APS Coding Problem Set 1 (Bats Problem)

2023-08-22

Load the CSV data file 'bats.csv' into the dataframe 'dfBat' and display the first few rows

```
dfBat= read.csv('Data/bats.csv')
head(dfBat)
```

```
##   Bat Gene.1 Gene.2 Gene.3 Gene.4 Gene.5 Ebola
## 1   1  FALSE  FALSE   TRUE   TRUE   TRUE  TRUE
## 2   2   TRUE  FALSE   TRUE  FALSE  FALSE FALSE
## 3   3   TRUE  FALSE   TRUE   TRUE   TRUE FALSE
## 4   4  FALSE   TRUE   TRUE   TRUE   TRUE  TRUE
## 5   5  FALSE  FALSE  FALSE   TRUE  FALSE FALSE
## 6   6   TRUE  FALSE  FALSE   TRUE  FALSE FALSE
```

Display the structure of the dataframe 'dfBat' and its column names

```
str(dfBat)
```

```
## 'data.frame':    99999 obs. of  7 variables:
##  $ Bat   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Gene.1: logi  FALSE TRUE TRUE FALSE FALSE TRUE ...
##  $ Gene.2: logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ Gene.3: logi  TRUE TRUE TRUE TRUE FALSE FALSE ...
##  $ Gene.4: logi  TRUE FALSE TRUE TRUE TRUE TRUE ...
##  $ Gene.5: logi  TRUE FALSE TRUE TRUE FALSE FALSE ...
##  $ Ebola : logi  TRUE FALSE FALSE TRUE FALSE FALSE ...
```

```
colnames(dfBat)
```

```
## [1] "Bat"    "Gene.1" "Gene.2" "Gene.3" "Gene.4" "Gene.5" "Ebola"
```

Create a vector 'gene_cols' containing the names of selected gene columns

```
gene_cols = c("Gene.1", "Gene.2", "Gene.3", "Gene.4", "Gene.5", "Ebola")
```

What is the chance of a random bat carrying the Ebola virus?

```
overall_ebola_chance = mean(dfBat$Ebola == TRUE)
ebola_chance = overall_ebola_chance*100
ebola_chance
```

```
## [1] 30.0793
```

For each gene, calculate the likelihood that it is expressed in a random bat.

```r
gene_likelihoods = sapply(gene_cols[1:5], function(col_name) {
  mean(dfBat[[col_name]]) * 100
})
gene_likelihoods
```

```
##   Gene.1   Gene.2   Gene.3   Gene.4   Gene.5
## 70.22770 30.07630 50.08950 80.16180 32.70533
```

Is the presence or absence of any of the genes indicative of a random bat potentially carrying the Ebola virus?

```r
significant_genes = character(0) # Initialize an empty vector to store gene names

for (gene_col in gene_cols[1:5]) {
  gene_present_ebola_chance = mean(dfBat$Ebola[dfBat[[gene_col]]] == TRUE)
  gene_absent_ebola_chance = mean(dfBat$Ebola[!dfBat[[gene_col]]] == TRUE)

  if (abs(gene_present_ebola_chance - overall_ebola_chance) > 0.1 ||
      abs(gene_absent_ebola_chance - overall_ebola_chance) > 0.1) {
    significant_genes = c(significant_genes, gene_col)
  }
}

significant_genes
```

```
## [1] "Gene.3" "Gene.4" "Gene.5"
```