 106   NA        NA        NA
## # ... with 101 more rows
```

Part 4

```
pop_data = read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Population.txt",
                      skip = 2)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
```

```
##   Year = col_double(),
##   Age = col_character(),
##   Female = col_double(),
##   Male = col_double(),
##   Total = col_double()
## )
```

```
death_data = read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt",
                        skip = 2, col_types = "dcddd")
```

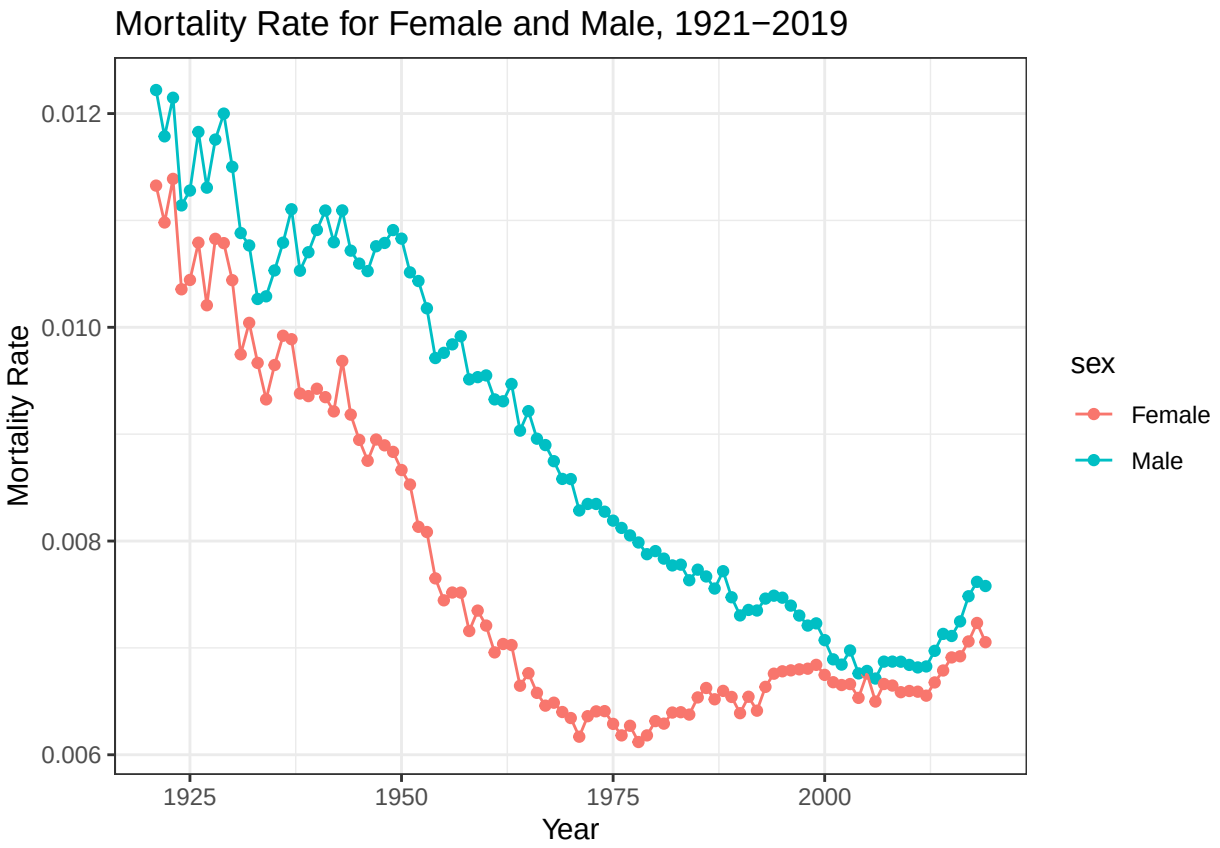
```
## Warning: 494 parsing failures.
## row    col               expected actual                                                  file
## 108 Female no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt'
## 109 Female no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt'
## 110 Female no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt'
## 110 Male   no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt'
## 110 Total  no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt'
## ... ...... ..................... ...... ..................................................
## See problems(...) for more details.
```

```
death_data = death_data|>
  select(-Total)|>
  pivot_longer(Female:Male, names_to = "sex", values_to = "mortality")

#death data has no 2020 data
#so we remove any data after 2020
#from the data set
pop_data = pop_data|>
  select(-Total)|>
  filter(Year < 2020)|>
  pivot_longer(Female:Male, names_to = "sex", values_to = "Population")


death_data$population = pop_data$Population

#Calculate the weight for the weighted average
death_data = death_data |>
  mutate(weight = mortality*population)


#Calculate weighted average

to_plot = death_data |>
  group_by(Year, sex) |>
  summarize(weighted_mean = sum(weight, na.rm = TRUE)/sum(population, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

```
#Plotting

to_plot |>
```

```
ggplot(aes(Year, weighted_mean, colour = sex))+
 geom_point()+
  geom_line()+
  labs(title = "Mortality Rate for Female and Male, 1921-2019")+
  ylab("Mortality Rate")+
  theme_bw()
```



We see that male mortality rate is higher than female mortality rate, but we also see that the difference shrunk as time went on.