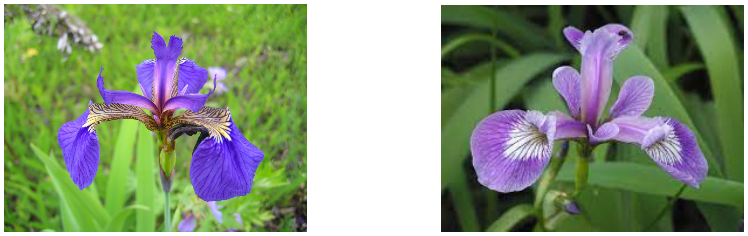
FISHER'S ANALYSIS OF IRIS DATA[*]

BY QIAN ZHAO [1]

[1] Department of Biomedical Data Science, Stanford University, qzhao1@stanford.edu

I use an analysis of the Iris data set to illustrate how to use the "rticles" package to create a reproducible manuscript.

1. The iris data. The Iris data, collected by Dr. E. Anderson, contains measurements of the flowers of fifty plants each of the two species *Iris setosa* and *I.versicolor*. Figure 1 shows pictures of the two species. The data includes four measurements: sepal length, sepal width, petal length, and petal width. A few rows of the data are shown in Table



(a) *Iris setosa*

(b) *I.versicolor*

Figure 1: Two iris species

TABLE 1

First few rows in the iris data

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa

2. Fisher linear discriminant analysis. In a 1936 article, Fisher (1936) considered the question: what linear function of the four measurements

$$X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4 \tag{2.1}$$

maximizes the *ratio* of the difference between the means to the standard deviation within species?

The observed means and their differences are shown in Table 2. We can also compute the sum of squares and products of deviation from specific means of each species (Table 3).

[*] Based on the article "The Use of Multiple Measurements in Taxonomic Problems" by R. A. Fisher (1936)

Keywords and phrases: reproducible manuscript, iris data.

TABLE 2
Observed means for two species and their difference (cm.)

Variable	Versicolor	Setosa	Difference
Sepal length	5.936	5.006	0.930
Sepal width	2.770	3.428	-0.658
Petal length	4.260	1.462	2.798
Petal Width	1.326	0.246	1.080

TABLE 3
Sums of squares and products of four measurements, within species (cm.2)

	Sepal length	Sepal width	Petal length	Petal Width
Sepal length	19.1434	9.0356	9.7634	3.2394
Sepal width	9.0356	11.8658	4.6232	2.4746
Petal length	9.7634	4.6232	12.2978	3.8794
Petal Width	3.2394	2.4746	3.8794	2.4604

The linear combination that maximizes D^2/S, where

$$D = \lambda_1 d_1 + \lambda_2 d_2 + \lambda_3 d_3 + \lambda_4 d_4, \tag{2.2}$$

where d_i are the differences between the species means and

$$S = \sum_{p=1}^{4} \sum_{q=1}^{4} \lambda_p \lambda_q S_{pq}, \tag{2.3}$$

is the solution to a set of linear equations

$$\begin{cases} S_{11}\lambda_1 + S_{12}\lambda_2 + S_{13}\lambda_3 + S_{14}\lambda_4 = d_1, \\ S_{21}\lambda_1 + S_{22}\lambda_2 + S_{23}\lambda_3 + S_{24}\lambda_4 = d_2, \\ S_{31}\lambda_1 + S_{32}\lambda_2 + S_{33}\lambda_3 + S_{34}\lambda_4 = d_3, \\ S_{41}\lambda_1 + S_{42}\lambda_2 + S_{43}\lambda_3 + S_{44}\lambda_4 = d_4. \end{cases} \tag{2.4}$$

For the iris data, the solution is (-0.03,-0.18,0.22,0.31).

REFERENCES

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7** 179-188.