Stat 6300 Project Yue Zhang, Jing-Yi Wu

Introduction:

We want to investigate the whether we can distinguish 3 species of Iris (Iris Setosa, Iris Versicolour and Iris Virginica). If we can find something different lengths or width of sepal and petals, it will become easier to categorize these three species of flowers.

Data collection:

We found the data from this website: <https://archive.ics.uci.edu/ml/datasets/Iris>

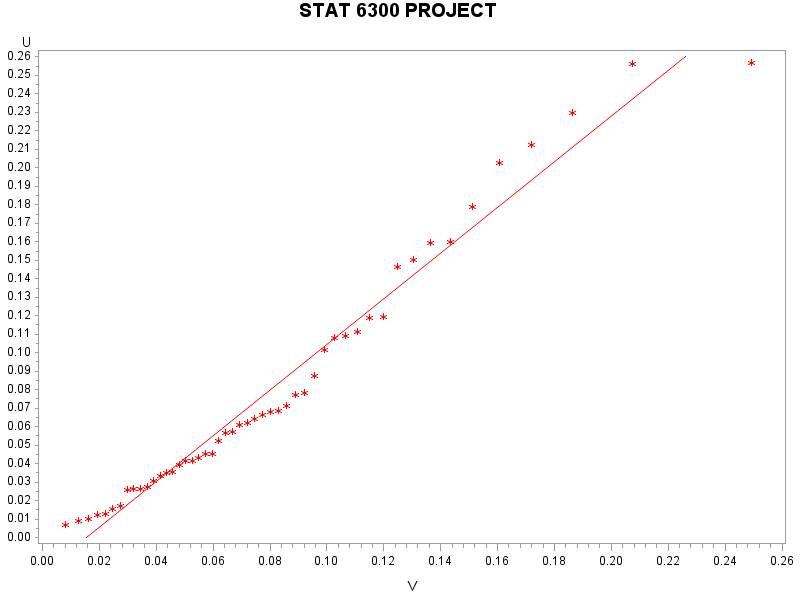
On each level of Iris flower, we have 50 instances. Four variables were measured from each sample: the length of sepals, the length of petals, the width of sepals and the width of petals, in centimeters. In summary, we have 150 observations with 4 variables and 3 groups.

Methods/results:

(1)Normality check (2) MANOVA

First of all, we need to check the multivariate normality for type1(Iris Setosa), type2(Iris Versicolour ) and type 3 (Iris Virginica)

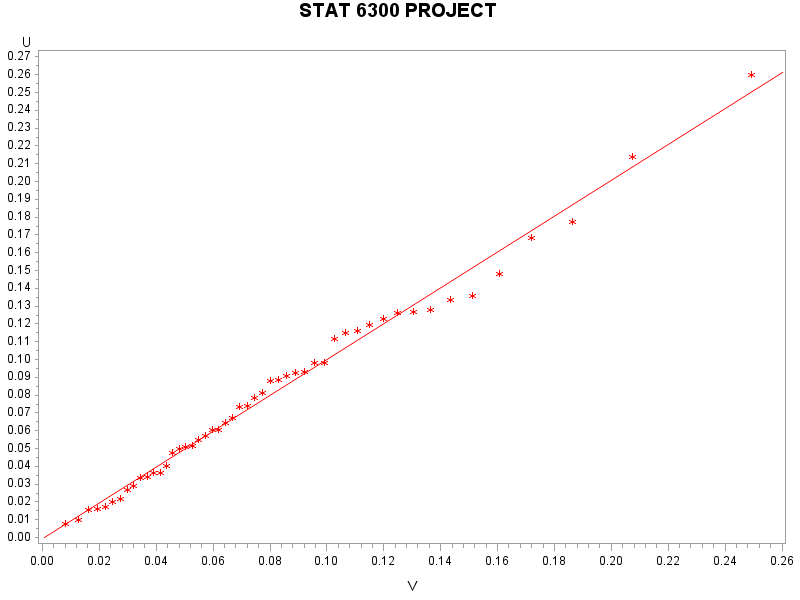
*Type 1(Iris Setosa):*



| **Pearson Correlation Coefficients, N = 50  Prob > |r| under H0: Rho=0** | | |
| --- | --- | --- |
|  | **U** | **V** |
| **U** | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.98679 | | <.0001 | |
| **V** | |  | | --- | | 0.98679 | | <.0001 | | |  | | --- | | 1.00000 | |  | |

For group 1, the plot seems normal.

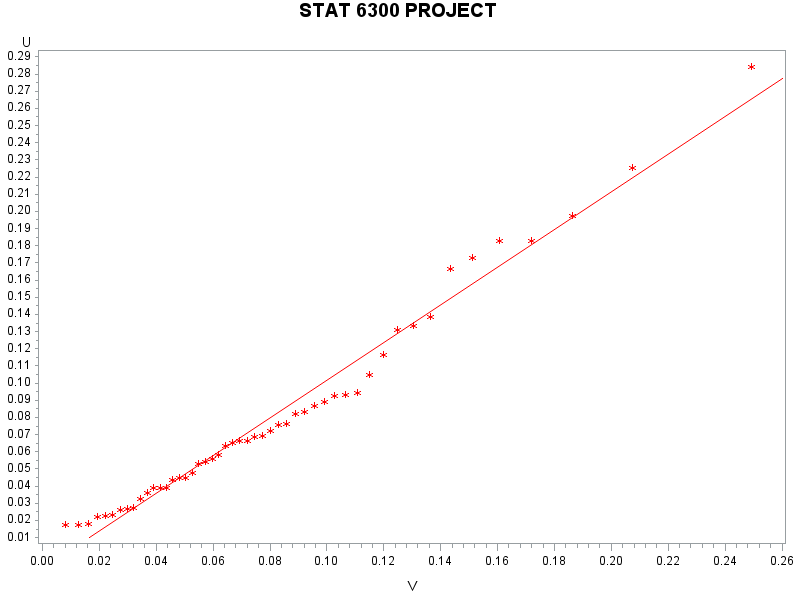
*Type 2 (Iris Versicolour):*



For group 2, the plot looks very normal.

| **Pearson Correlation Coefficients, N = 50  Prob > |r| under H0: Rho=0** | | |
| --- | --- | --- |
|  | **U** | **V** |
| **U** | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.99515 | | <.0001 | |
| **V** | |  | | --- | | 0.99515 | | <.0001 | | |  | | --- | | 1.00000 | |

*Type 3 (Iris Versicolour):*



For group3, the plot seems very normal

| **Pearson Correlation Coefficients, N = 50  Prob > |r| under H0: Rho=0** | | |
| --- | --- | --- |
|  | **U** | **V** |
| **U** | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.98962 | | <.0001 | |
| **V** | |  | | --- | | 0.98962 | | <.0001 | | |  | | --- | | 1.00000 | |  | |

We use one-way MANOVA :

| **Characteristic Roots and Vectors of: E Inverse \* H, where H = Type III SSCP Matrix for GRP E = Error SSCP Matrix** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Characteristic Root** | **Percent** | **Characteristic Vector V'EV=1** | | | |
| **Y1** | **Y2** | **Y3** | **Y4** |
| **32.1919292** | **99.12** | -0.06840592 | -0.12656121 | 0.18155288 | 0.23180286 |
| **0.2853910** | **0.88** | 0.00198791 | 0.17852670 | -0.07686357 | 0.23417227 |
| **0.0000000** | **0.00** | 0.10268742 | -0.19415509 | -0.22502879 | 0.37627520 |
| **0.0000000** | **0.00** | -0.24194505 | 0.10603485 | 0.10535376 | 0.00000000 |

| **MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall GRP Effect H = Type III SSCP Matrix for GRP E = Error SSCP Matrix  S=2 M=0.5 N=71** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Statistic** | **Value** | **F Value** | **Num DF** | **Den DF** | **Pr > F** |
| **Wilks' Lambda** | 0.02343863 | 199.15 | 8 | 288 | <.0001 |
| **Pillai's Trace** | 1.19189883 | 53.47 | 8 | 290 | <.0001 |
| **Hotelling-Lawley Trace** | 32.47732024 | 582.20 | 8 | 203.4 | <.0001 |
| **Roy's Greatest Root** | 32.19192920 | 1166.96 | 4 | 145 | <.0001 |

| **MANOVA Tests for the Hypothesis of No Overall GRP Effect H = Type III SSCP Matrix for GRP E = Error SSCP Matrix  S=2 M=0.5 N=71** | | |
| --- | --- | --- |
| **Statistic** | **Value** | **P-Value** |
| **Wilks' Lambda** | 0.02343863 | <.0001 |
| **Pillai's Trace** | 1.19189883 | <.0001 |
| **Hotelling-Lawley Trace** | 32.47732024 | <.0001 |
| **Roy's Greatest Root** | 32.19192920 | <.0001 |

Since the Eigenvalue is almost 100 %, we use Roy's test. F-value is 1166.96. P-value is <.0001 <0.05

We reject the null hypothesis. We conclude that sufficient evidence to indicate difference among variety at =0.05

Then we need to do the univariate test for 4 variables:

length of sepals, the length of petals, the width of sepals and the width of petals

(1) For Y1 length of sepals variable :

| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **GRP** | 2 | 63.21213333 | 31.60606667 | 119.26 | <.0001 |

T.S:F=119.26 P-value:<.0001 < 0.05

We reject null hypothesis. Sufficient evidence to indicate a difference in mean among 3 Types of Iris at α=0.05.

(2)For y2 length of petals variable:

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **GRP** | 2 | 11.34493333 | 5.67246667 | 49.16 | <.0001 |

T.S:F=49.16 P-value:<.0001 < 0.05

We reject null hypothesis. Sufficient evidence to indicate a difference in mean among 3 Types of Iris at α=0.05.

(3) For Y3 width of sepals variable :

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **GRP** | 2 | 437.1028000 | 218.5514000 | 1180.16 | <.0001 |

T.S:F=1180.16 P-value:<.0001 < 0.05

We reject null hypothesis. Sufficient evidence to indicate a difference in mean among 3 Types of Iris at α=0.05.

(4) For y4 width of petal variable :

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **GRP** | 2 | 80.41333333 | 40.20666667 | 960.01 | <.0001 |

T.S:F=960.01 P-value:<.0001 < 0.05

We reject null hypothesis. Sufficient evidence to indicate a difference among 3 Types of Iris at α=0.05.

We need to do the further analysis to know where the differences are.

We will use **Tuky’s multiple comparisons**

For y1 (length of sepals):

| **Means with the same letter are not significantly different.** | | | |
| --- | --- | --- | --- |
| **Tukey Grouping** | **Mean** | **N** | **GRP** |
| A | 6.5880 | 50 | 3 |
|  |  |  |  |
| B | 5.9360 | 50 | 2 |
|  |  |  |  |
| C | 5.0060 | 50 | 1 |

For y2(length of petals ):

| **Means with the same letter are not significantly different.** | | | |
| --- | --- | --- | --- |
| **Tukey Grouping** | **Mean** | **N** | **GRP** |
| A | 3.42800 | 50 | 1 |
|  |  |  |  |
| B | 2.97400 | 50 | 3 |
|  |  |  |  |
| C | 2.77000 | 50 | 2 |

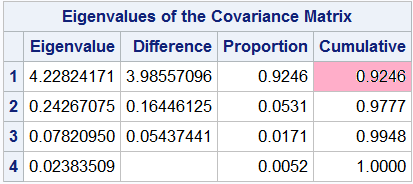
For y3 (width of sepals):

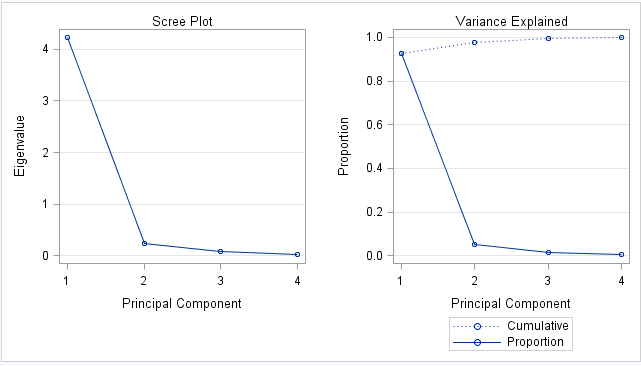
| **Means with the same letter are not significantly different.** | | | |
| --- | --- | --- | --- |
| **Tukey Grouping** | **Mean** | **N** | **GRP** |
| A | 5.55200 | 50 | 3 |
|  |  |  |  |
| B | 4.26000 | 50 | 2 |
|  |  |  |  |
| C | 1.46200 | 50 | 1 |

For y4(width of petals):

| **Means with the same letter are not significantly different.** | | | |
| --- | --- | --- | --- |
| **Tukey Grouping** | **Mean** | **N** | **GRP** |
| A | 2.02600 | 50 | 3 |
|  |  |  |  |
| B | 1.32600 | 50 | 2 |
|  |  |  |  |
| C | 0.24600 | 50 | 1 |

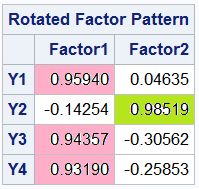
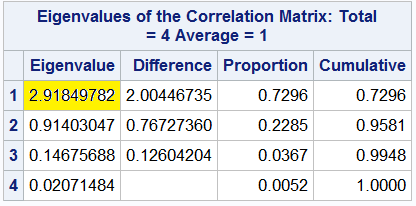
Principal component analysis:





We can see that the first eigenvalue takes up more than 80% out of total. Also, from the Scree plot, we keep all PC's to that left of the elbow. That would be only the first one.

Factor Analysis:



From the SAS output, we can find that there are only 1 eigenvalue greater than 1.

From rotated factor pattern, we put y1(length of sepals), y3(width of sepals) and y4(width of petals) into Facotor1, and we put y2(length of petals) into Factor2.



When we look at the screen plot, it's hard to see the elbow. However, according to the previous data. Only first eigenvalue is larger than1. So, we choose m as 1. We will use one factor.

Since we only have one factor now, we don't need to run MANOVA. We only need to check Factor1 to see where the difference is.Howver, we still need to check if factor one and factor 2 are independent. We can the correlation between factor1 and factor2 is zero, which means they are independent.

| **Pearson Correlation Coefficients, N = 150  Prob > |r| under H0: Rho=0** | | |
| --- | --- | --- |
|  | **Factor1** | **Factor2** |
| **Factor1** | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.00000 | | 1.0000 | |
| **Factor2** | |  | | --- | | 0.00000 | | 1.0000 | | |  | | --- | | 1.00000 | |  | |

For Factor 1:y1=length of sepals , y3= width of sepals and y4= width of petals

| **Means with the same letter are not significantly different.** | | | |
| --- | --- | --- | --- |
| **Tukey Grouping** | **Mean** | **N** | **GRP** |
| A | 1.04164 | 50 | 3 |
|  |  |  |  |
| B | 0.08379 | 50 | 2 |
|  |  |  |  |
| C | -1.12544 | 50 | 1 |

From the output, we can see the Type3 is significant higher than type2, and type2 is significant than type1.

(below no need! )

For factor 2: y1=length of petals.

| **Means with the same letter are not significantly different.** | | | |
| --- | --- | --- | --- |
| **Tukey Grouping** | **Mean** | **N** | **GRP** |
| A | 0.7132 | 50 | 1 |
|  |  |  |  |
| B | -0.0762 | 50 | 3 |
|  |  |  |  |
| C | -0.6370 | 50 | 2 |

We can see that type 1 is higher than type3, and type3 is higher than type2 for factor 2.