ADS2 Group Exercise ICA

Group 3

3/4/2022

Contents

Load the libraries

```
library(ggplot2)
library(tidyverse)
library(paletteer)
```

Load the Dataset

```r
malaria <- read.csv("malaria.csv")
head(malaria, 3)
```

```
##   measure                 location  sex         age   cause metric year
## 1  Deaths East Asia & Pacific - WB Both     Under 5 Malaria Number 2000
## 2  Deaths East Asia & Pacific - WB Both  5-14 years Malaria Number 2000
## 3  Deaths East Asia & Pacific - WB Both 15-49 years Malaria Number 2000
##         val     upper     lower
## 1 1873.7482  4692.839  741.3761
## 2  806.3444  1972.721  334.7347
## 3 4450.4837 10826.707 1905.1511
```

Clean the Data

Firstly, we choose to explore the dataset and clean the data, since this step will benefit following analysis.

Missing values

Deal with NA values and empty entries

```r
anyNA(malaria)
```

```
## [1] FALSE
```

```r
sum(malaria[ , ] == "")
```

```
## [1] 0
```

There is no missing values in the dataset. So we do not need to remove lines.

Duplicates in the data

Check for duplicates

```r
which(duplicated(malaria))
```

```
## integer(0)
```

There is no duplicate in the dataset. So we do not need to remove lines.

Typos and Naming schemes

Test the consistency of naming scheme of some columns and whether there are some typos

See the structure of the whole dataset first

```
str(malaria)
```

```
## ’data.frame':    700 obs. of  10 variables:
##  $ measure : chr  "Deaths" "Deaths" "Deaths" "Deaths" ...
##  $ location: chr  "East Asia & Pacific - WB" "East Asia & Pacific - WB" "East Asia & Pacific - WB" "
##  $ sex     : chr  "Both" "Both" "Both" "Both" ...
##  $ age     : chr  "Under 5" "5-14 years" "15-49 years" "50 to 74 years" ...
##  $ cause   : chr  "Malaria" "Malaria" "Malaria" "Malaria" ...
##  $ metric  : chr  "Number" "Number" "Number" "Number" ...
##  $ year    : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
##  $ val     : num  1873.7 806.3 4450.5 2057 68.3 ...
##  $ upper   : num  4693 1973 10827 4932 161 ...
##  $ lower   : num  741.4 334.7 1905.2 872.4 29.9 ...
```

```
# location column
table(malaria$location)
```

```
##
##       East Asia & Pacific - WB      Europe & Central Asia - WB
##                            100                             100
##  Latin America & Caribbean - WB Middle East & North Africa - WB
##                            100                             100
##                  North America                South Asia - WB
##                            100                             100
##        Sub-Saharan Africa - WB
##                            100
```

```
malaria$location <- gsub(' - WB', '', malaria$location) # delete ' - WB' part
table(malaria$location)
```

```
##
##       East Asia & Pacific      Europe & Central Asia
##                       100                        100
##  Latin America & Caribbean Middle East & North Africa
##                       100                        100
##             North America                 South Asia
##                       100                        100
##        Sub-Saharan Africa
##                       100
```

```
# age column
table(malaria$age)
```

```
##
##    15-49 years     5-14 years 50 to 74 years        75 plus        Under 5
##            140            140            140            140            140
```

```
malaria[malaria$age == '50 to 74 years', 'age'] <- '50-74 years' # replace 'to' with '-'
table(malaria$age)
```

```
##
## 15-49 years  5-14 years 50-74 years     75 plus     Under 5
##         140         140         140         140         140
```

Factoring *age* column

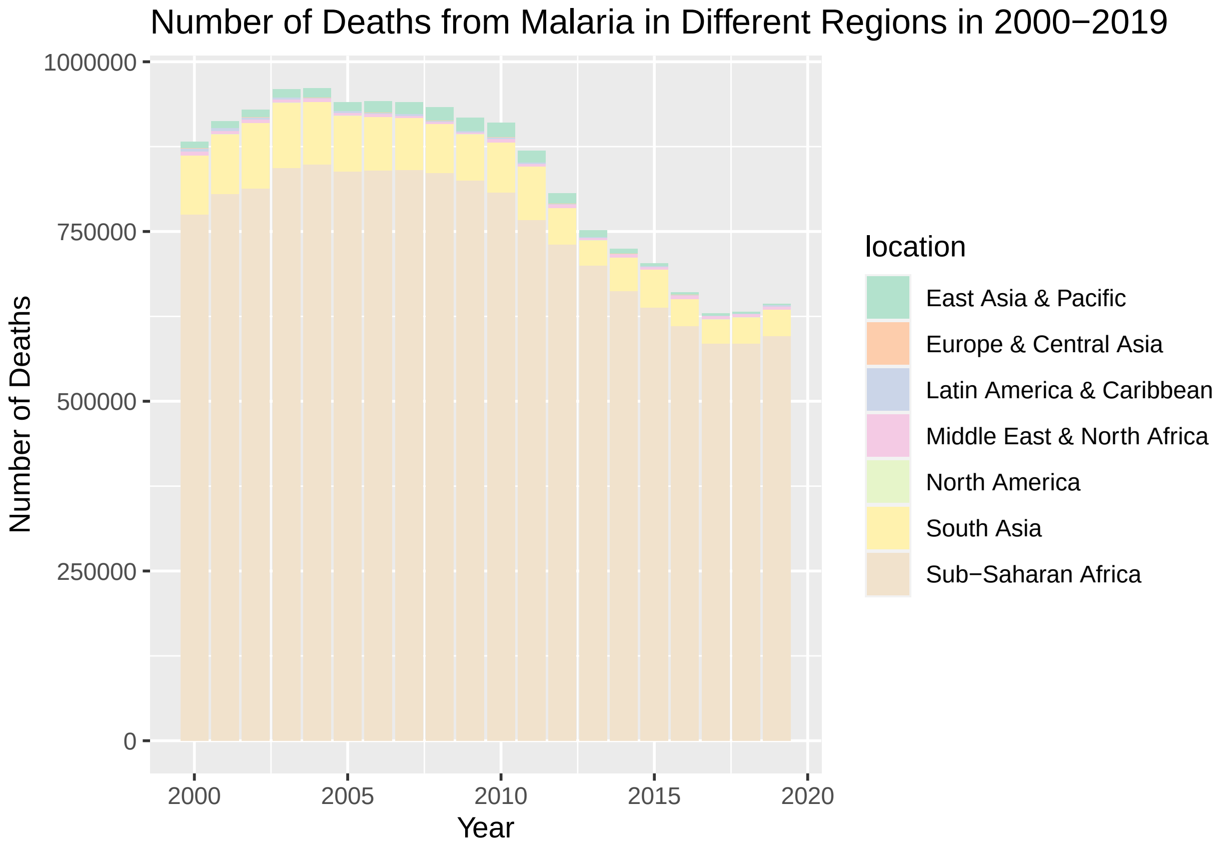
After checking the whole structure of this dataset, we want to turn *age* column into factor, in order to benefit the following plotting step.

```
malaria <- malaria %>% mutate(age = factor(age,
           levels = c('Under 5','5-14 years', '15-49 years','50-74 years','75 plus')))
```

Outliers & Strange Patterns

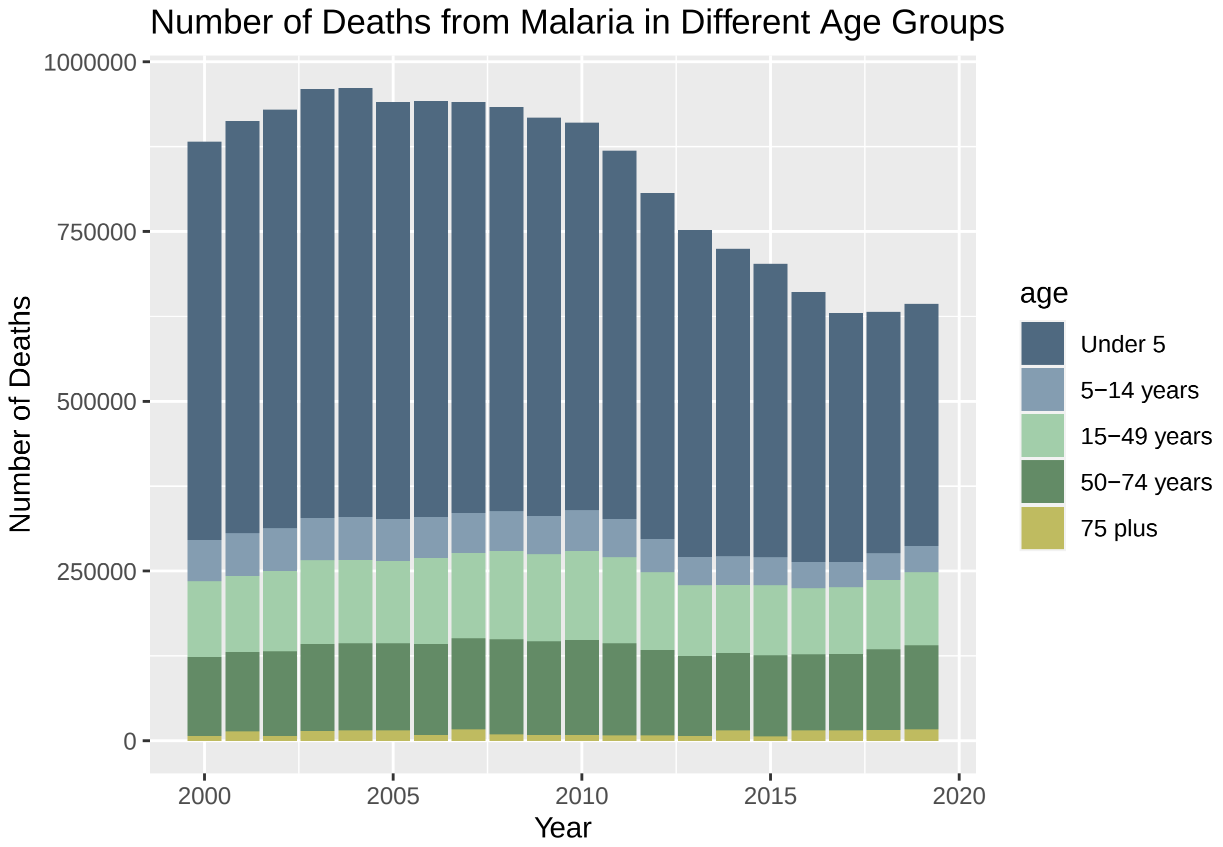
We wonder whether the data has outstanding outliers or strange distribution patterns. So we decide to plot the number of cases against different factors to explore and examine.

```
# deaths vs location
min_year <- min(malaria$year)
max_year <- max(malaria$year)

ggplot(malaria) +
  geom_bar(stat='identity', aes(x=year, y=val, fill=location)) +
  scale_fill_paletteer_d("RColorBrewer::Pastel2") +
  xlab("Year") +
  ylab("Number of Deaths") +
  ggtitle(paste("Number of Deaths from Malaria ", "in Different Regions in ",
                min_year,"-", max_year, sep = ''))
```



We can see that the great majority of deaths of malaria happened in Sub-Saharan Africa, followed by South Asia. However, with no additional information about the malaria cases, we decided to leave the data unchanged.

```
# deaths vs age
ggplot(malaria) +
  geom_bar(stat='identity', aes(x=year, y=val, fill=age)) +
  scale_fill_paletteer_d("ggthemes::Miller_Stone") +
  xlab("Year") +
  ylab("Number of Deaths") +
  ggtitle("Number of Deaths from Malaria in Different Age Groups")
```



We can see that the majority of deaths from malaria happened in age group “Under 5”. However, with no additional information about the malaria cases, we decided to leave the data unchanged.

Part 1: Exploring the data

In this part, we will answer the questions given in the guidance.

Question 1 Plotting the number of deaths

Question: Plot the number of deaths from malaria between 2000 and 2019 for each of the age groups for the East Asia and Pacific region. What age group seems to have the highest number of cases, and why do you think that is?

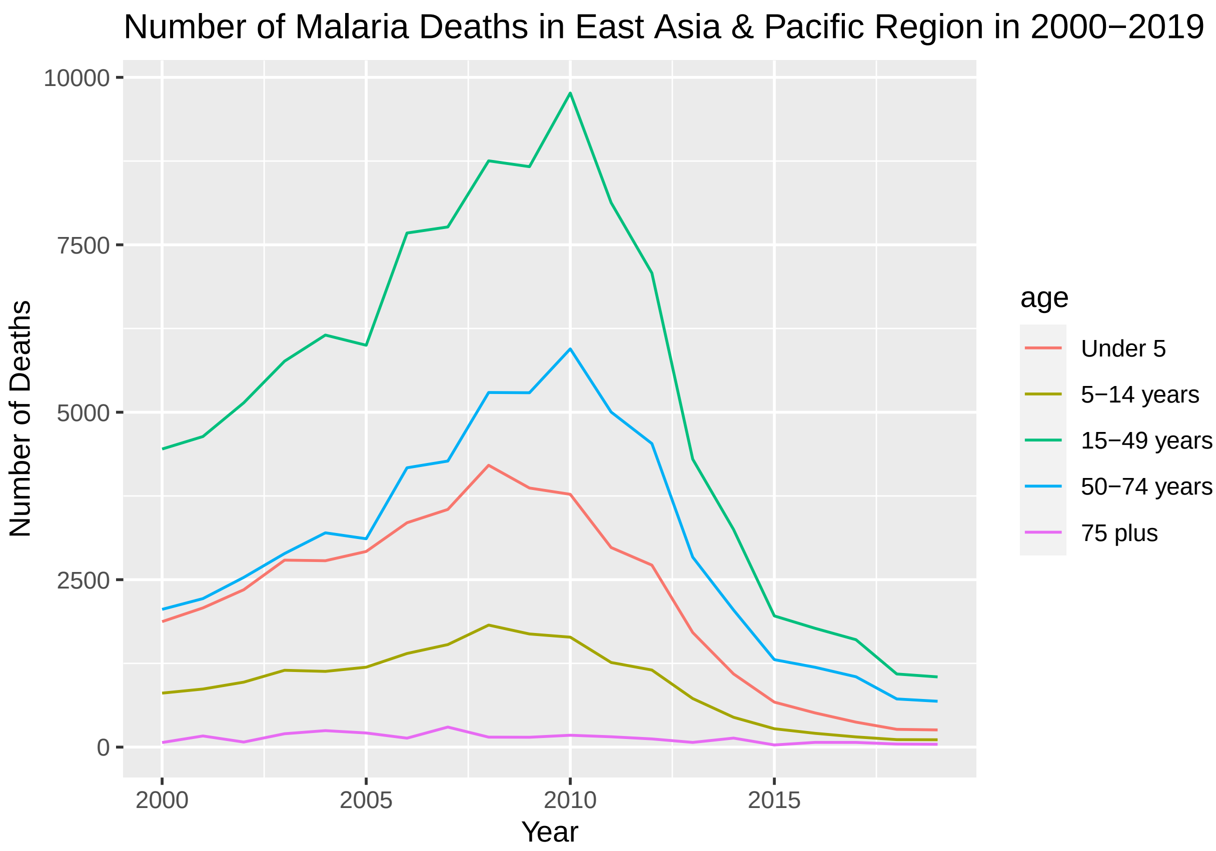
To solve this problem, firstly we need to extract data of the East Asia and Pacific region from the original dataset.

```
subregion <- malaria %>%
  filter(location == 'East Asia & Pacific') %>%
  mutate(age=factor(age, levels=c('Under 5','5-14 years',
                                  '15-49 years','50-74 years','75 plus')))
head(subregion, 2)
```

```
##   measure            location  sex       age   cause metric year       val
## 1  Deaths East Asia & Pacific Both    Under 5 Malaria Number 2000 1873.7482
## 2  Deaths East Asia & Pacific Both 5-14 years Malaria Number 2000  806.3444
##      upper    lower
## 1 4692.839 741.3761
## 2 1972.721 334.7347
```

Then we can plot the data depending on the time and age groups.

```
ggplot(subregion, aes(x = year, y = val, group = age)) +
  geom_line(aes(color = age)) +
  xlab("Year") +
  ylab("Number of Deaths") +
  ggtitle(paste("Number of Malaria Deaths ", "in East Asia & Pacific Region in ",
                min_year, "-", max_year, sep = ''))
```



As we can see from the results, it seems that the age group “15-49 years” always have the highest number of cases which died from malaria in every year between 2000 and 2019.

It may because that this group have the largest population compared with other age groups. Also, it is possible that this age group is optimal labour force, so they are more likely to get in touch with other people, so it have higher probability to get infected by malaria.

Question 2 Total number of malaria cases

Question: In which year was the total number of malaria cases (across all regions and age groups) the highest? In which year was it the lowest?

To solve this problem, we can transformed the data set, summarized number of deaths grouped by year.

```
malaria_year <- malaria %>%
  group_by(year) %>%
  summarise(val=sum(val))
head(malaria_year, 3)
```

```
## # A tibble: 3 x 2
##    year     val
##   <int>   <dbl>
## 1  2000 882060.
## 2  2001 912710.
## 3  2002 929197.
```

```
malaria_max <- malaria_year[malaria_year$val == max(malaria_year$val), 'year']
malaria_min <- malaria_year[malaria_year$val == min(malaria_year$val), 'year']

# Conclusion 1
print(paste("Total number of malaria cases was highest in", malaria_max))
```

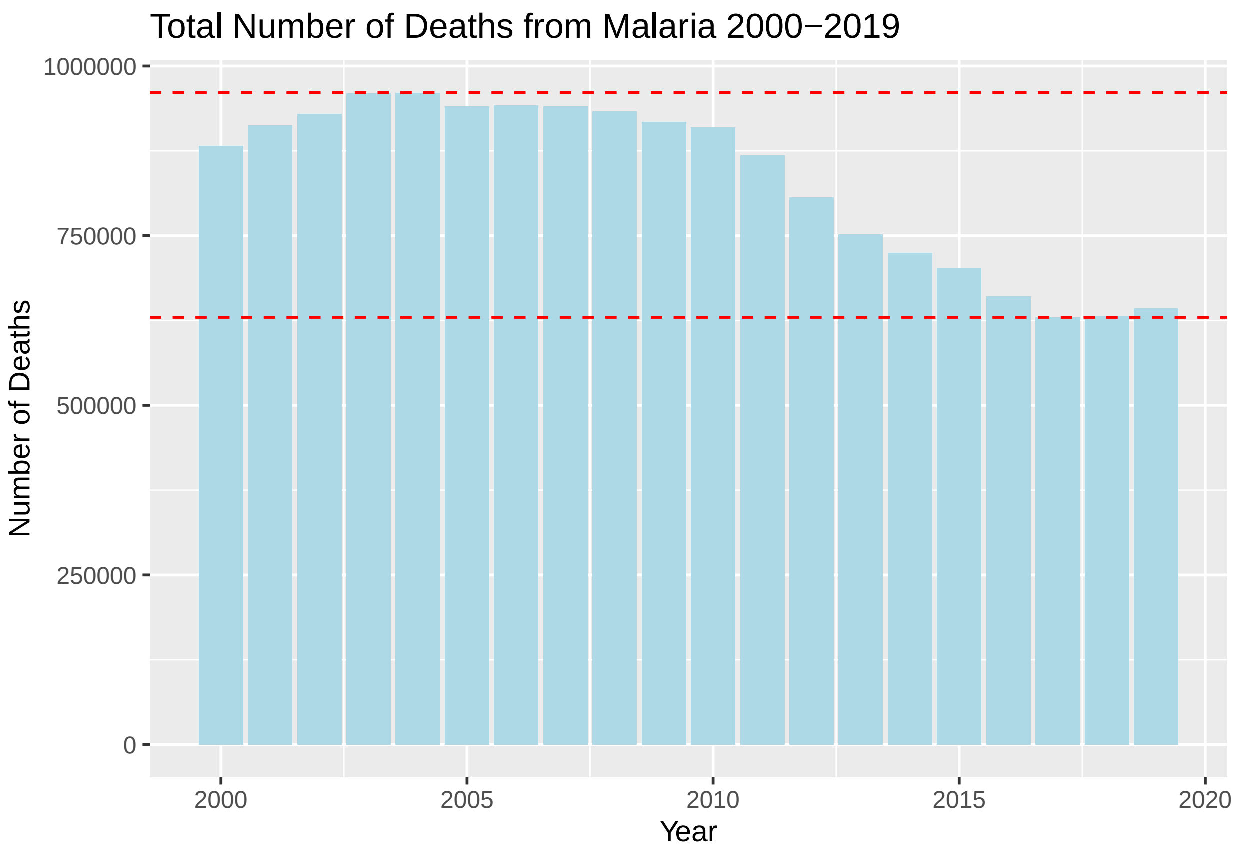
```
## [1] "Total number of malaria cases was highest in 2004"
```

```
# Conclusion 2
print(paste("Total number of malaria cases was lowest in", malaria_min))
```

```
## [1] "Total number of malaria cases was lowest in 2017"
```

This can also be verified by visualizing the number of total death each year.

```
ymax <- malaria_year[malaria_year$year == malaria_max$year, "val"]
ymin <- malaria_year[malaria_year$year == malaria_min$year, "val"]

ggplot(malaria_year) +
  geom_bar(stat='identity', aes(x=year, y=val), fill="lightblue") +
  xlab("Year") +
  ylab("Number of Deaths") +
  ggtitle(paste("Total Number of Deaths from Malaria ",
                min_year, "-", max_year, sep = '')) +
  geom_hline(col = "red", linetype = "dashed", yintercept = ymax$val) +
  geom_hline(col = "red", linetype = "dashed", yintercept = ymin$val)
```



Question 3 Percentage of deaths in certain region

Question: What percentage of total Malaria deaths in 2010 happened in the Latin America and Carribean region?

To solve this problem, we can transformed the data set, summarized number of deaths in 2010 grouped by location.

```
malaria_location <- malaria %>%
  filter(year == "2010") %>%
  group_by(location) %>%
  summarise(val=sum(val))
head(malaria_location, 2)
```

```
## # A tibble: 2 x 2
##   location                     val
##   <chr>                      <dbl>
## 1 East Asia & Pacific   21306.
## 2 Europe & Central Asia     0.936
```

Then we can choose total Malaria deaths happened in the Latin America and Carribean region.

```
malaria_latin_cari <- malaria_location %>%
  filter(location=='Latin America & Caribbean')
malaria_latin_cari
```

```
## # A tibble: 1 x 2
##   location                    val
##   <chr>                     <dbl>
## 1 Latin America & Caribbean 2675.
```

```
total_val <- sum(malaria_location$val)
total_val
```

```
## [1] 909816.9
```

```
# Conclusion
print(paste("About ", round(100*malaria_latin_cari$val/total_val, digits = 2),
            "% of total Malaria deaths in 2010 happened ",
            "in the Latin America and Carribean region.",
            sep = ""))
```

```
## [1] "About 0.29% of total Malaria deaths in 2010 happened in the Latin America and Carribean region.
```

Part 2: Ask our own question

In this part, we will ask one question that we wonder, and choose a suitable method to use the data provided to answer it.

The question we want to ask

Question: Is there a significant difference in the percentage of people dying from malaria each year in the Middle East and North Africa?

This percentage refers to the number of deaths of that region in that year/the total number of deaths of that region in all years.

Why we are interested in this question: TO be continued…..

Problem sovling process

Since we want to test whether there is a real difference between two samples, we have to formulate our hypothesis first.

Null Hypothesis: There is no significant difference of percentage of malaria deaths between any two years in Middle East & North Africa region.

Alternate Hypothesis: The differences of percentage of malaria deaths in Middle East & North Africa region are significant at least in two years.

Step1: Before using any statistical tests, we want to first plotted the percentage of deaths each year in Middle East & North Africa.

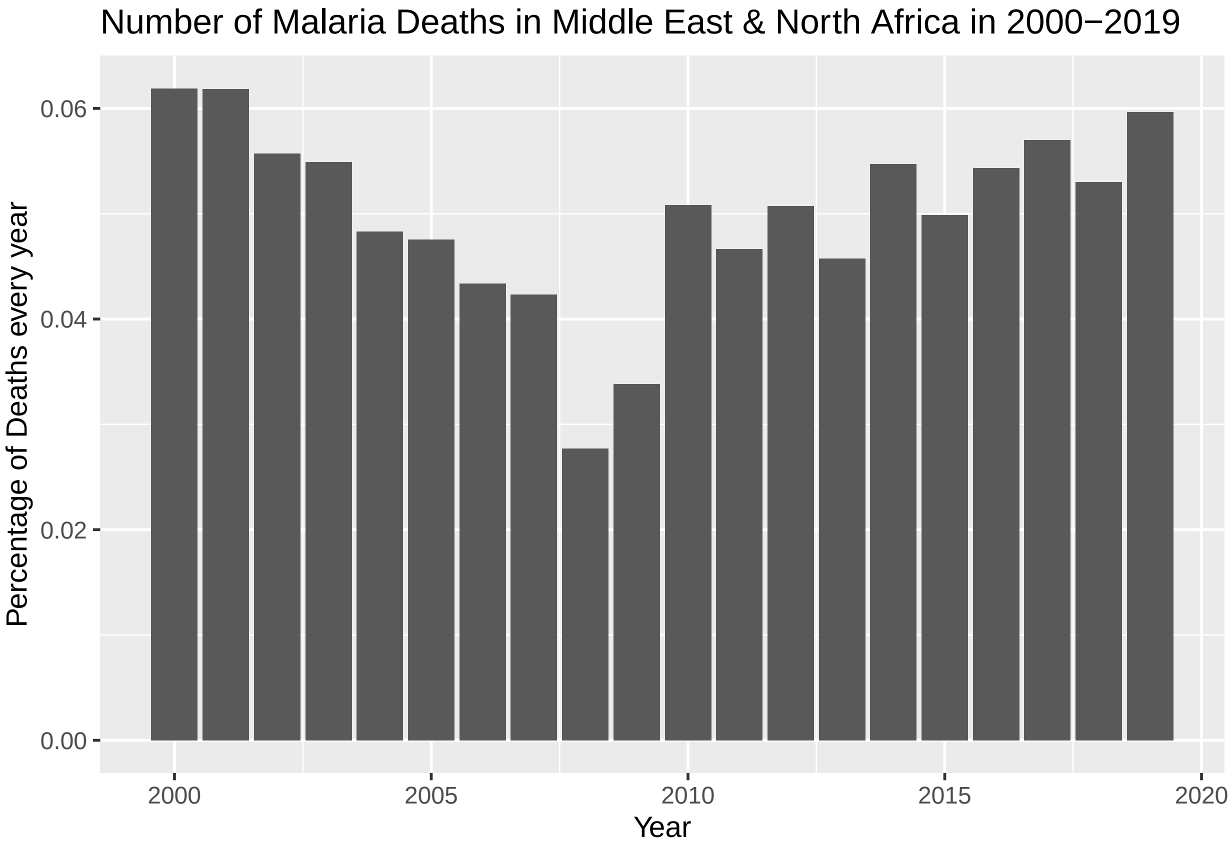
```r
# Malaria deaths (integer) in Middle East & North Africa
malaria_ME_NA <- malaria %>% filter(location == 'Middle East & North Africa') %>%
                 group_by(year) %>% summarise(val=sum(val))

malaria_ME_NA$val <- as.integer(malaria_ME_NA$val) # change to integer format
head(malaria_ME_NA, 3)
```

```
## # A tibble: 3 x 2
##    year   val
##   <int> <int>
## 1  2000  5385
## 2  2001  5378
## 3  2002  4848
```

```r
total_death <- sum(malaria_ME_NA$val) # total death over the years
total_death
```

```
## [1] 86985
```

```r
# plot the percentage of deaths each year
p5 <- ggplot(malaria_ME_NA) + geom_bar(stat = 'identity', aes(x=year, y=val/total_death))
p5 <- p5 + xlab('Year') + ylab('Percentage of Deaths every year')
p5 <- p5 + ggtitle(paste("Number of Malaria Deaths in Middle East & North Africa in ",
                min_year, "-", max_year, sep = ''))
p5
```



From the plot we can roughly observe a decreasing trend of the percentages of deaths. But wait, is there really a difference between, for example, percentages of death in 2005 and 2010, or is this kind of variation just caused by chance?

Step2: To address this question, we decided to adopt a bootstrap test.

The type of bootstrapping method we choose is *Case resampling* approach.

```
# Bootstrapping for each year
quantiles.total <- c()
for (y in malaria_ME_NA$year) {
  current_death <- malaria_ME_NA %>% filter(year == y) %>% select(val) %>% .$val
  boot.year <- c(rep(1, current_death), rep(0, total_death-current_death))
  year.boot.1000 <- replicate(100, mean(sample(boot.year, size = total_death, replace = T)))
  quantiles.total <- c(quantiles.total, quantile(year.boot.1000, c(0.025, 0.975)))
}
head(quantiles.total, 5)
```

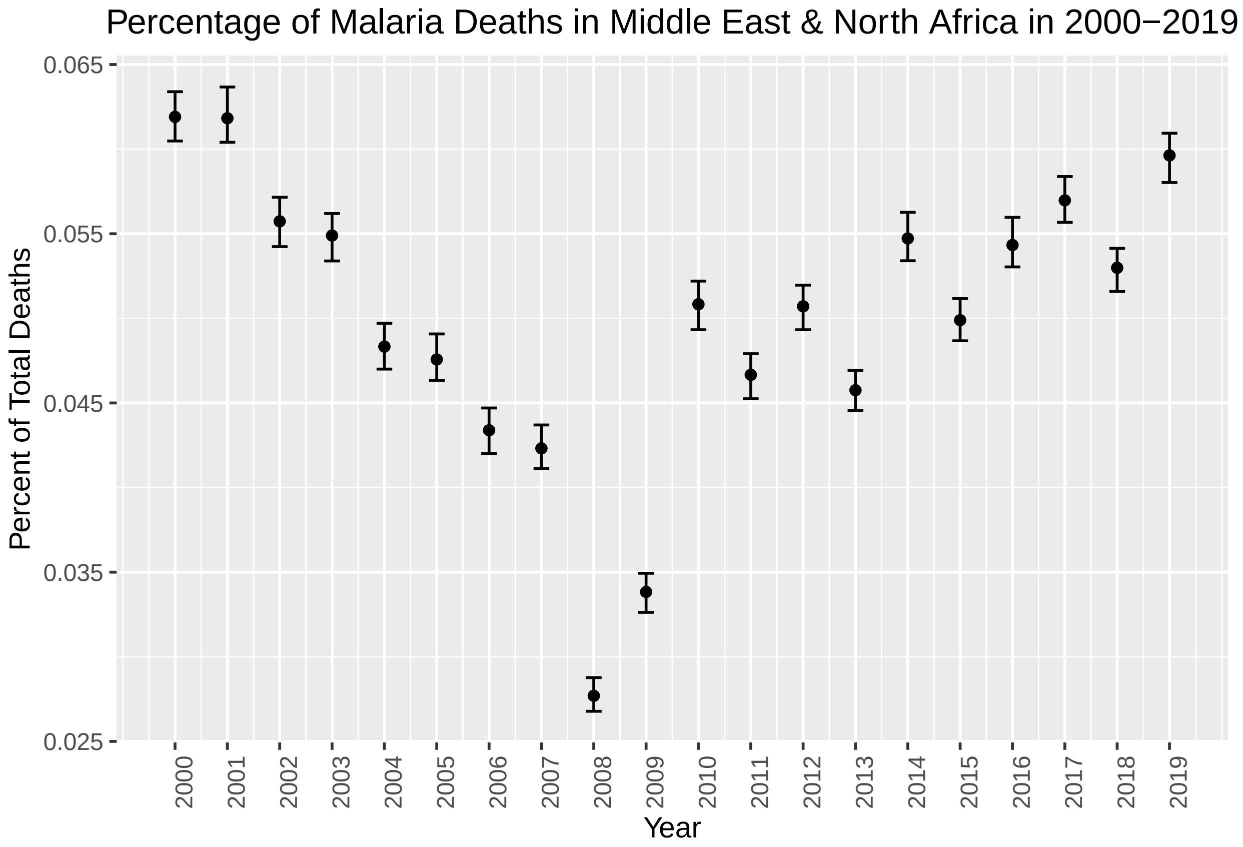
```
##       2.5%      97.5%       2.5%      97.5%       2.5% 
## 0.06048313 0.06339426 0.06040610 0.06367391 0.05423751
```

```
# Write down the bootstrapping results
matrix.quantile <- matrix(data = quantiles.total, nrow = length(quantiles.total)/2,
                          ncol = 2, byrow = T)

df.quantile <- as.data.frame(matrix.quantile)
df.quantile$year <- malaria_ME_NA$year
df.quantile$perc <- malaria_ME_NA$val/total_death
head(df.quantile)
```

```
##           V1         V2 year       perc
## 1 0.06048313 0.06339426 2000 0.06190723
## 2 0.06040610 0.06367391 2001 0.06182675
## 3 0.05423751 0.05715784 2002 0.05573375
## 4 0.05339340 0.05619676 2003 0.05489452
## 5 0.04700293 0.04971087 2004 0.04833017
## 6 0.04633730 0.04908030 2005 0.04757142
```

Step3: Visualize the 95% confidence interval data to check the differences between groups

```
# Visualization
ggplot(df.quantile) + geom_point(aes(x=year, group=year, y=perc)) +
  geom_errorbar(aes(x=year, ymin=V1, ymax=V2), width=0.3) +
  xlab('Year') +
  ylab('Percent of Total Deaths') +
  ggtitle(paste("Percentage of Malaria Deaths in Middle East & North Africa in ",
                min_year, "-", max_year, sep = '')) +
  scale_x_continuous(breaks = seq(min_year, max_year, 1)) +
  theme(plot.title = element_text(hjust=0.5),
        axis.text.x = element_text(angle=90, hjust=1))
```



Interpretion of results

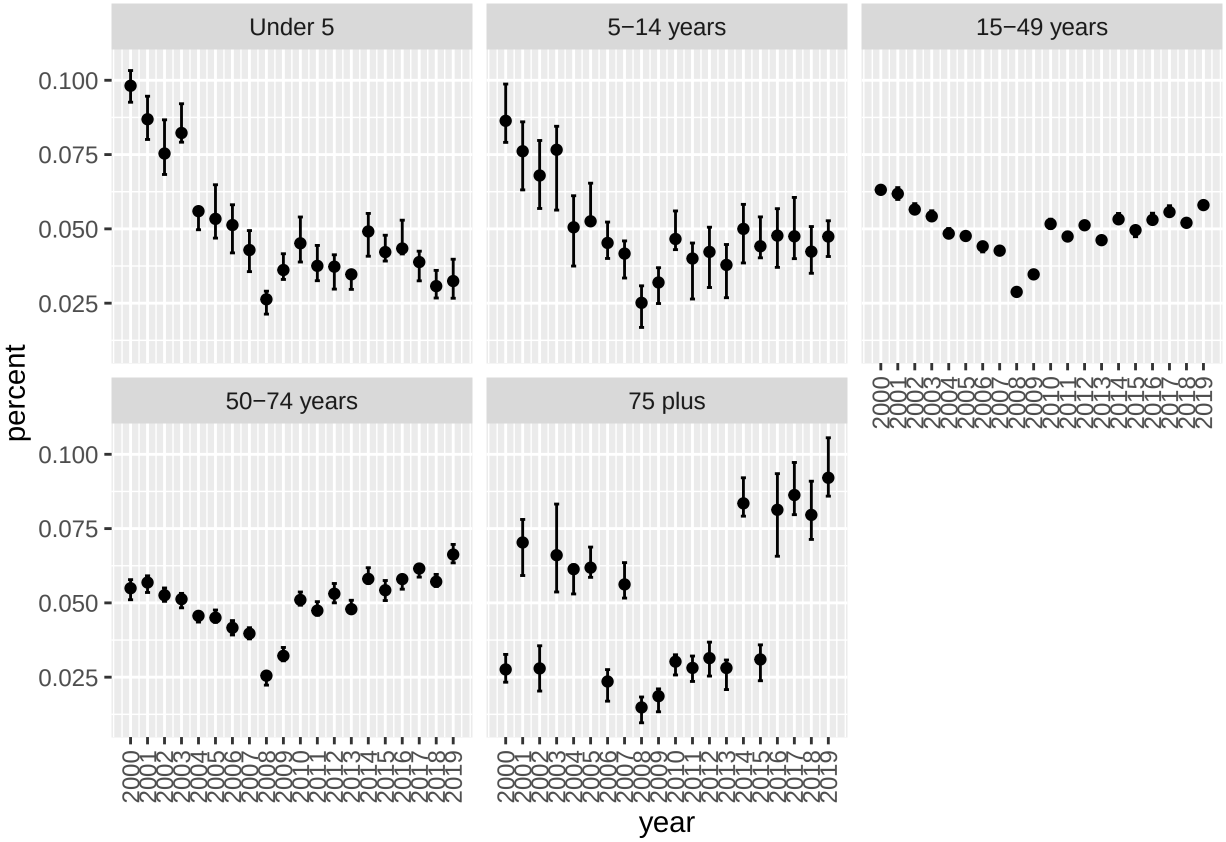
As we can see from above plot, we have sufficient evidence to reject H0. That is, the differences in percentage deaths from malaria in the Middle East & North Africa region are significant **between some two years.**

We would expect the 95% confidence intervals drawn above to be non-overlapping for groups which are significantly different from each other, such as in 2008 and 2009. Therefore, there is a significant difference for the percentage of Malaria deaths between 2008 and 2009. Additionally, there is no significant difference between 2000 and 2001 since they have overlaps during the 95% confidence intervals.

Method 2

```r
malaria_SSA <- malaria %>% filter(location == 'Middle East & North Africa')
malaria_SSA$year <- as.numeric(malaria_SSA$year)
age_list <- unique(malaria_SSA$age)
malaria_age_list <- list()
for(index in 1:length(age_list)){
  malaria_age_list[[index]] <- filter(malaria_SSA, age == age_list[index])
}
total_death_list <- malaria_SSA %>% group_by(age) %>% summarise(val=sum(val))

ages <- c()

years <- c()
percents <- c()
uppers <- c()
```

```
lowers <- c()

for(i in 1:length(malaria_age_list)){
  data <- malaria_age_list[[i]]
  ages <- c(ages, data$age)
  total_death <- total_death_list$val[i]
  data$percent <- data$val/total_death
  for(y in data$year){
    years <- c(years, y)
    percents <- c(percents, data[data$year == y, 'percent'])
    current_death <- data %>% filter(year == y) %>% .$val
    boot_pool <- c(rep(1, current_death), rep(0, total_death-current_death))
    boot_results <- c()
    for(i in 1:10){
      boot_sample <- sample(boot_pool, length(boot_pool), T)
      boot_results <- c(boot_results, mean(boot_sample))
    }
    uppers <- c(uppers, quantile(boot_results, 0.975))
    lowers <- c(lowers, quantile(boot_results, 0.025))
  }
}

overall_result <- data.frame('age'=age_list[ages],
                             'year'=years,
                             'percent'=percents,
                             'upper'=uppers,
                             'lower'=lowers)

ggplot(overall_result) +
  geom_point(aes(x=year, y=percent)) +
  geom_errorbar(aes(x=year, ymin=lowers, ymax=uppers), width=0.3) +
  facet_wrap(~age) +
  scale_x_continuous(breaks = seq(min_year, max_year, 1)) +
  theme(axis.text.x = element_text(angle=90, hjust=1, vjust=0.5))
```


Under 5
5−14 years
15−49 years
50−74 years
75 plus
percent
year
0.100
0.075
0.050
0.025
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019