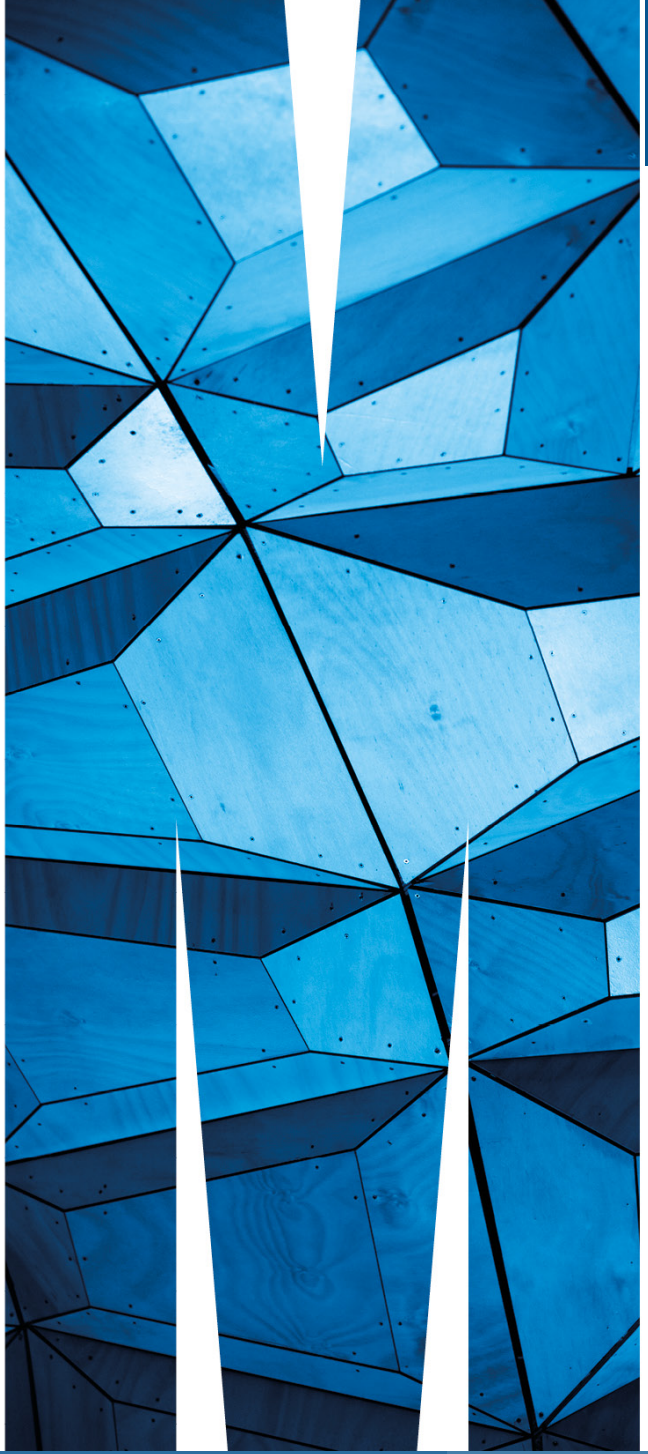
MONASH University

MONASH BUSINESS SCHOOL

# ETC3250: Classification with LDA

Week 8, class 1

Professor Di Cook, Econometrics and Business Statistics



GROUP OF EIGHT AUSTRALIA

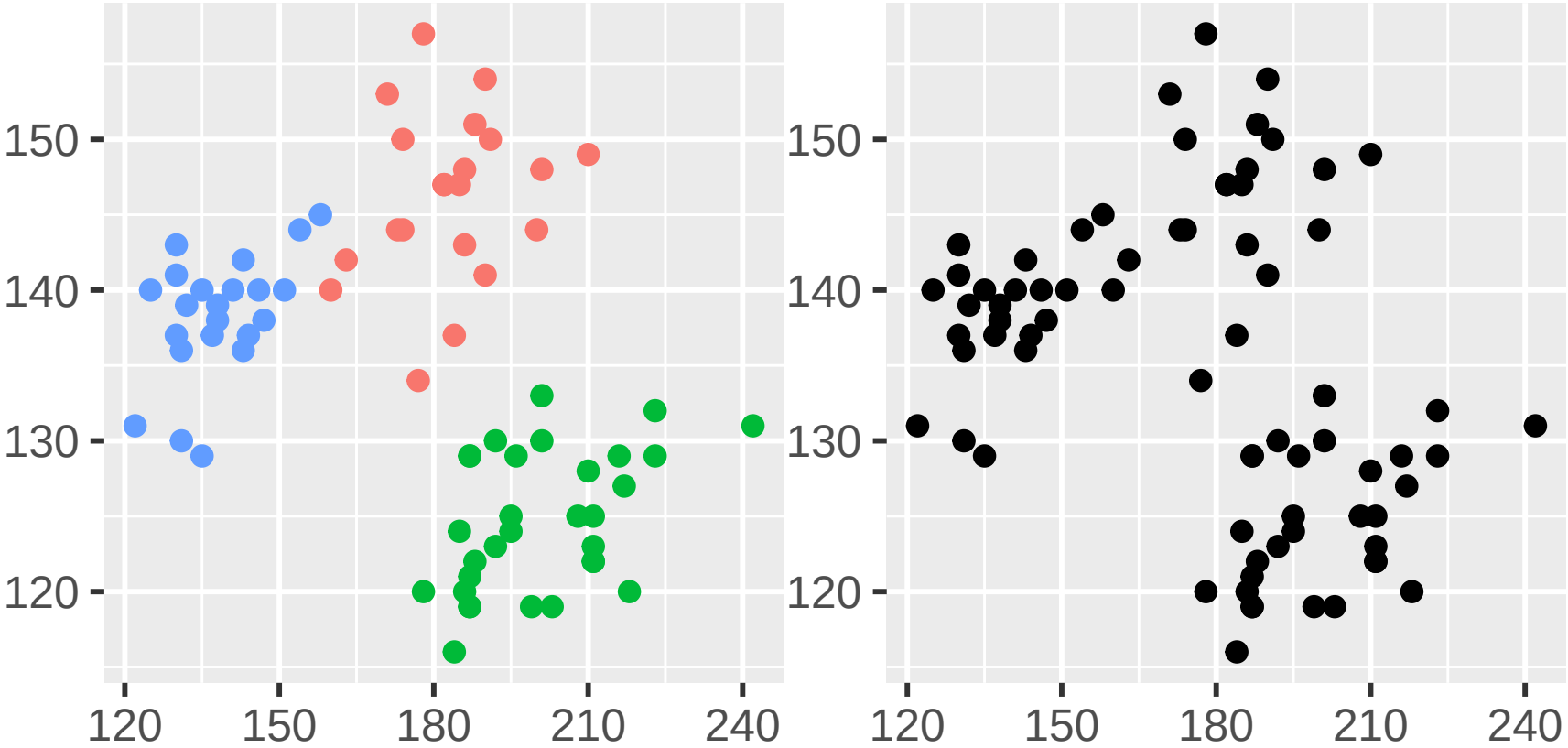
AACSB ACCREDITED EFMD EQUIS ACCREDITED ASSOCIATION OF MBAS ACCREDITED

# What is classification?

- Supervised classification includes multivariate techniques finding a rule for separating observations/cases into known classes, and using this rule to classify new observations.
- The process starts with a training sample, that is the full data set with known classes. Typically the variables that will be used to generate the classification rule are easy/cheap to measure, but the class is more difficult to measure. It is important to be able to classify new observations using variables that are easy to measure.

# Supervised vs Unsupervised

Unsupervised classification, also called cluster analysis, differs from supervised in that the classes are not known ahead of time, and need to be discovered first and labelled.

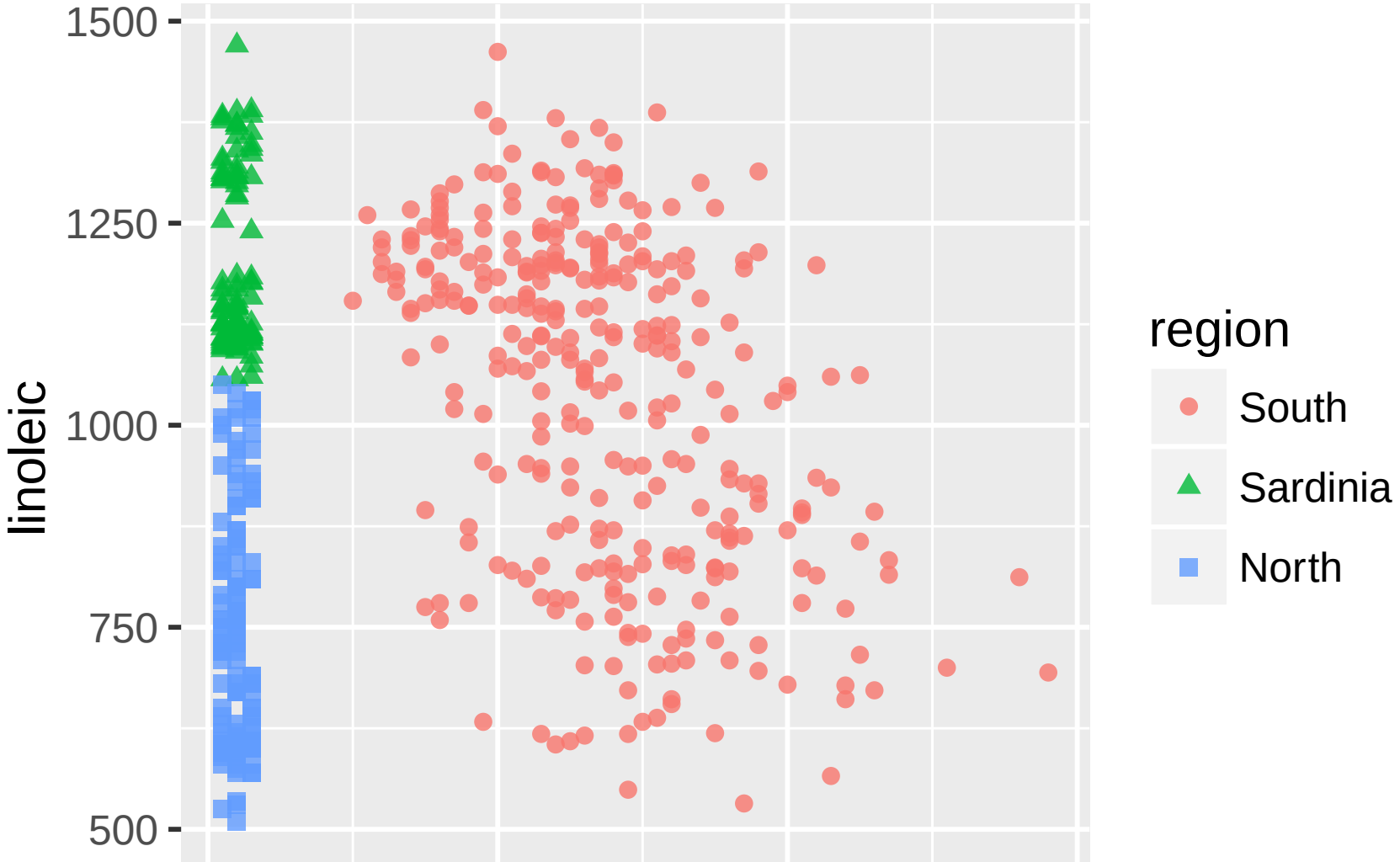


# Example 1: Olive oils


- Example data: Fatty acid composition of Italian olive oils.
- Three growing regions: labelled 1, 2, 3 (class variable)
- 8 fatty acids, % in the sample x 100.

# Question

- Two variables shown. How would you draw boundaries, so that if you received a new assayed sample you would be able to confidently predict which region produced the oil?



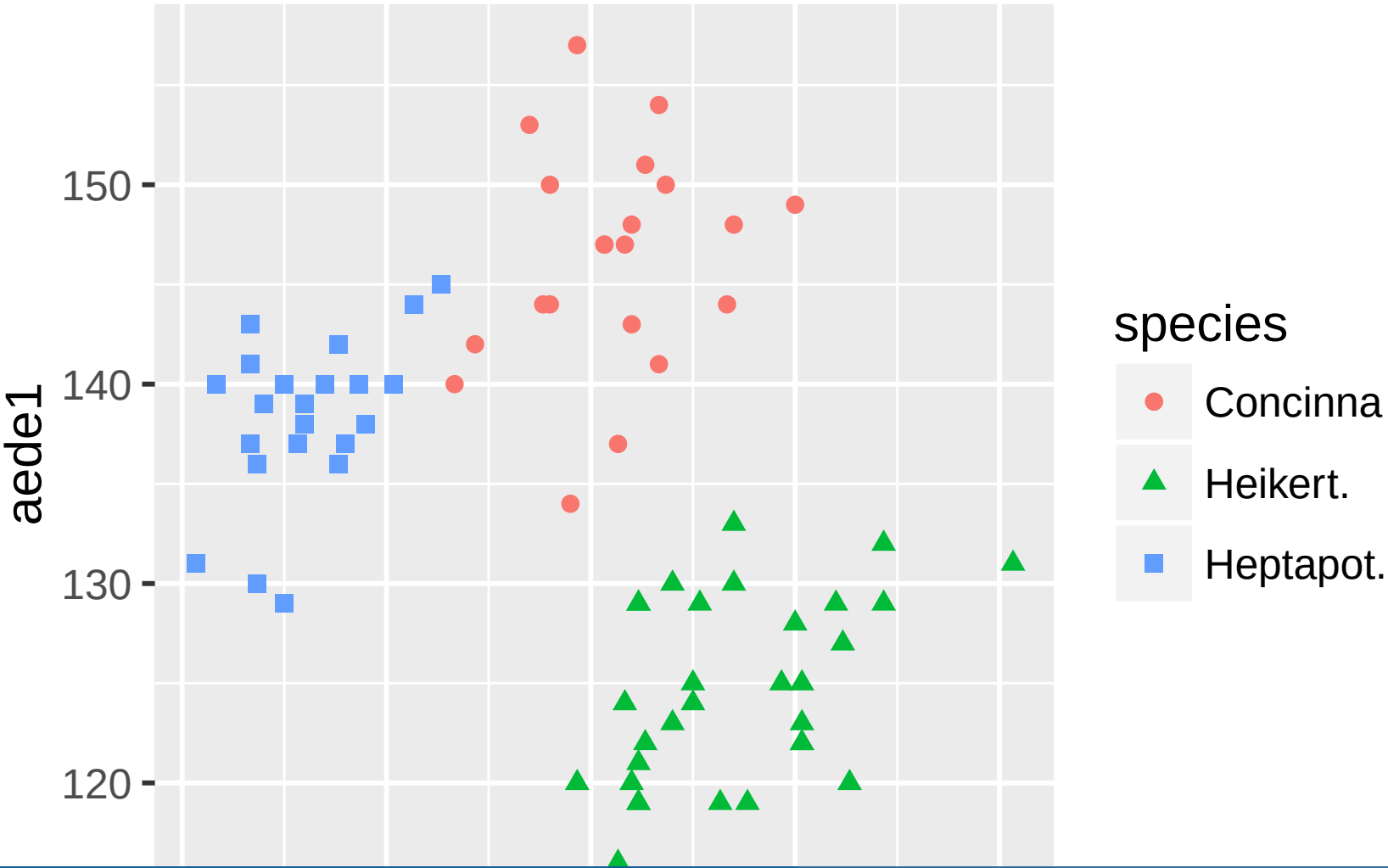
# Example 2: Beetles

(Copyright 2005 Jim McClarin)

- Example data: Beetles
- 3 species (class variable)
- 6 physical measurements

# Question

- Two variables shown. How would you draw boundaries, so that if you found a new specimen you would be able to confidently predict the species?



# Example 3: Australian crabs

(Andrei Nikulinsky @clusterpod)

- Example data: Crabs
- 2 species, 2 sexes (class variable)
- 5 physical measurements

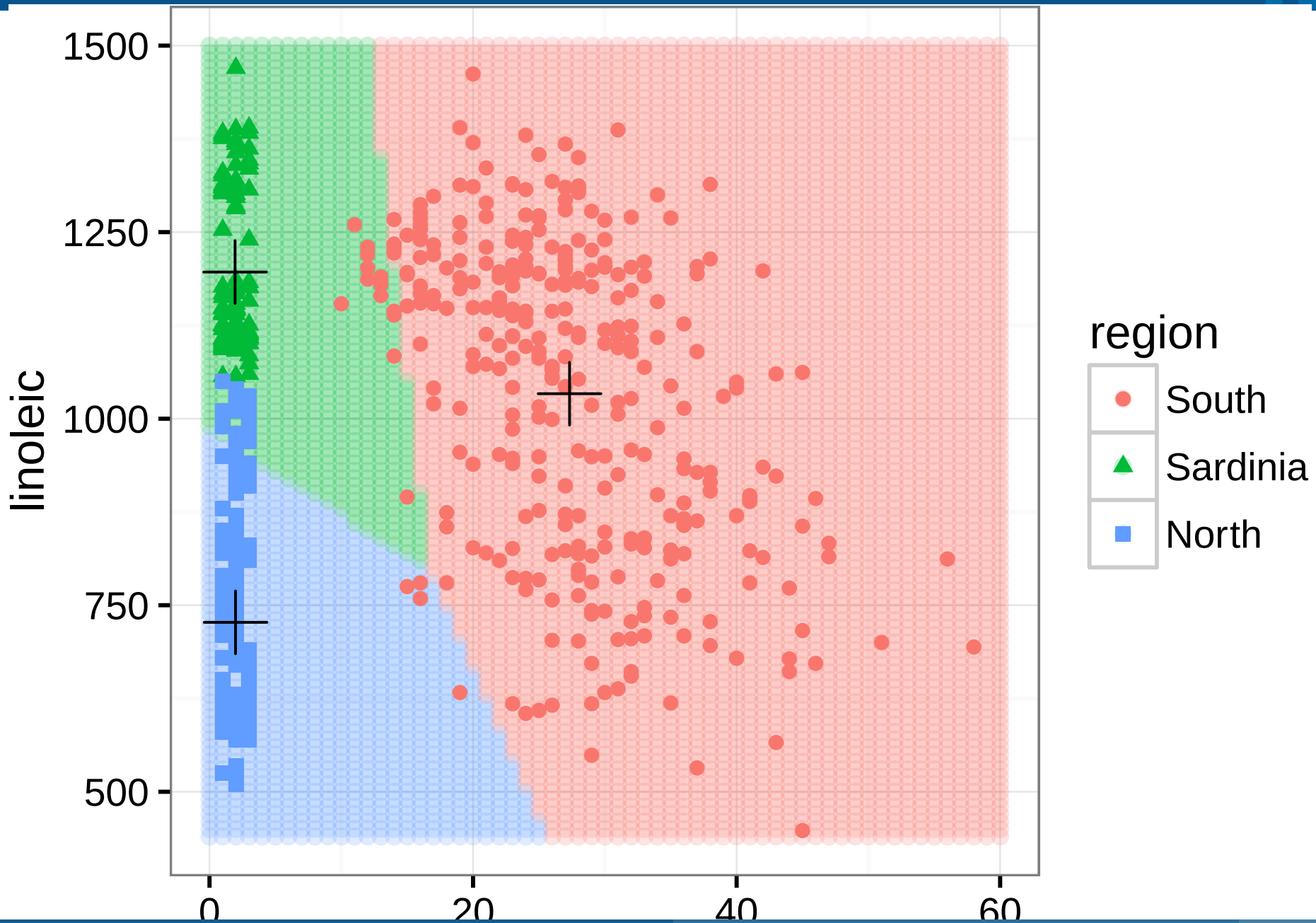
# Question

- Two variables shown. How would you draw boundaries, so that if you found a new specimen you would be able to confidently predict two sexes?

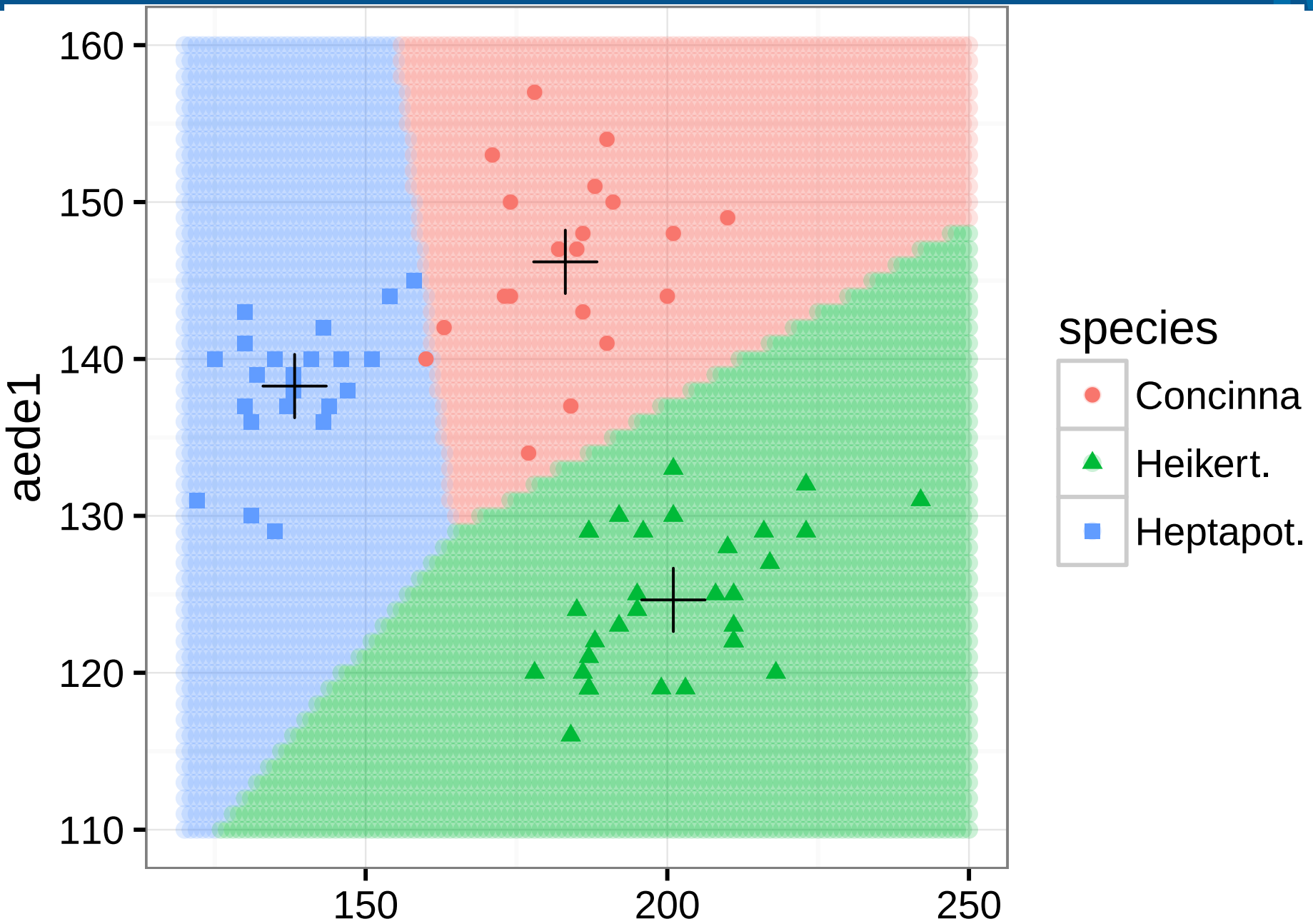
# Simple approach

- Calculate the mean of each class
- Calculate the distance between the new observation and each of the means
- Predict the new observation into the class of the closest mean

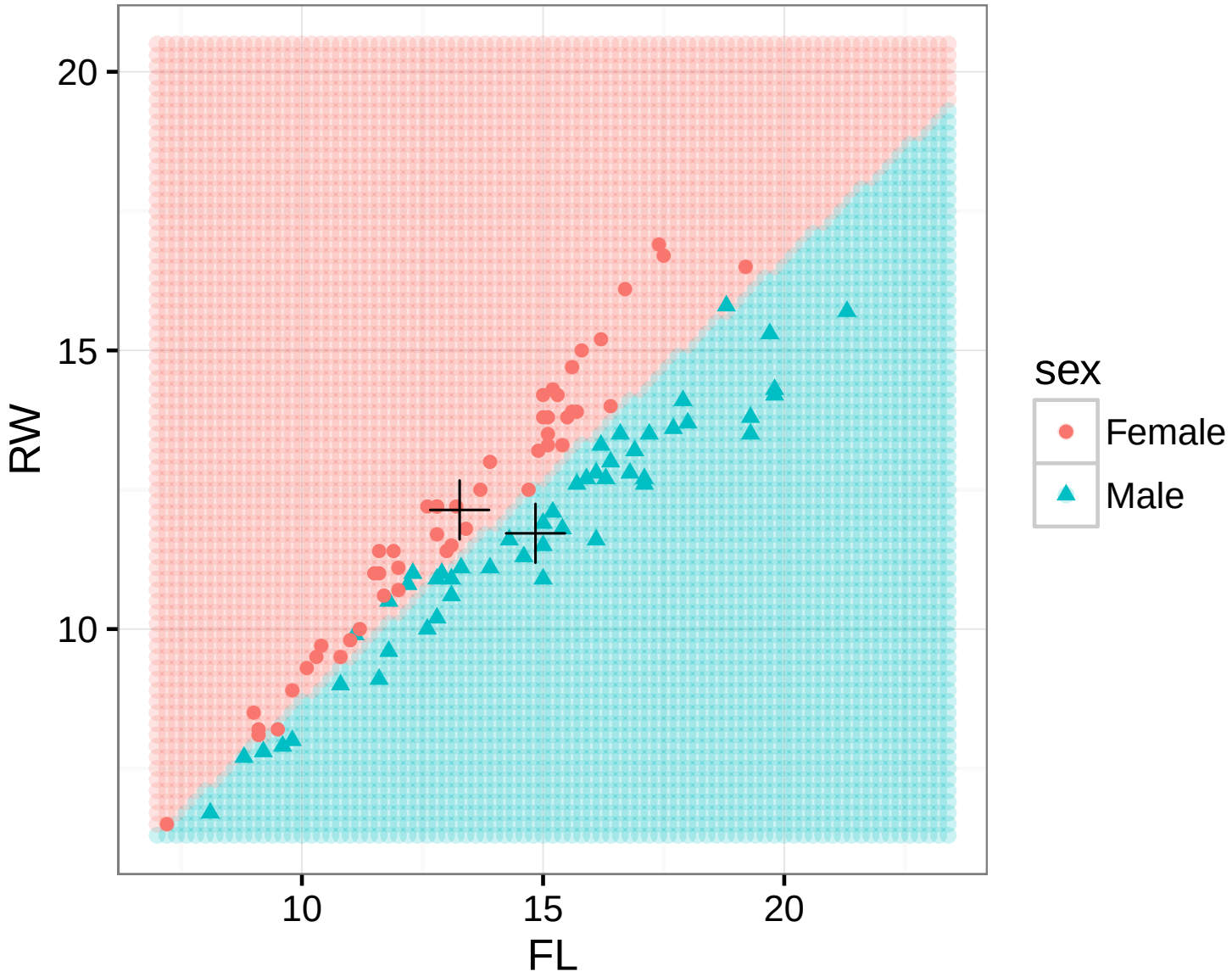
# Boundaries for Olive oils



# Boundaries for Beetles



# Boundaries for Crabs



# YOUR TURN

- How close do the LDA boundaries match what you designed?
- What is different?
- Why does it differ?

# LDA Rule for TWO groups, ONE variable

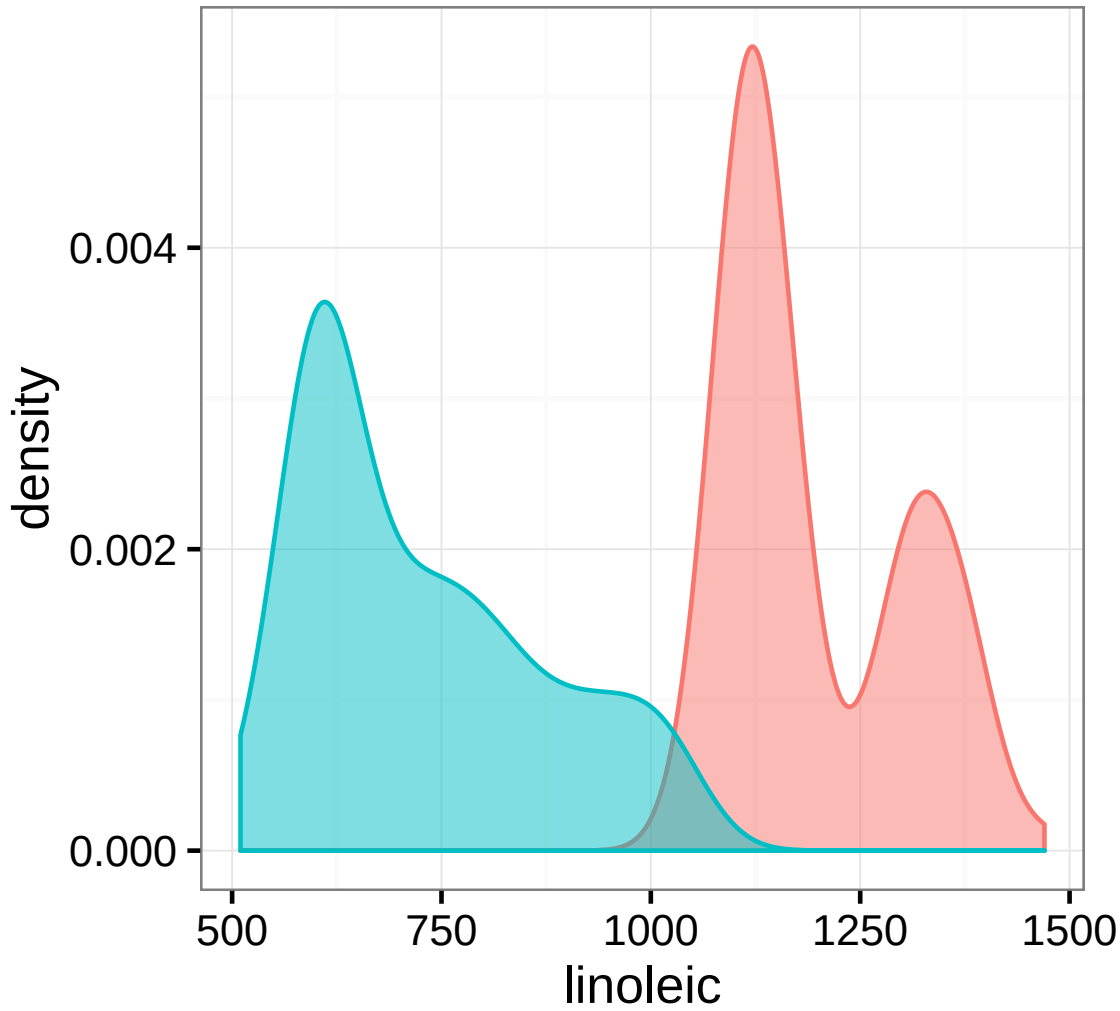
This method is called *linear discriminant analysis* (LDA). If there is only ONE VARIABLE and TWO GROUPS, then the LDA Rule would be:
*For a new observation, $x_0$, assign it to group 2 if*

$$x_0 - \frac{\bar{x}_1 + \bar{x}_2}{2} \geq 0 \quad (\ OR\ x_0 \geq \frac{\bar{x}_1 + \bar{x}_2}{2})$$

*otherwise assign it to group 1.*
This assumes that group 1 has the smaller mean.

# Example: olive oils

- One variable, two groups
- Where does the boundary go?



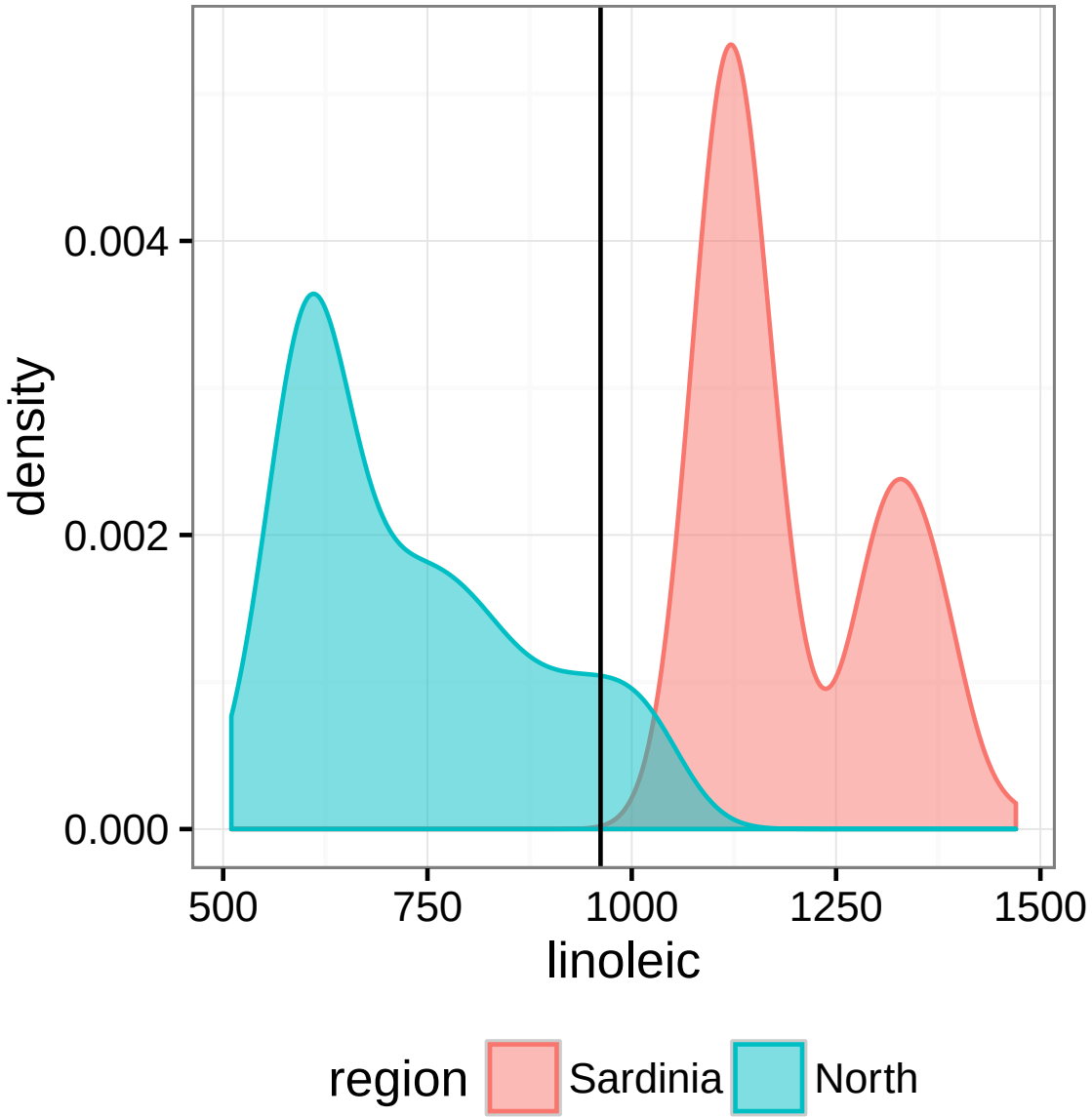
# LDA Rule

```
##           linoleic
## Sardinia 1196.5306
## North     727.0331
```

$\frac{\bar{x}_1+\bar{x}_2}{2} = 961.7818624$

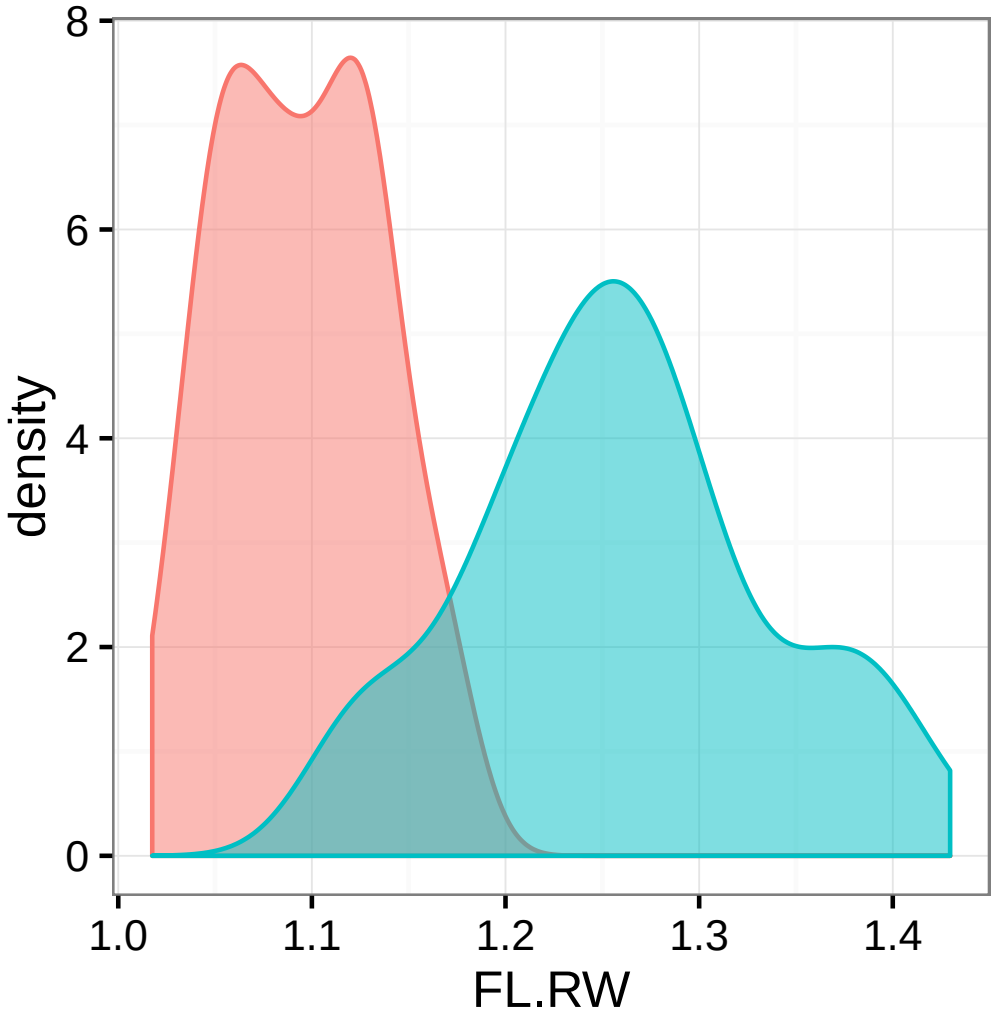
- What is the rule??
- To what group would a sample with 10.5% linoleic acid belong?
- What about a sample with 7.5% linoleic acid content?

# LDA Boundary



# Example: crabs

- One variable (ratio of FL and RW), two sexes
- Where does the boundary go?



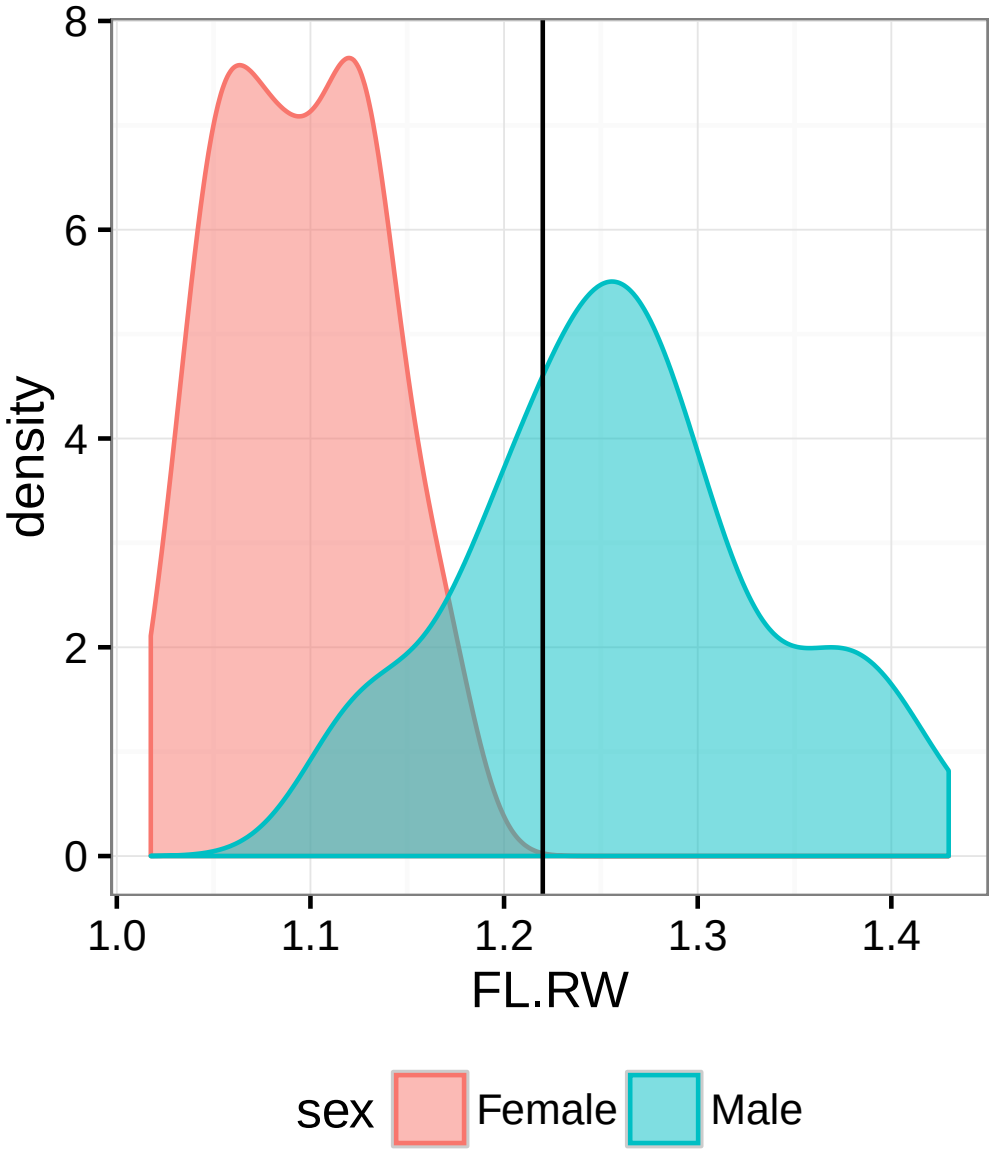
# LDA Rule

```
##             FL.RW
## Female 1.094717
## Male   1.259568
```

$\frac{\bar{x}_1+\bar{x}_2}{2} = 1.1771427$

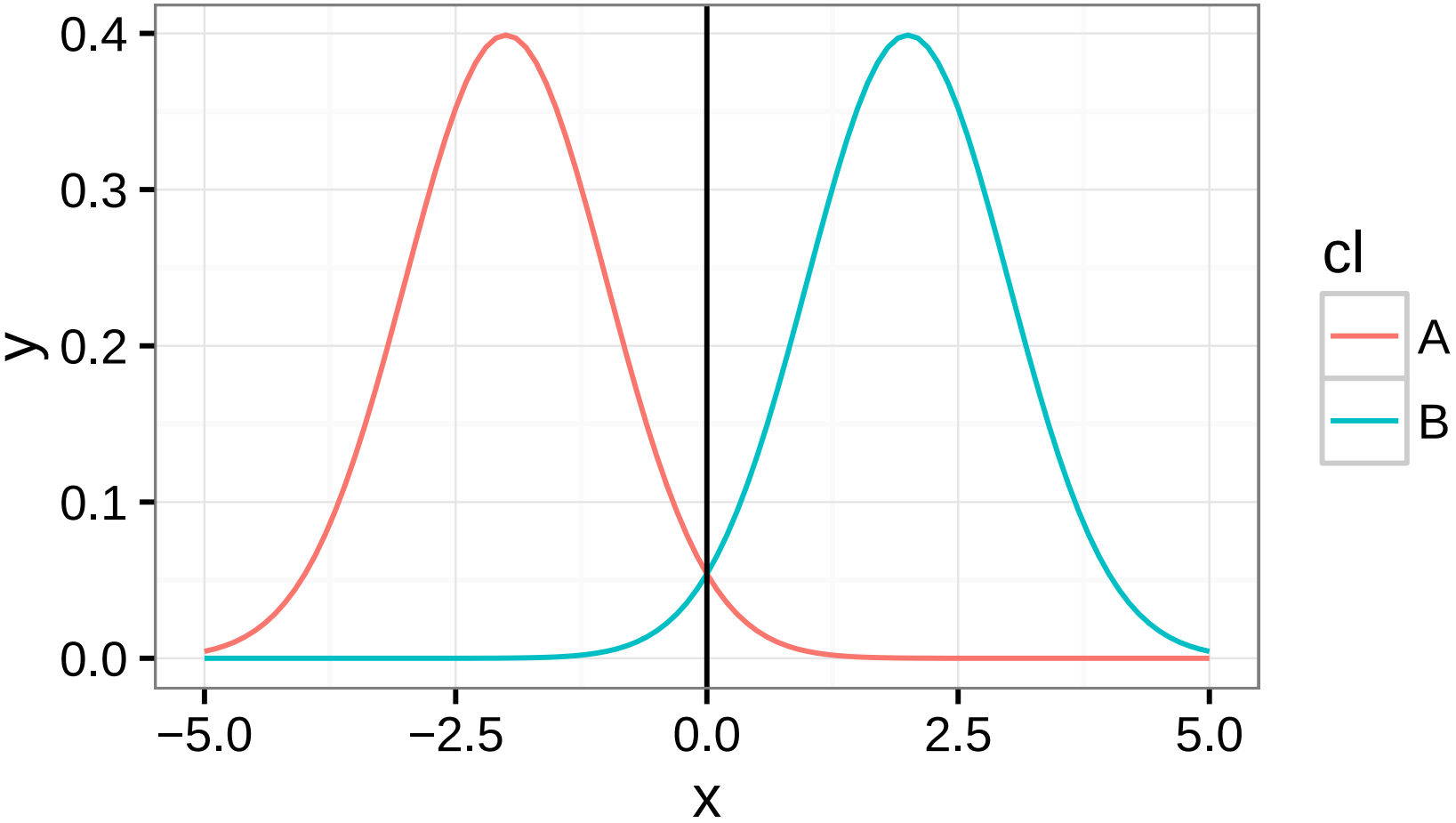
- What is the rule??

# LDA Boundary



# Theory

In a perfect world, if we assume we have two samples from two normal distributions with different means but same variance, then the LDA rule is exactly perfect.



# More than one variable, only two groups

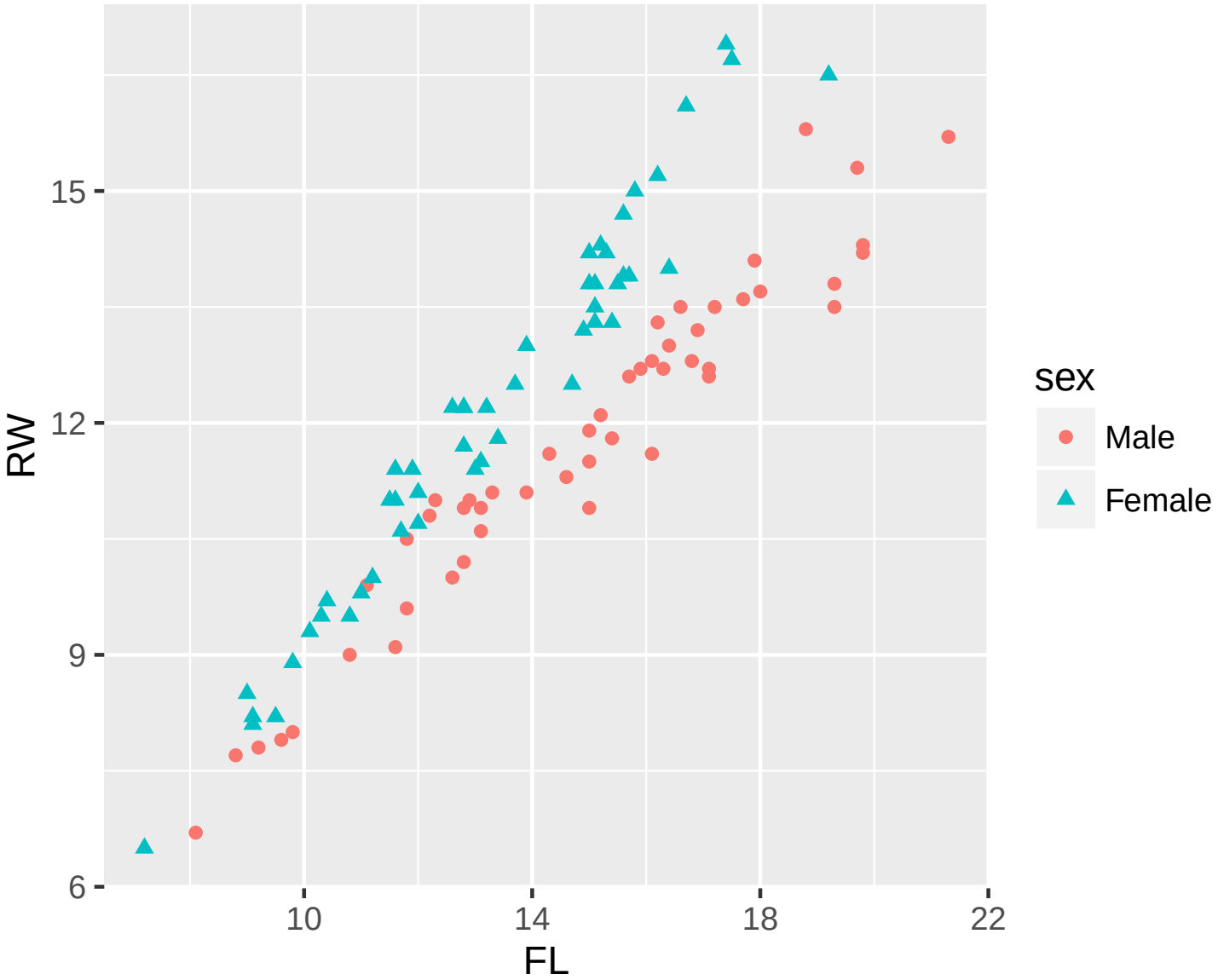
- With two groups project the data into 1D, and then compute means and compare

*For a new observation, $x_0$, assign it to group 2 if*

$$(\bar{x}_1 - \bar{x}_2)^T S_p^{-1}(x_0 - \frac{\bar{x}_1 + \bar{x}_2}{2}) \geq 0$$

*otherwise assign it to group 1.*
$S_p$ is the pooled variance-covariance matrix.

# Let's do it



# Summary statistics

```
## Source: local data frame [2 x 3]
##
##      sex    FL    RW
##   (fctr) (dbl) (dbl)
## 1   Male  14.8  11.7
## 2 Female  13.3  12.1

##       FL   RW
## FL 10.26 6.54
## RW  6.54 4.46

##      FL   RW
## FL 6.91 6.28
## RW 6.28 5.95
```

# Pooled variance-covariance

$$S_p = \frac{(n_1 - 1)S_1}{n_1 + n_2 - 2} + \frac{(n_2 - 1)S_2}{n_1 + n_2 - 2}$$

$S_p =$

```
##       FL   RW
## FL 8.58 6.41
## RW 6.41 5.20
```

$S_p^{-1} =$

```
##        [,1]  [,2]
## [1,]   1.47 -1.81
## [2,] -1.81   2.42
```

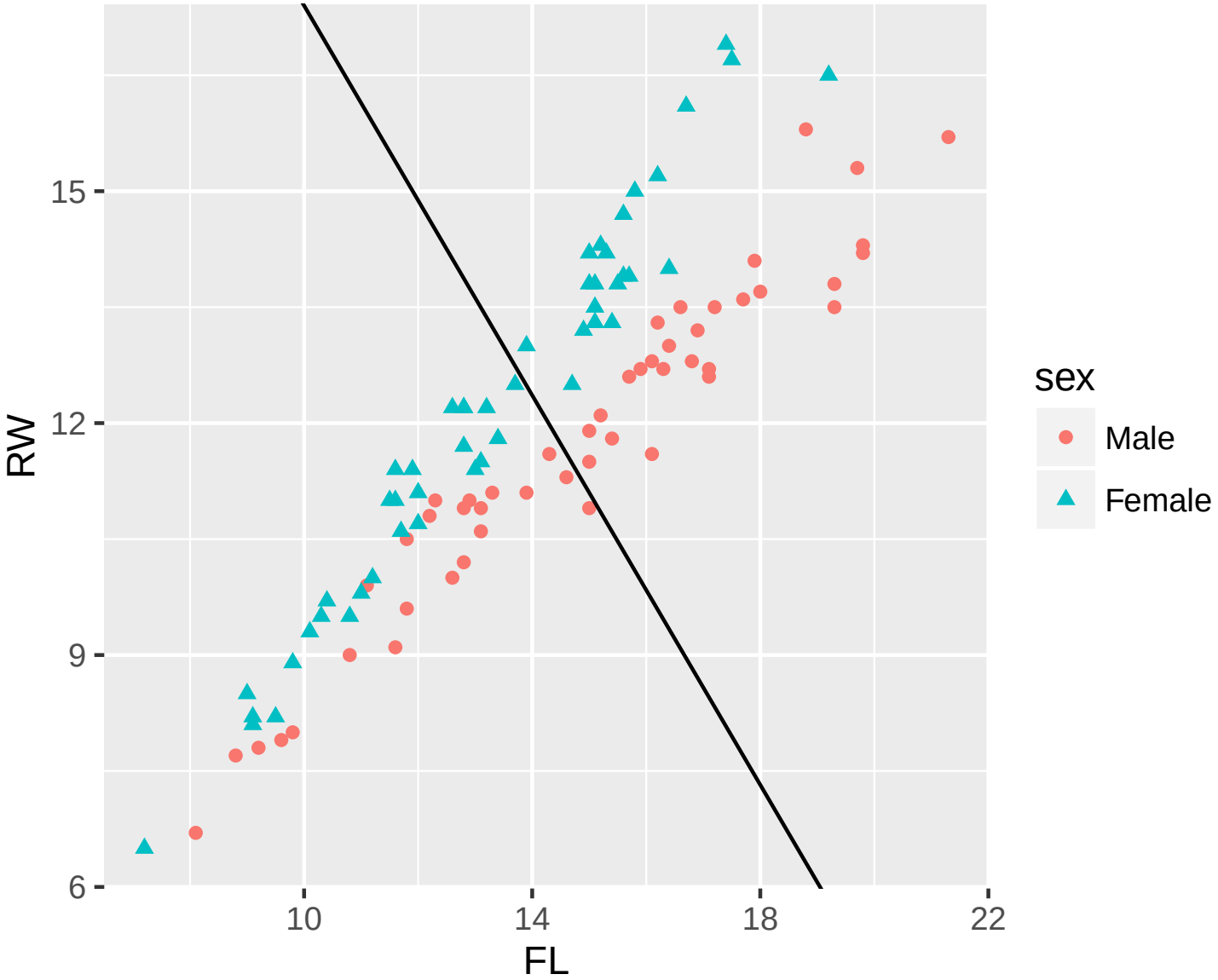
# Linear combination

$(\bar{x}_1 - \bar{x}_2)^T S_p^{-1} =$

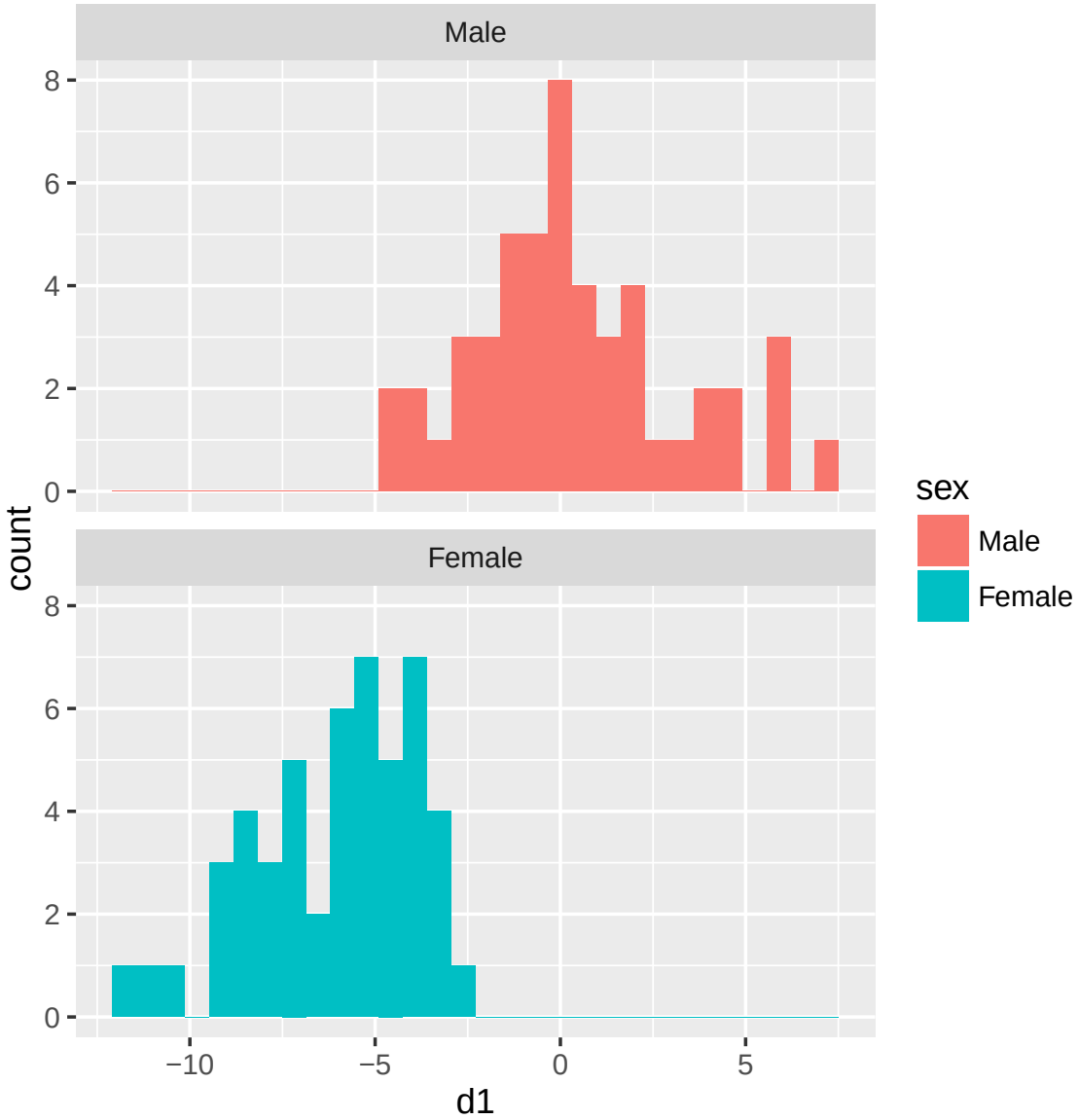
```
##          [,1]   [,2]
## [1,] 3.07 -3.86
```

New variable: 3.066xFL-3.859xRW
Low-dimensional space (discriminant space): 3.066xFL-3.859xRW=0

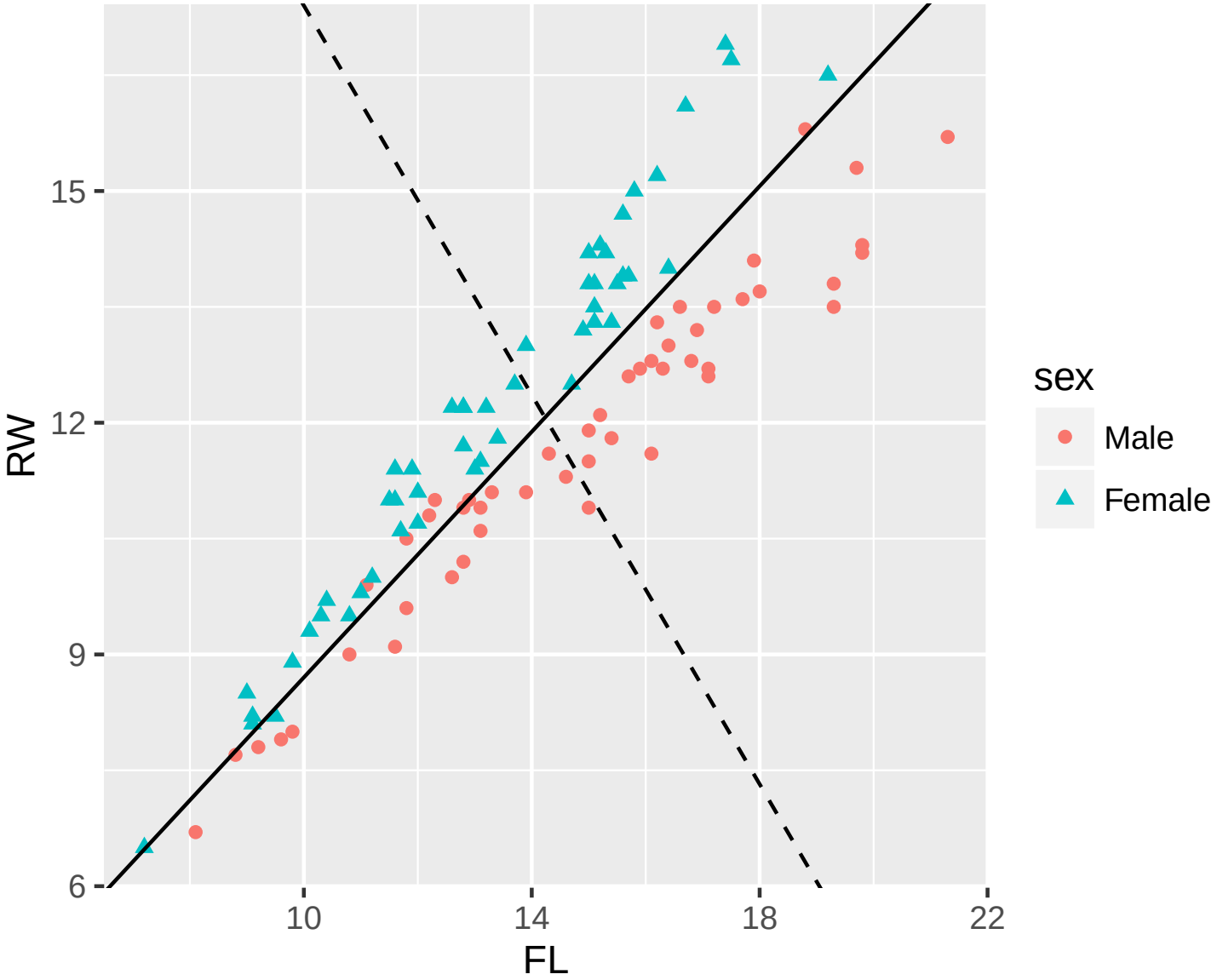
# Plot it - discriminant space



# Plot it - discriminant space



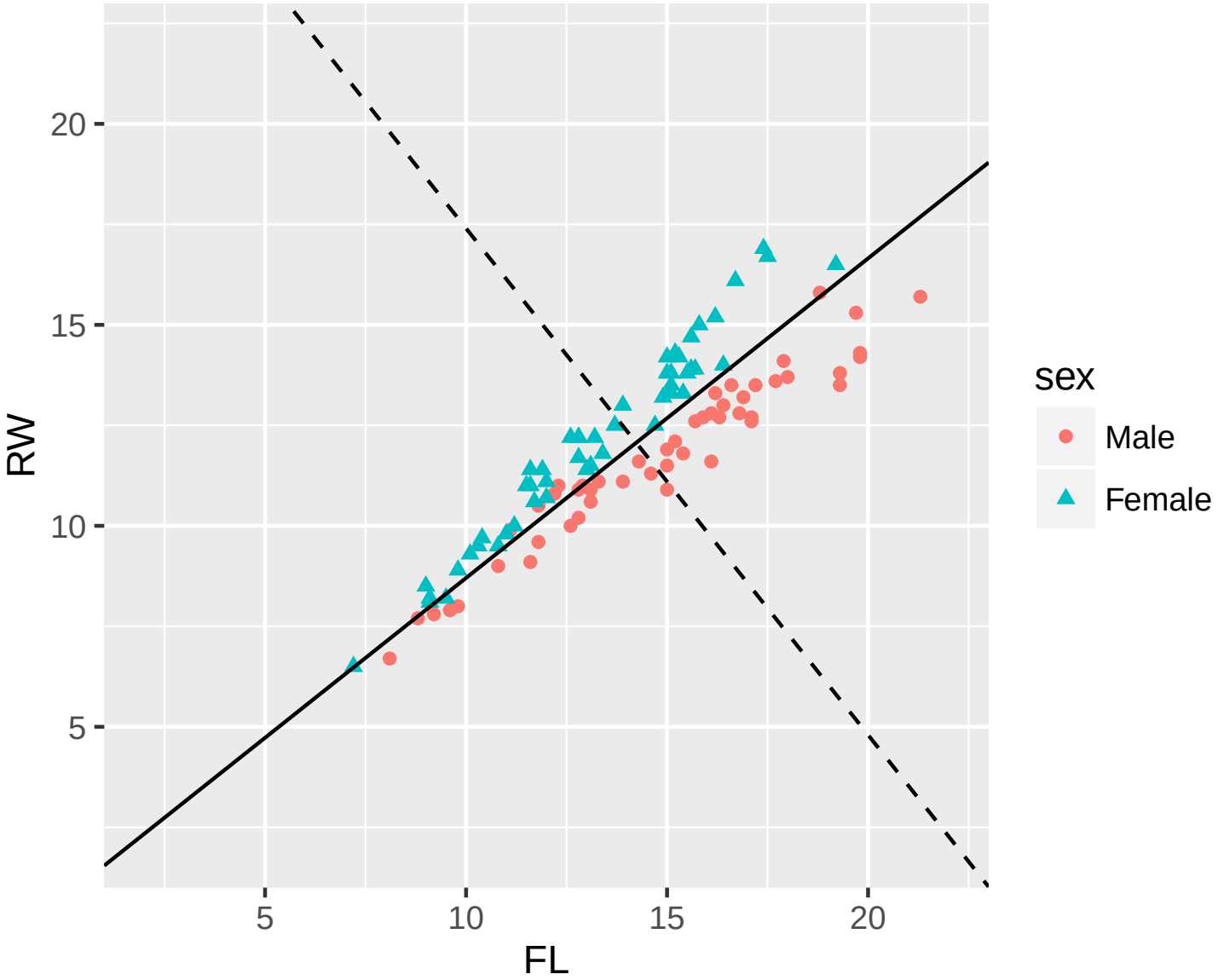
# Plot it - boundary



# ???

These lines are orthogonal !!!

# Plot it - boundary



# Accuracy and Error

- Error from the model is the proportion of misclassified cases to toal number of cases.
- It is important to look at this for each class also.
- Accuracy is 1-error

# Example: Crabs

```
##
##            Male Female
##    Male      45      5
##    Female     1     49
```

- Overall error $= 6/100 = 0.06$
- Males $= 5/50 = 0.10$
- Females $= 1/50 = 0.02$

# Training, validation and test error

In practice, it is important to use a separate data set to compute the error, in order to get the likely error that would be made with future samples.

# Splitting your data

1. Decide on % in each of training, validation and test sets
2. Use this % to select cases within each class, to preserve the % by class
3. Generate random numbers to select cases, and keep these (or the seed)

Why?

# Example

Say, use 50, 25, 25 %
Training: 2, 6, 10, 14, 15, 20, 21, 22, 25, 26, 27, 28, 29, 31, 32, 33, 34, 36, 37, 38, 40, 42, 47, 49, 50, 51, 52, 55, 56, 57, 58, 59, 61, 62, 64, 65, 71, 73, 76, 79, 80, 82, 84, 86, 88, 92, 93, 97, 98, 100
Validation: 1, 3, 7, 8, 11, 12, 17, 18, 19, 30, 39, 43, 44, 53, 54, 60, 63, 66, 69, 74, 78, 81, 85, 87, 89, 90
Test: 3, 5, 7, 8, 12, 13, 18, 19, 23, 24, 41, 43, 44, 45, 46, 48, 53, 54, 60, 63, 67, 68, 69, 70, 72, 74, 75, 77, 81, 83, 85, 90, 91, 94, 95, 96, 99

# Calculate test error

Training error:

```
## [1] 0.06
```

Validation error:

```
## [1] 0.12
```

Test error:

```
## [1] 0.08
```

# Incorporating prior probabilities

Sometimes it can be more costly to misclassify members on one group than those of the other group, e.g. malignant vs benign tumor. Address this by assigning prior probabilities $p_1, p_2 (p_1 + p_2 = 1)$ for each group.

$$(\bar{x}_1 - \bar{x}_2)^T S_p^{-1} (x_0 - \frac{\bar{x}_1 + \bar{x}_2}{2}) \geq ln \frac{p_2}{p_1}$$

The effect is to shift the boundary away from the group with the highest prior probability.

# More than two groups

The basic principle *allocate the new observation to the group that has the closest mean* holds.
Rearrange the equations:

$$(\bar{x}_1 - \bar{x}_2)^T S_p^{-1}(x_0 - \frac{\bar{x}_1 + \bar{x}_2}{2}) \geq ln\frac{p_2}{p_1}$$

$$\bar{x}_1^T S_p^{-1} x_0 + \bar{x}_1^T S_p^{-1}\bar{x}_1 - (\bar{x}_2^T S_p^{-1} x_0 - \bar{x}_2^T S_p^{-1}\bar{x}_2 \geq (ln(p_2) - ln(p_1))/2$$

$$\bar{x}_1^T S_p^{-1} x_0 + \bar{x}_1^T S_p^{-1}\bar{x}_1 - ln(p_2) \geq \bar{x}_2^T S_p^{-1} x_0 - \bar{x}_2^T S_p^{-1}\bar{x}_2 - ln(p_1)/2$$

# Discriminant equations/functions

For $j = 1, ..., g$

$$\bar{x}_j^T S_p^{-1} x_0 + \bar{x}_j^T S_p^{-1} \bar{x}_j - ln(p_2)$$

Calculate this for every group and allocate the new observation to the group which yields the largest value.

# Canonical coordinates

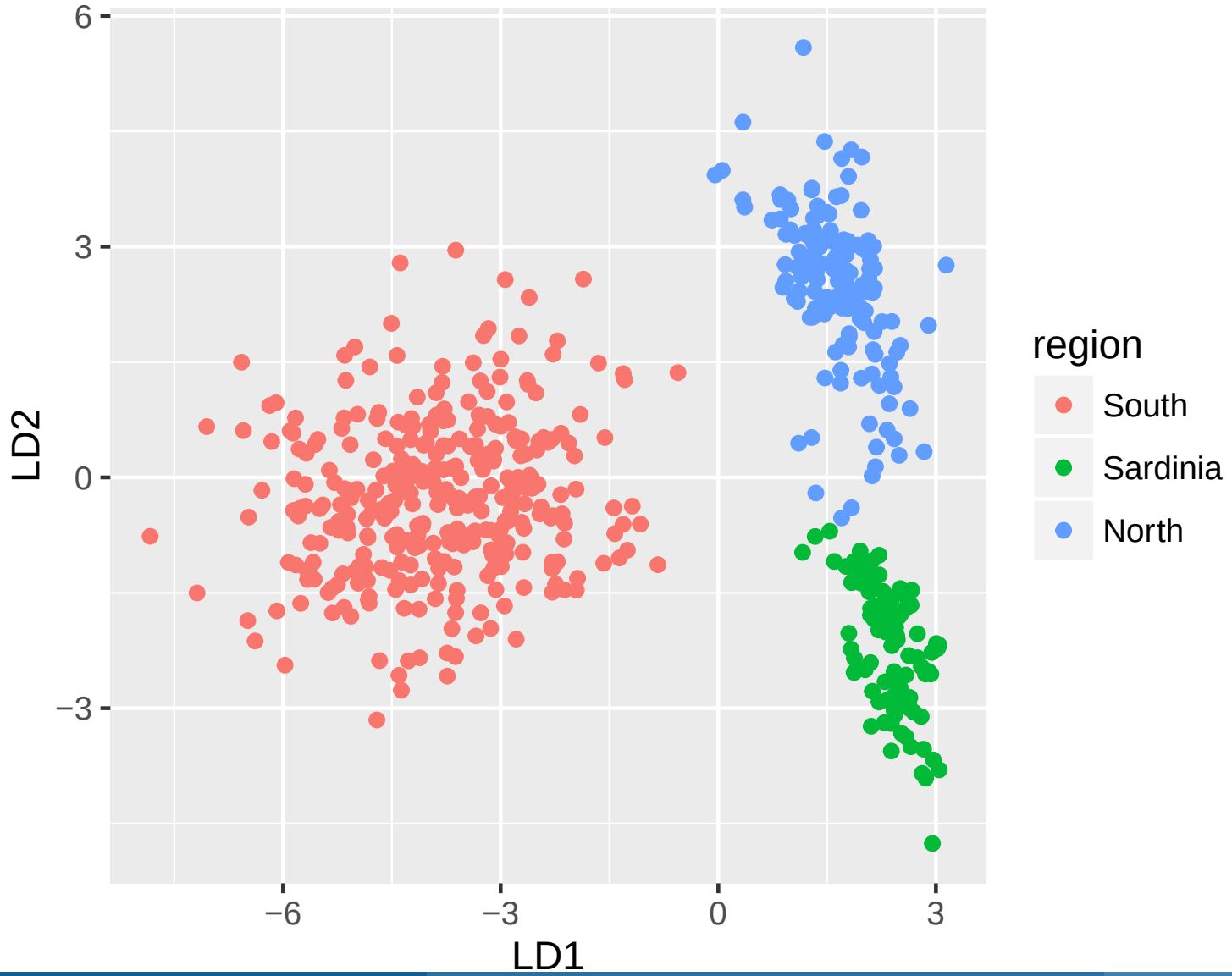
One of the major advantages of LDA is that a low-dimensional space of best separation can be found. This is called the discriminant space, and is defined by a new set of variables called canonical variables. Define, B=between groups sum of squares, W=within group sum of squares

$$B = \sum_{i=1}^{g} n_i \bar{X}_i - \bar{X}\bar{X}_i - \bar{X}^T W = \sum_{i=1}^{g} (n_i - 1) S_i$$

Compute an eigendecomposition of $W^{-1}B$ to get the discriminant space.

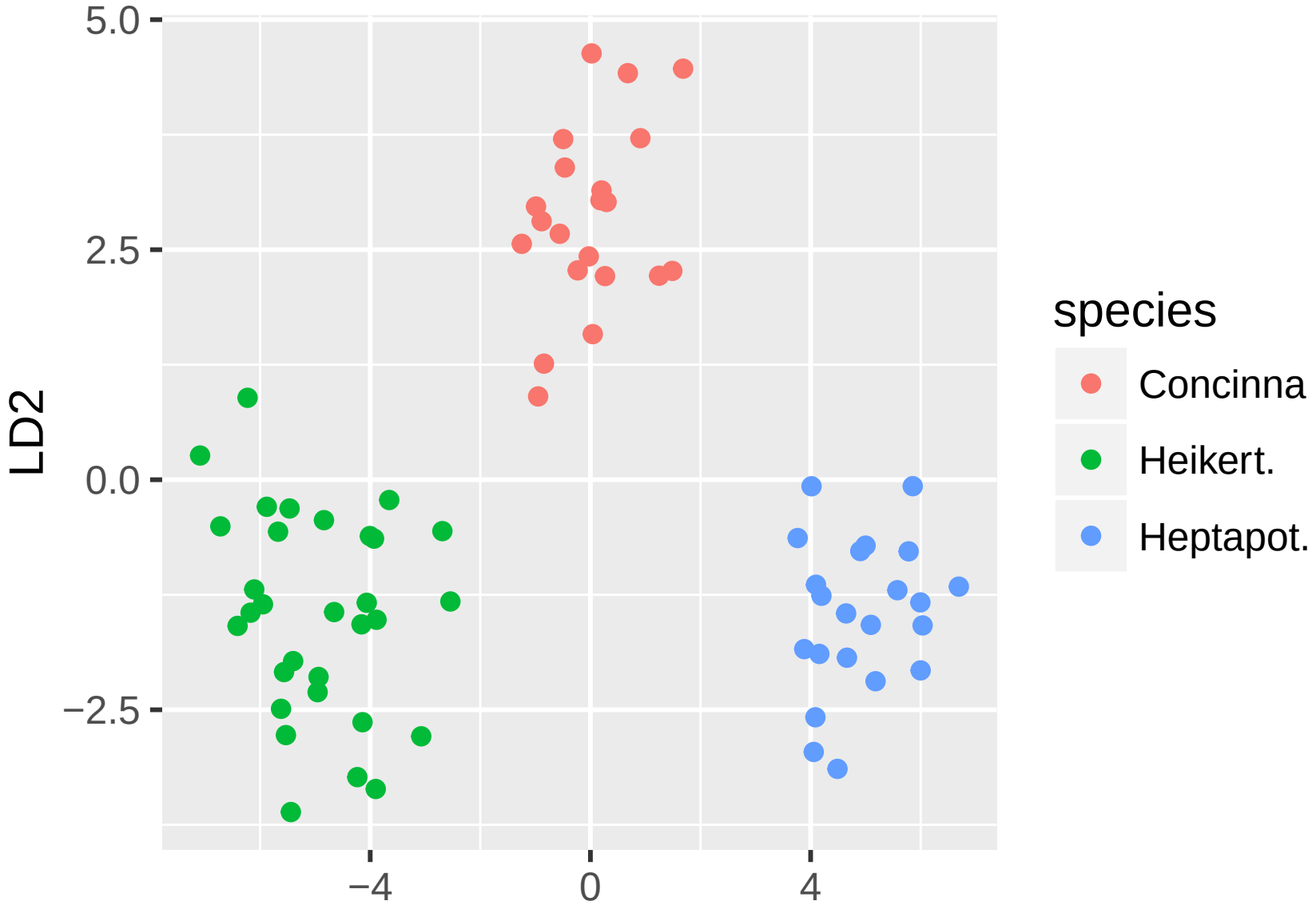
# Example: olive oils

- All 8 variables, and 3 groups.



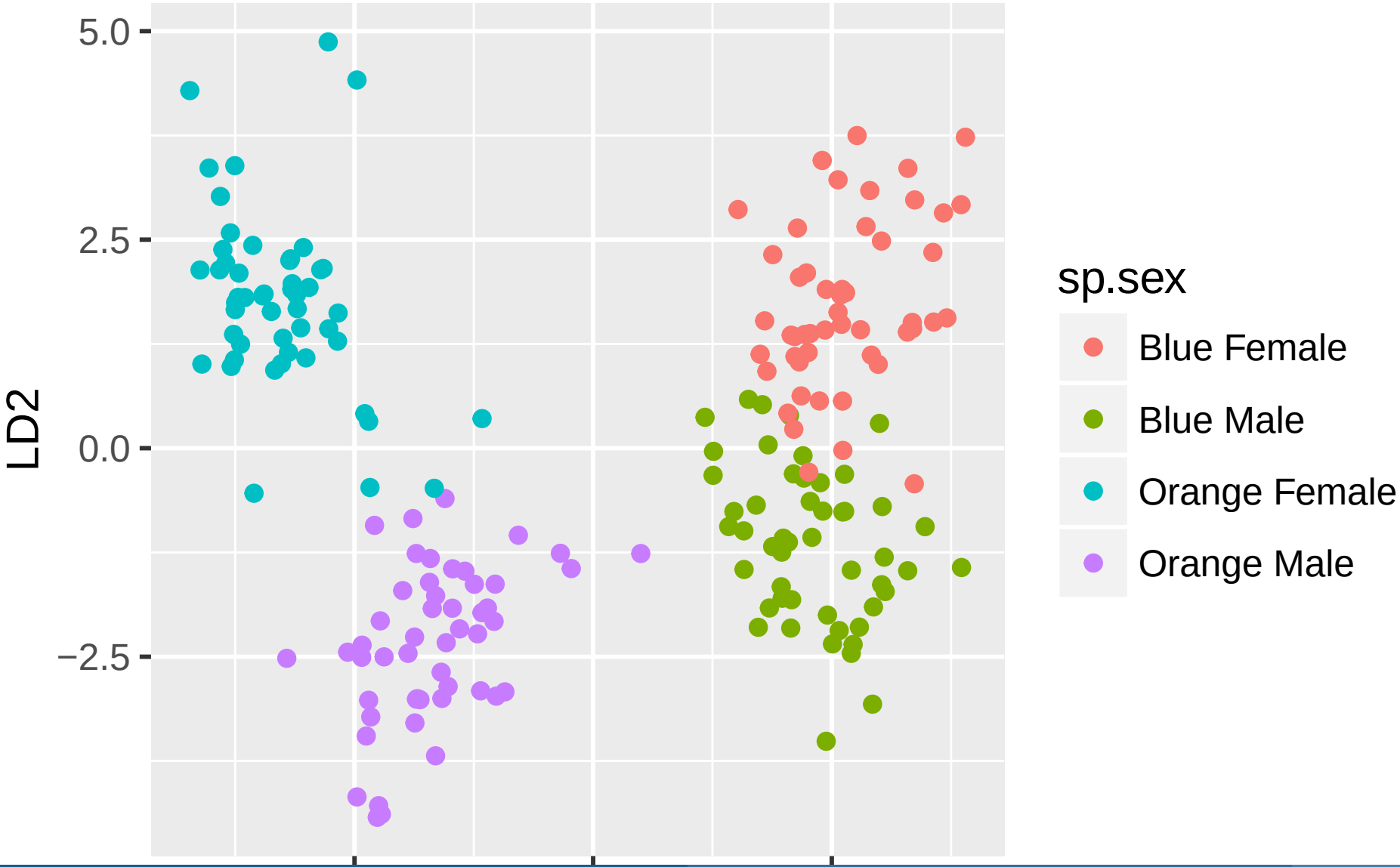
# Example: flea

- All 6 variables, and 3 groups.



# Example: crab

- All 5 variables, and 4 groups.



# Assumptions for LDA

- Variance-covariance for each group is the same! Pretty strict assumption! (Homogeneity, homoskedastic)
- Samples from from multivariate normal populations, with (different) means

# Quadratic discriminant analysis

When the equal variance-covariance assumption isn't satisfied but the population could still be considered to be normal, the rule would change from linear to quadratic:

*Allocate a new observation $x_0$ to group 1 if*

$$-\frac{1}{2}x_0^T(S_1^{-1}-S_2^{-1})x_0+(\bar{x}_1^T S_1^{-1}-\bar{x}_2^T S_2^{-1})x_0-$$

$$\frac{1}{2}ln\frac{|S_1|}{|S_2|}+(\bar{x}_1^T S_1^{-1}\bar{x}_1-\bar{x}_2^T S_2^{-1}\bar{x}_2)\geq ln(\frac{p_2}{p_1})$$

*otherwise allocate to group 2*

# Share and share alike