

Exercise 5

Tutorial Exercise.

1. The following R output relates to a discriminant analysis for two species of iris.

```

> ivers=as.matrix(iris[51:100,1:4])
> ivirg=as.matrix(iris[101:150,1:4])
> iversm=apply(ivers,2,mean)
> ivirgm=apply(ivirg,2,mean)
> iversm
Sepal.Length Sepal.Width Petal.Length Petal.Width
           5.936           2.770           4.260           1.326
> ivirgm
Sepal.Length Sepal.Width Petal.Length Petal.Width
           6.588           2.974           5.552           2.026
> a=solve(S)%*(iversm-ivirgm)
> a
           [,1]
Sepal.Length  3.556303
Sepal.Width   5.578621
Petal.Length -6.970128
Petal.Width  -12.386041
> (iversm-ivirgm)%*a
           [,1]
[1,] 14.21889
> anova(lm(c(rep(0,50),rep(1,50))~rbind(ivers,ivirg)[:3,4]+rbind(ivers,ivirg)))
Error: syntax error in "anova(lm(c(rep(0,50),rep(1,50))~rbind(ivers,ivirg)[:3,4]+rbind(ivers,ivirg)))"
> anova(lm(c(rep(0,50),rep(1,50))~rbind(ivers,ivirg)[,3:4]+rbind(ivers,ivirg)))
Analysis of Variance Table

Response: c(rep(0, 50), rep(1, 50))
          Df Sum Sq Mean Sq F value    Pr(>F)
rbind(ivers, ivirg)[, 3:4]  2 17.9939  8.9970 158.199 < 2.2e-16 ***
rbind(ivers, ivirg)         2  1.6033  0.8017  14.096 4.357e-06 ***
Residuals                   95  5.4028  0.0569

> sum(c(as.matrix(ivers)%*a)>rep(((iversm+ivirgm)/2)%*a,50))
[1] 48
> sum(c(as.matrix(ivirg)%*a)>rep(((iversm+ivirgm)/2)%*a,50))
[1] 1

```

- (a) Test the hypothesis that the two species have the same mean vectors.

- (b) Determine the discriminant function from the output.
- (c) Using the fitted discriminant rule determine how many of the 100 observations are misclassified.
- (d) Test the hypothesis that petal variables (variables 3 and 4) contain all the information necessary to discriminate between the species.

2. Observations on two responses are collected for each of 3 treatments:

Treatment 1: (6,7), (5,9), (8,6), (4,9), (7,9)

Treatment 2: (3,3), (1,6), (2,3)

Treatment 3: (2,3), (5,1), (3,1), (2,3)

For Treatment i calculate $\bar{\mathbf{x}}_i^T$, $S_i, i = 1, 2, 3$. Next calculate $\bar{\mathbf{x}}^T$ and the pooled sample covariance matrix S . Finally calculate the between groups sum of squares and products matrix B .

3. A fashion shop estimates the percentage increase in monthly sales after the launch of a promotion campaign. A simple random sample of $n = 10$ branches is drawn from $N = 256$ branches and the monthly sales in thousand dollars and related information are given below:

Branch i	1	2	3	4	5	6	7	8	9	10
Monthly sales after y_i	65	109	60	124	128	104	65	61	49	56
Monthly sales before x_i	52	100	60	128	104	98	48	64	96	48
Increase $d_i = y_i - x_i$	13	9	0	-4	24	6	17	-3	-47	8

$$\sum_i x_i = 798, \sum_i y_i = 821, \sum_i x_i^2 = 71,028, \sum_i y_i^2 = 75,765, \sum_i x_i y_i = 71,672,$$

$$\bar{d} = 2.3, s_d^2 = 377.3444$$

The shop records show that the total monthly sales for all branches before the launch of the campaign was $X = 20,500$ thousand dollars.

- (a) Estimate the *percentage increase* in monthly sales after the launch of the promotion campaign and its standard error using
 1. the *Ratio* estimator and
 2. the estimator $\hat{R}' = \frac{\bar{y}}{\bar{X}}$ estimators. [1.028822, 0.076021; 1.025249, 0.117988]]
- (b) Estimate the average monthly sales after the launch of the promotion campaign and its standard error using the two estimators in (a). [82.38614, 6.087601]
- (c) Estimate the average monthly sales after the launch of the promotion campaign and its standard error using a new estimator defined as

$$\hat{Y}_d = \bar{X} + \bar{d}$$

where \bar{X} is the average monthly sales of all branches before the campaign and \bar{d} is the sample mean of the increases d_i in monthly sales, after the campaign. [82.37813, 6.021664]

- (d) Given that the correlation coefficient $\hat{\rho}_{y,x} = 0.7854$, compare the three estimators.

Computer Exercise

1. Data on 9 mandible measurements (in mm) for samples drawn from different species of dog are stored in **canine1** (modern Thai dogs) and **canine2** (golden jackals).
 - (a) Calculate Fisher's linear discriminant function based on the variables X_1, \dots, X_5 and obtain the allocation rule for dogs based on this function.
 - (b) Estimate the misclassification probabilities based on these data.
 - (c) Obtain a discriminant rule using a multiple regression approach, checking that this is equivalent to that obtained earlier.
 - (d) Test the hypothesis that X_1, X_2, X_3 and X_5 do not contribute any information on the discrimination between the two groups in addition to that provided by X_4 .