Week_4_Lab_assignment_own_dataset

Upol Chowdhury

2024-09-20

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(readr)
```

```
data <- read_csv("G:/datasets_for_biostat/Data_Set.csv")
```

```
## New names:
## Rows: 373 Columns: 31
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (29): Timestamp, Data source location, Gender, Condition of Cough, sore ... dbl
## (1): Age lgl (1): ...31
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...31`
```

```
spec(data)
```

```
## cols(
##   Timestamp = col_character(),
##   `Data source location` = col_character(),
##   Age = col_double(),
##   Gender = col_character(),
##   `Condition of Cough` = col_character(),
##   `sore throat` = col_character(),
##   Wheezing = col_character(),
##   Chills = col_character(),
##   `Abdominal pain` = col_character(),
##   Vomiting = col_character(),
##   Fever = col_character(),
##   Headache = col_character(),
##   Nausea = col_character(),
##   Tiredness = col_character(),
##   Malaise = col_character(),
```

```
##   `Body ache` = col_character(),
##   Anorexia = col_character(),
##   `Shortness of breath` = col_character(),
##   Convulsion = col_character(),
##   `Weight loss` = col_character(),
##   `Face condition` = col_character(),
##   `Fauces condition` = col_character(),
##   `Chest pain` = col_character(),
##   Shivering = col_character(),
##   Sweating = col_character(),
##   `Shoulder pain` = col_character(),
##   `Herpes labialis found` = col_character(),
##   `Periodic asthmatic suffering` = col_character(),
##   `Nocturnal episode of dyspnea` = col_character(),
##   `Disease type` = col_character(),
##   ...31 = col_logical()
## )
```

Conditional Probability Lab

```
n <- nrow(data)
n
```

```
## [1] 373
```

```
names(data)
```

```
##  [1] "Timestamp"                    "Data source location"
##  [3] "Age"                          "Gender"
##  [5] "Condition of Cough"           "sore throat"
##  [7] "Wheezing"                     "Chills"
##  [9] "Abdominal pain"               "Vomiting"
## [11] "Fever"                        "Headache"
## [13] "Nausea"                       "Tiredness"
## [15] "Malaise"                      "Body ache"
## [17] "Anorexia"                     "Shortness of breath"
## [19] "Convulsion"                   "Weight loss"
## [21] "Face condition"               "Fauces condition"
## [23] "Chest pain"                   "Shivering"
## [25] "Sweating"                     "Shoulder pain"
## [27] "Herpes labialis found"        "Periodic asthmatic suffering"
## [29] "Nocturnal episode of dyspnea" "Disease type"
## [31] "...31"
```

```
table(data$Gender)
```

```
##
## Female   Male
##    184    189
```

```r
#proportion
prop.table(table(data$Gender))
```

```
##
##    Female      Male
## 0.4932976 0.5067024
```

```r
names(data)[names(data) == "Disease type"] <- "disease_type"


names(data)
```

```
##  [1] "Timestamp"                    "Data source location"
##  [3] "Age"                          "Gender"
##  [5] "Condition of Cough"           "sore throat"
##  [7] "Wheezing"                     "Chills"
##  [9] "Abdominal pain"               "Vomiting"
## [11] "Fever"                        "Headache"
## [13] "Nausea"                       "Tiredness"
## [15] "Malaise"                      "Body ache"
## [17] "Anorexia"                     "Shortness of breath"
## [19] "Convulsion"                   "Weight loss"
## [21] "Face condition"               "Fauces condition"
## [23] "Chest pain"                   "Shivering"
## [25] "Sweating"                     "Shoulder pain"
## [27] "Herpes labialis found"        "Periodic asthmatic suffering"
## [29] "Nocturnal episode of dyspnea" "disease_type"
## [31] "...31"
```

Summary of the disease types

```r
head(data$disease_type)
```

```
## [1] "COPD"             "Bronchial asthma" "COPD"             "Bronchial asthma"
## [5] "COPD"             "Bronchial asthma"
```

```r
table(data$disease_type)
```

```
##
## Bronchial asthma             COPD             none        Pneumonia
##               81              102               91               99
```

```r
data$disease_category <- ifelse(data$disease_type %in% c("COPD", "Bronchial asthma", "Pneumonia"),
                                "Respiratory", "Other")


table(data$disease_category)
```

```
##
##       Other Respiratory
##          91         282
```

Create a contingency table

```
cont_table <- table(data$Gender, data$disease_category)
print(cont_table)
```

```
##
##          Other Respiratory
##   Female    47         137
##   Male      44         145
```

Calculate marginal probabilities

```
total <- sum(cont_table)
total
```

```
## [1] 373
```

```
p_male <- sum(cont_table["Male",]) / total
p_male
```

```
## [1] 0.5067024
```

```
p_female <- sum(cont_table["Female",]) / total
p_female
```

```
## [1] 0.4932976
```

```
p_respiratory <- sum(cont_table[,"Respiratory"]) / total
p_respiratory
```

```
## [1] 0.7560322
```

```
p_other <- sum(cont_table[,"Other"]) / total
p_other
```

```
## [1] 0.2439678
```

Calculate Conditional Probabilities

```
p_respiratory_male <- cont_table["Male","Respiratory"] / sum(cont_table["Male",])
p_respiratory_male
```

```
## [1] 0.7671958
```

```
p_respiratory_female <- cont_table["Female","Respiratory"] / sum(cont_table["Female",])
p_respiratory_female
```

```
## [1] 0.7445652
```

Calculate prevalence, sensitivity, and specificity

I have created a binary outcome: Respiratory disease (1) vs. No respiratory disease (0) (used help from google for the syntax)

```
data$respiratory_disease <- ifelse(data$disease_type %in% c("COPD", "Bronchial asthma", "Pneumonia"), 1
```

I have created a binary predictor: Wheezing present (1) vs. Wheezing absent (0) (used help from google for the syntax)

```
data$wheezing_present <- ifelse(data$Wheezing %in% c("Low", "Moderate", "High"), 1, 0)
```

Create a contingency table

```
cont_table <- table(data$wheezing_present, data$respiratory_disease)
print("Contingency Table:")
```

```
## [1] "Contingency Table:"
```

```
print(cont_table)
```

```
##
##       0   1
##   0  85  92
##   1   6 190
```

It provides a direct view of the relationship between wheezing and respiratory disease in my dataset.

Calculate prevalence

```
prevalence <- sum(data$respiratory_disease) / nrow(data)
prevalence
```

```
## [1] 0.7560322
```

Calculate sensitivity (proportion of those with wheezing among those with respiratory disease)

```
sensitivity <- cont_table["1", "1"] / sum(cont_table[, "1"])
sensitivity
```

```
## [1] 0.6737589
```

Calculate specificity (proportion of those without wheezing among those without respiratory disease)

```
specificity <- cont_table["0", "0"] / sum(cont_table[, "0"])
specificity
```

```
## [1] 0.9340659
```

Simulation approach

```
# Set population size to number of rows in the dataset
population.size <- nrow(data)

# Create empty vectors to store results
disease.status <- vector("numeric", population.size)
test.result <- vector("numeric", population.size)



# Assign disease status
disease.status <- sample(c(0,1), size = population.size,
                         prob = c(1 - prevalence, prevalence),
                         replace = TRUE)

# Assign test result
for(k in 1:population.size){
  if(disease.status[k] == 0){
    test.result[k] = sample(c(0,1), size = 1,
                            prob = c(specificity, 1 - specificity))
  }
  if(disease.status[k] == 1){
    test.result[k] = sample(c(0,1), size = 1,
                            prob = c(1 - sensitivity, sensitivity))
  }
}
#create matrix of disease status and test result (part c)
disease.status.and.test.result = cbind(disease.status, test.result)
#create a table of test result by disease status
addmargins(table(test.result, disease.status))
```

```
##             disease.status
## test.result   0   1 Sum
##          0   78  88 166
##          1    5 202 207
##          Sum  83 290 373
```

Calculate PPV

```
ppv = sum(test.result[disease.status == 1])/sum(test.result)
ppv
```

```
## [1] 0.9758454
```

Calculate NPV

```
npv = sum(test.result == 0 & disease.status == 0) / sum(test.result == 0)
npv
```

```
## [1] 0.4698795
```