Lab-4 assignment using Arrhythmia genetic associaiton

Kasra Vand

2024-09-19

Introduction

This document analyzes the arrhythmia dataset to demonstrate conditional probability and positive predictive value (PPV) calculations. We'll use two categorical variables: var_class (SNP vs. Other) and clinical_significance (Pathogenic vs. Non-pathogenic).

Data Preparation and Analysis

First, we'll read the data, categorize our variables, and calculate initial probabilities.

```
# Read the data
arrhythmia <- read.csv("arrhythmia.csv")

# Categorize clinical_significance
arrhythmia$clinical_bin <- ifelse(
  arrhythmia$clinical_significance %in% c("pathogenic", "likely pathogenic"),
                                    "Pathogenic", "Non-pathogenic")

# Categorize var_class
arrhythmia$var_class_bin <- ifelse(arrhythmia$var_class == "SNP", "SNP", "Other")

# Create a contingency table
cont_table <- table(arrhythmia$var_class_bin, arrhythmia$clinical_bin)
print(cont_table)
```

```
##
##         Non-pathogenic Pathogenic
##   Other             27         79
##   SNP              359         55
```

```
# Calculate probabilities
total <- sum(cont_table)
p_snp <- sum(cont_table["SNP",]) / total
p_other <- sum(cont_table["Other",]) / total
p_pathogenic <- sum(cont_table[,"Pathogenic"]) / total
p_non_pathogenic <- sum(cont_table[,"Non-pathogenic"]) / total

# Print probabilities
cat("Probability of SNP:", p_snp, "\n")
```

```
## Probability of SNP: 0.7961538
```

```
cat("Probability of Other:", p_other, "\n")
```

```
## Probability of Other: 0.2038462
```

```r
cat("Probability of Pathogenic:", p_pathogenic, "\n")
```

```
## Probability of Pathogenic: 0.2576923
```

```r
cat("Probability of Non-pathogenic:", p_non_pathogenic, "\n")
```

```
## Probability of Non-pathogenic: 0.7423077
```

Data Simulation

Now, we'll use the probabilities calculated from our original data to simulate a larger dataset. This allows us to demonstrate the conditional probability concepts.

```r
# Parameters
population_size <- 10000
p_snp <- 0.8870056

# Simulate var_class
var_class_sim <- sample(c("SNP", "Other"), size = population_size,
                        prob = c(p_snp, 1 - p_snp), replace = TRUE)

# Simulate clinical_significance
clinical_sim <- vector("character", population_size)

# Note; here the probabilities are set based on the previous data analysis
for(k in 1:population_size) {
  if(var_class_sim[k] == "SNP") {
    clinical_sim[k] <- sample(c("Pathogenic", "Non-pathogenic"), size = 1,
                              prob = c(0.1428571, 0.8571429))
  } else {
    clinical_sim[k] <- sample(c("Pathogenic", "Non-pathogenic"), size = 1,
                              prob = c(0.7058824, 0.2941176))
  }
}

# Create simulated data frame
sim_data <- data.frame(var_class = var_class_sim, clinical_significance = clinical_sim)

# View results
table(sim_data$var_class, sim_data$clinical_significance)
```

```
##
##         Non-pathogenic Pathogenic
##   Other            338        797
##   SNP             7613       1252
```

Probability Calculations

Finally, we'll calculate prevalence, sensitivity, specificity, PPV, and NPV using our simulated data.

```r
# Calculate prevalence, sensitivity, and specificity
prevalence <- sum(sim_data$clinical_significance == "Pathogenic") / nrow(sim_data)
sensitivity <- sum(sim_data$var_class == "SNP" & sim_data$clinical_significance == "Pathogenic") /
               sum(sim_data$clinical_significance == "Pathogenic")
specificity <- sum(sim_data$var_class == "Other" & sim_data$clinical_significance == "Non-pathogenic") /
               sum(sim_data$clinical_significance == "Non-pathogenic")
```

```r
# Calculate PPV and NPV
ppv <- (sensitivity * prevalence) /
       (sensitivity * prevalence + (1 - specificity) * (1 - prevalence))
npv <- (specificity * (1 - prevalence)) /
       ((1 - sensitivity) * prevalence + specificity * (1 - prevalence))

# Print results
cat("Prevalence:", prevalence, "\n")
```

```
## Prevalence: 0.2049
```

```r
cat("Sensitivity:", sensitivity, "\n")
```

```
## Sensitivity: 0.6110298
```

```r
cat("Specificity:", specificity, "\n")
```

```
## Specificity: 0.04251038
```

```r
cat("PPV:", ppv, "\n")
```

```
## PPV: 0.1412296
```

```r
cat("NPV:", npv, "\n")
```

```
## NPV: 0.2977974
```

Conclusion

This analysis demonstrates the use of conditional probability and PPV calculations on the arrhythmia dataset. We categorized var_class and clinical_significance, simulated a larger dataset based on the original probabilities, and calculated key metrics including prevalence, sensitivity, specificity, PPV, and NPV.

The results provide insights into the relationship between genetic variations (SNP vs. Other) and their clinical significance in the context of arrhythmia. I actually think that these metrics can be valuable for understanding the predictive power of genetic markers for pathogenic conditions related to arrhythmia.