1.1) What is the goal of the study?

The goal is to determine which of the five preoperative variables are predictive of nodal involvement in prostate cancer patients.

1.2) What are the input variables and what is the output variable?

Input Variables

- age: Age at diagnosis.
- acid: Level of serum acid phosphatase (x100).
- x-ray: Xray reading.
- grade: Pathology reading.
- stage: Tumor stage by palpation.

Output variable:

- nodes: Nodal involvement (0 = no involvement, 1 = involvement).

1.3) Why can't we use the linear regression model to search for significant risk factors associated with nodal involvement?

Linear regression is inappropriate because the outcome is binary.

1. Download the data set from CANVAS into your computer, import the data into R, and use a logistic regression model to find the relationship between nodal involvement (nodes=1) and its risk factors.

```
Call:
glm(formula = nodes ~ age + acid + Xray + grade + stage, family = binomial,
    data = data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.06180    3.45992   0.018   0.9857
age         -0.06926    0.05788  -1.197   0.2314
acid         0.02434    0.01316   1.850   0.0643 .
Xray         2.04534    0.80718   2.534   0.0113 *
grade        0.76142    0.77077   0.988   0.3232
stage        1.56410    0.77401   2.021   0.0433 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 48.126  on 47  degrees of freedom
AIC: 60.126

Number of Fisher Scoring iterations: 5
```

2. Identify preoperative variables that are significant predictors of nodal involvement at 0.05 significance level.

We have two preoperative variables that are significant: `stage` and `Xray `.

Randomly divide the data set into a training set (containing 30 data points) and a test set (containing the rest of the data) and check the contents of the two new data sets.

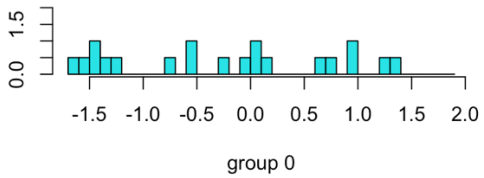
```
> n <- nrow(data)
> train_indices <- sample(1:n, 30)
> train_data <- data[train_indices, ]
> test_data <- data[-train_indices, ]
```

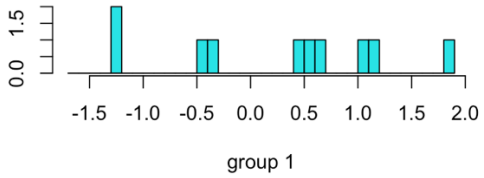
3. Use the training data to run a linear discriminant analysis on nodal involvement using the preoperative variable *age* and *acid*, then interpret the output.

The LDA output shows that, based on the training data, about 66.7% of patients are expected to have no nodal involvement while 33.3% are expected to be involved, with patients showing nodal involvement having lower mean age (57.70 vs. 60.15) and higher mean acid levels (74.4 vs. 68.5). The coefficients indicate that lower age (negative coefficient for age) and higher acid levels (positive coefficient for acid) increase the discriminant score, suggesting these factors contribute to predicting nodal involvement

```
Call:
lda(nodes ~ age + acid, data = train_data)

Prior probabilities of groups:
        0         1
0.6666667 0.3333333

Group means:
    age acid
0 60.15 68.5
1 57.70 74.4

Coefficients of linear discriminants:
             LD1
age  -0.14371713
acid  0.01815344
```

4. Plot the linear discriminant analysis result. What do you find?





The graph shows minimal overlap between two classes based on the linear discriminant

5. Apply the linear discriminant classifier to the test data set and find the test error rate.

```
> lda_pred <- predict(lda_model, newdata = test_data)
> mean(lda_pred$class != test_data$nodes)
[1] 0.3913043
```

6. Use the training data to construct a quadratic discriminant classifier on nodal involvement using the preoperative variable *age* and *acid*, then interpret the results.

```
Call:
qda(nodes ~ age + acid, data = train_data)

Prior probabilities of groups:
        0         1
0.6666667 0.3333333

Group means:
    age acid
0 60.15 68.5
1 57.70 74.4
```

In this QDA model, the prior probabilities indicate that the training data has twice as many patients without nodal involvement (class 0) as with involvement (class 1), hence 0.6667 vs. 0.3333. The group means show that patients without nodal involvement tend to be slightly older (60.15 vs. 57.70) and have lower acid levels (68.5 vs. 74.4) compared to those with nodal involvement, suggesting younger age and higher acid levels may be associated with nodal involvement under this quadratic discriminant model.

7. Apply the trained quadratic classifier to the test data, find and interpret the test error rate.

```
> qda_pred <- predict(qda_model, newdata = test_data)
> mean(qda_pred$class != test_data$nodes)
[1] 0.4782609
```

This output indicates that the quadratic discriminant analysis classifier misclassified about 47.8% of the test data, implying that the model's predictive accuracy is around 52.2%. In other words, nearly half of the patients in the test set were incorrectly assigned to the nodal-involvement or no-involvement groups based on only age and acid levels.

```
Call:
qda(nodes ~ stage + Xray, data = train_data)

Prior probabilities of groups:
        0         1
0.6666667 0.3333333

Group means:
  stage Xray
0   0.2 0.05
1   0.6 0.40
> qda_pred <- predict(qda_model, newdata = test_data)
> mean(qda_pred$class != test_data$nodes)
[1] 0.2608696
```

Additionally, we can observe that when the model is based on the significant preoperative variables Xray and stage, the error rate is considerably lower.