Analysis

Aazar and Nadia

2024-05-01

```
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
```

```
data = read.csv("diabetes_prediction_dataset.csv")
data <- data %>%
  mutate(weight_status = case_when(
    bmi < 18.5 ~ "Underweight",
    bmi >= 18.5 & bmi < 25 ~ "Healthy Weight",
    bmi >= 25 & bmi < 30 ~ "Overweight",
    bmi >= 30  ~ "Obesity"
  ))
head(data)
```

```
##   gender age hypertension heart_disease smoking_history   bmi HbA1c_level
## 1 Female  80            0             1           never 25.19         6.6
## 2 Female  54            0             0         No Info 27.32         6.6
## 3   Male  28            0             0           never 27.32         5.7
## 4 Female  36            0             0         current 23.45         5.0
## 5   Male  76            1             1         current 20.14         4.8
## 6 Female  20            0             0           never 27.32         6.6
##   blood_glucose_level diabetes  weight_status
## 1                 140        0     Overweight
## 2                  80        0     Overweight
## 3                 158        0     Overweight
## 4                 155        0 Healthy Weight
## 5                 155        0 Healthy Weight
## 6                  85        0     Overweight
```

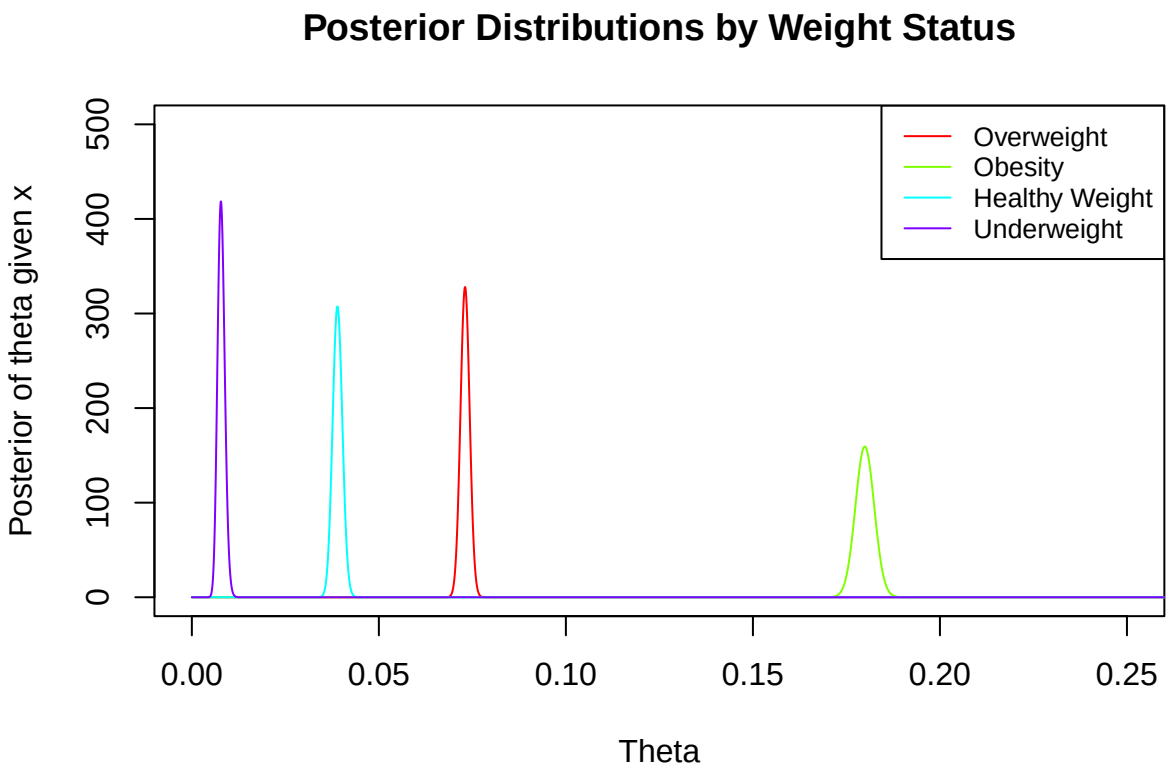
We use a Beta-Binomial Model

- The choice of alpha and beta is 3 because our prior belief is that theta is 0.5 for all the groups. A randomly chosen person, regardless of any group has a 50 percent chance of having diabetes.

```
categories = c('Overweight','Obesity', 'Healthy Weight', 'Underweight')

# Create an empty plot
plot(NULL, NULL, xlim=c(0, 0.25), ylim=c(0, 500), xlab="Theta", ylab="Posterior of theta given x", main=

for (category in categories){
  diabetesunder <- data %>%
    filter(weight_status == category)

  data_summary <- table(diabetesunder$weight_status, diabetesunder$diabetes)
  n = sum(data_summary[category, ]) # Total patients underweight
  x = data_summary[category, "1"]

  set.seed(202289)

  alpha_prior = 3; beta_prior = 3;
  theta =seq(0,1,0.00001)

  prior = dbeta(theta, alpha_prior, beta_prior)

  # likelihood
  likelihood = dbinom(x, n, theta)

  # posterior
  alpha_post = alpha_prior + x
  beta_post = beta_prior + n - x

  # Plot posterior distribution for each category
  lines(theta, dbeta(theta, shape1=alpha_post, shape2=beta_post), col=rainbow(length(categories))[which

  cat( category,": ", qbeta(c(0.025, 0.975), alpha_post, beta_post) , "\n")

}
```

```
## Overweight :  0.07069319 0.07546188
## Obesity :  0.1750528 0.1848662
## Healthy Weight :  0.03646063 0.04154822
## Underweight :  0.006114506 0.009869316
```

```
# Add legend
legend("topright", legend=categories, col=rainbow(length(categories)), lty=1, cex=0.8)
```



** Now that we have generated the posterior distributions, it is time to look at the data and see how close the evidence is to the posterior distributions.

```
data %>%
  group_by(weight_status) %>%
  summarize(
    sum_diabetes = sum(diabetes, na.rm = TRUE),
    percent_diabetic = round((sum_diabetes / n()) * 100,2)
  ) %>%
  arrange(desc(sum_diabetes))
```

```
## # A tibble: 4 x 3
##   weight_status  sum_diabetes percent_diabetic
##   <chr>                 <int>            <dbl>
## 1 Obesity                4233            18.0
## 2 Overweight             3340             7.3
## 3 Healthy Weight          863             3.88
## 4 Underweight              64             0.75
```

- Since we have a good amount of data, the data plays an overwhelmingly bigger role in determining the posterior distribution.