9705 FINAL PROJECT

By Michael Jagdharry, John Makhijani, Juna Iafelice, and Colin Brence

MAY 20, 2021

BARUCH COLLEGE

Table of Contents

REPORTS

Hotelling T2 Test

In Major League Baseball (MLB), a major change was implemented in 1973 when the American League (AL) voted to implement a designated hitter (DH). Normally, all nine players of a team are required to bat in order. The change here is the DH, a player who bats in place of the pitcher anytime the pitcher was supposed to hit. The DH does not play defense, hence the term designated hitter. Half the MLB teams are in the AL and the other half are in the National League (NL). Interestingly, the NL has not implemented a DH, instead claiming they would continue to play the game the way it was supposed to be played with the pitcher hitting when it was their turn to bat. Since the inception of baseball until today, the debate rages on as to whether each league should play with or without a DH. What we will explore is if, statistically, there is an advantage to the players on a team to have a DH. To do so we will compare certain player hitting statistics of AL and NL players.

For this analysis, we used the hitters dataset. The dataset is from Carnegie Mellon University and contains hitting statistics from the 1986 season and salary information for 1987. This dataset is also used to predict player salaries based on a variety of statistics. For the purpose of this analysis we will only be using the hitting statistics: hits, home runs, runs and walks.

Since the AL and NL are playing the same game only with the one difference of having a DH and not having a DH, we will divide the statistics of the hitters dataset into 2 groups, the NL and AL. Group 1 in the dataset represents AL while group 2 represent NL. We will compare two mean vectors with a 2 sample Hotelling T2 test using SAS. We will be testing the null hypothesis that the hitting statistics of the AL are equal to the hitting statistics of the NL. If we reject the null, we will know that having the DH makes a difference.

To explain the technique, the Hotelling T2 test is a multivariate analog of the univariate two sample t-test, but for p pairs of variables instead of 1 pair. Let ni denote the number of observations for the ith sample, i $\epsilon\{1,2\}$. We assume the ni observations of p variables are independently sampled and come from multivariate normal distributions with common covariance matrix. Specifically, sample 1 and 2 are distributed as $N_p(\boldsymbol{\mu_1}, \boldsymbol{\Sigma})$ and $N_p(\boldsymbol{\mu_2}, \boldsymbol{\Sigma})$ respectively, in contrast to being $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, as in the univariate case. Since we will obtain different sample covariance matrices S1 and S2 for the two samples, we will estimate the common covariance matrix by the pooled covariance matrix. This is defined as the average of the two sample covariance matrices weighted by their degrees of freedom, that is $\boldsymbol{S_p} = \frac{\boldsymbol{S_1}(n_1-1)+\boldsymbol{S_2}(n_2-1)}{n_1+n_2-2}$. We wish to test the null hypothesis H_0: $\boldsymbol{\mu_1} = \boldsymbol{\mu_2}$, against the alternative hypothesis H_1: $\boldsymbol{\mu_1} \neq \boldsymbol{\mu_2}$. To be clear about notation, $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ are px1 vectors containing the means for each of the p variables in the two samples. We could naively of course perform p univariate two sample t tests. But the disadvantage here is that it does not control for what is called the family-wise error rate. This is the probability of rejecting at least one of the null hypotheses H_{0j}: $\mu_{1j} = \mu_{2j}$, for $1 \leq j \leq p$. To illustrate this, suppose the p variables in our multivariate normal distributions are independent. Then the family wise error rate can be calculated as $1 - (1 - \alpha)^p$. For p > 1, this value is always greater than α, and increases with the number of variables p. Performing a single multivariate test eliminates this issue, keeping the family-wise error rate at α.

Before proceeding with the analysis, we would like to check the truth of our assumptions. Testing for multivariate normality is quite difficult. However, a consequence of multivariate normality is that each variable is univariate normal. This is something we can check using QQPlots, which are included in the Appendix. We see that Hits, Runs, and Walks are mostly on the QQLine (Fig 1-1, 1-3, 1-4), so we can conclude these populations are normal. This does not appear to be true for AL HmRuns (Fig 1-2).

To check the assumption of equal covariance matrices, we can use Bartlett's Test for Homogeneity of Covariance Matrices (Fig 1-5). The null hypothesis for this test is that $\Sigma_1 = \Sigma_2$, with

alternative hypothesis $H_A: \Sigma_1 \neq \Sigma_2$. We obtain a p=value of .0017 under the null hypothesis, providing sufficient evidence for us to reject it. Therefore, the assumption of equal covariance matrices is also unsupported. In conjunction with the previous findings that not all of our variables are normally distributed, we have sufficient evidence to conclude that our data is not multivariate normal with common covariance matrix. But assuming it were, the following is how the analysis would proceed.

Per SAS the Hotelling T2 statistic is 17.66 (Fig 1-6). Since $T^2 > T^2_{.05}(4, 320) = 9.817$, we reject H_0 and conclude that $\boldsymbol{\mu_1 \neq \mu_2}$. Therefore, with the mean hitting statistics of the AL not equal to that of the NL, we can conclude that the designated hitter does make a difference. Common sense tells us, that the hitting statistics of players in the AL are better than that of the players NL because a DH is a better hitter than a pitcher so players in the AL will get better opportunities to hit more often and with better opportunities like hitting with runners on base.

MANOVA

We will perform MANOVA on the diamonds dataset. This is a dataset that consists of 8 variables: Cut, Carat, Depth, Table, Price, X, Y, and Z. Let us conceive of the diamond as a mathematically geometric object. The *cut* of the diamond describes the arrangement of the faces (flat surfaces) of this object. These faces are a consequence of how the diamond is literally cut. Now the particular arrangement of faces (the cut) alters how light entering the diamond refracts through it, thereby altering the diamond's luminosity (how shiny it is). As such, there exist five categories for cut in our dataset: 1=Fair, 2=Good, 3=Very Good, 4=Premium, 5=Ideal. This produces five populations of diamonds in our dataset. The different cut qualities of diamonds can appear almost indistinguishable in the store to an untrained eye. We wish to determine if we can statistically classify the differences between these cuts.

Here are brief descriptions of the remaining variables. A *carat* is a unit of mass equal to 200 milligrams, and thus measures the weight of the diamond, with range 0.2 to 5.01. The *depth* of a diamond is its height (in millimeters) measured from the culet, the bottom tip, to the table, which is the flat, top surface. Define the total width of the diamond to be the widest measurable width. Then the *table* variable in our dataset is the ratio of the width of the diamond's table compared to the total width of the diamond, taken as a percentage, with range 43 to 95. The *price* is self-explanatory, with range $326 to $18,823. The *x, y*, and *z* variables measure the length, width, and depth in millimeters respectively, having ranges 0 to 10.74, 0 to 58.9, and 0 to 8.06.

MANOVA is a multivariate version of ANOVA, so to explain MANOVA let us first review ANOVA. The purpose of ANOVA, or Analysis of Variance, is to check whether observations in different groups come from the same population. Here we assume that each such population is normally distributed with their own population mean, but a common variance. We also assume these distributions are independent. When we move to MANOVA, these assumptions take on their multivariate form. Namely, we now have p-dimensional observation vectors, and we assume each multivariate observation comes from one of k multivariate normal populations. We assume these populations have their own population mean vector but a common covariance matrix. And like the univariate case, we assume the populations are again independent. We desire to test the (null) hypothesis that the mean vectors for each of the k populations are all equal. There are four tests to do so: Wilk's Lambda, Lawley-Hotelling, Pillai's Trace, and Roy's Greatest Root. We conservatively decide to reject the null hypothesis if all four tests fail. That is, if their test statistics exceed their critical values, or equivalently if their p-values are lower than the decided significance (alpha) level. We could of course naively do a series of k ANOVA tests. In doing so however, the joint probability of rejecting a true null hypothesis increases with each additional test. Therefore, by performing one MANOVA test, this error rate will be equal to the specified significance level (and thereby minizmied).

Before proceeding with the analysis, we will first check if the assumptions for MANOVA hold. Since it is difficult to assess multivariate normality, we will instead check for univariate normality amongst each variable. We can do this by evaluating visually quantile-quantile plots of the variables, which plot quantiles of the sample data against quantiles of the normal data. The idea is that there should be a linear relationship between these quantiles if the data is truly normal, with the ideal relationship shown on the graph as the QQLine. If many points end up on this line, we have good evidence for univariate normality. The second assumption we will check is that of equal covariance matrices. This can be carried out using a Chi-Square test known as Bartlett's Test, with null hypothesis H_0: $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$, and alternative hypothesis H_A: at least one inequality in H_0.

The QQPlots for depth, table, price, x, and y, corresponding to Figures 2-2, 2-3, 2-4, 2-5, and 2-6 respectively, do not illustrate very serious deviations from the QQLine. However, carat and z in Figures 2-1 and 2-7 respectively appear to deviate significantly from the QQLine, and therefore we conclude these variables are not normal. Moving on to check homogeneity of covariance matrices, Bartlett's Test outputs a p-value of <.0001 in Figure 2-8, which is very

strong evidence to conclude the matrices are heterogenous. We therefore have good reason to believe that our k groups are not multivariate normal nor do they have common covariance matrices. Nonetheless, we will proceed with the analysis as if the MANOVA premises are true.

In the following, we delineate the four MANOVA tests, the null and alternative hypothesis, their decision rules, whether we performed any interpolation of the table values (such as for values of s or m which did not appear on the tables), and whether tests are conservative due to for example using a smaller degrees of freedom from the table than we actually had in the data. SAS output is shown in Figure 2-9.

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$
H_A: At least one inequality in H_0

Wilks' Test: Reject H_0 if $\Lambda = 0.367 < \Lambda_\alpha(p, \nu H, \nu E) = \Lambda_{0.05}(7, 4, 2860) = 0.959$. Hence, we reject H_0.

Lawley-Hotelling Test
Reject H_0 if $\left(\frac{v_E}{v_H}\right) U^{(s)} > U_{.05}^{(s)}$. If $p > v_H$, we use $(v_H;\ p;\ v_E + v_H - p)$ for $(p;\ v_H;\ v_E)$, which is our case. Therefore we have our table value as $U_{.05}^{(s)} = U_{.05}^{(s)}$ (4,7,2857). $\left(\frac{v_E}{v_H}\right) U^{(s)} > U_{.05}^{(s)}$ (4,7,2857).
Since 7 is not on the table for v_H, we will interpolate between 6 and 8. This yields
$\left(\frac{v_E}{v_H}\right) U^{(s)} = (2857/7)(1.29) = 514.26 > 5.921 = U_{.05}^{(s)}$. Therefore we reject H_0.

Pillai Test
Reject H_0 if $V^{(s)} > V_\alpha{}^{(s)}(s.m, N)$ This yields $0.79 > V_{.05}(4,1,280) = 0.761$. 280 does not show up as a value for N on the table, so we use the closest number which is N=20. Observing that the table value is inversely proportional to N, the true N=280 which would yield a much smaller Pillai critical value. Using N=20 therefore expands the rejection region from what it should truly be, thereby yielding a conservative test. Since this conservate form of the test results in rejecting H_0, we have even greater confidence about doing so for the current case where N=280.

Roy's Test
Reject H_0 if $\theta > \theta_\alpha(s, m, N)$. $\theta = {}^{0.82}/_{1.82} = 0.45$. $\theta_{.05}(4,1,280) = 0.047$.
280 does not show up as a value for N on the table, so we use the closest number which is N=240.
Since 0.45 > 0.047, we reject the null. By the same reasoning as the above for Pillai's Test, this yields again a conversative estimate which us greater confidence about rejecting H_0.

All four MANOVA tests yield the same conclusion. Therefore, we can safely conclude that the mean vectors for the five cut populations across the 7 variables (carat, depth, table, price, x, y, z) are different. That is, while perhaps not perceivable to the untrained eye, there exist statistically significant differences among the various cuts of diamonds.

Discriminant and Classification Analysis

We chosen to perform Linear Discriminant Analysis on the Iris dataset. The iris dataset contains fifty observations from three species of Iris, Setosa, Versicolor and Virginica. Each is a multidimensional observation consisting of four variables, length and width of the iris' sepals, and length and width of the iris' petals.

Linear/Quadratic Discriminant Analysis aims to classify observations into known groups. In this case we have the 150 iris observations. This technique attempts to answer the question, supposing we have a new observation belonging to an unknown group, how do we determine this unknown group? That is, given that we have prior information about the correct classification of many previous observations, how should we use this knowledge to inform our new classification?

For linear discriminant analysis, we will do this by first supposing each group in our data come from a multivariate normal population with common covariance matrix. We will then estimate the population mean vector for each group as that group's sample mean vector, and the common covariance matrix as the pooled covariance matrix across all groups. For a given observation vector, we finally classify it as belonging to the group which gives the smallest Mahalanobis distance to that group's mean vector. If instead the covariance matrices are not all equal, we perform Quadratic Discriminant Analysis, in which the Mahalanobis distance to the ith group now uses the ith group's covariance matrix instead of the pooled covariance matrix.

A common method for evaluating the performance of this classification technique is to perform leave one out cross validation. This is a method in which we train the model on all but one observation, and then test it on that left out observation, calculating a squared error. We then leave out a different observation and repeat this process for all observations. We can then average the squared errors to get an estimate of the test mean squared error. However, it would be even better to split the data into a training and test set, then train the model on the former and evaluate it on the latter. Since the test data are unseen to the model, this provides a more objective measure of the test error, and so we employ this method in addition to resubstitution and cross validation.

A test set was created from the Iris using stratified random sampling to extract 5 observations randomly from each of the three groups of flowers, Setosa, Versicolor and Virginica. The remaining observations constitute the training set and are used to construct the classification model. Since there are an equal number of observations per group, we will assume equal priors (that is, equal probability of belonging to a group, without any other information). Bartlett's test for homogenous covariance matrices provides strong evidence that the within group covariance matrices are unequal (Figure 3-1). Therefore, we should proceed with Quadratic Discriminant Analysis. From the "Generalized Squared Distance to species" table (Figure 3-2), we see that the distances between means for groups 2 and 3 are very close at about 5.55, compared to the distances between groups 1&2 and 1&3 which are 106.21 and 149.89 respectively. If we are judging these flowers purely by their sepal and petal dimensions, these generalized squared distances imply that the Setosa species is most easily distinguished from the Versicolor and Virginica species, since its mean vector has the farthest distances from these two. In contrast, the Versicolor and Virginica species are the two least distinguishable species, as their mean vectors are quite close to each other. This foreshadows that any misclassifications will most likely be between Versicolors and Virginicas. Looking at the classification output, this is exactly what each of the resubstitution, cross validation, and test tables show (Figures 3-3, 3-4, 3-5). The resubstitution table reports only one error, a Virginica misclassified as a Versicolor, yielding an overall error rate of only .74%. But we will not heed the resubstitution table as it is overly optimistic. The cross validation confusion table shows two errors: one Versicolor are misclassified a Virginica, and vice versa. This yields an overall cross validation error rate of 1.48%, and therefore an accuracy of 98.52%, which is very good. However, when we apply the model to the 15 unseen test data points, we get two Versicolors misclassified as Virginicas, yielding a test error rate of 13.33%.

Let us instead take a nonparametric approach, relaxing the assumption that the data from each group come from a multivariate normal distribution with its own group mean vector. That is, let us take

the k nearest neighbors approach, in which we must specify some neighborhood around each data to judge by majority vote its group membership. Doing so with k=5, k=10, or k=15, we see that each of these choices of flexibility yield the same cross validation error rate of .74% and test error rates of 13.33% (Figures 3-6, 3-7, 3-8). Therefore, in any of these three cases, the accuracy is at least as great as the parametric approach. In addition, for each of these test errors, it is always 2 Versicolors that are misclassified as Virginicas, as was the case for the parametric approach. This suggests that these two misclassifications may be the same two pesky observations throughout all tests.

As a further analysis we can examine the discriminant functions $z_i = \boldsymbol{a}_i^T \boldsymbol{y}$ that have the best separation amongst each other. For multivariate data, the coefficients $\boldsymbol{a_i}$ turn out to be the eigenvectors of the matrix $E^{-1}H$ as it was defined for MANOVA. We will obtain $rank(H) = \min(k-1, p) = \min(3-1, 4) = 2$ such eigenvectors. The relative importance of these eigenvectors is determined by their corresponding eigenvalue proportions, that is $\lambda_i / \sum_j \lambda_j$. From Figure 3-9, we see that the first eigenvalue takes up 98.89% of the total proportion, so a single discriminant function sufficiently describes species separation. The choice is now to select between the standardized or raw eigenvectors; we could pick the former if the variables are commensurate, that is measured on the same scale and with comparable variances. The variables are indeed measured on the same scale, as they have the same units. However, the variances are not commensurate – for Sepal Length, Sepal Width, Petal Length, and Petal Width they are respectively 0.6856935, .1899794, 3.1162779, and 0.5810063. We observe that Petal Length has a dominating variance, for example being approximately 16 times the variance of Sepal Width. Therefore, we will use the standardized coefficients. SAS outputs these as the Pooled Within-Class Standardized Canonical Coefficients (See Figure 3-10), which simplifies to $\boldsymbol{a}^* = diag\left(\boldsymbol{S}_{pl}\right)^{1/2} \boldsymbol{a}$, in accordance with Eq (8.17) of Rencher. We see that the variable Petal Length has the highest coefficient magnitude for the first discriminant function (Figure 3-11). Therefore, it contributes most to the separation of the groups.

Principal Component Analysis

Our dataset describes Pima Indian women who have diabetes. Pima Indians are a group of Native Americans living in central and southern Arizona (as well as northwestern Mexico). The Pima Indians have the highest rate of type 2 diabetes in the world. It is hypothesized this is so due to governmental influences having dramatically changed their way of life and their environment. Specifically, the traditional farming economy has been significantly reduced, and their consumption of processed foods has increased. They have in addition become less physically active. Being very similar to each other genetically as well as experienced the same shift in environment, this minimizes their external variable influence, making them ideal for scientific investigation. As such, these Indians have frequently been used in studies of diabetes.

The predictors for this dataset are: number of pregnancies, blood glucose levels, blood pressure, skin thickness, insulin, BMI, Diabetes Pedigree Function (this represents how likely a given subject is to get Diabetes Melitus by extrapolating from their relatives' history), and Age. We would like to determine whether this data can be simplified into fewer dimensions, thereby also simplifying the analyses for any interested research groups. But how can it be possible to do this?

Suppose our data has p variables, and we have n p-dimensional observations of these variables. We can conceive of these observations as a cloud of points in p-dimensional space R^p. It may be possible to project these points onto a lower dimensional subspace R^k for some $k < p$ while maximally preserving information about the variation amongst these points. That is, if $k < p$ transformed variables (z_1, z_2, ... , z_k) can describe the variation in our data almost as well as p variables, we can opt for these transformed variables.

There are p projections onto any the subspaces R^1, R^2, ... , up to R^p, and to remain general suppose we picked the k^{th} such subspace R^k. Now suppose that we denote the ith largest eigenvalue of the covariance matrix as λ_i, and the corresponding eigenvector as e_i. Then it turns out that the k-dimensional subspace that when projected onto attains maximal separation amongst the projected points in fact has as its basis vectors the eigenvectors e_1, e_2, ... up to e_k. This is quite remarkable, as we have performed a change of basis from the standard Euclidean basis vectors to this new eigenvector set which reduces dimensionality while preserving as much information about variability as possible. This procedure is called Principal Component Analysis (PCA).

We would like to perform PCA on this dataset to determine whether its dimensionality can be reduced for simpler analysis. The variables in this data have very different units and scales - when this is the case, a potential issue can arise. To illustrate, suppose one set of observations is measured on a different scale than a second set. For example, consider the measurements {1,2,3} vs. {1000,2000,3000}. In fact, these could be the same measurements, where the first is measured in meters, and the second in millimeters. But although the observations are identical, the latter sample has higher variance than the first! This is how differences in scale can misconstrue the relative degrees of variability between different variables. Standardizing these samples will eliminate this problem. We can do this by using the correlation matrix instead of the covariance matrix. In this way, variances for the principal components can be fairly assessed.

Doing so, the principal components obtained are shown in Figure 4-1. Figure 4-2 contains information about the proportions of variance explained.

A natural question is, how many principal components should we retain? There are three commonly used methods to make this judgement. The first is to define a threshold cumulative proportion of variance explained by the principal components, say 80%, and take the smallest number of components which satisfies or exceeds this threshold. The second method is to take the number of components whose eigenvalue is greater than the average eigenvalue across all principal components. For the correlation matrix, the sum of the eigenvalues is number of variables p, so the average is $p/p = 1$, and hence we would keep any principal components whose eigenvalue is greater than 1. The third is to use a scree plot, which is a line graph with proportion of variance explained on the y-axis and principal component number on the x-axis. This will be a decreasing function of the principal component number. The idea is

to visually assess if there exists a significant drop-off and where it ends on the graph. The name comes from the image of a steep mountain meeting its more gently sloped base, where one would find scree. So we would take the number of components which constitute the steep part of the plot.

We would need to take 5 principal components to reach an 80% cumulative proportion of variance explained (Figure 4-2). In contrast, the eigenvalues for the first three principal components here are greater than 1. The scree plot has a sharp fall after the second component and another smaller fall after the sixth, so there is not a clear visual break in the graph here. To simplify matters we will use the second method and retain three components, Prin1, Prin2, and Prin3. This will explain 60% of the total variance.

The coefficients of the components inform us of the variables these components represent (Figure 4-1). For example, we see that Prin2's coefficients are highest for pregnancies and age - therefore we can say Prin2 represents these variables. Prin1 does not have one or two dominating coefficients and is more homogenous. It represents Glucose, SkinThickness, Insulin, and BMI. Prin3 coefficients are highest in Blood Pressure and DiabetesPedigreeFunction, and therefore represent these variables. Using these components to summarize our variables, we have reduced the dimensionality of our 8-dimensional dataset to a 3-dimensional dataset while retaining 60% of the variance explained. Hopefully this simplified dataset can aid researchers in their understanding of diabetes as it prevails in the Pima Indian community.

APPENDIX

Hotelling T^2 Test Figures

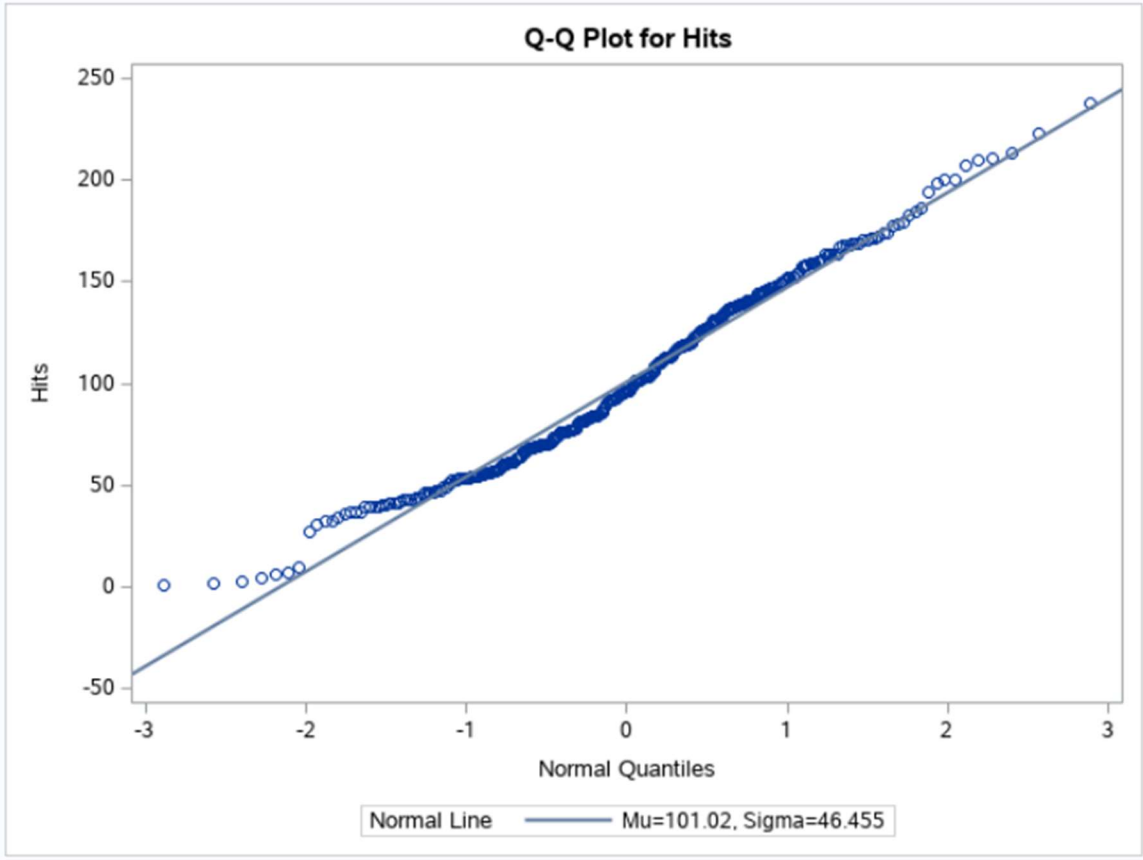

Figure 1-1

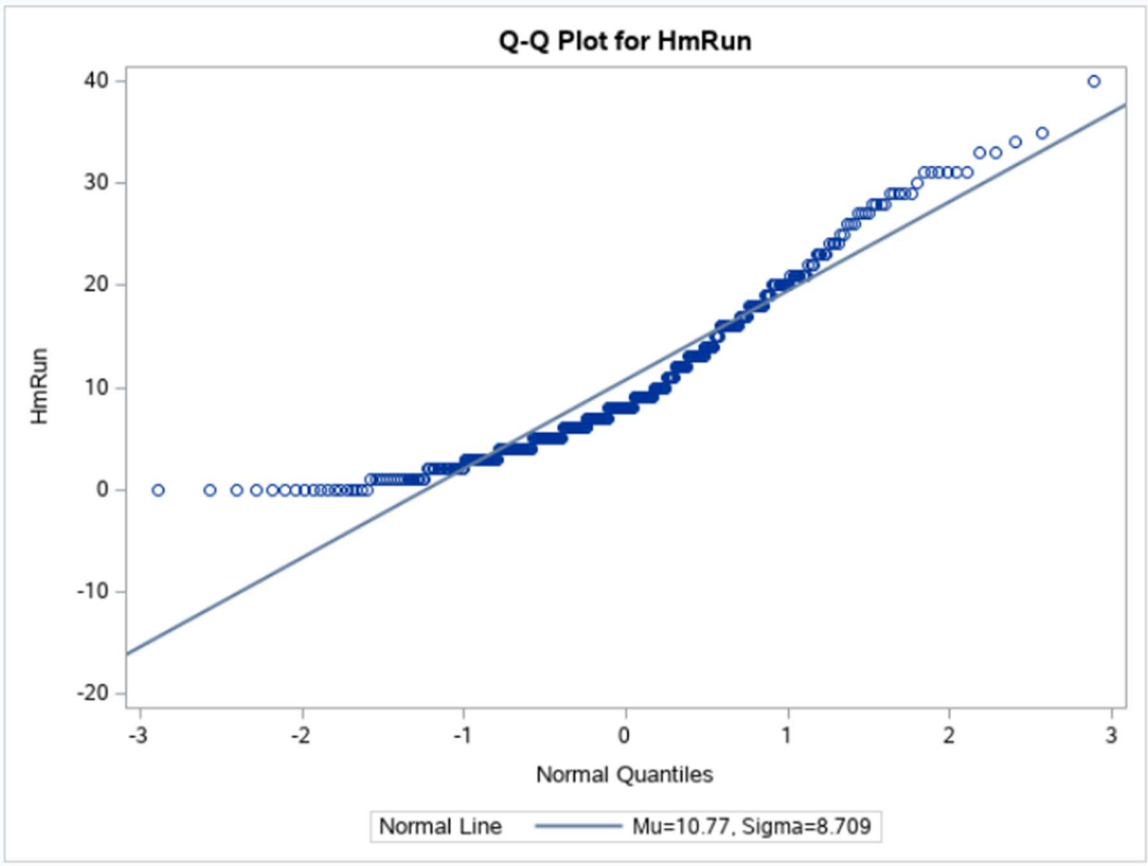
Q-Q Plot for HmRun
HmRun
40
30
20
10
0
-10
-20
-3
-2
-1
0
1
2
3
Normal Quantiles
Normal Line
Mu=10.77, Sigma=8.709

Figure 1-2

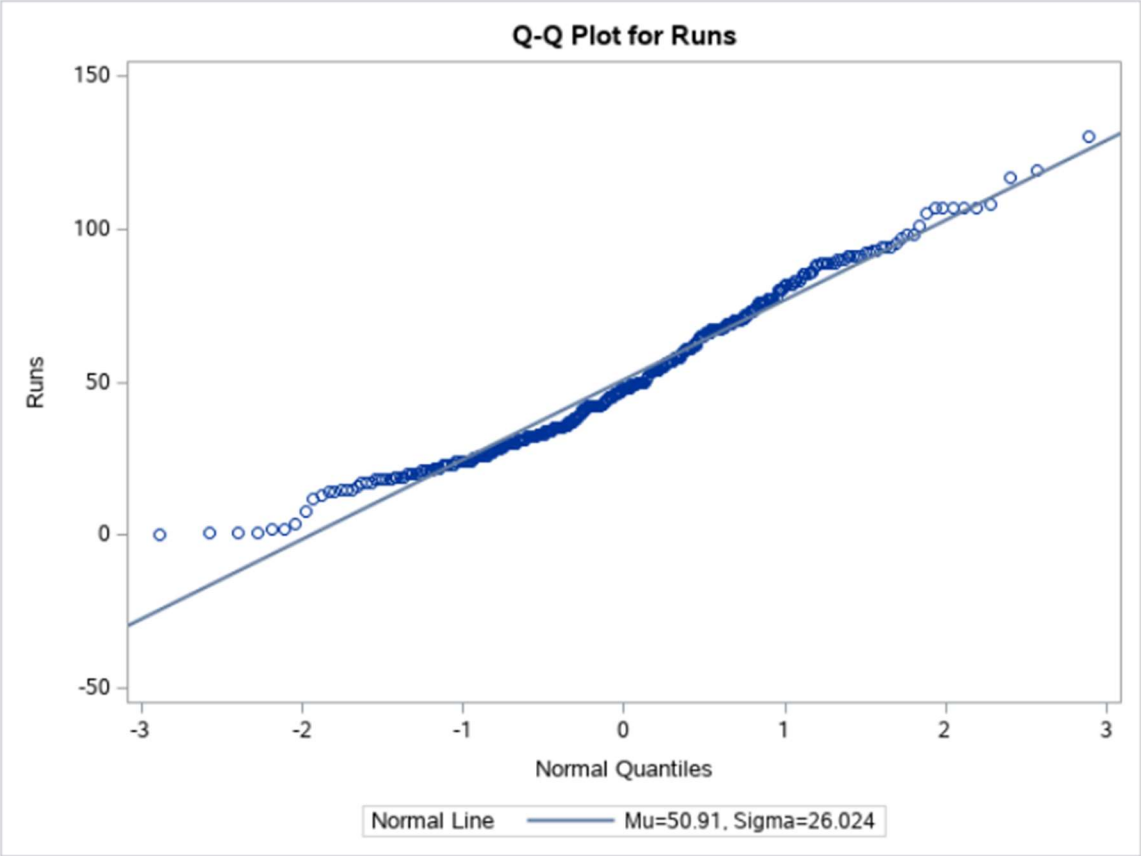
Q-Q Plot for Runs
150
100
50
0
-50
Runs
-3
-2
-1
0
1
2
3
Normal Quantiles
Normal Line
Mu=50.91, Sigma=26.024

Figure 1-3

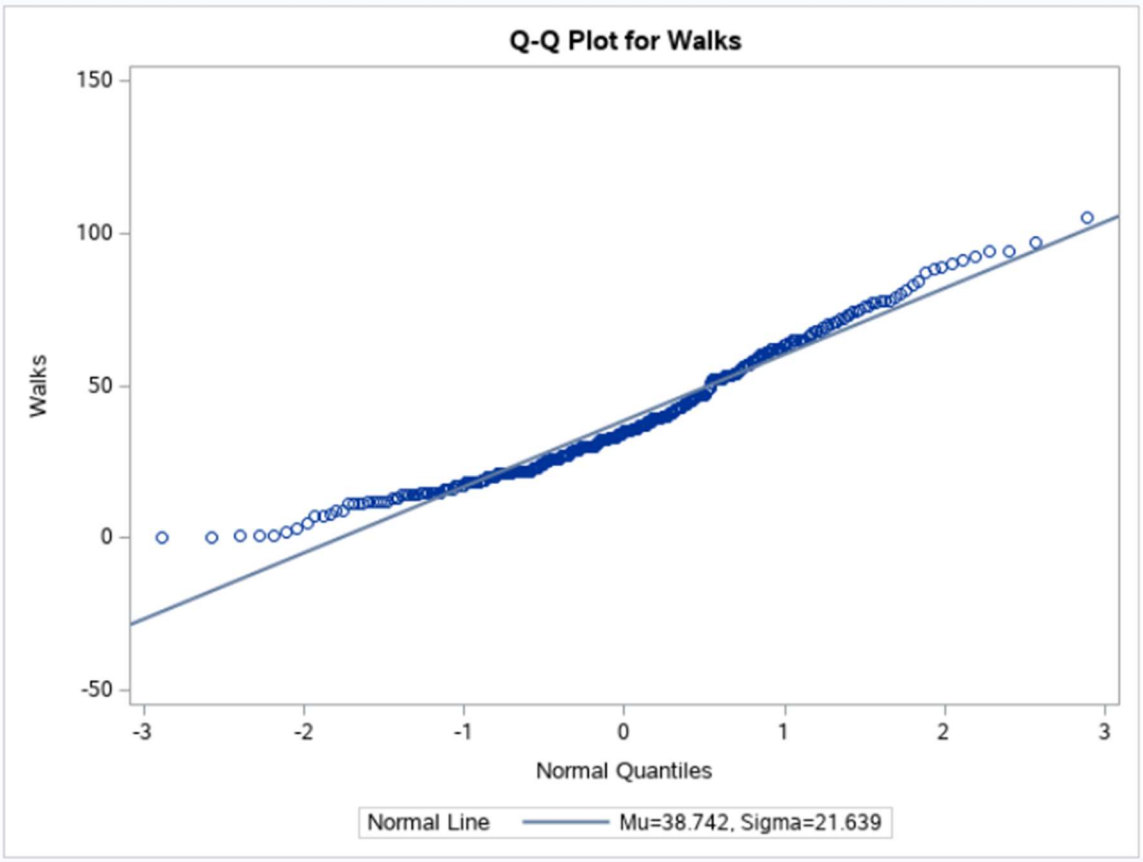

Figure 1-4

Hotelling T2 Test on Hitters

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
28.115852	10	0.0017

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

Figure 1-5

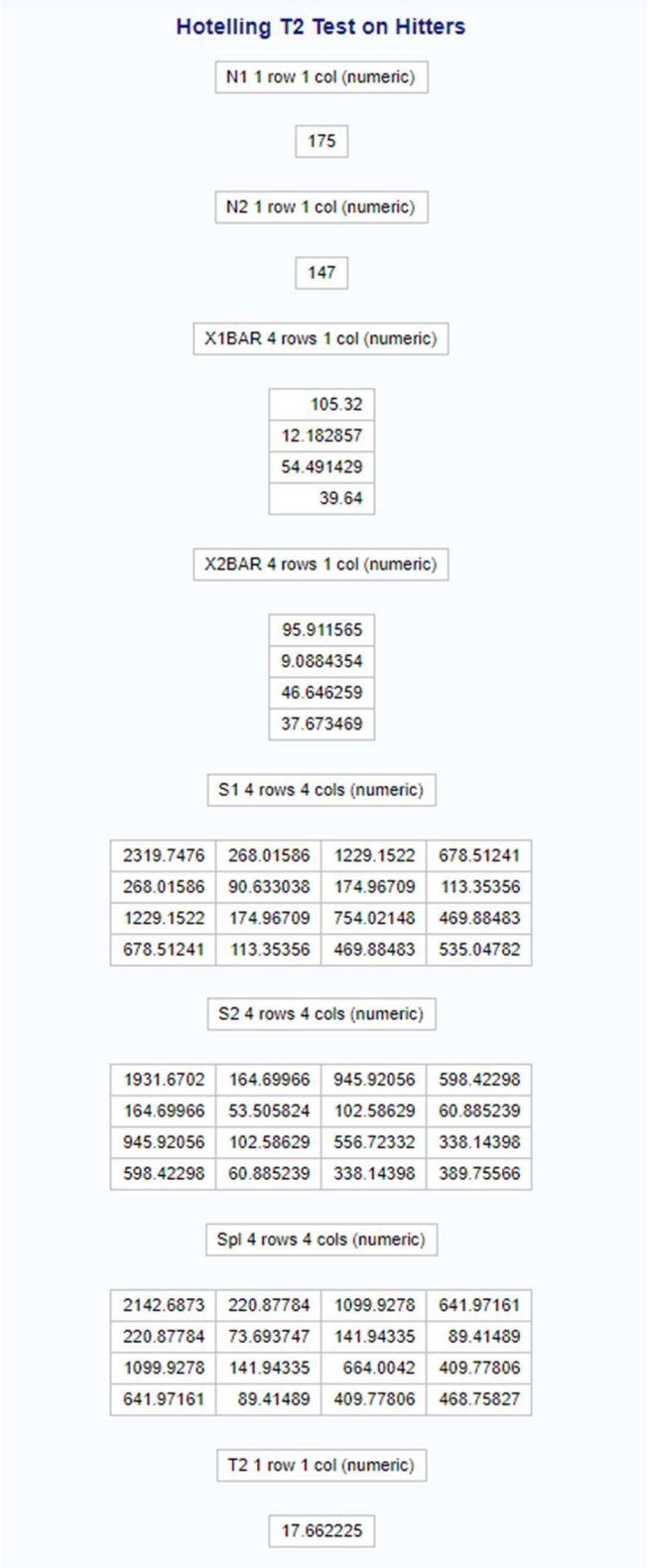
Hotelling T2 Test on Hitters
N1 1 row 1 col (numeric)
175
N2 1 row 1 col (numeric)
147
X1BAR 4 rows 1 col (numeric)
105.32
12.182857
54.491429
39.64
X2BAR 4 rows 1 col (numeric)
95.911565
9.0884354
46.646259
37.673469
S1 4 rows 4 cols (numeric)
2319.7476 268.01586 1229.1522 678.51241
268.01586 90.633038 174.96709 113.35356
1229.1522 174.96709 754.02148 469.88483
678.51241 113.35356 469.88483 535.04782
S2 4 rows 4 cols (numeric)
1931.6702 164.69966 945.92056 598.42298
164.69966 53.505824 102.58629 60.885239
945.92056 102.58629 556.72332 338.14398
598.42298 60.885239 338.14398 389.75566
Spl 4 rows 4 cols (numeric)
2142.6873 220.87784 1099.9278 641.97161
220.87784 73.693747 141.94335 89.41489
1099.9278 141.94335 664.0042 409.77806
641.97161 89.41489 409.77806 468.75827
T2 1 row 1 col (numeric)
17.662225

Figure 1-6

MANOVA Figures

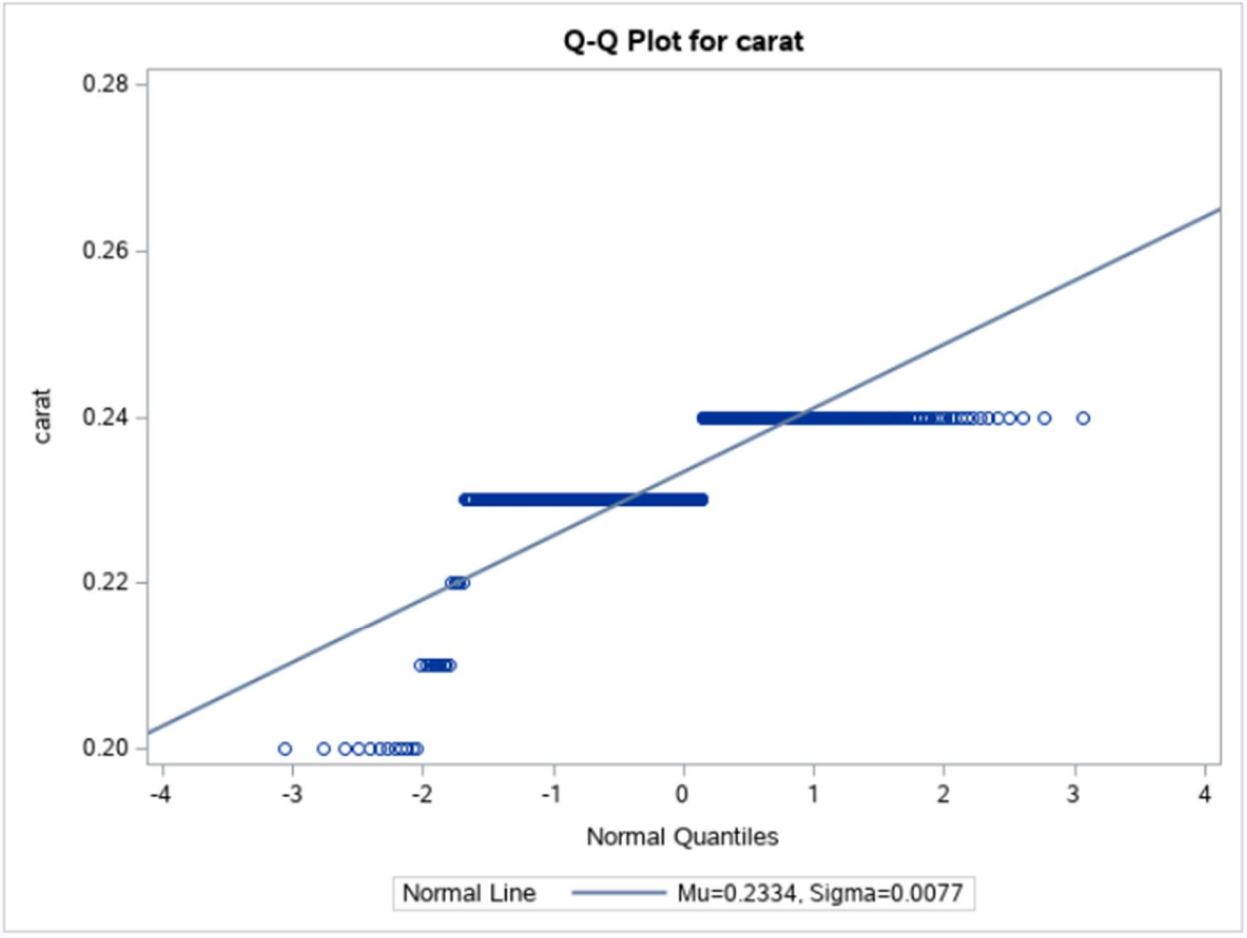

Figure 2-1

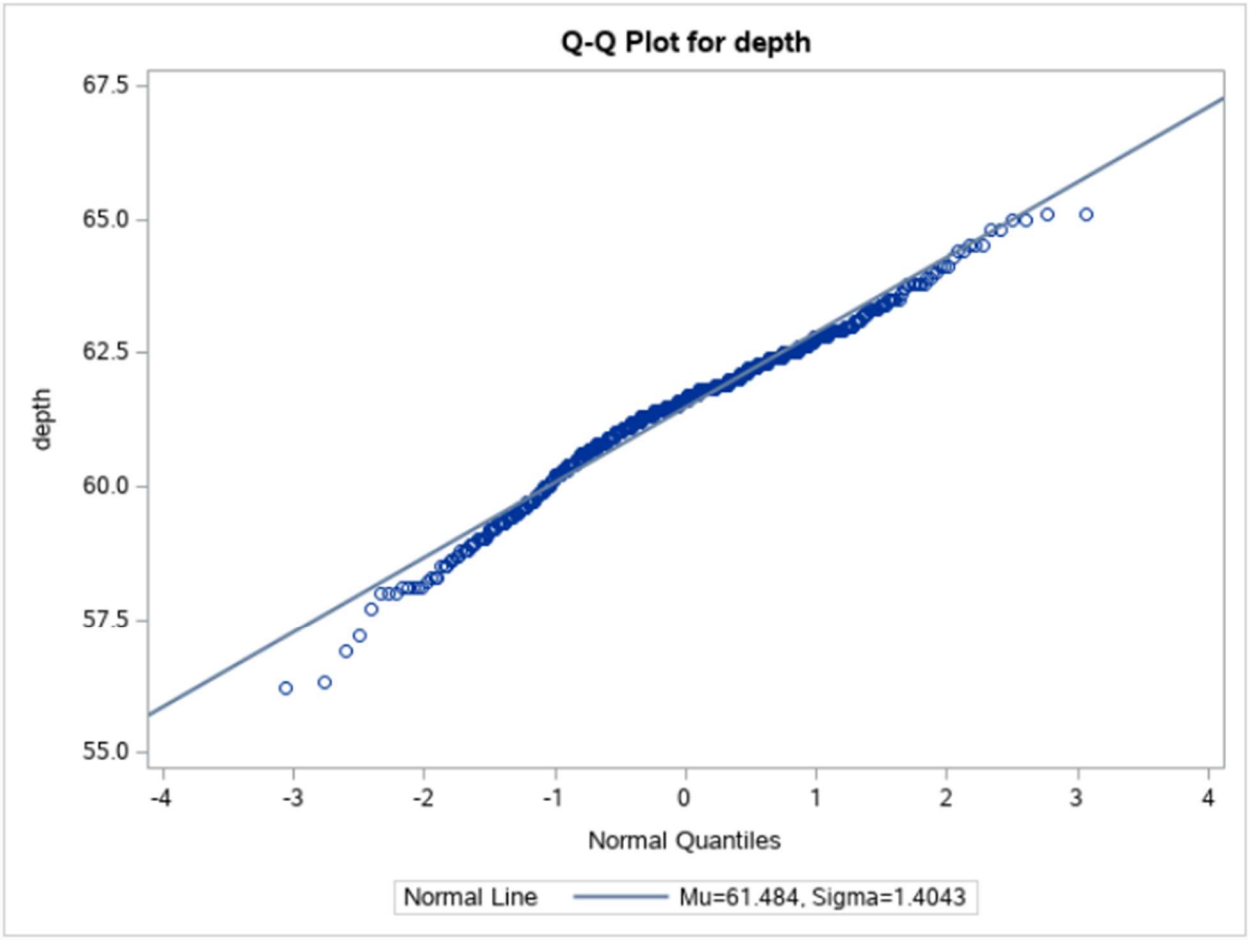
Q-Q Plot for depth
67.5
65.0
62.5
60.0
57.5
55.0
depth
-4
-3
-2
-1
0
1
2
3
4
Normal Quantiles
Normal Line
Mu=61.484, Sigma=1.4043

Figure 2-2

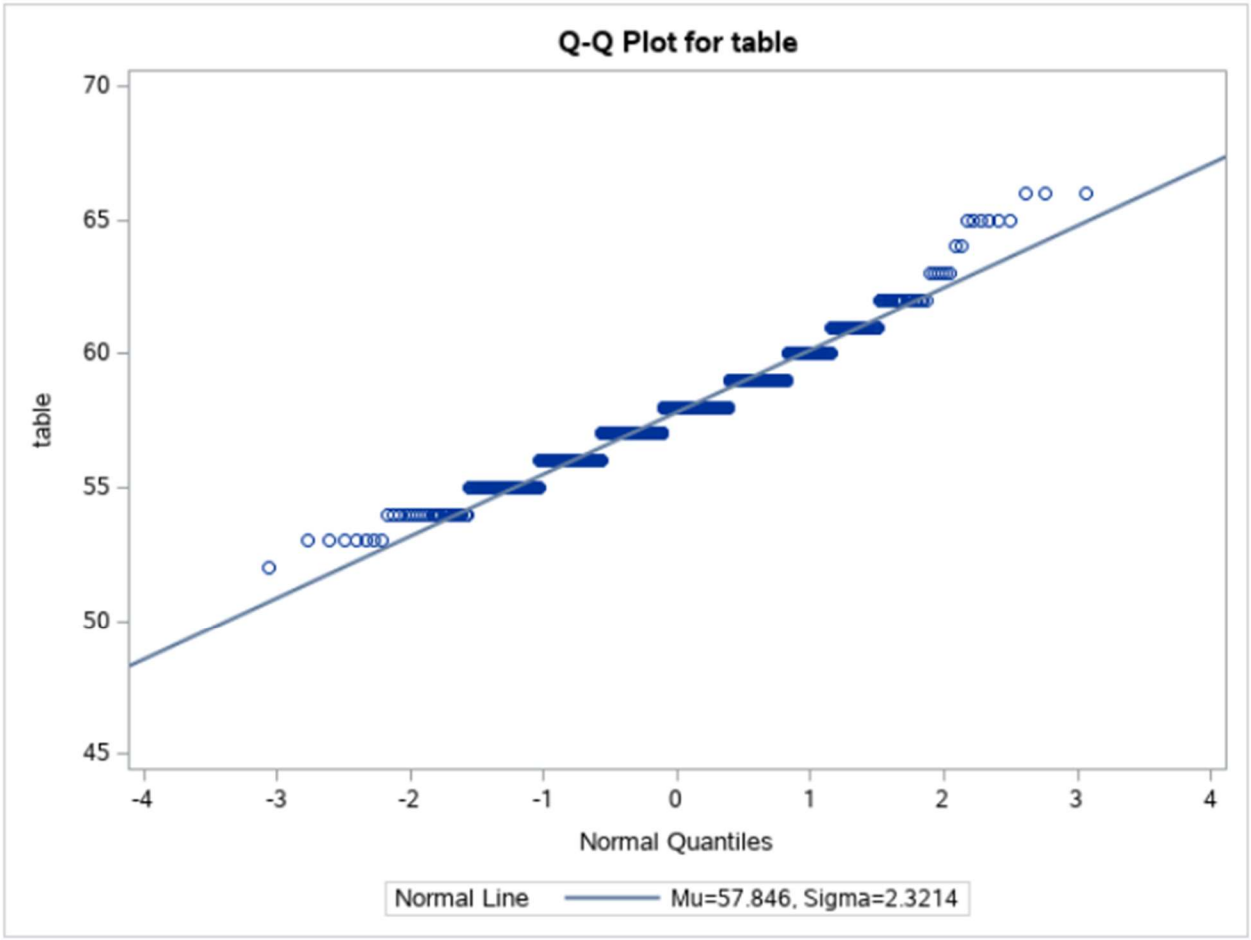
Q-Q Plot for table
70
65
60
55
50
45
table
-4
-3
-2
-1
0
1
2
3
4
Normal Quantiles
Normal Line
Mu=57.846, Sigma=2.3214

Figure 2-3

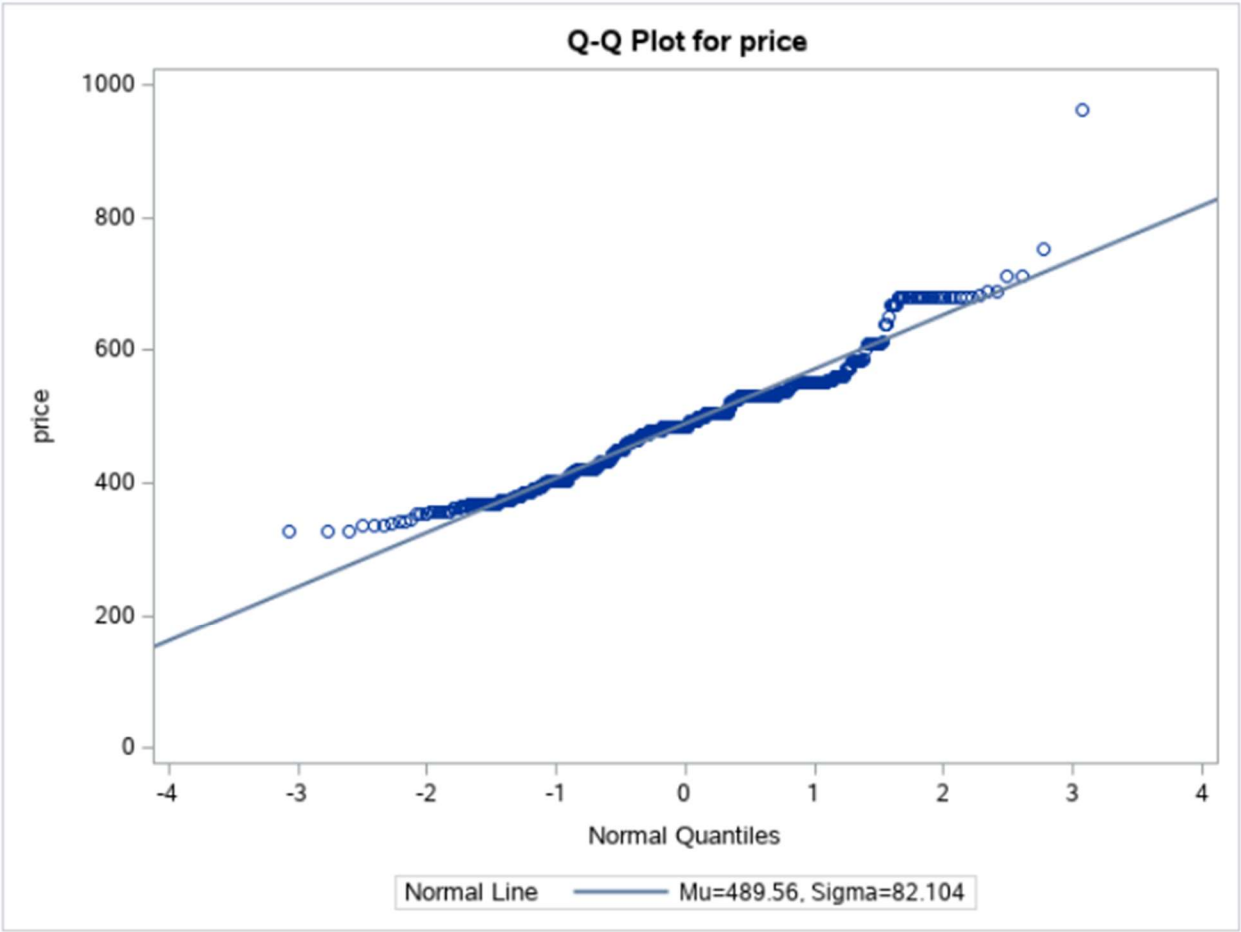
Q-Q Plot for price
price
1000
800
600
400
200
0
-4
-3
-2
-1
0
1
2
3
4
Normal Quantiles
Normal Line
Mu=489.56, Sigma=82.104

Figure 2-4

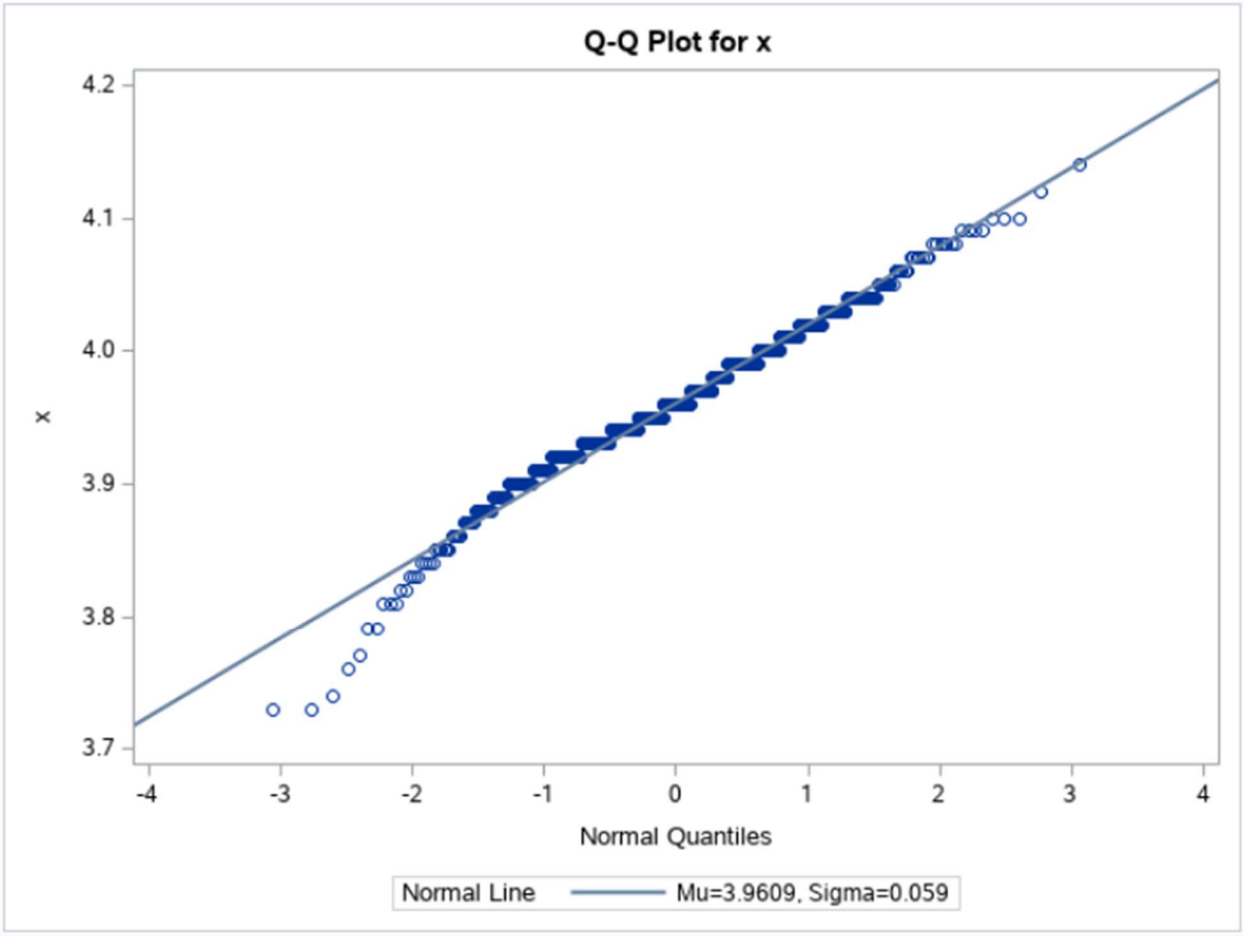
Q-Q Plot for x
4.2
4.1
4.0
3.9
3.8
3.7
x
-4
-3
-2
-1
0
1
2
3
4
Normal Quantiles
Normal Line
Mu=3.9609, Sigma=0.059

Figure 2-5

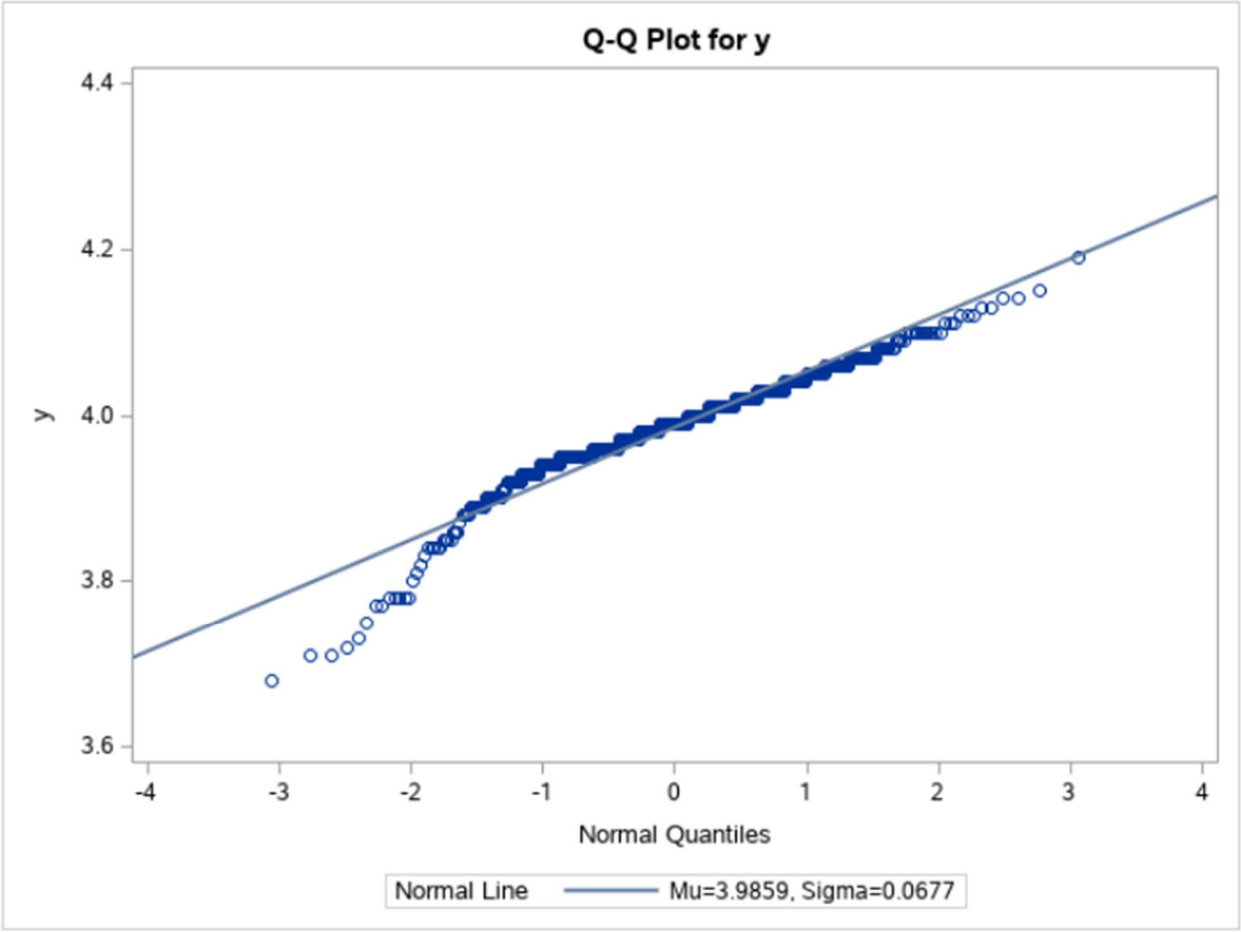
Q-Q Plot for y
4.4
4.2
4.0
3.8
3.6
y
-4
-3
-2
-1
0
1
2
3
4
Normal Quantiles
Normal Line
Mu=3.9859, Sigma=0.0677

Figure 2-6

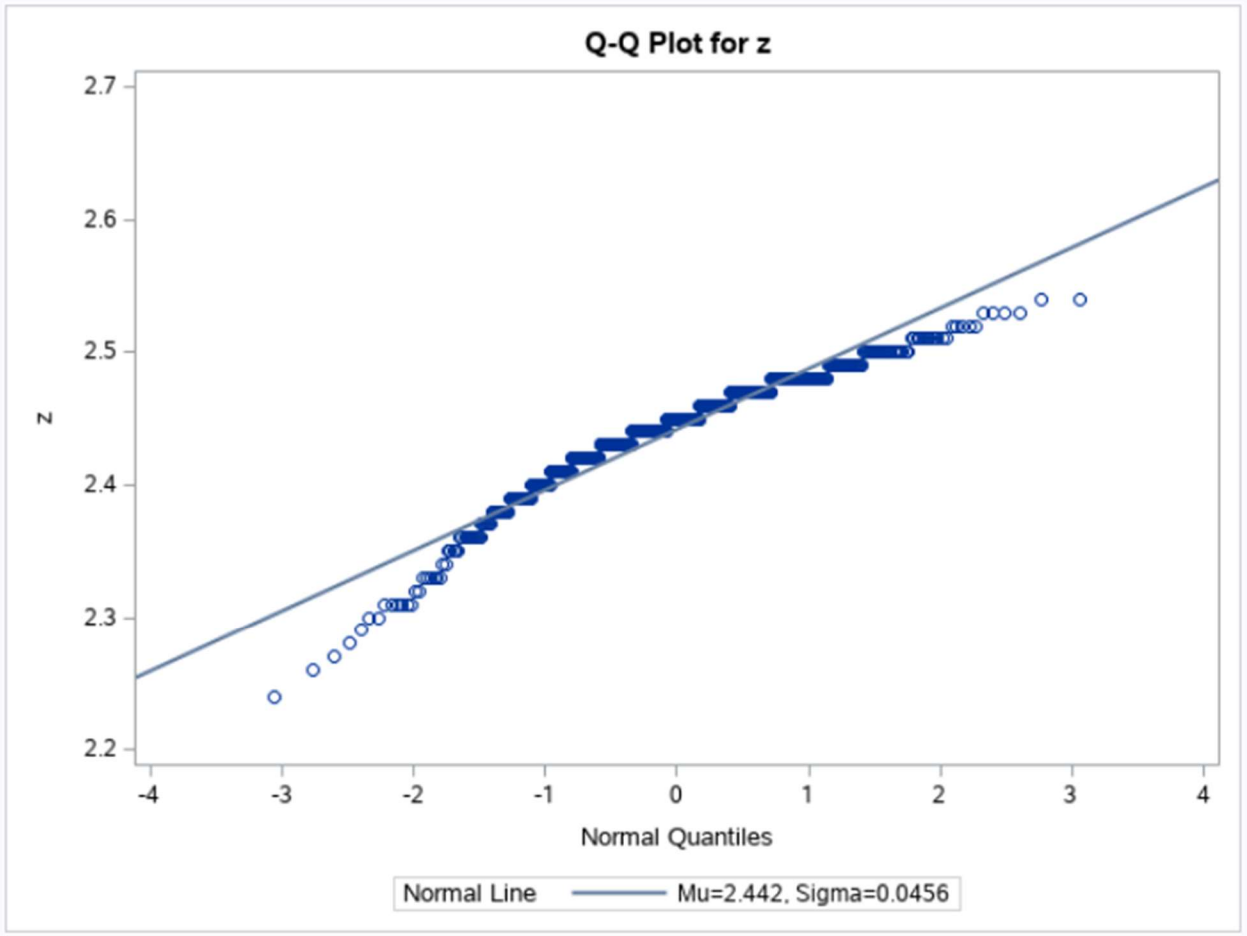
Q-Q Plot for z
2.7
2.6
2.5
2.4
2.3
2.2
z
-4
-3
-2
-1
0
1
2
3
4
Normal Quantiles
Normal Line
Mu=2.442, Sigma=0.0456

Figure 2-7

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
238.994118	112	<.0001

Figure 2-8

MANOVA Tests for the Hypothesis of No Overall cut Effect
H = Type III SSCP Matrix for cut
E = Error SSCP Matrix

S=4 M=1 N=280

Statistic	Value	P-Value
Wilks' Lambda	0.36751135	<.0001
Pillai's Trace	0.79300126	<.0001
Hotelling-Lawley Trace	1.29818827	<.0001
Roy's Greatest Root	0.82004972	<.0001

Figure 2-9

Discriminant and Classification Analysis Figures

Discriminant Analysis of Iris Data

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
127.978971	20	<.0001

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

Figure 3-1

Generalized Squared Distance to species			
From species	1	2	3
1	-12.98897	106.21219	149.89594
2	275.85652	-11.09714	5.55829
3	647.03299	11.32451	-8.87359

Figure 3-2

Discriminant Analysis of Iris Data

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.IRIS
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into species				
From species	1	2	3	Total
1	45 100.00	0 0.00	0 0.00	45 100.00
2	0 0.00	45 100.00	0 0.00	45 100.00
3	0 0.00	1 2.22	44 97.78	45 100.00
Total	45 33.33	46 34.07	44 32.59	135 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for species				
	1	2	3	Total
Rate	0.0000	0.0000	0.0222	0.0074
Priors	0.3333	0.3333	0.3333	

Figure 3-3

Discriminant Analysis of Iris Data

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.IRIS
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into species				
From species	1	2	3	Total
1	45 100.00	0 0.00	0 0.00	45 100.00
2	0 0.00	44 97.78	1 2.22	45 100.00
3	0 0.00	1 2.22	44 97.78	45 100.00
Total	45 33.33	45 33.33	45 33.33	135 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for species				
	1	2	3	Total
Rate	0.0000	0.0222	0.0222	0.0148
Priors	0.3333	0.3333	0.3333	

Figure 3-4

Discriminant Analysis of Iris Data

The DISCRIM Procedure
Classification Summary for Test Data: WORK.TEST
Classification Summary using Quadratic Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	15
Number of Observations Used	15

Number of Observations and Percent Classified into species				
From species	1	2	3	Total
1	5 100.00	0 0.00	0 0.00	5 100.00
2	0 0.00	3 60.00	2 40.00	5 100.00
3	0 0.00	0 0.00	5 100.00	5 100.00
Total	5 33.33	3 20.00	7 46.67	15 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for species				
	1	2	3	Total
Rate	0.0000	0.4000	0.0000	0.1333
Priors	0.3333	0.3333	0.3333	

Figure 3-5

Discriminant Analysis of Iris Data

The DISCRIM Procedure
Classification Summary for Test Data: WORK.TEST
Classification Summary using 5 Nearest Neighbors

Observation Profile for Test Data	
Number of Observations Read	15
Number of Observations Used	15

Number of Observations and Percent Classified into species				
From species	1	2	3	Total
1	5 100.00	0 0.00	0 0.00	5 100.00
2	0 0.00	3 60.00	2 40.00	5 100.00
3	0 0.00	0 0.00	5 100.00	5 100.00
Total	5 33.33	3 20.00	7 46.67	15 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for species				
	1	2	3	Total
Rate	0.0000	0.4000	0.0000	0.1333
Priors	0.3333	0.3333	0.3333	

Figure 3-6

Discriminant Analysis of Iris Data

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.IRIS
Cross-validation Summary using 10 Nearest Neighbors

Number of Observations and Percent Classified into species				
From species	1	2	3	Total
1	45 100.00	0 0.00	0 0.00	45 100.00
2	0 0.00	45 100.00	0 0.00	45 100.00
3	0 0.00	1 2.22	44 97.78	45 100.00
Total	45 33.33	46 34.07	44 32.59	135 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for species				
	1	2	3	Total
Rate	0.0000	0.0000	0.0222	0.0074
Priors	0.3333	0.3333	0.3333	

Figure 3-7

Discriminant Analysis of Iris Data

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.IRIS
Cross-validation Summary using 15 Nearest Neighbors

Number of Observations and Percent Classified into species				
From species	1	2	3	Total
1	45 100.00	0 0.00	0 0.00	45 100.00
2	0 0.00	45 100.00	0 0.00	45 100.00
3	0 0.00	1 2.22	44 97.78	45 100.00
Total	45 33.33	46 34.07	44 32.59	135 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for species				
	1	2	3	Total
Rate	0.0000	0.0000	0.0222	0.0074
Priors	0.3333	0.3333	0.3333	

Figure 3-8

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
Eigenvalue	Difference	Proportion	Cumulative
32.9733	32.6047	0.9889	0.9889
0.3686		0.0111	1.0000

Figure 3-9

Let $\mathbf{V}$ be the matrix with the eigenvectors $\mathbf{v}_i$ that correspond to nonzero eigenvalues as columns. The raw canonical coefficients are calculated as follows

$$\mathbf{R} = {\mathbf{S}_p}^{-1/2}\mathbf{V}$$

The pooled within-class standardized canonical coefficients are

$$\mathbf{P} = \mathrm{diag}(\mathbf{S}_p)^{1/2}\mathbf{R}$$

And the total sample standardized canonical coefficients are

$$\mathbf{T} = \mathrm{diag}(\mathbf{S}_{xx})^{1/2}\mathbf{R}$$

Figure 3-10

Pooled Within-Class Standardized Canonical Coefficients		
Variable	Can1	Can2
SepalLength	-0.538395347	-0.141989419
SepalWidth	-0.447606953	0.843484850
PetalLength	1.002244631	-0.266445204
PetalWidth	0.581317026	0.491560401

Figure 3-11

PCA Figures

Eigenvectors								
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
pregnancies	0.128432	0.593786	-.013087	0.080691	0.475606	-.193598	0.588790	0.117841
Glucose	0.393083	0.174029	0.467923	-.404329	-.466328	-.094162	0.060153	0.450355
BloodPressure	0.360003	0.183892	-.535494	0.055986	-.327953	0.634116	0.192118	-.011296
SkinThickness	0.439824	-.331965	-.237674	0.037976	0.487862	-.009589	-.282213	0.566284
Insulin	0.435026	-.250781	0.336709	-.349944	0.346935	0.270651	0.132010	-.548621
BMI	0.451941	-.100960	-.361865	0.053646	-.253204	-.685372	0.035366	-.341518
DiabetesPedigreeFunction	0.270611	-.122069	0.433189	0.833680	-.119810	0.085784	0.086091	-.008259
Age	0.198027	0.620589	0.075248	0.071201	0.109290	0.033357	-.712085	-.211662

Figure 4-1

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.09437995	0.36316980	0.2618	0.2618
2	1.73121014	0.70158027	0.2164	0.4782
3	1.02962987	0.15410083	0.1287	0.6069
4	0.87552904	0.11318466	0.1094	0.7163
5	0.76234439	0.07971600	0.0953	0.8116
6	0.68262839	0.26281221	0.0853	0.8970
7	0.41981618	0.01535413	0.0525	0.9494
8	0.40446205		0.0506	1.0000

Figure 4-2

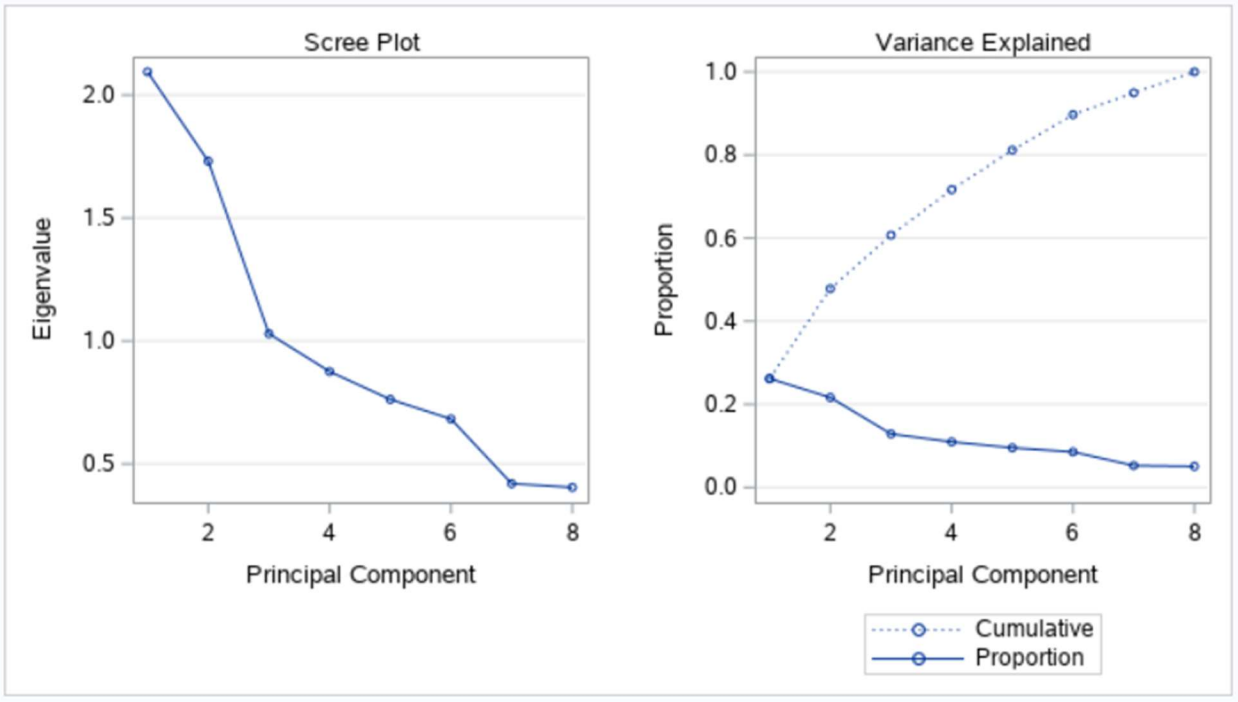
Scree Plot
Eigenvalue
2.0
1.5
1.0
0.5
2
4
6
8
Principal Component
Variance Explained
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
2
4
6
8
Principal Component
Cumulative
Proportion

Figure 4-3

SAS CODE

Hotelling T2 Test

```
DATA hitters;
  INFILE '/folders/myfolders/SAS Examples/9705 Final
Project/john_hitters.dat';
  INPUT League $ Hits HmRun Runs Walks; run;

/* Code to separate data into Al and NL groups */
DATA AL;
   SET hitters;
   IF League = "N" THEN DELETE;
   RUN;
DATA NL;
   SET hitters;
   IF League = "A" THEN DELETE;
   RUN;
PROC PRINT DATA = AL; PROC PRINT DATA = NL; RUN;

*Checking normality of the 4 variables per AL and NL groups;
title "QQPlots of AL variables";
proc univariate data=AL;
   qqplot Hits HmRun Runs Walks/ normal(MU=EST SIGMA=EST);
  run;
title "QQPlots of NL variables";
proc univariate data=NL;
   qqplot Hits HmRun Runs Walks /normal(MU=EST SIGMA=EST);
  run;
*/

*Testing Normality of the Variables;
title "QQPlots";
proc univariate data=hitters;
   qqplot Hits HmRun Runs Walks /normal(MU=EST SIGMA=EST);
  run;

*Testing Homogeneity of Covariance Matrices;
*(See section titled "Test of Homogeneity of Within Covariance Matrices");
proc discrim data=hitters pool=test;
class League;
var Hits HmRun Runs Walks;
run;

TITLE 'Hotelling T2 Test on Hitters';
PROC IML;
  USE hitters;
  READ ALL VAR {Hits HmRun Runs Walks} INTO X;
  X1 = X[1:175,];
  X2 = X[176:322,];
  RESET PRINT;
  N1 = NROW(X1);
  N2 = NROW(X2);
  X1BAR = 1/N1*X1`*J(N1,1);
  X2BAR = 1/N2*X2`*J(N2,1);
  S1 = 1/(N1-1)*X1`*(I(N1)-1/N1*J(N1))*X1;
  S2 = 1/(N2-1)*X2`*(I(N2)-1/N2*J(N2))*X2;
```

```
  Spl = 1/(N1+N2-2)*((N1-1)*S1+(N2-1)*S2);
  T2 = N1*N2/(N1+N2)*(X1BAR-X2BAR)`*INV(Spl)*(X1BAR-X2BAR);
RUN;
```

MANOVA

```
data diamonds; *Load Data;
  infile "/folders/myfolders/SAS Examples/9705 Final
Project/diamonds2dat.dat";
  input cut carat depth table price x y z;
run;
*Cut Labels:
1 = Fair
2 = Good
3 = Very Good
4 = Premium
5 = Ideal;

PROC GLM;
  CLASS cut;
  MODEL carat depth table price x y z = cut;
  MANOVA H=cut/PRINTE PRINTH mstat=exact;
RUN;

*Assessing normality of populations via QQPlots;
proc univariate data=diamonds;
   qqplot carat depth table price x y z/ normal(MU=EST SIGMA=EST);
run;

*Testing Homogeneity of Covariance Matrices;
*(See section titled "Test of Homogeneity of Within Covariance Matrices");
proc discrim data=diamonds pool=test;
class cut;
var carat depth table price x y z;
run;
```

Discriminant and Classification Analysis

```
data iris; *Load Data;
  infile "/folders/myfolders/SAS Examples/iris_train.dat";
  input species SepalLength SepalWidth PetalLength PetalWidth;
run;
*Species Labels:
1 = Setosa
2 = Versicolor
3 = Virginica;
data test;
 infile "/folders/myfolders/SAS Examples/iris_test.dat";
 input species SepalLength SepalWidth PetalLength PetalWidth;
run;

proc discrim data=iris pool=test list crossvalidate testdata=test;
*pool=no : means we are assuming different covariance matrices per group and
therefore are using QDA;
*pool=yes: means we are assuming equal     covariance matrices per group and
therefore are using LDA;
*pool=test: means we are using Bartlett's test to check for equal covariance
matrices, then performing QDA or LDA based on the result.

*METHOD=NORMAL | NPAR
determines the method to use in deriving the classification criterion.
When you specify METHOD=NORMAL, a parametric method based on a multivariate
normal distribution within each class is used to derive a linear or quadratic
discriminant function.
The default is METHOD=NORMAL. When you specify METHOD=NPAR, a nonparametric
method is used and you must also specify either the K= or R= option.;

class species;
var SepalLength SepalWidth PetalLength PetalWidth;
title 'Discriminant Analysis of Iris Data';
run;

proc discrim data=iris method=npar k=5 crossvalidate noclassify
testdata=test;
*k=5, 10, or 15 all yield the same cross validation and test error rates;
class species;
var SepalLength SepalWidth PetalLength PetalWidth;
title 'Discriminant Analysis of Iris Data';
run;

proc discrim data=iris method=npar k=10 crossvalidate noclassify
testdata=test;
class species;
var SepalLength SepalWidth PetalLength PetalWidth;
title 'Discriminant Analysis of Iris Data';
run;

proc discrim data=iris method=npar k=15 crossvalidate noclassify
testdata=test;
class species;
var SepalLength SepalWidth PetalLength PetalWidth; run;
```

```
title 'Discriminant Analysis of Iris Data';
proc candisc data=iris out=iris_cand; class species; run;
proc print data=iris_cand; run;
```

Principal Component Analysis

```
DATA DIABETES;
 INFILE '/folders/myfolders/SAS Examples/9705 Final
Project/pima_diabetes.dat';
 INPUT pregnancies Glucose BloodPressure SkinThickness Insulin BMI
DiabetesPedigreeFunction Age;

PROC PRINCOMP DATA= DIABETES;
 VAR pregnancies Glucose BloodPressure SkinThickness Insulin BMI
DiabetesPedigreeFunction Age;
RUN;
```

DATA SOURCES:

Hotelling T2 Test: Hiters Dataset

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R.* <https://gist.github.com/keeganhines/59974f1ebef97bbaa44fb19143f90bad>

MANOVA: Diamonds Dataset

The diamonds data was exported from the ggplot2 library in R using the write.table() function.
Can also be found here <https://www.kaggle.com/shivam2503/diamonds>

Discriminant and Classification Analysis: Iris Dataset

The iris data was exported from the base library in R using the write.table() function.
It is originally from the following source:
Fisher,R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936);

PCA: Pima Indian Diabetes Dataset

<https://www.kaggle.com/uciml/pima-indians-diabetes-database>