Project 3

2024-05-15

```r
rm(list = ls())

# directory
setwd("C:/Users/rahul/OneDrive/Desktop/Notes/SS-24/ICS/Project_3")

# Load required libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(epitools)


# Load the dataset
dataset <- read.csv("Aspirin.csv", sep=";", header=TRUE)[,-1]

names(dataset)[3] <- "Smoking"

## Convert it to as factor
dataset$Group <- as.factor(dataset$Group)
dataset$HeartAttack <- as.factor(dataset$HeartAttack)


dataset$Smoking[dataset$Smoking == "Yes"] <- "Smoker"
dataset$Smoking[dataset$Smoking == "No"] <- "Non-Smoker"

dataset$Smoking <- as.factor(dataset$Smoking)




str(dataset)
```

```
## 'data.frame':    20021 obs. of  4 variables:
##  $ Group      : Factor w/ 2 levels "Aspirin","Placebo": 1 1 2 2 1 1 1 2 2 2 ...
##  $ HeartAttack: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Smoking    : Factor w/ 2 levels "Non-Smoker","Smoker": 1 1 1 1 1 1 2 1 2 2 ...
```

```
##  $ Age        : int  63 64 61 62 61 62 62 63 61 63 ...
```

```
## Since only 5 unique values
unique(dataset$Age)
```

```
## [1] 63 64 61 62 65
```

```
# hence can be converted in factor too
dataset$Age <- as.factor(dataset$Age)

# Summary statistics
summary(dataset)
```

```
##      Group       HeartAttack       Smoking      Age
##  Aspirin: 9987   No :19230   Non-Smoker:15992   61:3953
##  Placebo:10034   Yes:  791   Smoker    : 4029   62:3991
##                                                 63:3989
##                                                 64:4107
##                                                 65:3981
```

```
dim(dataset)
```

```
## [1] 20021     4
```

```
# Display the data in a table format
head(dataset)
```

```
##     Group HeartAttack    Smoking Age
## 1 Aspirin          No Non-Smoker  63
## 2 Aspirin          No Non-Smoker  64
## 3 Placebo          No Non-Smoker  61
## 4 Placebo          No Non-Smoker  62
## 5 Aspirin          No Non-Smoker  61
## 6 Aspirin          No Non-Smoker  62
```

Group causing Heart attack

```
##-------------- Contingency Table --------------------------

# Freq. dist.
( table.group <- table(dataset$Group, dataset$HeartAttack) )
```

```
##
##             No  Yes
##   Aspirin 9755  232
##   Placebo 9475  559
```

```
# Freq. dist. + Sum
( table.group.sum <- addmargins(table.group) )
```

```
##
##              No   Yes   Sum
##   Aspirin  9755   232  9987
##   Placebo  9475   559 10034
##   Sum     19230   791 20021
```

```
# rel. freq.
( table.rel.group <- prop.table(table.group) )
```

```
##
##                    No        Yes
##   Aspirin 0.48723840 0.01158783
##   Placebo 0.47325308 0.02792068
```

```
# rel. freq. + SUm
( table.rel.group.sum <- addmargins(table.rel.group) )
```

```
##
##                    No        Yes        Sum
##   Aspirin 0.48723840 0.01158783 0.49882623
##   Placebo 0.47325308 0.02792068 0.50117377
##   Sum     0.96049148 0.03950852 1.00000000
```

```
## Proportion of heart attack in Aspirin (the rate)
(232/9987)*100
```

```
## [1] 2.32302
```

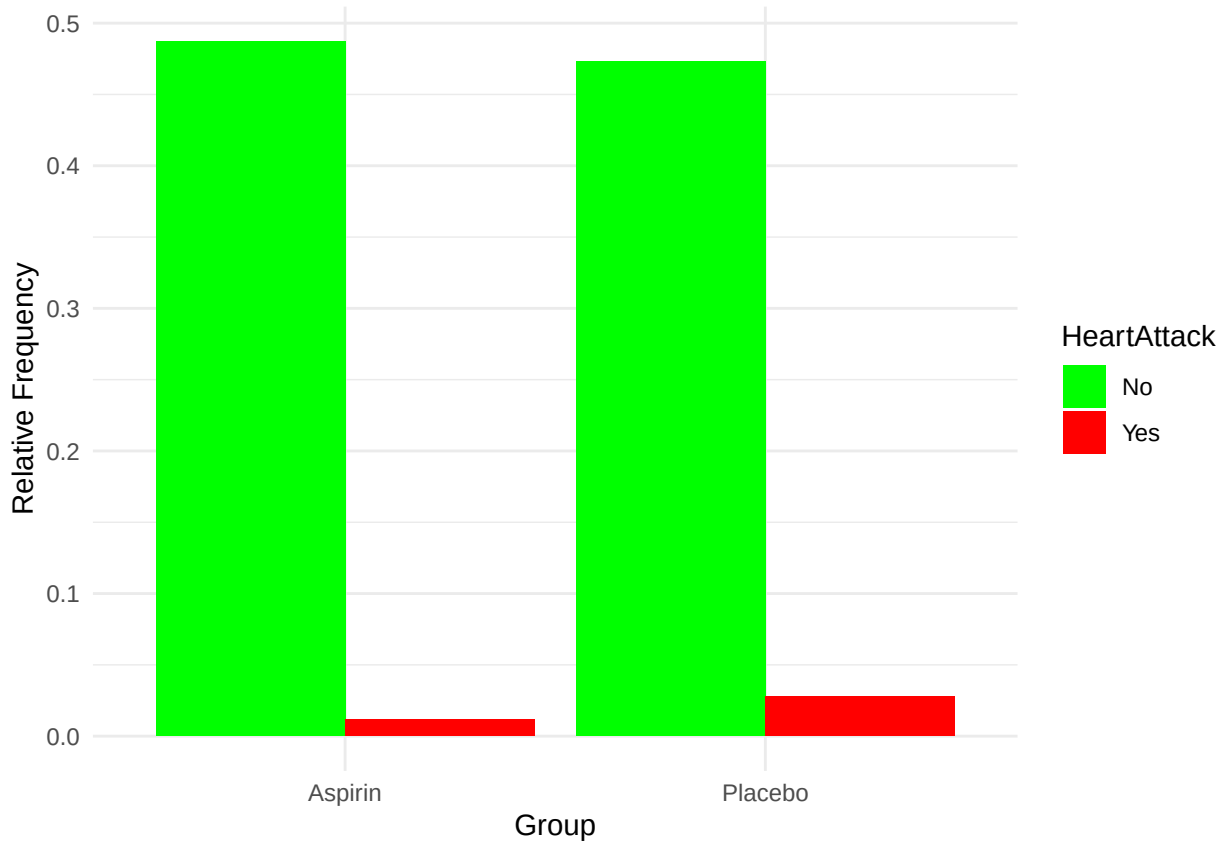
```
(559/10034)*100
```

```
## [1] 5.571058
```

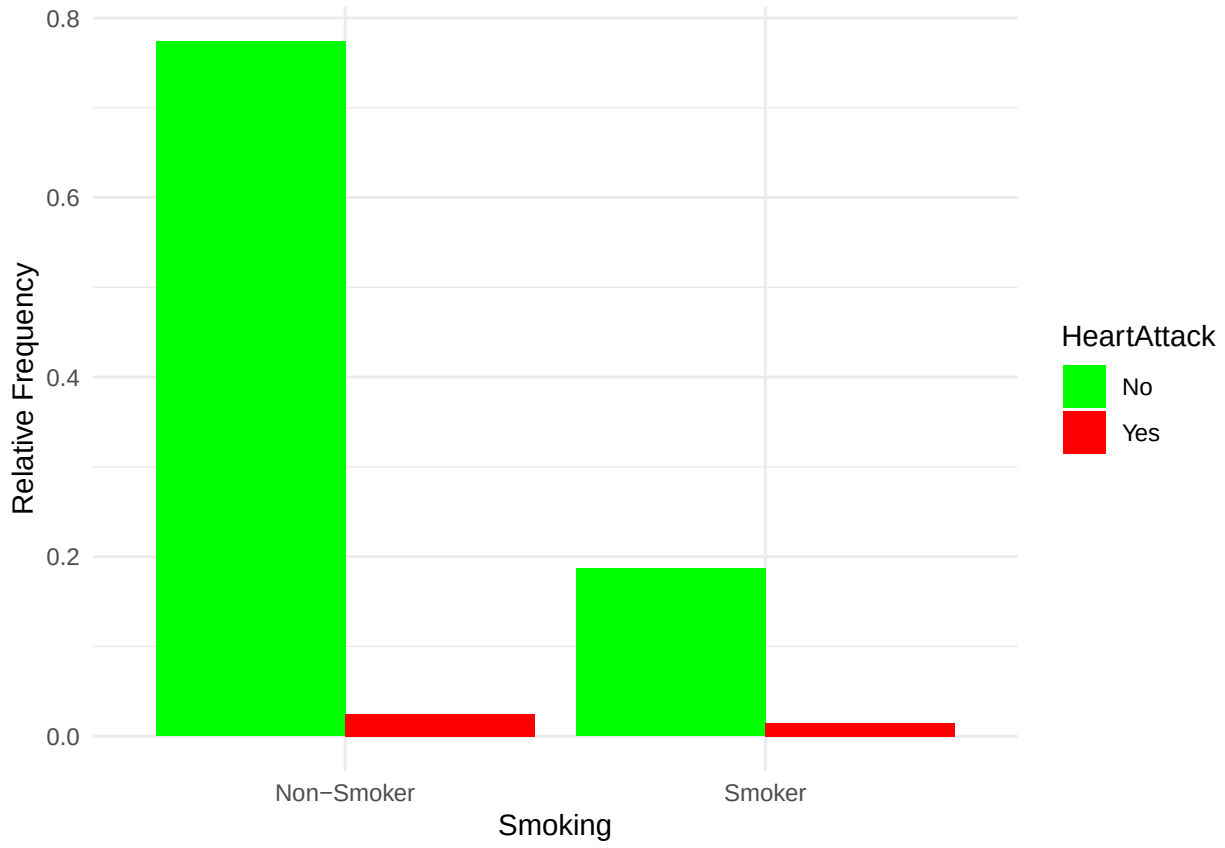
```
##-------------- Joint Bar Plots ------------------------

# Group barplot
ggplot(data = dataset, aes(x = Group, fill = HeartAttack)) +
  geom_bar(aes(y = (..count..)/sum(..count..)), position = "dodge") +
  labs(y = "Relative Frequency") +
  scale_fill_manual(values = c("green", "red")) +
  theme_minimal()
```

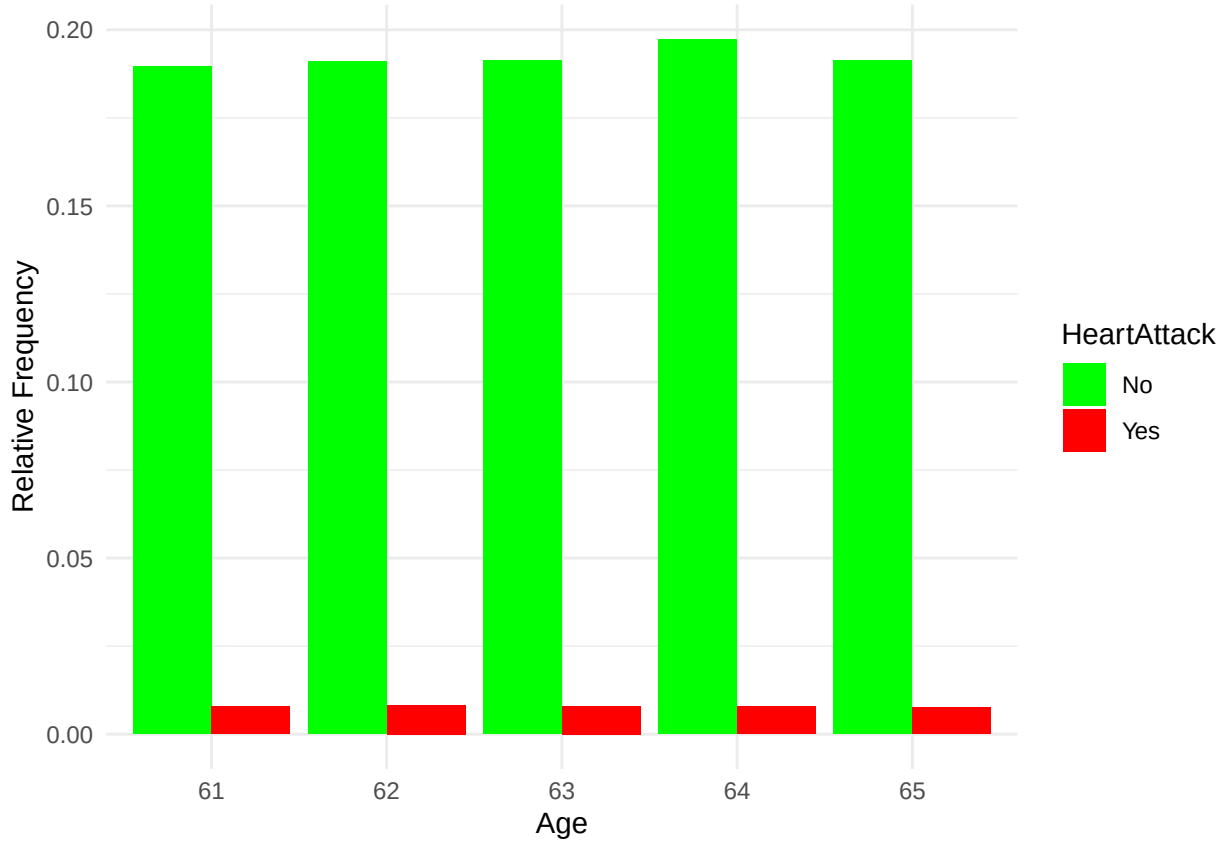
```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
# Smoking barplot
ggplot(data = dataset, aes(x = Smoking, fill = HeartAttack)) +
  geom_bar(aes(y = (..count..)/sum(..count..)), position = "dodge") +
  labs(y = "Relative Frequency") +
  scale_fill_manual(values = c("green", "red")) +
  theme_minimal()
```




```
# Age barplot
ggplot(data = dataset, aes(x = Age, fill = HeartAttack)) +
  geom_bar(aes(y = (..count..)/sum(..count..)), position = "dodge") +
  labs(y = "Relative Frequency") +
  scale_fill_manual(values = c("green", "red")) +
  theme_minimal()
```




```
## Risk Ratio for Group ========================================
RR_group <- epitools::riskratio(table.group, rev = "rows",
                                correction = FALSE, method = "wald",
                                conf.level = 0.95)$measure
RR_group
```

```
##          risk ratio with 95% C.I.
##           estimate    lower     upper
##   Placebo  1.00000       NA        NA
##   Aspirin  0.41698 0.358704 0.4847237
```

```
## Odds Ratio
OR_group <- epitools::oddsratio(table.group, rev = "rows",
                                correction = FALSE, method = "wald",
                                conf.level = 0.95)$measure
OR_group
```

```
##          odds ratio with 95% C.I.
##            estimate     lower     upper
##   Placebo 1.0000000        NA        NA
##   Aspirin 0.4031142 0.3450062 0.4710091
```

```
## Chi square test for independence
chisq.test(table.group, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
```

```
## data:  table.group
## X-squared = 139.15, df = 1, p-value < 2.2e-16
```

Smoking and Heart attack

```
# Freq. dist. Contingency Table
( table.smoking <- table(dataset$Smoking, dataset$HeartAttack) )
```

```
##
##                  No   Yes
##   Non-Smoker 15495   497
##   Smoker      3735   294
```

```
# Freq. dist. Contingency Table Sum
( table.smoking.sum <- addmargins(table.smoking) )
```

```
##
##                  No   Yes   Sum
##   Non-Smoker 15495   497 15992
##   Smoker      3735   294  4029
##   Sum        19230   791 20021
```

```
# rel. freq. Contingency Table
( table.rel.smoking <- prop.table(table.smoking) )
```

```
##
##                       No        Yes
##   Non-Smoker 0.77393737 0.02482393
##   Smoker     0.18655412 0.01468458
```

```
# rel. freq. Contingency Table SUm
( table.rel.smoking.sum <- addmargins(table.rel.smoking) )
```

```
##
##                       No        Yes        Sum
##   Non-Smoker 0.77393737 0.02482393 0.79876130
##   Smoker     0.18655412 0.01468458 0.20123870
##   Sum        0.96049148 0.03950852 1.00000000
```

```
## Proportion of heart attack in (the rate)
(497/15992)*100 #Non-Smoker
```

```
## [1] 3.107804
```

```
(294/3732)*100 #Smoker
```

```
## [1] 7.877814
```

```
## Risk ratio for Smoker ======================================
RR_smoking <- epitools::riskratio(table.smoking, rev = "rows",
                                  correction = FALSE, method = "wald",
                                  conf.level = 0.95)$measure
RR_smoking
```

```
##             risk ratio with 95% C.I.
##              estimate     lower     upper
##   Smoker     1.000000        NA        NA
##   Non-Smoker 0.425896 0.3702538 0.4899002
```

```r
## Odds Ratio
OR_smoking <- epitools::oddsratio(table.smoking, rev = "rows",
                                  correction = FALSE, method = "wald",
                                  conf.level = 0.95)$measure
OR_smoking
```

```
##             odds ratio with 95% C.I.
##               estimate    lower     upper
##   Smoker     1.0000000       NA        NA
##   Non-Smoker 0.4074817 0.351226 0.4727478
```

```r
## Chi square test for independence
chisq.test(table.smoking, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table.smoking
## X-squared = 148.84, df = 1, p-value < 2.2e-16
```