Weekly Exercise - Week 3

The following function, which is almost identical the the function from weekly exercise 1, creates data sets from a population.

```
gen_data <- function(n) {
  p <- 15
  n1 <- n2 <- n/2
  cov_1 <- diag(rep(1,p)) + 0.2
  x_class1 <- mvrnorm(n1, mu = rep(3,p), Sigma = cov_1)
  x_class2 <- mvrnorm(n2, mu = rep(2,p), Sigma = cov_1)
  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 logistic regression 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 logistic regression and obtain their test set balanced accuracy, for both training sets (Hint: you might want to reuse the function you created during the exercises). 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?

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