Weekly Exercise - Week 2

The following function creates data sets from a population (see also weekly exercise 1).

```
gen_data <- function(n) {
  p <- 3
  n1 <- n2 <- n/2
  cov_1 <- diag(rep(1,p)) + 0.2
  cov_2 <- cov_1
  cov_2[1,2] <- cov_2[2,1] <- cov_2[1,2] + 0.5
  x_class1 <- mvrnorm(n1, mu = rep(3,p), Sigma = cov_1)
  x_class2 <- mvrnorm(n2, mu = rep(2,p), Sigma = cov_2)
  x <- rbind(x_class1, x_class2)
  y <- rep(c(1,2), c(n1, n2))
  df <- as.data.frame(cbind(x,y))
  names(df) <- c(paste0("x", 1:p), "y")
  return(df)
}
```

Use this function to generate two training sets (size 50, and $10,000$) and a test set of size $10,000$.

1. Do you expect QDA or LDA to perform better on the small training set? What about on the large training set? Explain using bias and variance. Note that we refer to performance on the test set.
2. Train LDA and QDA and obtain their test set accuracy for both training sets. To eliminate randomness, repeat this process at least 100 times and obtain the average accuracies (across repetitions). Are the obtained numbers in line with your expectations? If not, what could explain the discrepancy?
3. OPTIONAL: What is the Bayes classifier in this case conceptually? What is the formula?
4. OPTIONAL: Implement the Bayes classifier and obtain its test set performance. How close is it to the average performance of QDA and the large training set? Is this a surprise?

Generate and upload one pdf using either `RMarkdown` or `Python Notebook` including both your code and the textual answers.