UNIT 21 INFERENCE-III

Structure **Page No**

21.1 INTRODUCTION

Let X be a $p \times 1$ vector random variable having $N_p(\mu, \Sigma)$ distribution. Let $X(p \times n)$ be a data matrix observed from a random sample of size n. The population distribution involves as parameters p components of mean μ and $\frac{1}{2}p(p+1)$ components of variance-covariance Σ. For these parameters, minimum $2^{p(p+3)} - 1$ null hypothesis can be formulated. These null hypotheses can specify the values of a subset of parameters. In this unit, we shall consider the problem of testing hypothesis about the mean μ under both the situations when variance-covariance matrix Σ is known in Section 21.2 and when it is not known in Section 21.3.

In Section 21.4, we shall introduce standardized sum test. The test for equality of mean vector will be discussed in Section 21.5.

Objectives

After reading this unit, you should be able to:

- formulate the null and alternative hypothesis for mean vector when Σ is known or unknown;
- derive a test statistic for testing the hypothesis of mean vectors;
- apply the tests to the given data.

21.2 TEST FOR μ WHEN Σ IS KNOWN

Let $X(p \times n)$ be a sample of size n drawn from a Normal population $N_p(\mu, \Sigma)$ where Σ is known. Based on observed X, we wish to test the null hypothesis $H_0 : \mu = \mu_0$ against an alternative hypothesis $H_1 : \mu \neq \mu_0$ (a specified vector of means). In Unit 19, we have seen that the likelihood function for the sample observation is

$$L(\mu, \Sigma \mid X) = \frac{|\Sigma^{-1}|^{n/2}}{(2\pi)^{np/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)' \Sigma^{-1}(X_i - \mu)\right] \quad (1)$$

If we assume that the null hypothesis is true, i.e. $\mu = \mu_0$, then likelihood is

$$L(\mu_0, \Sigma \mid X) = \frac{|\Sigma^{-1}|^{n/2}}{(2\pi)^{np/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu_0)' \Sigma^{-1}(X_i - \mu_0)\right] \quad (2)$$

Let us first carry out the likelihood ratio test.

We know that the likelihood ratio test is based on statistic λ that is the ratio of maximum of likelihood function under the null hypothesis and over the whole space. That is

$$\lambda = \frac{\max L(\mu_0, \Sigma \mid X)}{\max\limits_{\mu} L(\mu, \Sigma \mid X)} \tag{3}$$

In Unit 19, we have derived that $L(\mu, \Sigma \mid X)$, when Σ is known, attains maximum for $\mu = \bar{X}$ (sample mean vector). The numerator has all the parameters specified and thus its maximum is $L(\mu_0, \Sigma \mid X)$ as given in Eqn.(2). Substituting $\bar{X}$ in place of μ in Eqn.(1), we get the denominator. Therefore

$$\lambda = \frac{L(\mu_0, \Sigma \mid X)}{L(\bar{X}, \Sigma \mid X)} = \exp\left[-\frac{1}{2}\left\{\sum_{i=1}^{n}\left[(X_i - \mu_0)' \Sigma^{-1} (X_i - \mu_0) - (X_i - \bar{X})' \Sigma^{-1} (X_i - \bar{X})\right]\right\}\right]$$

But $\sum_{i=1}^{n}\left[(X_i - \mu_0)' \Sigma^{-1} (X_i - \mu_0) - (X_i - \bar{X})' \Sigma^{-1} (X_i - \bar{X})\right]$

$$= \sum_{i=1}^{n}\left[(X_i - \bar{X} + \bar{X} - \mu_0)' \Sigma^{-1} (X_i - \bar{X} + \bar{X} - \mu_0) - (X_i - \bar{X})' \Sigma^{-1} (X_i - \bar{X})\right]$$

$$= \sum_{i=1}^{n}(X_i - \bar{X})' \Sigma^{-1} (X_i - \bar{X}) + \sum_{i=1}^{n}(X_i - \bar{X})' \Sigma^{-1} (\bar{X} - \mu_0) + \sum_{i=1}^{n}(\bar{X} - \mu_0)' \Sigma^{-1} (X_i - \bar{X})$$

$$+ \sum_{i=1}^{n}(\bar{X} - \mu_0)' \Sigma^{-1} (\bar{X} - \mu_0) - \sum_{i=1}^{n}(X_i - \bar{X})' \Sigma^{-1} (X_i - \bar{X}).$$

The first term cancels with the last term, second and third terms reduce to zero and the fourth term is $n(\bar{X} - \mu_0)' \Sigma^{-1} (\bar{X} - \mu_0))$ and, hence

$$\lambda = \exp\left[-\frac{1}{2}\{n(\bar{X} - \mu_0)' \Sigma^{-1} (\bar{X} - \mu_0)\}\right]. \tag{4}$$

Therefore,

$$-2\log\lambda = n(\bar{X} - \mu_0)' \Sigma^{-1} (\bar{X} - \mu_0). \tag{5}$$

It may be recalled at this stage that likelihood ratio test rejects the null hypothesis if $\lambda \le \lambda_\alpha$ where λ_α is chosen such that $P[\lambda \le \lambda_\alpha \mid H_0]$ is equal to α (the pre-specified value of size of the test). It may be noted from Eqn.(5) that $\lambda \le \lambda_\alpha$ is equivalent to $n(\bar{X} - \mu_0)' \Sigma^{-1} (\bar{X} - \mu_0) \ge C_\alpha (= -2\log\lambda_\alpha)$. We know that $n(\bar{X} - \mu_0)' \Sigma^{-1} (\bar{X} - \mu_0)$ has a chi-square distribution with p degrees of freedom and, hence $C_\alpha = \chi^2_{\alpha, p}$ (tabulated value of chi-square with p degrees of freedom at specified level of significance α). Therefore, the test procedure for testing the above-mentioned hypothesis would be to reject the null hypothesis $H_0 : \mu = \mu_0$ if $n(\bar{X} - \mu_0)' \Sigma^{-1} (\bar{X} - \mu_0)$ is greater than $\chi^2_{\alpha, p}$ the tabulated value of chi-square with p degrees of freedom at α level of significance.

Let us now discuss another test known as the union and intersection test.

Consider an arbitrary linear combination $a'X$ of X (Here this X is different from data matrix X), where a is a non-null real column vector of order p. We know that $a'X$ is also normally distributed as univariate normal distribution with mean $a'\mu$ verses variance $a'\Sigma a$. Now the multivariate hypotheses $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ can be transformed to univariate hypotheses $H_{0a} : a'\mu = a'\mu_0$ and $H_{1a} : a'\mu \neq a'\mu_0$. The test statistics for testing this hypothesis is

$$Z^2(a) = na'\left(\bar{X} - \mu_0\right)\left(\bar{X} - \mu_0\right)' a / a'\Sigma a.$$

Moreover, the null hypothesis is rejected if calculated value of $Z^2(a)$ is greater than $\chi^2_{\alpha,\, p}$. It may be noted here that if null hypothesis $H_0 : \mu = \mu_0$ is true, $H_{0a} : a'\mu = a'\mu_0$ must be true for all non-zero choices of a. Hence, the multivariate hypothesis H_0 can be written as the intersection of H_{0a} for all a i.e. $H_0 = \bigcap_a H_{0a}$. Therefore, the null hypothesis H_0 will be accepted if the univariate hypothesis H_{0a} is accepted for all the values of a and the α -level acceptance region for multivariate H_0 will be the intersection of acceptance regions of H_{0a} i.e. $\bigcap_a \left[Z^2(a) \leq \chi^2_{\alpha,\, p}\right]$. It implies that for acceptance region maximum of $Z^2(a)$ is less than $\chi^2_{\alpha,\, p}$. Thus the test statistics for the multivariate null hypothesis will be $\max_a Z^2(a)$. However, it can be proved that $\max_a Z^2(a) = (\bar{X} - \mu_0)'\Sigma^{-1}(\bar{X} - \mu_0)$ which follows chi-square distribution with p degrees of freedom. Thus, the test statistics is same as obtained in likelihood ratio test.

Remark: The likelihood ratio and union intersection tests are same.

Example 1: The length of Centrum (x_1) and width of Centrum (x_2) of a sample of 12 fishes of Serrandae family were observed as given in Table 1.

Table 1

i	(x_1)	(x_2)
1	7.5	6.7
2	6.8	6.2
3	8.5	7.1
4	5.8	6.0
5	5.2	5.8
6	7.0	6.2
7	8.2	7.5
8	6.9	7.3
9	7.4	6.8
10	8.4	7.3
11	7.6	7.0
12	9.2	7.8

A researcher claims that the mean length and width of Centrum of fishes belonging to Serrandae family is 8.94 and 6.76, respectively with a variance-covariance matrix of x_1, x_2 as

$$\Sigma = \begin{bmatrix} 13.2496 & 9.0418 \\ 9.0418 & 7.6176 \end{bmatrix}.$$

Test this claim at 5% level of significance.

Solution: We wish to test whether the above-mentioned data confirms his claim about mean length and width of Centrum of Sarrendae fishes.

Assuming that the length and width of Centrum for Serrandae fishes follow bi-variate normal distribution $N(\mu, \Sigma)$. The null hypothesis for the present problem would be $H_0 : \mu = [8.94, 6.76]' (= \mu_0$ say) against the alternative hypothesis $H_A : \mu \neq [8.94, 6.76]'$.

The population covariance matrix Σ is known as claimed by researcher. Hence, Chi-square test discussed in Eqn. (5) can be used. The value of test statistics under H_0 is

$$\chi_p^2 = n(\bar{x} - \mu_0)' \Sigma^{-1} (\bar{x} - \mu_0)$$

The sample mean $\bar{x}$ for the above data

$$\bar{x} = [7.375, 6.808]'$$

Thus,

$$\chi_p^2 = 12[-1.565, 0.048] \begin{bmatrix} 0.3972 & -0.4715 \\ -0.4715 & 0.6909 \end{bmatrix} \begin{bmatrix} -1.565 \\ 0.048 \end{bmatrix}$$

$$= 12.54$$

Since $p = 2$, the upper point of chi-square for $\alpha = 0.05$ is 10.60 which is less than the calculated value thus we reject H_0 at 5% level of significance and conclude that the data contradicts the claim of the researcher.

Now, try the following exercise.

E1) A sample of ten industrial corporations was considered for the pairs of observations of their sales (x_1) and profits (x_2). The observations are given in the Table 2.

Table 2

Corporation No.	x_1 (Rs. in Lakhs)	x_2 (Rs. in Lakhs)
1	40	8
2	42	10
3	34	6
4	16	6
5	50	10
6	24	4
7	37	6
8	42	8
9	25	7
10	20	5

The expected mean vector and variance-covariance matrix is $\mu = \begin{bmatrix} 40 \\ 10 \end{bmatrix}$ and $\begin{bmatrix} 13 & 10 \\ 10 & 6 \end{bmatrix}$. Test whether the sample confirms its truth ness of mean vector at 5% level of significance.

So far we discuss the test for μ when Σ is known, now we shall discuss the test for μ when Σ is unknown in the following.

21.3 TEST FOR μ WHEN Σ IS UNKNOWN

Let $X(n \times p)$ is a sample of size n drawn from a Normal population $N_p(\mu, \Sigma)$ where μ and Σ both are unknown. Based on observed X, we wish to test the null hypothesis $H_0 : \mu = \mu_0$ against an alternative $H_1 : \mu \neq \mu_0$ (a specified vector of means). Again, we shall consider both the tests i.e. likelihood ratio test and union intersection test for this case. The likelihood function for the sample is given in Eqn.(1) and under the null hypothesis, it reduces to Eqn.(2).

Firstly, we shall discuss the likelihood ratio test.

As mentioned in Sec. 21.2, the likelihood ratio test statistic is the ratio of maximum of likelihood function under the null hypothesis and over the whole parameter space, i.e., the likelihood ratio for the present case is defined as

$$\lambda = \frac{\max\limits_{\Sigma} L\left(\mu_{0,} \Sigma \mid X\right)}{\max\limits_{\mu,\Sigma} L\left(\mu, \Sigma \mid X\right)}. \tag{6}$$

and,

$$-2\log\lambda = 2[\max_{\mu,\Sigma} \log L\left(\mu, \Sigma \mid X\right) - \max_{\Sigma} \log L(\mu_0, \Sigma \mid X)]. \tag{7}$$

If population mean μ is known to be equal to μ_0, the maximum likelihood estimate of Σ is $[S + (\bar{X} - \mu_0)(\bar{X} - \mu_0)']$ where S is the sample variance-covariance matrix defined as $n^{-1}X'X - \bar{X}\,\bar{X}'$. However, if μ and Σ both are unknown, the maximum likelihood estimates of μ and Σ are, $\bar{X}$ and S respectively. Therefore,

$$\max_{\Sigma} \log L(\mu_0, \Sigma \mid X) = -\frac{n}{2}[p\log 2\pi + \log|S| + \log\{1 + (\bar{X} - \mu_0)'S^{-1}(\bar{X} - \mu_0)\} + p] \tag{8}$$

and

$$\max_{\mu,\Sigma} \log L(\mu, \Sigma \mid X) = -\frac{n}{2}[p\log 2\pi + \log|S| + p]. \tag{9}$$

Substituting the values from Eqn.(8) and Eqn.(9) in Eqn.(7), we get

$$-2\log\lambda = n\log[1 + (\bar{X} - \mu_0)'S^{-1}(\bar{X} - \mu_0)] \tag{10}$$

Now it is easy to verify that $\lambda \leq \lambda_\alpha$ is equivalent to $(\bar{X} - \mu_0)'S^{-1}(\bar{X} - \mu_0) \geq C_\alpha$. In Unit 20, we have derived that $(n-1)(\bar{X} - \mu_0)'S^{-1}(\bar{X} - \mu_0) = T^2$ follows Hotelling's T^2-distribution with $(p, (n-1))$ degrees of freedom. It has also been pointed out that this statistic can be converted in F-statistic by defining $F = (n-p)T^2/[(n-1)p]$ which has $(p, n-p)$ degrees of freedom. Therefore, the test procedure would be to reject the null hypothesis if calculated value of $(n-p)(n-1)(\bar{X} - \mu_0)'S^{-1}(\bar{X} - \mu_0)/[(n-1)p]$ is greater than tabulated value of F with $(p, n-p)$ degrees of freedom and at specified level of significance.

As we discussed the case when Σ is known, now we shall test μ when Σ is unknown by union and intersection test.

We have already seen that if we consider a linear combination $a'X$, it follows univariate normal distribution with mean $a'\mu$ and variance $a'\Sigma a$. Similarly, the multivariate hypothesis $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ can be transformed to univariate hypothesis $H_{0a} : a'\mu = a'\mu_0$ and $H_{1a} : a'\mu \neq a'\mu_0$. The test statistics for testing this hypothesis is well-known univariate t-statistic defined for the present case as

$$t(a) = \frac{a'(\bar{X} - \mu_0)\sqrt{n}}{\sqrt{a'Sa}}.$$

Thus, the acceptance region of level α would be $t^2(a) \le t^2_{\alpha/2,\,n-1}$. Following the argument given in Eqn.(2), the acceptance region for the multivariate hypothesis H_0 would be $\bigcap_a [t^2(a) \le t^2_{\alpha/2,\,n-1}]$ i.e. $\max . t^2(a) \le t^2_{\alpha/2,\,n-1}$. However, $\max . t^2(a)$ is $(n-1)(\bar{X} - \mu_0)' S^{-1} (\bar{X} - \mu_0) = T^2$ which is a Hotelling's T^2 statistic. You may note that this test is same as likelihood ratio test.

Example 2: Let us reconsider the problem given in Example 1. Suppose we wish to test the researcher's claim about mean length and width of Centrum for Serrandae fishes to be 8.94 and 6.76 respectively at 5% level of significance and no information is available about the population covariance matrix.

Solution: The null hypothesis and alternative hypothesis are same as given in Example 1, i.e.

$$H_0 : \mu = [8.94 \quad 6.76]' \quad (= \mu o \text{ say})$$
$$H_0 : \mu \ne [8.94 \quad 6.76]$$

The sample covariance matrix for the data can be easily calculated as:

$$S = \begin{bmatrix} 1.919 & 0.5877 \\ 0.5877 & 0.3741 \end{bmatrix}$$

The test statistic T^2 under H_0 can be calculated as

$$T^2(p,\, n-1) = (n-1)\,(\bar{X} - \mu_0)'\, S^{-1}\, (\bar{X} - \mu_0)$$

$$= 11[-1.565, 0.048]' \begin{bmatrix} 3.72224 & -5.8478 \\ -5.8478 & 11.8599 \end{bmatrix} \begin{bmatrix} -1.565 \\ 0.148 \end{bmatrix}$$

$$= 110.25$$

and $$F_{(p,n-p)} = \frac{n-p}{(n-1)p} T^2(p,\, n-1)$$

$$= \frac{12-2}{11 \times 2} \times 110.25 = 50.11.$$

The tabulated value F as (2,10) degrees of freedom at 0.05 is 4.10 which is much smaller than calculated value 50.11. Hence, we reject H_0 at 5% level and conclude that researcher's claim is not true in the light of observed data.

Now try the following exercise.

E2) Test the hypothesis for the problem given in E1) if Σ is not known.

The classical multivariate tests such as Hotelling's T^2 test are not feasible if $p > n$. The classical tests do not keep their nominal level of significance if a procedure for selecting variables or compressing data is applied as a preparatory step [Lauter *et al*(1996)], O'Brien (1984) and Tang et al (1993) suggested the effective alternatives. These tests also do not exactly maintain the significance level, especially in cases with small sample sizes. To avoid the problem, Lauter *et al* (1996), Lauter (1996), Lauter *et al* (1998) have suggested some stable multivariate tests.

In the following section, we shall discuss the standardized sum test.

21.4 STANDARDIZED SUM TEST [SS TEST]

Let $X_i' = [x_{i1}, x_{i2}, \cdots, x_{ip}]$ $(i = 1, 2, \cdots, n)$ be n independent p-dimensional observation vectors distributed according to normal distribution $N_p(\mu, \Sigma)$ with unknown parameters μ, Σ. We develop a test of $H_0 : \mu : 0$, against $H_1 : \mu \neq 0$.

Let us arrange the n individual vectors X_i into a $p \times n$ matrix X, where $X = (x_{ji}) \sim N_{p \times n}(\mu 1n, \Sigma \otimes I_n)$.
where 1_n is the column vector of n ones and I_n is identity matrix of order $(n \times n)$.

Define $Z = (z_1, z_2, \ldots, z_n) = (d_1, d_2, \ldots, d_p)X = d'X$, Z_i representing the weighted linear combination of p components of observation X_i with respective $d_1, d_2, \cdots, d_p$ for $Z_i (i = 1, 2, \ldots, n)$ are called the individual scores of n observations. The vector d of the weights has to be a unique function of the $p \times p$ matrix

$$XX' = \left(\sum_{i=1}^{n} x_{ji} x_{ik} \right) \quad j, k = 1, 2, \ldots, p$$

satisfying $d'X \neq 0$ with probability 1.

Under the above condition and under the null hypothesis H_0, the distribution of Z has a spherical shape in n-dimensional space. This shape is not, in general, the shape of a normal distribution. From the sphericity, the statistic

$$T = \sqrt{n}\, \overline{Z} / s_z$$

where $\overline{Z} = (1/n) \sum_{i=1}^{n} Z_i$ and $S_z^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \overline{Z})^2$ has the exact 't'-distribution with $(n-1)$ d.f. The test using this 't' is an exact α-level test. Lauter et al (1996) have shown that this SS test can be more effective than Hotelling's one sample T^2-test in many situations. In fact, the SS test is only suitable if the expected deviation from the hypothesis H_0 have same direction for all p variables.

For practical application purposes of SS test, the vector d is obtained in such a way that each variable is standardized by the square root of the corresponding diagonal elements of XX', that is

$$d = [\operatorname{diag}(XX')]^{-1/2} 1_p .$$

Thus the value of Z_i has the form

$$Z_i = \sum_{j=1}^{p} \frac{x_{ji}}{\sqrt{i \sum_{k=1}^{n} k x_j^2}} , \quad i = 1, 2, \ldots, n .$$

Example 3: A random sample of 20 families of low economic status from a slum area was selected to test whether or not the parents of the children of the area have spent any time for education. The number of years (in completed years) spent for education is given in Table 3.

Table 3

Family Sr. No.	No. of years spent for education by	
	Father	Mother
1.	5	0
2.	10	5
3.	12	4
4.	14	14
5.	0	0
6.	10	0
7.	5	5
8.	0	0
9.	8	4
10.	12	10
11.	10	5
12.	8	0
13.	0	0
14.	5	0
15.	0	0
16.	6	0
17.	0	0
18.	0	0
19.	0	0
20.	6	0

Solution: Let $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ be the mean number of years spent for education by parents.

It may be noted that $H_0 : \mu = 0$. Let us apply the above-mentioned SS test, considering $H_A : \mu \neq 0$. For the above data

$$\overline{X} = [5.4 \quad 2.35]' \; S = \begin{bmatrix} 21.84 & 13.36 \\ 13.36 & 14.63 \end{bmatrix} \text{and} \quad XX' = \begin{bmatrix} 1020 & 521 \\ 521 & 403 \end{bmatrix}$$

Hence, $d = \left[\frac{1}{\sqrt{1021}}, \frac{1}{\sqrt{403}}\right]' = [0.0313 \quad 0.0498]'$

And Z_i : 0.1565, 0.562, 0.5748, 1.1354, 0.00, 0.313, 0.4055, 0.00,

0.4496, 0.8736, 0.562, 0.1565, 0.00, 0.1565, 0.00, 0.1878

0.00, 0.00, 0.00 and 0.1878.

Thus $\overline{Z} = 0.28605$, $S_z^2 = 0.1045$, $S_z = 0.3233$.

Moreover $t = \dfrac{\sqrt{20} \times 0.28605}{0.3233} = 3.96$.

Tabulated value of t with 19 d.f. at 5% and 1% level of significances are 2.093 and 2.861, respectively.

Therefore, we reject H_0 even at 1% level of significance.

Try the following exercises.

E3) Calculate T^2 and F for the data given in Table 3.

E4) Find the value of t for the data given in E1).

In the following section, we shall discuss the test for equality of mean vectors.

21.5 TEST FOR EQUALITY OF MEAN VECTORS

Let $X_i(n_i \times p), i = 1, 2, \ldots, k$, be data matrices drawn from $N_p(\mu_i, \Sigma_i), i = 1, 2, \ldots, k$, respectively. Based on the sample information X_i's, we wish to test the null hypothesis about the equality of mean vectors μ_i's. First, we will consider the two-sample problem under the various assumptions about the covariance matrices. First, we will consider the case that covariance matrices are known which is given in the following subsection. In the next subsection, we will discuss the case when covariance matrices are equal but unknown.

Case I: Test for Equality of Two Mean Vectors when Covariance Matrices are Known

Let $X_1(n_1 \times p)$ and $X_2(n_2 \times p)$ be the random sample of sizes n_1 and n_2, respectively drawn from multivariate normal populations. If $\bar{X}$ denotes the sample mean for i^{th} sample, then it is unbiased estimate of corresponding mean vector μ_i and $\bar{X}_i \sim N_p(\mu_i, n_i^{-1}\Sigma_i), i = 1, 2$. Let us define a new variable $\bar{X}_d = \bar{X}_1 - \bar{X}_2$. Naturally, $\bar{X}_d$ will follow a multivariate normal distribution with mean vector $\mu = \mu_1 - \mu_2$ and covariance matrix $\Sigma_c = n_1^{-1}\Sigma_1 + n_2^{-1}\Sigma_2$. Therefore, $(\bar{X}_d - \mu)'\Sigma_c^{-1}(\bar{X}_d - \mu)$ will have a chi-square distribution with p degrees of freedom. Under the null hypothesis $H_0: \mu_1 = \mu_2$ i.e. $\mu = \mu_1 - \mu_2 = 0$, it reduces to $\bar{X}_d\Sigma_c^{-1}\bar{X}_d$ and shall follow a chi-square distribution. Hence, we reject null hypothesis H_0 if calculated value of $\bar{X}_d\Sigma_c^{-1}\bar{X}_d$ is greater than the tabulated value of chi-square with p degrees of freedom and at specified level of significance.

Remark: If $\Sigma_1 = \Sigma_2 = \Sigma$ (say), $\Sigma_c = (n_1^{-1} + n_2^{-1})\Sigma$. Hence, the chi-square statistics will now take the form $(n_1^{-1} + n_2^{-1})\,(\bar{X}_d - \mu)'\Sigma^{-1}(\bar{X}_d - \mu)$ and under H_0 it reduces to $(n_1^{-1} + n_2^{-1})\,\bar{X}_d\Sigma^{-1}\bar{X}_d$.

Case II: Test for Equality of Two Mean Vectors when Covariance Matrices are Equal and Unknown

Let $X_1(n_1 \times p)$ be a data matrix from $N_p(\mu_1, \Sigma_1)$ and $X_2(n_2 \times p)$ be another data matrix from $N_p(\mu_2, \Sigma_2)$. Assume that $\Sigma_1 = \Sigma_2 = \Sigma$ (unknown). Then Mahalanobis distance is defined as

$$\Delta^2 = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2).$$

The value of Δ^2 cannot be calculated, as the parameters μ_1, μ_2 and Σ are not known. However, Mahalanobis Distance for sample observations is given by

$$D^2 = (\bar{X}_1 - \bar{X}_2)'\, S_u^{-1}(\bar{X}_1 - \bar{X}_2),$$

where, $\bar{X}_i (i = 1, 2)$ is the mean of i^{th} sample and

$$S_u = \frac{n_1 S_1 + n_2 S_2}{n_1 + n_2 - 2}.$$

Here S_i is the sample covariance matrix of i^{th} sample $(i = 1, 2)$. It may be noted here that for the given situation, $\bar{X}_1$ and $\bar{X}_2$ are the unbiased estimators of μ_1 and μ_2

respectively. Similarly, n_1S_1 and n_2S_2 are maximum likelihood estimators of Σ and the pooled unbiased estimator of Σ is S_u. Further $\bar{X}_1$, $\bar{X}_2$, S_1 and S_2 are independently distributed.

It may be recalled that if we denote $\bar{X}_d = \bar{X}_1 - \bar{X}_2$, then it follows multivariate normal distribution with mean vector $(\mu_1 - \mu_2)$ and covariance matrix $(n_1^{-1} + n_2^{-1})\Sigma$, under the null hypothesis $H_0 : \mu_1 = \mu_2$, $\bar{X}_d \sim N_p(0, (n_1^{-1} + n_2^{-1})\Sigma)$. Since the samples are independent, n_1S_1 and n_2S_2 are independently distributed, following Wishart distribution, namely $n_iS_i \sim W_p(\Sigma, n_i - 1)$ $(i = 1, 2)$. Therefore,

$(n_1^{-1} + n_2^{-1})(n_1S_1 + n_2S_2) \sim W_p[(n_1^{-1} + n_2^{-1})\Sigma, (n_1 + n_2 - 2)]$ and is independent of $\bar{X}_d$.

From the definition of Hotelling's T^2-statistics, we get that

$$(n_1 + n_2 - 2)\bar{X}_d' [(n_1^{-1} + n_2^{-1})(n_1S_1 + n_2S_2)]^{-1}\bar{X}_d$$

is distributed as $T^2(p, n-2)$. We may note that $D^2 = \bar{X}_d' S_u^{-1}\bar{X}_d$ and

$(n_1 + n_2 - 2)\bar{X}_d'[(n_1^{-1} + n_2^{-1})(n_1S_1 + n_2S_2)]^{-1}\bar{X}_d = (n_1^{-1} + n_2^{-1})^{-1}\bar{X}_d'S_u^{-1}\bar{X}_d$.

Hence, D^2 can be transformed to T^2 by the relation

$$T^2(p, n-2) = \frac{n_1n_2}{n}D^2,$$

where $n = n_1 + n_2$. We know that $T^2(p, m) = \frac{mp}{m-p+1}F_{p,\, m-p+1}$ where $F_{a,\,b}$ denote the F-statistic with (a, b) degrees of freedom. Therefore, finally, the significance of the hypothesis $H_0 : \mu_1 = \mu_2$ is tested by the statistic

$$F = \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p}T^2(p, n_1 + n_2 - 2)$$

$$= \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)}\frac{n_1n_2}{n_1 + n_2}(\bar{X}_1 - \bar{X}_2)' S_u^{-1} (\bar{X}_1 - \bar{X}_2)$$

which follows F-distribution with $(p, n_1 + n_2 - p - 1,)$ degrees of freedom. Therefore, the test procedure would be to reject the null hypothesis H_0 if calculated value of the above statistics is greater than tabulated value of F-statistics at specified level of significance and above-mentioned degree of freedom.

Example 4: The following data show the number of ever born children and dead children to a number of couples belonging to low and medium socio economic status:

Low Socio Economic		Medium Socio Economic	
Number of Children		Number of Children	
10	3	2	0
3	0	5	0
5	0	4	1
2	0	6	1
12	2	3	0
1	0	4	0
8	1	5	1
7	2	10	2
4	0	8	3
2	0	6	1

1	0	7	0
5	1	6	0
4	0	5	0
6	2	4	0
7	1	8	1
4	1	7	2
8	2	3	0
3	1	2	0
6	1	3	0
5	0	1	0
		2	0
		4	1
		5	1
		4	0
		6	0
		9	0
		6	0
		5	0

Test the hypothesis that average ever born children and dead children to low and medium status couples are equal, assuming that the number of ever born and dead children follow bi-variate normal population with covariance matrices equal but unknown. Thus, we wish to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ assuming that $\Sigma_1 = \Sigma_2 = \Sigma$ (unknown). Here $n_1 = 20$, $n_2 = 28$, $p = 2$.
From the data given above, we can calculate

$$\overline{X}_1 = \begin{bmatrix} 5.15 \\ 0.85 \end{bmatrix} \quad \overline{X}_2 = \begin{bmatrix} 4.821 \\ 0.500 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} 8.127 & 2.072 \\ 2.072 & 0.827 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 4.146 & 1.017 \\ 1.017 & 06.07 \end{bmatrix}$$

Therefore $S_\mu = \dfrac{n_1 S_1 + n_2 S_2}{n_1 + n_2} = \begin{bmatrix} 60.57 & 1.520 \\ 1.520 & 0.729 \end{bmatrix}$

and $S_\mu^{-1} = \begin{bmatrix} 0.3463 & -0.7220 \\ -0.7220 & 2.8772 \end{bmatrix}$

$$D^2 = (\overline{X}_1 - \overline{X}_2)\, S_\mu^{-1} (\overline{X}_1 - \overline{X}_2) = 0.193$$

Hence $T^2 = \dfrac{n_1 n_2}{n} D^2 = \dfrac{20 \times 28}{48} \times 0.193 = 2.25$

$$F = \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2 = \frac{48 - 2 - 1}{(48 - 2)2} \times 2.25 = 1.10.$$

The tabulated value $F(2, 45)$ at 5% level of significance is 3.27. Since calculated value is less than tabulated value at 5% level of significance, we conclude that data provide no evidence for rejection of hypothesis.

Case III: Standardized Sum Test [SS Test]

Let $X_1(n_1 \times p)$ and $X_2(n_2 \times p)$ be two data matrices drawn from $N_p(\mu_1, \Sigma_1)$ and $N_p(\mu_2, \Sigma_2)$ respectively. The problem is to test the hypothesis

$$H_0 : \mu_1 = \mu_2$$

Let us assume that $\Sigma_1 = \Sigma_2 = \Sigma$ (say). Let S_u be pooled unbiased estimator of Σ, where

$$S_u = \frac{n_2 S_1 + n_2 S_2}{n_1 + n_2 - 2}.$$

Here $S_i (i = 1, 2)$ is the sample covariance matrix of i^{th} sample. Let $\bar{X}_i$ be the sample mean of i^{th} sample. Following the discussion given in sub-section 21.4.1, we can easily see that under H_0 the statistic

$$t = \sqrt{a}\frac{(\bar{X}_1 - \bar{X}_2)' d}{\sqrt{d' S_u d}}$$

has exactly Student's 't' distribution with $(n-2)$ degrees of freedom [Lauter *et al* (1998)]. Here $d = d(W)$ may be a function of the total sums of product matrix

$$W = (n-2) S_u + a(\bar{X}_1 - \bar{X}_2)(\bar{X}_1 - \bar{X}_2)'$$

where, $n = n_1 + n_2$, $a = \dfrac{n_1 n_2}{n_1 + n_2}$

The matrix W can also be written as

$$W = (X - \bar{X})' (X - \bar{X}) = X'X - \bar{X}'\bar{X}$$

where, $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$, $\bar{X} = n^{-1} 1_n 1_n' X$.

Here 1_n is the vector of n ones.

In practice, SS test is done using pooled sums of product matrix W. From W, d is defined by

$$D = [\text{Diag}(W)]^{-1/2} 1_p.$$

Let $\mu_i' = (\mu_{i1}, \cdots, \mu_{ip})$, $i = 1, 2$ be the two population mean vectors.

According to Lauter *et al* (1998) this SS test is sensible if $\mu_{1j} - \mu_{2j}$ $[j = 1, 2, \ldots, p]$ are all assumed to be positive or all negative. However, if any one of $\mu_{1j} - \mu_{2j}$ changes the direction, the test is not suitable. If all the variables have identical univariate properties, the SS test attains the highest efficiency.

Example 5: Let us apply the test on the data given in Example 1. Here

$$W = (n_1 + n_2 - 2) S_u = \begin{bmatrix} 278.622 & 69.92 \\ 69.92 & 33.534 \end{bmatrix}.$$

Therefore $d = (\text{diag}(W))^{-1/2} 1_p = \begin{bmatrix} 0.0599 & 0 \\ 0 & 0.1727 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$$= \begin{bmatrix} 0.0599 \\ 0.1727 \end{bmatrix}$$

$$d' S_u d = [0.0599 \quad 0.1727] \begin{bmatrix} 6.057 & 1.520 \\ 1.520 & 0.729 \end{bmatrix} \begin{bmatrix} 0.0599 \\ 0.1727 \end{bmatrix}$$

$= 0.0749$

$$t = \frac{(\bar{X}_1 - \bar{X}_2)'d}{\sqrt{d'S_u d}} \frac{n_1 n_2}{n_1 + n_2} = 1.00.$$

Since tabulated value of t with 46 d.f. at 5%, level of significance is 2.101, which is greater than calculated value of t, we conclude that data provide no evidence for rejection of null hypothesis. Therefore, we may accept it.

Case IV: Test of Equality of Two Mean Vectors when Covariance Matrices are Unequal

Let $X_1(n_1 \times p)$ and $X_2(n_2 \times p)$ be two data matrices, where $X_i[i = 1, 2]$ is a random sample from $N_p(\mu_i, \Sigma_i)$. We need to test the significance of the null hypothesis

$$H_0 : \mu_1 = \mu_2$$

under the condition, that $\Sigma_1 \neq \Sigma_2$ and both are unknown.

The test of such a hypothesis in univariate case is known as Fisher-Behrens problem. Yao (1965) has suggested a test statistic to test the hypothesis. The basis of his suggestion is based on the solution of Fisher-Behrens problem.

Let $\bar{X}_i$ denote the sample mean vector for the data matrix $X_i(n_i \times p)$ and S_i denote the sample covariance matrix $(i = 1, 2)$. It is well known that $\bar{X}_i \sim N_p(\mu_i n_i^{-1}\Sigma)$ and $S_i \sim W_p(n_i^{-1}\Sigma_i, n_i - 1)\ i = 1, 2$. Let $\bar{X}_d = \bar{X}_1 - \bar{X}_2$, $U_i = S_i/(n_i - 1)$, $U = U_1 + U_2$ and $\Sigma = (\Sigma_1/n_1) + (\Sigma_2/n_2)$. It is easy to verify that U is and unbiased estimator of Σ. Under $H_0 \bar{X}_d \sim N_p(0, \Sigma)$. Let $fU \sim W_p(\Sigma, f)$, where $\bar{X}_d$ and f are independent and the value of f is to be selected in such a way that for any p-dimensional vector α

$$fa'Ua \sim a'\Sigma a\chi_f^2.$$

Then $W_a = t_\alpha^2(f) = (a'\bar{X}_d)^2 / a'Ua$

Here $t_\alpha^2(f)$ is Student's t-statistic with f d.f.

Thus, according to UIT, it can be written that

$$\max_\alpha W_a = W_\alpha^* = \bar{X}_d' U^{-1} \bar{X}_d \sim T^2(p, f)$$

where for the maximum the value of α is $a^* = U^{-1}\bar{X}_d$. Welch has shown that if fU does not always follow $W_p(\Sigma, f)$ even then $W_a \sim t_\alpha^2(f_a)$ where

$$\frac{1}{f_a} = \frac{1}{(n_1 - 1)} \left(\frac{\alpha' U_1 \alpha}{\alpha' U \alpha} \right)^2 + \frac{1}{(n_2 - 1)} \left(\frac{\alpha' U_2 \alpha}{\alpha' U \alpha} \right)^2.$$

Now, replacing α by α^* it can be said that $\bar{X}_d U^{-1}$ is approximately distributed as $T^2(p\ f^*)$, where

$$\frac{1}{f_a} = \frac{1}{(n_1 - 1)} \left(\frac{\bar{X}_d' U^{-1} U_1 U^{-1} \bar{X}_d}{\bar{X}_d' U^{-1} \bar{X}_d} \right)^2 + \frac{1}{(n_2 - 1)} \left(\frac{\bar{X}_d' U^{-1} U_2 U^{-1} \bar{X}_d}{\bar{X}_d' U^{-1} \bar{X}_d} \right)^2.$$

Remark: The solution of multivariate extension of Fisher-Behren problem can also be obtained as per suggestion of Bennet (1951). Let us assume that $n_1 < n_2$. Now consider the data matrix X_2 to be partitioned as

$$X_2 = \begin{bmatrix} X_{21} \\ X_{22} \end{bmatrix},$$

Here X_{21} is $(n_1 \times p)$ and X_{22} is $((n_2 - n_1) \times p)$. Let us define Z as a linear combination of X_1 and X_2 such that

$$Z = X_1 - X_{21}B - \frac{1}{n_2} X_{22}\, 1_{(n_2-n_1)n_1}$$

where $B = \sqrt{\frac{n_1}{n_2}}\left(I - \frac{1}{n_1} 1_{n_1 n_1}\right) + \frac{1}{n_2} 1_{n_1 n_1}$ and 1_{nn} is a $(n \times n)$ matrix of elements unity.

The test statistic for the null hypothesis $H_0 : \mu_1 = \mu_2$ is

$$n_1(n_1 - 1)\bar{Z}' S_z^{-1} \bar{Z} \sim T^2(p, n_1 - 1)$$

where $S_z = V\left(I - \frac{1}{n_1} 1_{n_1 n_1}\right)V'$ and $V = X_1 - \sqrt{\frac{n_1}{n_2}} X_{21}$. The above T^2-statistics is transformed to F-statistics by defining

$$F = \frac{(n_1 - p)T^2(p, n_1 - 1)}{(n_1 - 1)p} \sim F_{p,\, n_1 - p}.$$

The problem can also be solved using the technique of James (1954).

Example 6: Consider the following summary data

$$\bar{X}_1 = [27.41 \quad 778.04 \quad 782.68]'$$

$$\bar{X}_2 = [31.59 \quad 776.25 \quad 769.82]$$

$$S_1 = \begin{bmatrix} 20.76 & -27.15 & -165.38 \\ -24.15 & 3548.82 & 2286.51 \\ -165.38 & 2286.51 & 4006.82 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 19.21 & -20.24 & -112.41 \\ -20.24 & 3660.04 & 2847.54 \\ -112.41 & 2847.54 & 3854.43 \end{bmatrix}$$

Solution: We wish to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ assuming that $\Sigma_1 \neq \Sigma_2$. We may calculate $d = [-4.18 \;\; 1.79 \;\; 12.86]'$

$$U_2 = \begin{bmatrix} 0.77 & -1.01 & -6.13 \\ -1.01 & 131.44 & 84.69 \\ -6.13 & 84.69 & 148.38 \end{bmatrix}$$

$$U_3 = \begin{bmatrix} 0.71 & -0.75 & -4.16 \\ -0.75 & 135.56 & 105.46 \\ -4.16 & 105.46 & 142.76 \end{bmatrix}$$

$$U = \begin{bmatrix} 1.48 & -1.76 & -10.29 \\ -1.76 & 267.00 & 190.15 \\ -10.29 & 190.15 & 291.14 \end{bmatrix}$$

$$T^2(p, f^*) = d^1\ U^{-1}\ d$$

$$= \begin{bmatrix} -4.18 & 1.79 & 12.86 \end{bmatrix} \begin{bmatrix} 1.0593 & -0.0368 & 0.0615 \\ -0.0368 & 0.0083 & -0.0067 \\ 0.0615 & -0.0067 & 0.0100 \end{bmatrix} \begin{bmatrix} -4.18 \\ 1.79 \\ 12.86 \end{bmatrix}$$

$$= 13.8190.$$

Hence $p = 3f^* = 54$. Therefore

$$F(p, f^* - p + 1) = \frac{(f^* - p + 1)}{pf^*} T^2\ (p, f^*)$$

$$F_{(3,\,52)} = 4.44.$$

Tabulated value of F with (3,52) degrees of freedom at 5% level of significance is 3.73. Since the calculated value is greater than the tabulated value at 5% level of significance, we may reject H_0 and conclude that the two mean vectors differ significantly.

Now try the following exercise.

E5) Two samples of size 50 bars and 60 bars were taken from the lots produced by method 1 and method 2. Two characteristics X_1 = lather and X_2 = mildness were measures. The summary statistics for bars produced by methods 1 and 2 is given by

$$\overline{X}_1 = \begin{bmatrix} 8 \\ 4 \end{bmatrix}, \qquad S_1 = \begin{bmatrix} 2 & 1 \\ 1 & 5 \end{bmatrix}$$

$$\overline{X}_2 = \begin{bmatrix} 10 \\ 4 \end{bmatrix}, \qquad S_2 = \begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix}$$

Test at 5% level of significance whether $\mu_1 = \mu_2$ or not.

Now let us summarize the unit.

21.6 SUMMARY

In this unit, we have covered the concepts of testing of hypothesis regarding Population Mean vector under following situations:

1. Test for mean vector for one sample case when population covariance is known.
2. Test for mean vector for one sample case when population covariance is unknown.
3. Test for equality of mean vectors for two independent sample cases when population covariances are known.
4. Test for equality of mean vectors for two independent sample case when population covariances are equal but unknown.
5. Test for equality of mean vectors for two independent sample case when population covariances are unequal and unknown,

21.7 SOLUTIONS/ANSWERS

E1) $H_0 : \mu = [40 \quad 10]'$ (say μ_0)

$H_A : \mu \neq [40 \quad 10]'$.

Here Σ is known. Hence

$$\chi_P^2 = n\left(\overline{X} - \mu_0\right)' \Sigma^{-1}\left(\overline{X} - \mu_0\right)$$

from the given data

$$\overline{X} = [33 \quad 7]'$$

$$\chi_p^2 = 10[-7 \quad -3]\left(-\frac{1}{22}\right)\begin{bmatrix} 6 & -10 \\ -10 & 13 \end{bmatrix}\begin{bmatrix} -7 \\ -3 \end{bmatrix}$$

$$= 4.1$$

for p = 2, the tabulated value of χ^2 is 10.60.

Since $\chi^2_{p,tab.} < \chi^2_{p,cal.}$ therefore, we cannot reject H_0 at 5% level of significance.

E2) If Σ is not given, then the sample covariance matrix for the data is

$$S = \begin{bmatrix} 112 & 15.9 \\ 15.9 & 3.6 \end{bmatrix}.$$

The test statistic,

$$T^2(p, n-1) = (n-1)\left(\overline{X} - \mu_0\right)' S^{-1}\left(\overline{X} - \mu_0\right)$$

$$= 9[-7 \quad -3]\begin{bmatrix} .024 & -0.106 \\ -0.106 & 0.747 \end{bmatrix}\begin{bmatrix} -7 \\ -3 \end{bmatrix}$$

$$= 30.78$$

$$F_{(p,\, n-p)} = \frac{n-p}{(n-1)p} T^2(p, n-1)$$

$$= \frac{10-2}{9 \times 2} \times 30.78$$

$$= 13.68$$

The tabulated value F(2, 8) at $\alpha = 0.05$ is 4.46, which is less than the calculated value 13.68. Therefore, we reject H_0 at 5% level of significance.

E3) $T^2 = (n-1)\left(\overline{X} - \mu_0\right)' S^{-1}\left(\overline{X} - \mu_0\right)$

Here, $\mu_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$$F = \frac{n-p}{(n-1)p} T^2(p, n-1).$$

E4) Using SS Test, we get

$$\overline{X} = [7 \quad 3]'$$

$$S = \begin{bmatrix} 112 & 15.9 \\ 15.9 & 3.6 \end{bmatrix}$$

$$XX' = \begin{bmatrix} 12010 & 2469 \\ 2469 & 526 \end{bmatrix}$$

$$d = \left[\frac{1}{\sqrt{12010}} \quad \frac{1}{\sqrt{526}} \right]' = [109.6, 22.9]'$$

And

$Z_i = 4567, 4832.2, 3863.8, 1891, 5709, 2722, 4192.6,.4786.4, 2900.3, 2306.5.$

$\bar{Z} = 3776.8$

$$t = \frac{\sqrt{n}\bar{Z}}{S_Z}.$$

E5) $H_0 : \mu_1 = \mu_2$

and $H_1 : \mu_1 \neq \mu_2$

$\Sigma_1 = \Sigma_2 = \Sigma$ (not known)

from the data given,

$$S\mu = \frac{n_1 S_1 + n_2 S_2}{n_1 + n_2} = \frac{1}{110}\begin{bmatrix} 220 & 110 \\ 110 & 610 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 1 \\ 1 & 5.56 \end{bmatrix}$$

$$S\mu^{-1} = \begin{bmatrix} 0.556 & -0.1 \\ -0.1 & 0.2 \end{bmatrix}$$

$$D^2 = (\bar{X}_1 - \bar{X}_2) S\mu^{-1} (\bar{X}_1 - \bar{X}_2)$$

$$= \begin{bmatrix} -2 \\ 0 \end{bmatrix}' \begin{bmatrix} 0.556 & -0.1 \\ -0.1 & 0.2 \end{bmatrix}_{2\times 2} \begin{bmatrix} -2 \\ 0 \end{bmatrix}_{2\times 1} = 2.2$$

$$T^2 = \frac{n_1 n_2}{n} D^2 = 60$$

$$F = \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2$$

$$= \frac{117}{118 \times 2} \times 60 = 29.50$$

21.8 PRACTICAL ASSIGNMENT

Session 6

1. Write a programme in 'C' language to test the equality of mean vectors when covariance matrices are equal.

UNIT 22 INFERENCE-IV

Structure **Page No**

22.1 INTRODUCTION

Let $X(n \times p)$ be a data matrix observed for a random sample of size n from a multivariate normal population $N_p(\mu, \Sigma)$ where μ is the population mean vector and Σ is the population covariance matrix. In Unit 21, we have studied certain tests for testing the hypothesis regarding the population mean vector under the situation that covariance matrix is known as well as when it is unknown. We have also discussed tests for testing the equality of two mean vectors when two independent samples are available from multivariate normal populations. You may recall that the tests described there do depend on the assumption about the covariance matrices, e.g. whether the covariance matrices are known, unknown but equal, and unequal and unknown. Therefore, before performing the test for the mean vectors, we have to test certain hypotheses regarding the covariance matrices. The different cases of these will be discussed in Section 22.2, 22.3 and 22.4. In Section 22.5, we shall discuss the test for independence of components of multivariate random variables. In addition to the testing of hypotheses about the covariance matrices, another interest would be to test whether a number of independent samples available to us have come from same multivariate normal population or not i.e., we would like to test the homogeneity of covariance matrices which is done in Section 22.6 and the complete homogeneity of samples is done in Section 22.7. Tests for correlation matrices may be another interest which we shall discuss in Section 22.8. In this unit, we shall consider all these problems and corresponding tests would be developed.

Objectives

After reading this unit, you should be able to:

- formulate the null and alternation hypothesis for covariance matrices;
- derive a test statistics for testing the hypothesis regarding covariance matrices;
- use the tests for given data;
- test the homogeneity of covariance matrices;
- test for a correlation matrix.

22.2 TEST FOR EQUALITY OF ALL VARIANCES AND COVARIANCES

Let $X(n \times p)$ is a sample of size n drawn from a Normal population $N_p(\mu, \Sigma)$. Suppose we wish to test a null hypothesis that all variances and covariances are equal, i.e. matrix Σ is of a special form as given below:

$$\sigma^2 \begin{bmatrix} 1 & \rho & .... & \rho \\ \rho & 1 & .... & \rho \\ .... & .... & .... & .... \\ \rho & \rho & .... & 1 \end{bmatrix} = \sum_0 \quad \text{(say)} \tag{1}$$

The alternative hypothesis is that the covariance matrix $\Sigma \neq \Sigma_0$ is a general positive definite matrix.

We know that the likelihood function for the sample observation is

$$L(\mu, \Sigma|X) = \frac{\left|\Sigma^{-1}\right|^{n/2}}{(2\pi)^{np/2}} \exp\left[-\frac{1}{2}\sum_{\alpha=1}^{n}(X_\alpha - \mu)' \Sigma^{-1}(X_\alpha - \mu)\right]. \tag{2}$$

Let us denote by $X_{\alpha i}$, the i^{th} $(i = 1, 2, \dots p)$ component of the α^{th} observation $(\alpha = 1, 2, \dots n)$ in the sample, $X_\alpha = (X_{\alpha 1}, \cdots, X_{\alpha p})'$, $\mu = [\mu_1, \mu_2, \dots \mu_p]'$ and $A = ((Aij)) = \Sigma^{-1}$. Thus, Eqn. (2) can be rewritten as

$$L(\mu, \Sigma|X) = \frac{|A|^{n/2}}{(2\pi)^{np/2}} \exp\left[-\frac{1}{2}\sum_{\alpha=1}^{n}\sum_{i,j=1}^{p} A_{ij}(X_{\alpha i} - \mu_i)(X_{\alpha j} - \mu_j)\right] \tag{3}$$

Under the null hypothesis $H_0 : \Sigma = \Sigma_0$, the matrix A should have a specific form having all its diagonal terms equal and similarly all its non-diagonal terms also equal to each other i.e.

$$A = \begin{bmatrix} A_{11} & A_{12} & .... & A_{1p} \\ A_{21} & A_{22} & .... & A_{2p} \\ .... & .... & .... & .... \\ A_{p1} & A_{p2} & .... & A_{pp} \end{bmatrix} = \begin{bmatrix} a & b & .... & b \\ b & a & .... & b \\ .... & .... & .... & .... \\ b & b & .... & a \end{bmatrix} \tag{4}$$

where $a = \dfrac{1 + (p-2)\rho}{\sigma^2(1-\rho)(1+(p-1)\rho}$ and $b = \dfrac{-\rho}{\sigma^2(1-\rho)(1+(p-1)\rho}$.

We know that the likelihood ratio test (LRT) is based on the ratio of maximum of likelihood function under the null hypothesis and over the whole space respectively i.e. the likelihood ratio for the present case is defined as

$$\lambda = \frac{\max\limits_{\mu, \Sigma_0} L(\mu, \Sigma|X)}{\max\limits_{\mu, \Sigma} L(\mu, \Sigma|X)} \tag{5}$$

The denominator of on right hand side of Eqn. (5) i.e. $\max\limits_{\mu,\Sigma} L(\mu, \Sigma|X)$, can be obtained by choosing the values of μ_i and A_{ij} such that it maximizes r.h.s. of Eqn. (3) and, hence, the values of μ_i and A_{ij} must satisfy the following $(p^2 + 3p)/2$ equations.

$$\frac{\partial L(\mu, \Sigma|X)}{\partial \mu_i} = 0 \quad i = 1, 2, \cdots, p \tag{6}$$

$$\frac{\partial L(\mu, \Sigma|X)}{\partial A_{ij}} = 0, i, j, = 1, 2, \ldots, p(i \le j). \tag{7}$$

These equations are respectively

$$n\left[\sum_{j=1}^{p} A_{ij}(\bar{X}_j - \mu_j)\right] L(\mu, \Sigma|X) = 0, \quad i = 1, 2, \ldots, p \tag{8}$$

$$\left[(n/2)A^{ij} - (1/2)\sum_{\alpha=1}^{n}(X_{\alpha i} - \mu_i)(X_{\alpha j} - \mu_j)\right]$$

$$L(\mu, \Sigma|X) = 0, i, j, = 1, 2, \ldots, p(i \le j) \tag{9}$$

where A^{ij} is the element in the i^{th} row and j^{th} column of $A^{-1} = \Sigma$, i.e., $A^{ij} = \rho i j \sigma_i \sigma_j$ and $\bar{X}_i = (1/n)\sum_{\alpha=1}^{n} X_{\alpha i}$. The solution of Eqn. (8) and Eqn. (9) is

$$\mu_i = \bar{X}_i \; i = 1, 2, \ldots, p \tag{10}$$

$$A^{ij} = s_{ij} \text{ or } A_{ij} = s^{ij} \; i, j, = 1, 2, \ldots, p(i \le j) \tag{11}$$

where $s_{ij} = (1/n)\sum_{\alpha-1}^{n}(X_{\alpha i} - \bar{X}_i)(X_{\alpha i} - \bar{X}_j), ((s_{ij})) = S$ (hence Eqn. (11) can also be written as $\Sigma = S$) and $((s_{ij}))^{-1} = ((s^{ij}))$. Substituting the values from Eqn.(10) and Eqn.(11) in Eqn.(3) and noting that the exponent in Eqn. (3) reduces to $\left(\frac{n}{2}\sum_{i,j=1}^{p} s^{ij}s_{ij}\right)$, which in turn reduces to $(-1/2)np$, since $\sum_{i=1}^{p} s^{ij}s_{ij} = 1$ for each value of j, we obtain

$$\max_{\mu,\Sigma} L(\mu, \Sigma|X) = \frac{\exp(-np/2)}{|S|^{n/2}(2\pi)^{np/2}}. \tag{12}$$

In order to obtain the numerator of right hand side of Eqn. (5) i.e. $\max_{\mu,\Sigma_0} L(\mu, \Sigma|X)$, we find μ_i's and A_{ij}'s so as to maximize the likelihood function in Eqn. (3) subject to the condition that $\Sigma = \Sigma_0$ i.e. the matrix A has the form given in Eqn. (4). Noting that the determinant A reduces to $\left[(a-b)^{p-1}(a + (p-1)b)\right]$, Eqn. (3) reduces to the following form

$$L^*(\mu, \Sigma|X) = \frac{[(a-b)^{p-1}(a+(p-1)b)]^{n/2}}{(2\pi)^{np/2}}$$

$$\exp\left[-\frac{1}{2}\left\{a\sum_{\alpha=1}^{n}\sum_{i=1}^{p}(X_{\alpha i} - \mu_i)^2 + b\sum_{\alpha=1}^{n}\sum_{i \ne j=1}^{p}(X_{\alpha i} - \mu_i)(X_{\alpha j} - \mu_j)\right\}\right] \tag{13}$$

The values of μ_i's, a and b which maximize L* are obtained by solving the following $(p+2)$ equations;

$$\frac{\partial L^*(\mu, \Sigma|X)}{\partial \mu_i} = 0, \quad i = 1, 2, \ldots, p;$$

$$\frac{\partial L^*(\mu, \Sigma|X)}{\partial a} = 0$$

and

$$\frac{\partial L^*(\mu, \Sigma|X)}{\partial b} = 0 \tag{14}$$

These equations are respectively

$$\left[(a-b)\sum_{\alpha=1}^{n}(X_{\alpha i}-\mu_i)+b\sum_{\alpha=1}^{n}\sum_{j=1}^{p}(X_{\alpha j}-\mu_j)\right]L^*(\mu,\Sigma|X)=0,\ i=1,2,\cdots,p \tag{15}$$

$$\left[\frac{n}{2}\left\{\frac{p-1}{a-b}+\frac{1}{a+(p-1)b}\right\}-\frac{1}{2}\sum_{\alpha=1}^{n}\sum_{i=1}^{p}(X_{\alpha i}-\mu_i)^2\right]L^*(\mu,\Sigma|X)=0 \tag{16}$$

and

$$\left[-\frac{n}{2}\left\{\frac{p-1}{a-b}-\frac{p-1}{a+(p-1)b}\right\}-\frac{1}{2}\sum_{\alpha=1}^{n}\sum_{i\neq j=1}^{p}(X_{\alpha i}-\mu_i)(X_{\alpha j}-\mu_j)\right]L^*(\mu,\Sigma|X)=0. \tag{17}$$

The above equations can further be simplified and written as follows:

$$\left[(a-b)(\overline{X}_i-\mu_i)+b\sum_{j=1}^{p}(\overline{X}_j-\mu_j)\right]=0,\ i=1,2,\ldots,p \tag{18}$$

$$\left[\left\{\frac{p-1}{a-b}+\frac{1}{a+(p-1)b}\right\}-\left\{ps^2+\sum_{i=1}^{p}(\overline{X}_i-\mu_i)^2\right\}\right]=0 \tag{19}$$

and

$$\left[\left\{\frac{p-1}{a-b}-\frac{p-1}{a+(p-1)b}\right\}+rp(p-1)s^2+\left(\sum_{i\neq j=1}^{p}(\overline{X}_i-\mu_i)(\overline{X}_j-\mu_j)\right)\right]=0 \tag{20}$$

where $r=\dfrac{\sum_{i\neq j=1}^{p}s_{ij}}{(p-1)\sum_{i=1}^{p}s_{ii}}$ and $s^2=\dfrac{\sum_{i=1}^{p}s_{ii}}{p}$. Solving Eqn. (18), we get $\mu_i=\overline{X}_i$ and substituting $\overline{X}_i$ for μ_i in Eqn. (19) and Eqn. (20), these reduce to

$$\frac{p-1}{a-b}+\frac{1}{a+(p-1)b}=ps^2 \tag{21}$$

and

$$\frac{-1}{a-b}+\frac{1}{a+(p-1)b}=rps^2 \tag{22}$$

respectively. Solving these equations for a and b, we get

$$a=\frac{1+(p-2)r}{s^2(1-r)(1+(p-1)r)} \tag{23}$$

and

$$b=\frac{-r}{s^2(1-r)(1+(p-1)r)}. \tag{24}$$

Substituting the values of μ_i, a and b in Eqn. (13), we get

$$\max_{\mu,\Sigma_0} L(\mu,\Sigma|X) = \frac{\exp\left(-\frac{1}{2}np\right)}{\left[(s^2)^p (1-r)^{p-1}(1+(p-1)r)\right]^{n/2} (2\pi)^{np/2}}. \tag{25}$$

Thus, substituting the values from Eqn. (12) and Eqn. (25) in Eqn. (5), we get the likelihood ratio for testing the hypothesis $H_0 : \Sigma = \Sigma_0$ as

$$\lambda = \left[\frac{|S|}{(s^2)^p (1-r)^{p-1}(1+(p-1)r)}\right]^{n/2} \tag{26}$$

We know that under $H_0, -2\log\lambda$ is approximately distributed as chi-square for large n. Therefore, Wilks (1946) suggested that if

$$\frac{|S|}{(s^2)^p (1-r)^{p-1}(1+(p-1)r)} = T, \tag{27}$$

$-n\log T$ would follow approximately chi-square distribution with $\{p(p+1)/2-2\}$ degrees of freedom under the null hypothesis. However Box (1949,1950) has shown that the following test statistic follows chi-square distribution with $\{p(p+1)/2-2\}$ degrees of freedom for large n:

$$\chi^2 = -\left[(n-1) - \frac{p(p+1)^2(2p-3)}{6(p-1)(p^2+p-4)}\right]\log\left(\frac{nT}{n-1}\right) \tag{28}$$

Therefore , we reject the null hypothesis if the calculated value of the test statistics as given in Eqn. (28) is greater than the tabulated value of chi-square with $\{p(p+1)/2-2\}$ degrees of freedom and at specified level of significance.

Now, let us illustrate it in the following example.

Example 1: The body dimensions of a certain species have been recorded. The information on body length (L) and body weight (W) are given in Table 1:

Table 1

Body length L (in mm)	Body weight W (in gm)
41	3.3
45	2.1
45	2.6
50	2.5
50	2.9
53	5.8
56	7.1
35	2.5
40	2.5
61	8.7

Test the hypothesis that all variances are equal and all co-variances are equal in the variance-covariance matrix for the given data.

Solution: Here, $n = 10$ and $p = 2$

$$S = \begin{bmatrix} 56.422 & 13.398 \\ 13.398 & 4.838 \end{bmatrix}$$

$$|S| = 93.463$$

Now, $T = \frac{|S|}{(s^2)^p (1-r)^{p-1}(1+(p-1)r)}$

where $r = \frac{\sum_{i\neq j=1}^{p} s_{ij}}{(p-1)\sum_{i=1}^{p} s_{ii}}$ and $s^2 = \frac{\sum_{i=1}^{p} s_{ii}}{p}$

Therefore, $s^2 = 30.630$ and $r = 0.437$

$T = 0.1231$ and $nT/(n-1) = .1367$

Substituting above in Eqn.(28) which is again given below

$$\chi^2 = -\left[(n-1) - \frac{p(p+1)^2(2p-3)}{6(p-1)(p^2+p-4)}\right]\log\left(\frac{nT}{n-1}\right)$$

We get, $\chi^2 = -\left[9 - \frac{2(2+1)^2(2x2-3)}{6(2-1)(2^2+2-4)}\right]\log(0.1367) = -\frac{15}{2}\log(0.1367) = 6.481$.

Since the tabulated value of chi-square with $[p(p+1)/2-2]$ i.e. one degree of freedom at 5% level of significance is 3.84 which is less than the calculated value of the test statistics, we may reject the null hypothesis. Therefore, we may conclude that the body length and body weight for the given data are not having the same variances and same covariances.

Remark: One would like to test the hypothesis that all the components of mean vector are also equal along with the equality of variances and covariances i.e.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p (= \mu \text{ say}), \sigma_1 = \sigma_2 = \cdots = \sigma_p \text{ and } \rho_{12} = \rho_{13} = \cdots = \rho_{p-1,p}$$

The likelihood ratio for testing the above hypothesis is defined as follows

$$\lambda^* = \frac{\max_{H_0} L(\mu, \Sigma|X)}{\max_{\mu,\Sigma} L(\mu, \Sigma|X)} \quad . \tag{29}$$

The denominator of the above is same as obtained above in Eqn.(12) i.e.

$$\max_{\mu,\Sigma} L(\mu, \Sigma|X) = \frac{\exp(-np/2)}{|S|^{n/2}(2\pi)^{np/2}} . \tag{30}$$

Under the null hypothesis H_0 the likelihood takes the following form

$$L(\mu, \Sigma|X) = \frac{[(a-b)^{p-1}(a+(p-1)b)]^{n/2}}{(2\pi)^{np/2}}$$
$$\exp\left[-\frac{1}{2}\left\{a\sum_{\alpha=1}^{n}\sum_{i=1}^{p}(X_{\alpha i}-\mu)^2 + b\sum_{\alpha=1}^{n}\sum_{i\neq j=1}^{p}(X_{\alpha i}-\mu)(X_{\alpha j}-\mu)\right\}\right] \tag{31}$$

where a and b has been defined in Eqn.(4). The value of μ, a and b, which maximizes the above likelihood, is obtained by solving the following three equations:

$$\frac{\partial L(\mu, \Sigma|X)}{\partial \mu} = 0, \frac{\partial L(\mu, \Sigma|X)}{\partial a} = 0 \text{ and } \frac{\partial L(\mu, \Sigma|X)}{\partial b} = 0 , \tag{32}$$

These equations simplify as follows

$$\left[(a-b)\sum_{\alpha=1}^{n}\sum_{i=1}^{p}(X_{\alpha i}-\mu)+b\sum_{\alpha=1}^{n}\sum_{i=1}^{p}(X_{\alpha i}-\mu)\right]L(\mu,\Sigma|X)=0,\ i=1,2,\ldots,p \qquad (33)$$

$$\left[\frac{n}{2}\left\{\frac{p-1}{a-b}+\frac{1}{a+(p-1)b}\right\}-\frac{1}{2}\sum_{\alpha=1}^{n}\sum_{i=1}^{p}(X_{\alpha i}-\mu)^2\right]L(\mu,\Sigma|X)=0 \qquad (34)$$

and

$$\left[-\frac{n}{2}\left\{\frac{p-1}{a-b}-\frac{p-1}{a+(p-1)b}\right\}-\frac{1}{2}\sum_{\alpha=1}^{n}\sum_{i\neq j=1}^{p}(X_{\alpha i}-\mu)(X_{\alpha i}-\mu)\right]L(\mu,\Sigma|X)=0 \qquad (35)$$

Solving these equations, we get

$$\mu=\overline{X} \qquad (36)$$

$$a=\frac{1+(p-2)r_0}{s_0^2(1-r_0)(1+(p-1)r_0)} \qquad (37)$$

and

$$b=\frac{-r_0}{s_0^2(1-r_0)(1+(p-1)r_0)}. \qquad (38)$$

where,

$$\overline{X}=\frac{1}{np}\sum_{\alpha=1}^{n}\sum_{i=1}^{p}X_{\alpha i}$$

$$s_{0ij}=\frac{1}{n}\sum_{\alpha=1}^{n}(X_{\alpha i}-\overline{X})(X_{\alpha j}-\overline{X})=s_{ij}+(\overline{X}_i-\overline{X})(\overline{X}_j-\overline{X})$$

$$r_0=\frac{\sum_{i\neq j=1}^{p}s_{0ij}}{(p-1)\sum_{i=1}^{p}s_{0ii}}=\frac{\sum_{i\neq j=1}^{p}s_{ij}-\sum_{i=1}^{p}(\overline{X}_i-\overline{X})}{(p-1)\sum_{i=1}^{p}s_{ii}+\sum_{i=1}^{p}(\overline{X}_i-\overline{X})^2}$$

and

$$s_0^2=\frac{1}{p}\sum_{i=1}^{p}s_{0ii}.$$

Thus, we get

$$\max_{H_0} L(\mu,\Sigma|X)=\frac{\exp\left(-\frac{1}{2}np\right)}{\left[(s_0^2)^p(1-r_0)^{p-1}(1+(p-1)r_0)\right]^{n/2}(2\pi)^{np/2}} \qquad (39)$$

Substituting the values from Eqn. (30) and Eqn. (39) in Eqn. (29) we get the likelihood ratio for testing the hypothesis H_0 as

$$\lambda^*=\left[\frac{|S|}{(s_0^2)^p(1-r_0)^{p-1}(1+(p-1)r_0)}\right]^{n/2} \qquad (40)$$

and $-n\log T^*$, where

$$T^*=\frac{|S|}{(s_0^2)^p(1-r_0)^{p-1}(1+(p-1)r_0)}, \qquad (41)$$

is distributed as a chi-square with $\left(\frac{p(p+3)}{2}-3\right)$ degrees of freedom under the null hypothesis. Therefore null hypothesis is rejected if $-n \log T^*$ greater than tabulated x_2 value with $\left\{\frac{p(p+3)}{2}-3\right\}$ degrees of freedom.

Example 2: A group of 100 students appeared in three consecutive tests and the scores obtained by them are noted down to check whether the tests can be considered as similar. If the tests are similar, it is expected that the students' score in three tests should have same means, variances and covariances. Thus, the test discussed in the above remark can be applied. From the data, the various statistics needed for the test are calculated and these are given below:

$$\overline{X_1} = 10.9900\ \overline{X_2} = 10.9300\ \overline{X_3} = 11.2600$$

$$S = \begin{bmatrix} 16.8451 & 13.5493 & 14.5826 \\ 13.5493 & 18.1099 & 13.8056 \\ 14.5826 & 13.8056 & 17.7124 \end{bmatrix} \text{and} |S| = 545.5308$$

$$s_0^2 = 17.5764,\ |S| = 545.5308,\ r_0 = 0.7984$$

Since, $n = 100$ and $p = 3$, it may be verified that the value of the test statistic T^* given in Eqn. (41) comes out to be 0.9370. Further $-n \log T^* = 6.5072$ and the tabulated value of chi-square with 6 degrees of freedom at 5% level of significance is 12.59. The calculated value of the statistics is less than the tabulated value, therefore we may conclude that the data do not provide enough evidence for the rejection of null hypothesis and hence we may accept that the scores of the students in the three tests have same means, variances and covariances.

Now try the following exercise.

E1) The body dimensions of 4 different species of insects collected during February-March, 1993 have been recorded [Bhuyan and Nair (1995)]. The information on body length (L) and body weight (W) of 20 insects of each species are given below:

Specie I		Specie II		Specie III		Specie IV	
L(mm)	W(gm)	L(mm)	W(gm)	L(mm)	W(gm)	L(mm)	W(gm)
36	1.3	41	3.3	45	1.8	27	0.8
37	1.8	45	2.1	45	1.7	28	0.5
37	1.7	45	2.6	46	1.9	35	1.6
40	2.4	45	3.8	49	2.0	30	0.5
41	2.5	45	6.9	49	2.3	45	2.8
42	2.3	45	4.4	50	2.2	42	1.3
43	2.5	50	2.5	50	2.4	45	2.4
44	2.2	50	2.9	51	2.0	40	1.3
46	2.7	50	2.6	52	2.1	45	2.8
47	3.3	50	3.1	53	2.1	48	3.4
48	2.2	52	5.3	54	3.2	45	2.9
48	2.9	53	5.8	53	2.1	48	2.4
51	3.1	54	3.3	55	3.0	45	2.8
52	3.3	55	5.9	55	2.2	48	2.9
53	3.6	56	7.1	55	2.5	44	2.4
59	2.9	57	8.0	56	3.4	45	2.3
55	4.5	35	2.5	57	3.3	45	3.1
55	3.6	40	1.7	57	3.4	42	1.7
56	5.0	40	2.5	58	3.1	50	2.4
60	4.5	61	8.7	50	2.5	52	3.7

Use the above data to test the hypothesis that variances and covariances for three species are equal.

In the following section we shall discuss the test for covariance matrix to be equal to a given matrix.

22.3 TEST FOR COVARIANCE MATRIX TO BE EQUAL TO A GIVEN MATRIX

Let $X(nxp)$ is a sample of size n drawn from a Normal population $N_p(\mu, \Sigma)$. Suppose we wish to test the null hypothesis that covariance matrix is equal to the given matrix i.e. $H_0: \Sigma \neq \Sigma_0$, against the alternative hypothesis $H_A: \Sigma \neq \Sigma_0$. It is also assumed that vector μ is unknown.

The likelihood ratio for testing the above hypothesis is as follows

$$\lambda = \frac{\max\limits_{\mu,\Sigma_0} L(\mu, \Sigma|X)}{\max\limits_{\mu,\Sigma} L(\mu, \Sigma|X)} \tag{42}$$

The likelihood function for the sample observation is given in Eqn. (2) which can be rewritten as

$$L(\mu, \Sigma | X) = \frac{|\Sigma^{-1}|^{n/2}}{(2\pi)^{np/2}}\left[\exp -\frac{1}{2}\operatorname{tr}\Sigma^{-1}\left\{\sum_{\alpha=1}^{n}(X_\alpha - \mu)(X_\alpha - \mu)'\right\}\right] \tag{43}$$

Note: The trace of a (pxp) square matrix $A = ((Aij))$ is defined as tr $A = \sum_{i=1}^{p} A_{ii}$. The power of the exponential term in Eqn. (2) is 1×1, hence it will be equal to its trace. The following are the properties of trace:

$$\operatorname{tr}(A + B) = \operatorname{tr} A + \operatorname{tr} B$$

$$\operatorname{tr} AB = \operatorname{tr} BA$$

We know that the maximum likelihood estimator of μ is $\overline{X}$ and that of Σ is S, where S is the sample variance-covariance matrix. Hence it is easy to verify that

$$\max_{\mu,\Sigma_0} L(\mu, \Sigma | X) = \frac{|\Sigma_0^{-1}|^{n/2}}{(2\pi)^{np/2}}\exp\left[-\frac{n}{2}\operatorname{tr}\Sigma_0^{-1}S\right] \tag{44}$$

and

$$\max_{\mu,\Sigma} L(\mu, \Sigma | X) = \frac{|S^{-1}|^{n/2}}{(2\pi)^{np/2}}\exp\left[-\frac{np}{2}\right] \tag{45}$$

substituting Eqn. (44) and Eqn. (45) in Eqn. (42), we have

$$\lambda = \left|\Sigma_0^{-1}S\right|^{n/2}\exp\left[-\frac{n}{2}\operatorname{tr}\Sigma_0^{-1}S + \frac{np}{2}\right]$$

and hence

$$-2\log\lambda = -n\log\left|\Sigma_0^{-1}S\right| + n\operatorname{tr}\Sigma_0^{-1}S - np \tag{46}$$

This statistics is a function of the eigen values of $\Sigma_0^{-1}S$. Let a and g be the arithmetic and geometric mean of eigen values of $\Sigma_0^{-1}S$. Then tr $\Sigma_0^{-1}S = pa$ and $\left|\Sigma_0^{-1}S\right| = g^p$. Therefore Eqn. (47) can also be written as

$$-2\log\lambda = np(a - \log(g) - 1) \tag{47}$$

It may be recalled that the asymptotic distribution of $-2\log\lambda$ is chi-square distribution with $p(p+1)/2$ degrees of freedom and hence H_0 is rejected when $-2\log\lambda$ given by Eqn.(46) (or Eqn.(47)) is greater than tabulated χ^2 value with $\frac{p(p+1)}{2}$ degrees of freedom.

Example 3: The data given in Example 1 is a part of information. For the complete data set the population variance-covariance matrix may be as given below:

$$\Sigma_0 = \begin{bmatrix} 20.421 & 2.582 \\ 2.582 & 1.838 \end{bmatrix}.$$

Let us investigate whether or not the data given in the Example 1 confirms to the hypothesis that it has been drawn from a bivariate normal population having variance covariance matrix as Σ_0, i.e. we wish to test the hypothesis $H_0 : \Sigma = \Sigma_0$ against the alternative hypothesis $H_1 : \Sigma \neq \Sigma_0$. From the Example 1 you may note that

$$S = \begin{bmatrix} 56.422 & 13.398 \\ 13.398 & 4.838 \end{bmatrix}.$$

Hence,

$$\Sigma_0^{-1}S = \begin{bmatrix} 0.0595 & -0.0836 \\ -0.0836 & 0.6615 \end{bmatrix}\begin{bmatrix} 56.422 & 13.398 \\ 13.398 & 4.838 \end{bmatrix} = \begin{bmatrix} 2.237 & 0.3927 \\ 4.0947 & 2.0972 \end{bmatrix}.$$

The calculated value of the test statistic $-n\log + n\,\mathrm{tr}\,\Sigma_0^{-1}S - np\left|\Sigma_0^{-1}S\right|$ is, therefore obtained as $-10\log(3.0835) + 10(4.3352) - 10\times 2 = 22.8529$.

Since the calculated value of the test statistics which follows chi-square distribution under null hypothesis with $p(p+1)/2$ (3 in this case) degrees of freedom, is greater than the tabulated value (7.81) of chi-square with 3 degrees of freedom at 5% level of significance, we reject H_0. Hence we conclude that covariance matrix for the length and weight data differs from the assumed covariance matrix Σ_0.

Now try an exercise.

E 2) Let us consider the data given in E1) and test the hypothesis that

$$\Sigma = \begin{bmatrix} 19.63 & 1.46 \\ 1.46 & 4.31 \end{bmatrix}.$$

In the following section we shall discuss the test for the covariance matrix to be equal to constant multiple of a given matrix.

22.4 TEST FOR COVARIANCE MATRIX TO BE EQUAL TO CONSTANT MULTIPLE OF A GIVEN MATRIX

Let X $(n\times p)$ is a sample of size n drawn from a Normal population $N_p(\mu,\Sigma)$. Let us consider that we wish to test the null hypothesis that covariance is equal to constant multiple of a given matrix i.e. $H_0 : \Sigma = k\Sigma_0$ against the alternative hypothesis $H_A : \Sigma \neq k\Sigma_0$, where k is some unknown constant. It is also assumed that μ is unknown.

The likelihood ratio for testing the above hypothesis is as follows

$$\lambda = \frac{\max_{H_0} L(\mu, \Sigma|X)}{\max_{H_A} L(\mu, \Sigma|X)}. \tag{48}$$

The likelihood function for the sample observation is given in Eqn. (43). Therefore we have

$$\log L(\mu, \Sigma|X) = -\frac{np}{2}\log 2\pi - \frac{n}{2}\log|\Sigma| - \frac{n}{2}\text{tr}\{\Sigma^{-1}S + (\overline{X}-\mu)'\Sigma^{-1}(\overline{X}-\mu)\} \tag{49}$$

Under H_0 it reduces to

$$\log L(\mu, k\,\Sigma_0|X) = -\frac{np}{2}\log 2\pi - \frac{n}{2}\log$$

$$|k\Sigma_0| - \frac{n}{2}k^{-1}\text{tr}\{\Sigma_0^{-1}S + (\overline{X}-\mu)'\Sigma_0^{-1}(\overline{X}-\mu)\} \tag{50}$$

Here $\overline{X}$ is sample mean vector $\overline{X} = \left(\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_p\right)'$.

In order to obtain the numerator of Eqn. (48), we choose the values of μ and k such that Eqn. (50) is maximized. Such a value of μ and k can be obtained by solving the following equations:

$$\frac{\partial}{\partial \mu}\log L(\mu, k\Sigma_0|X) = nk^{-1}\Sigma_0^{-1}(\overline{X}-\mu) = 0 \tag{51}$$

and

$$\frac{\partial}{\partial k}\log L(\mu, k\Sigma_0|X) = \left(-\frac{n}{2}\right)\left[\left(\frac{p}{k}\right) + \frac{1}{k^2}\text{tr}\left\{\Sigma_0^{-1}S + (\overline{X}-\mu)'\Sigma_0^{-1}(\overline{X}-\mu)\right\}\right] = 0 \tag{52}$$

Solving Eqn. (51) and Eqn. (52) simultaneously we get $\mu = \overline{X}$ and $k = \text{tr}\ \Sigma_0^{-1}S/p$.
Hence

$$\max_{H_0} L(\mu, \Sigma|X) = \frac{1}{(2\pi)^{np/2}|\hat{k}\Sigma_0|^{n/2}}\exp\left[\left(-\frac{1}{2}np\right)\right] \tag{53}$$

And

$$\max_{H_A} L(\mu, \Sigma|X) = \frac{1}{(2\pi)^{np/2}|S|^{n/2}}\exp\left[\left(-\frac{1}{2}np\right)\right] \tag{54}$$

Substituting Eqn. (53) and Eqn. (54) in Eqn. (48), we have

$$\lambda = \frac{|S|^{n/2}}{|\hat{k}\Sigma_0|^{n/2}} = \frac{|\Sigma_0^{-1}S|^{n/2}}{(\hat{k})^{np/2}} \tag{55}$$

$$\text{Thus, } -2\log\lambda = np\log\left(\frac{\hat{k}}{|\Sigma_0^{-1}S|^{1/p}}\right) = np\log\left(\frac{\text{tr}\Sigma_0^{-1}S/p}{|\Sigma_0^{-1}S|^{1/p}}\right) = np\log(a/g) \tag{56}$$

where a and g are the arithmetic mean and geometric mean of eigen values of $\Sigma_0^{-1}S$, respectively. Since $-2\log\lambda$ is asymptotically distributed as chi-square with $(p-1)(p+2)/2$ degrees of freedom, we reject null hypothesis H_0 if calculated value of the statistics from Eqn. (56) is greater than the corresponding tabulated value of chi-square with $(p-1)(p+2)/2$ degrees of freedom at specified level of significance.

Example 4: In Example 3 we have seen that the data given in Example 1 do not confirm the assumption that it has come from the population with variance-covariance matrix

$$\Sigma_0 = \begin{bmatrix} 20.421 & 2.582 \\ 2.582 & 1.838 \end{bmatrix}$$

Let us now try to test whether or not the population variance-covariance matrix is a constant multiple of Σ_0. In other words, we wish to test the null hypothesis $H_0 : \Sigma = k\,\Sigma_0$ against the alternative hypothesis $H_1 : \Sigma \neq k\,\Sigma_0$. The test described above can be used for this purpose. It may be noted from Example 1 that the sample variance-covariance matrix is

$$S = \begin{bmatrix} 56.422 & 13.398 \\ 13.398 & 4.838 \end{bmatrix}.$$

Further note from Example 3 that

$$\Sigma_0^{-1}S = \begin{bmatrix} 2.237 & 0.3927 \\ 4.0947 & 2.0972 \end{bmatrix}.$$

Here, $n = 10$ and $p = 3$ and hence, the estimate of k

$$\hat{k} = p^{-1}\mathrm{Tr}(\Sigma_0^{-1}S) = 2.1671$$

$$\text{And} = \left|\Sigma_0^{-1}S\right|^{1/p} = 1.7559$$

The value of the test statistic $np\log\ \hat{k}\big/\left(\Sigma_0^{-1}S\right)^{1/p}$ which distributed as chi square with $(p-1)(p+2)/2$ degrees of freedom is $20\log(1.2341) = 1.870$.

Since the calculated value of the test statistics is less than the tabulated value (5.99) of chi square with 2 degrees of freedom at 5% level of significance, hence we cannot reject the null hypothesis. Therefore we may conclude that the population covariance is constant multiple of matrix Σ_0.

Now try the exercise.

E 3) Let us consider E 1) and test the hypothesis that variance-covariance matrix is constant multiple of $\begin{bmatrix} 19.63 & 1.46 \\ 1.46 & 4.31 \end{bmatrix}$.

Let us now discuss the test for independence of components of variables in the following section.

22.5 TEST FOR INDEPENDENCE OF COMPONENTS OF VECTOR RANDOM VARIABLE

Let $X(n \times p)$ is a sample of size n drawn from a Normal population $N_p(\mu, \Sigma)$. Suppose that each observed vector is partitioned in two parts, one consisting of p_1 and the other p_2 components such that $p = p_1 + p_2$. The likelihood function of X can also be split into two factors, one for parameters μ_1 and Σ_{11}, and an other for μ_2 and Σ_{22}, where, $\mu = \left[\mu_1' \ \mu_2'\right]$ and $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$. Let us consider that we wish to test that the sets of variables corresponding to first p_1 components are independent to the set of variables corresponding to remaining p_2 components of the observed vector i.e. $H_0 : \Sigma_{12} = 0$, against the alternative hypothesis $H_A : \Sigma_{12} \neq 0$.

Under $H_0, \Sigma_{12} = 0$, and the MLE of Σ is $\hat{\Sigma} = \begin{bmatrix} S_{11} & 0 \\ 0 & S_{22} \end{bmatrix}$.

Also $\hat{\mu} = \begin{bmatrix} \hat{\mu}_1', \hat{\mu}_2' \end{bmatrix}' = \overline{X}$. Here MLE of μ under both H_0 and H_A is $\overline{X}$. Comparing with discussion in earlier sections we now see that the LRT for the hypothesis depends on the eigenvalues of $\hat{\Sigma}^{-1} S$, i.e. $\Sigma^{-1}S = \begin{bmatrix} S_{11} & 0 \\ 0 & S_{22} \end{bmatrix}^{-1} \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} I & S_{11}^{-1}S_{12} \\ S_{22}^{-1}S_{21} & I \end{bmatrix}$

Here tr $\Sigma^{-1}S = p$ and hence the arithmetic mean of eigenvalues is one. The geometric mean is obtained as

$$g^p = \left|\hat{\Sigma}^{-1} S\right| = |S| / (|S_{11}||S_{22}|) = \left|S_{22} - S_{21}S_{11}^{-1}S_{12}\right| / |S_{22}|$$

Now, according to the LRT criterion, the test statistics (similar to Eqn.(47)) is

$$-2 \ln \lambda = -n \ln\left(\left|S_{22} - S_{21}S_{11}^{-1}S_{12}\right| / |S_{22}|\right) = -n \ln\left|I - S_{22}^{-1}S_{21} S_{11}^{-1}S_{12}\right|$$

$$= -n \ln \prod_{i=1}^{k} (1 - \lambda_i) \qquad (57)$$

where $k = \min(p_1, p_2)$ and λ_i's are the non zero eigenvalues of $S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}$. The test statistic can also be written in terms of correlation matrix i.e.

$$-2 \ln \lambda = -n \ln\left|I - R_{22}^{-1}, R_{21}R_{11}^{-1}R_{12}\right|.$$

Finally, the test statistic can be written as:

$$\lambda^{2/n} = \prod_{i=1}^{k} (I - \lambda_i).$$

Bartlett (1954) has shown that

$$-\left[n - \frac{1}{2}(p_1 + p_2 + 3)\right] \ln \lambda^{2/n} \sim x^2_{p_1 p_2}$$

for larger n such that $n - 1 \geq p_1 + p_2$.

If $p_1 = 1$ and $p_2 = p - 1$, the test statistic takes the form of

$-2 \ln \lambda = -n \ln(1 - R^2)$

where R is the multiple correlation coefficient between the first variable and the others.

Now try the following exercise.

E 4) The following information relate with fertility level and child mortality level along with some socio-economic variable of couples of child bearing ages in a country according to their socio-economic status:

Objects	No. of ever born children x_1	No. of dead children under age 5 x_2	Level of education Father (in Complete years) x_3	Mother x_4	Duration of marriage (in years) x_5
1.	10	3	5	0	18
2.	3	0	10	5	8
3.	5	0	12	4	15

4.	2	0	14	14	6
5.	12	2	0	0	17
6.	1	0	10	0	3
7.	8	1	5	5	15
8.	7	2	0	0	18
9.	4	0	8	4	10
10.	2	0	12	10	5
11.	1	0	10	5	3
12.	5	1	5	0	12
13.	4	0	0	0	11
14.	6	2	5	0	13
15.	7	1	0	0	10
16.	4	1	6	0	12
17.	8	2	0	0	15
18.	3	1	0	0	8
19.	6	1	0	0	14
20.	5	0	6	0	12
21.	2	0	17	12	5
22.	5	0	12	12	12
23.	4	1	10	10	8
24.	6	1	8	0	14
25.	3	0	12	10	8
26.	4	0	10	5	17
27.	5	1	6	6	8
28.	10	2	6	6	19
29.	8	3	6	6	12
30.	6	1	10	6	14
31.	7	0	10	0	15
32.	6	0	16	10	14
33.	5	0	16	12	13
34.	4	0	10	10	12
35.	8	1	0	0	15
36.	7	2	0	0	13
37.	3	0	0	0	8
38.	2	0	16	16	5
39.	3	0	16	16	10
40.	1	0	17	17	4
41.	2	0	17	17	6
42.	4	1	10	10	9
43.	5	1	10	10	18
44.	4	0	10	10	12
45.	6	0	8	8	15
46.	4	0	10	10	12
47.	6	0	12	12	17
48.	5	0	10	10	15
49.	4	0	12	12	10
50.	3	0	10	10	12
51.	7	1	10	10	11
52.	8	4	6	6	12
53.	2	0	17	17	6
54.	1	0	17	12	3
55.	4	0	10	10	12
56.	5	0	12	12	16
57.	3	0	10	10	10
58.	4	0	6	6	16
59.	6	1	8	0	14

60.	5	0	6	0	15
61.	1	0	12	12	4
62.	2	0	12	10	3
63.	1	0	12	12	2
64.	2	0	10	10	5

Test for independence of number of ever born children (x_1) and number of dead children (x_2) with mother's education (x_4) and duration of marriage (x_5) using above data.

Now we shall discuss the test of homogeneity of covariance matrices.

22.6 TEST OF HOMOGENEITY OF COVARIANCE MATRICES

Suppose there are K multivariate normal populations $N_p(\mu_i, \Sigma_i)$ $i = 1, 2, \cdots, k$. Let $X_i(n_i \times p)$ be data matrix from $N_p(\mu_i, \Sigma_i)$ $[i = 1,2,\ldots,k,]$ and we want to test the hypothesis that k covariance matrices are equal i.e. $H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$ against the alternative hypothesis that all the co-variance matrices are not equal i.e. $H_A : \Sigma_i \neq \Sigma_j$ for at least one $i \neq j$, $i, j = 1, 2, \ldots, k$.

Under H_0 and H_A the MLE of μ_i is $\overline{X_i}$.

Under H_0 the MLE of common covariance matrix Σ is $\frac{w}{n}$ where $n = \sum_{i=1}^{k} n_i$ and $w = \sum n_i S_i$, and S_i is the MLE of Σ_i under H_A. Then the test statistic for the null hypothesis is

$$-2 \ln \lambda = n \ln|S| - \sum n_i \ \ln|S_i|, \tag{58}$$

where $S = n^{-1}w$. This statistic has an asymptotic Chi-square distribution with $\frac{1}{2}p(p+1)(k-1)$ degrees of freedom. Box (1949) proposed M-statistic, where

$$M = (n-k)\ln|S_u| - \sum_{i=1}^{k}(n_i - 1)\ln|S_{iu}|,$$

$$S_u = \frac{nS}{n-k} \text{ and } S_{iu} = \frac{n_i S_i}{n_i - 1} \tag{59}$$

According to Box MC^{-1} has a Chi-Square distribution with $\frac{1}{2}p(p+1)(k-1)$ degrees of freedom, where

$$C^{-1} = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)}\left[\sum_{i=1}^{k}\left\{\frac{1}{n_i - 1} - \frac{1}{n-k}\right\}\right]. \tag{60}$$

If all n_i are equal, then

$$C^{-1} = 1 - \frac{(2p^2 + 3p - 1)(k+1)}{6(p+1)(n-k)k} \tag{61}$$

The above MC^{-1} statistic is known as Box's M-test statistic. The approximation of MC^{-1} statistic is good if each n_i, exceeds 20, and p and k do not exceed 5. For p and k greater than $5, n_i < 20$, Box has suggested an asymptotic F distribution [Pearson and Hartley (1972)]. Roy, Pillai and others have proposed alternative test procedures for $k = 2$.

Example 5: To illustrate the above test procedure let us consider the data on the body dimension of three species X_1, X_2 and X_3. The body length and body weight for these are given below:

X_1		X_2		X_3	
Body length L (mm)	Body weight W (gm)	Body length L (mm)	Body weight W (gm)	Body length L (mm)	Body weight W (gm)
41	3.3	45	1.8	27	0.8
45	2.1	46	1.9	35	1.6
45	2.6	49	2.3	30	0.5
50	2.5	50	2.4	45	2.8
50	2.9	52	2.1	42	1.3
53	5.8	53	2.1	40	1.3
56	7.2	54	3.2	48	3.4
35	2.5	53	2.1	45	2.9
40	2.5	58	3.1	44	2.4
61	8.7	50	2.5	42	1.7

We wish to test the homogeneity of variance-covariance matrices for the three species i.e., our null hypothesis is $H_0 : \Sigma_1 = \Sigma_2 = \Sigma_3$ against the alternative hypothesis H_A : At least two out of the three Σ_i 's differ. Here $k = 3, n_i = 10(i = 1,2,3)$ and $p = 2$. From the data it is easy to calculate

$$S_{1u} = \begin{bmatrix} 56.422 & 13.398 \\ 13.398 & 4.838 \end{bmatrix}$$

$$S_{2u} = \begin{bmatrix} 13.40 & 1.20 \\ 1.20 & 0.20 \end{bmatrix}$$

$$S_{3u} = \begin{bmatrix} 4.160 & 5.164 \\ 5.164 & 0.832 \end{bmatrix}$$

Hence, $|S_{1u}| = 93.463, |S_{2u}| = 1.24$ and $|S_{3u}| = 9.243$, $S_u = \begin{bmatrix} 112.982 & 20.762 \\ 20.762 & 5.270 \end{bmatrix}$ and hence,

$|S_u| = 164.3545$.

Substituting the above values in Eqn. (60), we get $M = 19.1535$ and from Eqn. (62), we get $C^{-1} = 0.8395$. The value of MC^{-1} is, therefore, obtained as 16.07944. The tabulated value of chi-square with 6 degrees of freedom at 5% level of significance is 12.59. Since, calculated value of the test statistics is greater than tabulated value, we may reject the null hypothesis at 5% level of significance and conclude that the covariance matrices are not homogeneous.

Try an exercise.

E5) Consider E1) and test the homogeneity of covariance matrixes for species I, II, III and IV between their length and weight.

In the following section, test of complete homogeneity is discussed.

22.7 TEST OF COMPLETE HOMOGENEITY

Let $X_i(n_i \times p)$ be data matrix from $N_p(\mu_i, \Sigma_i)[i = 1,2,\ldots,k]$. For such data matrices the LRT to test $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ has been discussed in the previous unit i.e. Unit 21 and the test statistics is

$$\lambda = \left|I + W^{-1}B\right|^{-n/2}$$

under the assumption that $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$. Again, for the hypothesis $H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$ the test statistic as discussed in Section 22.6 is

$$-2\ln\lambda_1 = (n-k)\ln|S_u| - \sum_{i=1}^{k}(n_i - 1)\ln|S_{iu}|$$

where $n = \sum_{i=1}^{k} n_i$. Now, the problem is to test the hypothesis

$H_0 : \mu_1 = \ldots. = \mu_k$ and $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$. The LRT statistic for this latter hypothesis is $-2\ln\lambda_2 = -2\ln\lambda_1 - 2\ln\lambda$

This statistic is asymptotically distributed as χ^2 with $\frac{1}{2}p(k-1)(p+3)$ degrees of freedom. On simplification the statistic stands as

$$-2\ln\lambda_2 = (n-k)\ln|W_1| - \sum_{i=1}^{k}(n_i - 1)\ln|S_{iu}| \qquad (62)$$

where $S_{iu} = \frac{n_i S_i}{n_i - 1}$ is the unbiased estimator of Σ_i and $W_1 = \frac{\Sigma(n_i - 1)S_{iu}}{n-k}$.

After this we shall discuss the test of correlation matrix.

22.8 TEST OF CORRELATION MATRIX

Let $X(n \times p)$ be a data matrix from $N_p(\mu, \Sigma)$. Assume that the population correlation matrix of p variable is P. We need to test of hypothesis $H_0 : P = I$ against $H_A : P \neq I$ The null hypothesis indicates that all the p variable is uncorrelated. Under this hypothesis, the mean and variance of each variable are estimated separately. The ML estimates provide $\hat{\mu} = \overline{X}$ and $\hat{\Sigma} = \text{diag}(s_{11} \ldots. s_{pp})$, where s_{ii} is the sample variance of i^{th} variable. Let the sample correlation matrix be R.

The above hypothesis is equivalent to $H_0 : \Sigma = \text{diag}(\sigma_{11}\, \sigma_{22} \ldots \sigma_{pp})$, a special form of the Σ matrix. The LRT statistic for such a hypothesis is equivalent that of $H_0 : \Sigma = \Sigma_0$. Now, using R matrix instead of S matrix, the LRT statistic for the required hypothesis is

$$-2\ln\lambda = -n\ln|R| \qquad (63)$$

where $-2\ln\lambda$ is asymptotically distributed as Chi-square with $\frac{1}{2}p(p-1)$ degrees of freedom. Box (1949) has shown that if n is replaced by $n' = n - \frac{1}{6}(2p+11)$, then chi-square approximation is improved.

Example 6: In an investigation of the relationship between four variables X_1, X_2, X_3 and X_4, the sample correlation matrix is obtained as given below:

$$R = \begin{bmatrix} 1.000 & 0.715 & -0.526 & 0.807 \\ 0.715 & 1.000 & -0.432 & 0.396 \\ -0.526 & -0.432 & 1.000 & -0.441 \\ 0.8.7 & 0.396 & -0.441 & 1.000 \end{bmatrix}.$$

Now, we want to test the hypothesis H_0 : The variables are pair wise uncorrelated against the alternative hypothesis H_A : Not all pairs of variables are not uncorrelated. It may be seen that $|R| = 0.098$.

Therefore, the value of the test statistics given in Eqn. (64), taking n as suggested by Box(1949), is

$$-2\ln\lambda = \left[64 - \frac{1}{6}(2x4 + 11)\right]\ln(0.098) = 141.30$$

The tabulated value of chi-square with 6 degrees of freedom at 5% level of significance is 12.59 is less than the calculated value. Therefore, we may reject the null hypothesis and conclude that the variables are correlated.

Try an exercise.

E6) Consider the data given in E 4) and test whether ever born children (x_1), number of dead children (x_2), mother's education (x_4) and duration of marriage (x_5) are uncorrelated or not.

Now let us summarize the unit.

22.9 SUMMARY

In this unit we have covered the following points:.

1. We tested the equality of all variances and covariances based on a multivariate random sample from normal population. That is, to investigate the significance of the difference between the variances and covariances.

   Level of Significance (α) and Critical Region is

   $$\chi^2 > \chi^2 a(\alpha) \text{ such that } P\{\chi^2 > \chi^2 a(\alpha)\} == \alpha \text{ and } d = \left\{\frac{1}{2}p(p+1) - 2\right\}$$

   Test Statistic

   $$\chi^2 = -\left[(n-1) - \frac{p(p+1)^2(2p-3)}{6(p-1)(p^2+p-4)}\right]\log\left(\frac{nT}{n-1}\right)$$

   where,

   $$T = \frac{|S|}{(s^2)^p(1-r)^{p-1}(1+(p-1)r)}$$

   The Statistic χ^2 follows χ^2 distribution with $\left\{\frac{1}{2}p(p+1) - 2\right\}$ degrees of freedom.

   Conclusion

   If $\chi^2 \le \chi^2_d(\alpha)$, we conclude that the data do not provide us any evidence against the null hypothesis H_0 and hence it may be accepted at $\alpha\%$ level of significance. Otherwise reject H_0 or accept H_1.

2. We tested for covariance matrix to be equal to a given matrix.

Level of Significance (α) and Critical Region is

$$\chi^2 > \chi^2 a(\alpha) \text{ such that } P\{\chi^2 > \chi^2 a(\alpha)\} = \alpha \text{ and } d = \left\{\frac{1}{2}p(p+1)\right\}$$

The Test Statistic is

$$\chi^2 = np(a - \log g - 1)$$

where $a = \text{tr } \Sigma_0^{-1}S/p$

and $g = \left|\Sigma_0^{-1}S\right|^{(1/p)}$

The Statistic χ^2 follows χ^2 distribution with $d = \left\{\frac{1}{2}p(p+1)\right\}$ degrees of freedom.

Conclusion

If $\chi^2 \le \chi^2 d(s\alpha)$, we conclude that the data do not provide us any evidence against the null hypothesis H_0 and hence it may be accepted at $\alpha\%$ level of significance. Otherwise reject H_0 or accept H_A.

3) We tested for covariance matrix to be equal to a constant multiple of given matrix.

Level of Significance (α) and Critical Region is

$$\chi^2 > \chi^2{}_d(\alpha) \text{ such that } P\{\chi^2 > \chi^2{}_d(\alpha)\} = \alpha \text{ and } d = \left\{\frac{1}{2}(p-1)(p+2)\right\}$$

The Test Statistic is

$$\chi^2 = np \log(a_0 / g_0)$$

where $a_0 = \hat{k}, g_0 = a_0 g$ and $g = \left|\Sigma_0^{-1}S\right|^{(1/p)}$

The Statistic χ^2 follows χ^2 distribution with $d = \left\{\frac{1}{2}(p-1)(p+2)\right\}$ degrees of freedom.

Conclusion

If If $\chi^2 \le \chi^2{}_d(\alpha)$, we conclude that the data do not provide us any evidence against the null hypothesis H_0 and hence it may be accepted at $\alpha\%$ level of significance. Otherwise reject H_0 or accept H_A.

4) We tested for independence of components of variable

Level of Significance (α) and Critical Region is

$$\chi^2 > \chi^2{}_d(\alpha) \text{ such that } P\{\chi^2 > \chi^2{}_d(\alpha)\} = \alpha \text{ and } d = p_1 p_2$$

Test Statistic

$$\chi^2 = -n \ln \prod_{i=1}^{k}(1-\lambda_i)$$

where $k = \min(p_1, p_2)$ and λ_i's are non zero eigen values of $S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}$.

The statistic χ^2 follows χ^2 distribution with $d = p_1 p_2$ degrees of freedom provided $n - 1 > p_1 + p_2$

Conclusion

If $\chi^2 \le \chi^2_d(\alpha)$, we conclude that the data do not provide us any evidence against the null hypothesis H_0 and hence it may be accepted at $\alpha\%$ level of significance. Otherwise reject H_0 or accept H_A.

5) We tested homogeneity of covariance matrices.

Level of Significance (α) and Critical Region is

$$\chi^2 > \chi^2_d(\alpha) \text{ such that } P\{\chi^2 > \chi^2_d(\alpha)\} = \alpha \text{ and } d = \frac{1}{2}(p(p+1)(k-1))$$

Test Statistic

$$\chi^2 = MC^{-1}$$

where $M = (n - k)\ln|S_u| - \sum_{i=1}^{k}(n_i - 1)\ln|S_{iu}|$

$$S_u = \frac{nS}{n-k} \qquad S_{iu} = \frac{n_i S}{n_i - k}$$

The Statistic χ^2 follows χ^2 distribution with $\frac{1}{2}(p(p+1)(k-1))$ degrees of freedom.

Conclusion

If $\chi^2 \le \chi^2_d(\alpha)$, we conclude that the data do not provide us any evidence against the null hypothesis H_0 and hence it may be accepted at $\alpha\%$ level of significance. Otherwise reject H_0 or accept H_A.

6) We tested for correlation matrix.

Level of Significance (α) and Critical Region is

$$\chi^2 > \chi^2_d(\alpha) \text{ such that } P\{\chi^2 > \chi^2_d(\alpha)\} = \alpha \text{ and } d = \frac{1}{2}(p(p-1))$$

Test Statistic

$$\chi^2 = -n\ln|R|$$

The Statistic χ^2 follows χ^2 distribution with $\frac{1}{2}(p(p-1))$ degrees of freedom.

Conclusion

If $\chi^2 \le \chi^2_d(\alpha)$, we conclude that the data do not provide us any evidence against the null hypothesis H_0 and hence it may be accepted at $\alpha\%$ level of significance. Otherwise reject H_0 or accept H_A.

22.10 SOLUTIONS/ANSWERS

E1) Here $n = 2, p = 2$. Using the formula given in Section 22.2, we may get

$$S = \begin{bmatrix} 55.422 & 6.192 \\ 6.192 & 0.948 \end{bmatrix}$$

$$s^2 = 28.125$$

$$s_r^2 = 6.192$$

$$|S| = 14.1992$$

$$r = 0.8542$$

$$\frac{nT}{n-1} = 0.0661$$

$$\chi^2 = 47.54.$$

Conclusion

Since $x^2 < x^2_{0.05,1}$, H_0 is rejected at 5% level of significance and concluded that body length and body weight of species 1 are not having same variances and covariances.

E2) $H_0 : \Sigma = \begin{bmatrix} 19.63 & 1.46 \\ 1.46 & 4.31 \end{bmatrix}$ against

$H_A : \Sigma \neq \begin{bmatrix} 19.63 & 1.46 \\ 1.46 & 4.31 \end{bmatrix}$

For the data given on species I, it may be verified that the sample variance covariance matrix $S = \begin{bmatrix} 52.65 & 5.8825 \\ 5.8825 & 0.9013 \end{bmatrix}$

$$\Sigma_0^{-1}S = \begin{bmatrix} 0.0522 & -0.0177 \\ -0.0177 & 0.2380 \end{bmatrix} \begin{bmatrix} 52.65 & 5.8825 \\ 5.8825 & 0.9013 \end{bmatrix} = \begin{bmatrix} 2.6552 & 0.2911 \\ 0.4681 & 0.1104 \end{bmatrix}$$

Hence the calculated value of test statistic can be obtained as

$$\chi^2 = 52.29.$$

Conclusion

Since calculated χ^2 is greater than tabulated χ^2 with 3 degrees of freedom at 5% level of significance therefore we reject the null hypothesis at 5% level of significance and conclude that covariance matrix for length and weight for species I differs from $\begin{bmatrix} 19.63 & 1.46 \\ 1.46 & 4.31 \end{bmatrix}$.

E 3) $H_0 : \Sigma = k\begin{bmatrix} 19.63 & 1.46 \\ 1.46 & 4.31 \end{bmatrix}$, k is unknown

against

$H_A : \Sigma \neq k\begin{bmatrix} 19.63 & 1.46 \\ 1.46 & 4.31 \end{bmatrix}$

For the data given on species I, it may be verified that the sample variance covariance matrix $S = \begin{bmatrix} 52.65 & 5.8825 \\ 5.8825 & 0.9013 \end{bmatrix}$

$$\Sigma_0^{-1}S = \begin{bmatrix} 2.6552 & 0.2911 \\ 0.4681 & 0.1104 \end{bmatrix}, n = 20 \text{ and } p = 2.$$

We may check that $\hat{k} = a_0 = p^{-1} tr \Sigma_0^{-1} S = 1.3773$ and go $= 0.3945$

Hence the calculated value of test statistic can be obtained as

$$\chi^2 = 50.01$$

Conclusion

Since calculated χ^2 is greater than tabulated χ^2 with 2 degrees of freedom at 5% level of significance therefore we reject the null hypothesis at 5% level of significance and conclude that covariance matrix for length and weight for species I is not a constant multiple of $\begin{bmatrix} 19.63 & 1.46 \\ 1.46 & 4.31 \end{bmatrix}$.

E4) $H_0 : \Sigma_{12} = 0$

against

$H_0 : \Sigma_{12} \neq 0$

where $\Sigma_{12} = \begin{bmatrix} \sigma_{14} & \sigma_{15} \\ \sigma_{24} & \sigma_{25} \end{bmatrix}$

For the data given above it may be verified that

$$S = \begin{bmatrix} 5.797 & 1.545 & -6.691 & 8.699 \\ 1.545 & 0.806 & -2.047 & 1.591 \\ -6.691 & -2.047 & 27.897 & -10.497 \\ 8.699 & 1.591 & -10.497 & 20.030 \end{bmatrix}$$

here $n = 64$ and $p_1 = p_2 = 2$.

$$S_{11} = \begin{bmatrix} 5.797 & 1.545 \\ 1.545 & 0.806 \end{bmatrix} S_{12} = \begin{bmatrix} -6.691 & 8.699 \\ -2.047 & 1.591 \end{bmatrix}$$

$$S_{21} = \begin{bmatrix} -6.691 & -2.407 \\ 8.699 & 1.591 \end{bmatrix} S_{22} = \begin{bmatrix} 27.897 & -10.497 \\ -10.497 & 20.030 \end{bmatrix}.$$

Therefore $S_{22}^{-1} S_{21} S_{11}^{-1} S_{12} = \begin{bmatrix} 0.12916 & -0.08974 \\ -0.4093 & 0.67171 \end{bmatrix}$.

The eigen values of this products matrix are $\lambda_1 = 0.7326, \lambda_2 = 0.0683$.

Thus, we have $\chi^2 = -\left[64 - \frac{1}{2}(2+2+3)\right] \ln(1-0.7326)(1-0.0683) = 84.09$

Conclusion

Since calculated χ^2 is greater than tabulated χ^2 with 4 degrees of freedom at 5% level of significance therefore we reject the null hypothesis at 5% level of significance and conclude that ever born children and no of dead children are related with the mother education and duration of marriage.

E5) Here $p = 2\, n_i = 20\, (i = 1,2,3,4)$ and $k = 4$ from the data we have

$$S_{1u} = \begin{bmatrix} 55.422 & 6.192 \\ 6.192 & 0.948 \end{bmatrix} S_{2u} = \begin{bmatrix} 44.366 & 9.928 \\ 9.928 & 4.536 \end{bmatrix}$$

$$S_{3u} = \begin{bmatrix} 15.79 & 1.81 \\ 1.81 & 0.318 \end{bmatrix} S_{4u} = \begin{bmatrix} 50.156 & 5.89 \\ 5.89 & 0.878 \end{bmatrix}$$

$$S_u = \begin{bmatrix} 41.4329 & 5.9553 \\ 5.9553 & 3.7244 \end{bmatrix}, |S_{1u}| = 14.1992$$

$|S_{2u}| = 102.679 |S_{3u}| = 1.7451$

$|S_{4u}| = 9.3449 |S_u| = 118.847$

$M = 171.66, C^{-1} = 0.9525$

Therefore $MC^{-1} = 163.51$

Conclusion

Since calculated MC^{-1} which is approximately distributed as χ^2 with 9 degrees of freedom is greater than tabulated χ^2 with 9 degrees of freedom at 5% level of significance therefore we reject the null hypothesis at 5% level of significance and conclude that the covariance matrixes for the four different species I, II, III and IV are not homogeneous.

E6) The sample correlation matrix can be easily obtained as follows

$$R = \begin{bmatrix} 1.000 & 0.715 & -0.526 & 0.807 \\ & 1.00 & -0.432 & 0.396 \\ & & 1.00 & -0.441 \\ & & & 1.00 \end{bmatrix}$$

$$|R| = 0.098$$

$$\chi^2 = -\left[64 - \frac{1}{6}(2x4 + 11)\ln(0.098)\right] = 141.30$$

Conclusion

Since calculated χ^2 with 6 degrees of freedom is greater than tabulated χ^2 with 6 degrees of freedom at 5% level of significance therefore we reject the null hypothesis at 5% level of significance and conclude that the variables are not uncorrelated.

22.11 PRACTICAL ASSIGNMENT

Session 7

1. Write a program in 'C' language to test for covariance matrix to be equal to a given matrix.

UNIT 23 APPLICATIONS OF MVN

Structure **Page No**

23.1 INTRODUCTION

The problem of classification occurs when an investigator makes a set of measurements on an individual and based on these measurements wishes to classify the individual into one of the several categories. The investigator cannot identify the individual with a category otherwise and for this purpose must use these measurements. However, he may identify a finite number of categories or populations from which the individual may have come. Each population can further be characterized by a probability distribution of the vector of measurements. Thus, the vector of measurements on the individual are considered as a random observation from one of these populations. The question is that for given individual with certain measurements, how to decide from which population did it come.

In some instances, the categories are specified beforehand in the sense that the probability distributions of the measurements are completely known. In other cases, the form of each distribution must be estimated from a sample from that population.

Let us give an example of a problem of classification. Prospective students applying for admission into college are given a battery of tests; the set of scores is a vector of measurements **x** on a student. The prospective student may be a member of one population consisting of those students who will successfully complete college training or, rather, have potentialities for successfully completing training, or he may be a member of the other population, those who will not complete the college course successfully. The problem is to classify a student applying for admission based on his scores on the entrance examination. Another example of classification may be identification of a skull as male skull or a female skull based on anthropological measurements like circumference, height, volume, etc. We shall begin the unit by defining the criterion for good classification in Section 23.2. We shall describe the procedure of classification into one of two populations with known probability distribution in Section 23.3, and in Section 23.4 with known multivariate normal populations.

In this unit, we shall also develop the theory of classification in general terms in Section 23.5 and then apply it to cases involving the normal distribution in Section 23.6.

In order to explain the application of the test to given data, practical have been provided in section

Objectives

After reading this unit, you should be able to:

- formulate a classification problem;
- choose a suitable classification rule;
- use the classification rule for given data to classify an individual into various possible groups.

23.2 CRITERION FOR GOOD CLASSIFICATION

In constructing a procedure of classification, it is desired to minimize the probability of misclassification, or, more specifically, it is desired to minimize on the average the bad effects of misclassification. Now let us make this notion precise. For convenience, we shall now consider the case of only two categories. Later we shall treat the more general case.

Suppose an individual is an observation from either population π_1 or population π_2. The classification of an observation depends on the vector of measurements $x' = (x_1, \ldots, x_p)$ on that individual. We set up a rule that if an individual is characterized by certain set of values of $x_1, \ldots, x_p$ then that individual will be classified as from π_1; if it has any other values of $x_1, x_2, \cdots, x_p$, then it is classified as from π_2.

We can think of an observation as a point in a p-dimensional space. We divide this space into two regions R_1 and R_2. If the observation x' falls in R_1, we classify it as coming from population π_1 and if it falls in R_2 we classify the observation as coming from population π_2.

Naturally, in using the above classification procedure, the statistician can make two types of misclassifications. The first may be when the individual is actually from π_1 but the statistician classifies it as coming from population π_2 and the second may be when it is from π_2 but the statistician classifies it as from π_1. It is needed at this stage to know the relative undesirability of these two types of misclassifications. Let the "cost" of the first type of misclassification be $C(2/1)(>0)$ and let the cost of second type of misclassification be $C(2/1)(>0)$. These costs may be measured in any kind of units. As we shall see later, it is only the ratio of the two costs that is important. Although the statistician may not know these costs in each case, they will often have at least a rough idea of them.

We shall consider the ways of defining "minimum cost" in two cases. In one case, we shall suppose that we have a priori probabilities of the two populations. Let the probability that an observation comes from population π_1 be q_1 and from population π_2 be q_2. The probabilistic properties of population π_1 are specified by a distribution function. For convenience we shall treat only the case that the distribution has a density, although the case of discrete probabilities lends itself to almost the same treatment. Let the density of population π_1 be $p_1(x)$. If we have classification procedure R with a region R_1 of classification as from π_1, the probability of correctly classifying an observation that actually is drawn from population π_1 is

$$P(1|1, R) = \int_{R_1} p_1(x)dx, \qquad (1)$$

where $dx = dx_1 \ldots dx_p$, and the probability of misclassification of observation from π_1 is

$$P(2|1,R) = \int_{R_2} p_1(x)dx. \tag{2}$$

where R_2 being complementary region of R_1.

Similarly, the probability of correctly classifying an observation from π_2 is

$$P(2|2,R) = \int_{R_2} p_2(x)\, dx \tag{3}$$

and the probability of misclassifying such an observation is

$$P(1|2,R) = \int_{R_1} p_2(x)\, dx \tag{4}$$

Since probability of drawing an observation from π_1 is q_1 therefore the probability of drawing an observation from π_1 and correctly classifying it is $q_1 P(1|1,R)$. Similarly, it shall be easy to see that the probability of drawing an observation from π_2 and correctly classifying it is $q_2 P(2|2,R)$ whereas $q_2 P(1|2,R)$ is the probability of drawing an observation from π_2 and misclassifying it.

Thus, the average or expected loss from costs of misclassification is

$$C(2|1)P(2|1,R)q_1 + C\,(1|2)P(1|2,R)q_2 \tag{5}$$

It is this average loss that we wish to minimize. That is, we want to divide our p-dimensional space space into regions R_1 and R_2 such that the expected loss is as small as possible. A procedure that minimizes Expression in Eqn.(5) for a given q_1 and q_2 is called a Bayes procedure.

The other case we shall treat is the case in which there are no known a priori probabilities. In this case the expected loss if the observation is from π_1 is

$$C(2/1)P(2/1,R) = r(1,R), \tag{6}$$

and the expected loss if the observation is from π_2 is

$$C(1/2)P(1/2,R) = r(2,R). \tag{7}$$

We do not know whether the observation is from π_1 or from π_2 and we do not know probabilities of these two instances.

A procedure R is at least as good as a procedure R^* if $r(1,R) \le r(1,R^*)$ and $r(2,R) \le r(2,R^*)$; R is better than R^*, if at least one of these inequalities is a strict inequality. Usually there is no procedure that is better than all other procedures or it is at least as good as all other procedures. A procedures R is called admissible if there is no procedure better than R; we shall be interested in the entire class of admissible procedures. It will be shown that under certain conditions this class is the same as the class of Bayes procedures. A class of procedures is complete if for every procedure outside the class there is one in the class, which is better and a class is called essentially complete if for every procedure outside the class there is one in the class, which is at least as good. A minimal complete class (if it exists) is a complete class such that no proper subset is a complete class; a similar definition holds for a minimal essentially complete class. Under certain conditions we shall show that the admissible class is minimal complete. To simplify the discussion we shall consider procedures the same if they only differ on sets of probability zero. In fact, throughout the next section we shall make statements which are meant to hold "except for sets of probability zero" without saying so explicitly.

A principle that usually leads to a unique procedure is the minimax principle. A procedure is minimax if the maximum of expected losses, r(i,R) i = 1, 2, is a minimum. From the conservative point of view, it may be considered as an optimum procedure.

Now, let us discuss the procedure of classification into one of two populations with known probability distributions.

23.3 PROCEDURE OF CLASSIFICATION INTO ONE OF TWO POPULATIONS WITH KNOWN PROBABILITY DISTRIBUTIONS

Let us suppose that we wish to classify an observation x into one of the two populations π_1 and π_2 having known probability distributions $p_1(x)$ and $p_2(x)$ respectively.

Case 1: When Prior Probabilities are Known

Let q_1 and q_2 be the prior probabilities of the observation belong to the population π_1 and π_2, respectively. Therefore the conditional probability that the observation x comes from the population π_i, when i = 1, 2 is

$$q_i\, p_i(x)/\{q_1\, p_1(x)+q_2\, p_2(x)\} \tag{8}$$

If we assume the cost of misclassification is same i.e. $C(1|2)=C(2|1)=1$, then the expected loss is

$$q_1 \int_{R_2} p_1(x)dx + q_2 \int_{R_1} p_2(x)dx\,. \tag{9}$$

In fact it is the probability of misclassification also; hence we wish to choose R_1 and R_2 which minimize it. In order to achieve it, it seems natural to classify x into R_1 if the conditional probability for π_1 is more than that of π_2, i.e.

$$\frac{q_1\, p_1(x)}{q_1\, p_1(x)+q_2\, p_2(x)} \geq \frac{q_2\, p_2(x)}{q_1\, p_1(x)+q_2\, p_2(x)} \tag{10}$$

Thus the classification rule defines the regions R_1 and R_2 as

$$R_1 : q_1\, p_1(x) \geq q_2\, p_2(x)$$

$$R_2 : q_1\, p_1(x) < q_2\, p_2(x) \tag{11}$$

It may be noted here that if $q_1\, p_1(x) = q_2\, p_2(x)$, then the point x could be classified as either in π_1 or in π_2.

It can be easily proved that Eqn.(11), defines the best procedure because for any procedure (R_1^*, R_2^*) the probability of misclassification is

$$q_1 \int_{R_2^*} p_1(x)dx + q_2 \int_{R_1^*} p_2(x)dx = \int_{R_2^*} [q_1p_1(x) - q_2p_2(x)]dx + q_2 \int p_2(x)dx \tag{12}$$

On the right hand side the second term is constant $q_2 > 0$ and therefore in order to minimise right hand side we should choose the points in R_2^* as those x for which $q_1p_1(x) - q_2p_2(x) < 0$, which are the points in R_2.

If we assume the cost of misclassification is not same i.e. $C(1|2) \neq C(2|1)$, we have seen in the previous section that the expected loss is

$$[C(2|1)q_1]\int_{R_2} p_1(x)dx + [C(1|2)q_2]\int_{R_1} p_2(x)dx. \quad (13)$$

Following the arguments given above, you can easily check that the classification rule now classifies the observation x into π_1 or π_2 according as it lies in the region R_1 or R_2 respectively, where R_1 or R_2 is defined below:

$$R_1 : [C(2|1)\, q_1]p_1(x) \geq [C(1|2)q_2]p_2(x)$$

$$R_2 : [C(2|1)\, q_1]p_1(x) < [C(1|2)q_2]p_2(x)$$

It can also be expressed as

$$R_1 : p_1(x)/p_2(x) \geq [C(1|2)q_2]/[C(2|1)q_1]$$

$$R_2 : p_1(x)/p_2(x) < [C(1|2)q_2]/[C(2|1)q_1] \quad (14)$$

It may be noted here that the procedure is unique except for the sets of probability zero, that is if

$$P([C(2|1)q_1]p_1(x) = [C(1|2)q_2]p_2(x) | \pi_i) = 0.$$

Case 2: When Prior Probabilities are Unknown

In many instances of classification the statistician cannot assign a priori probabilities to the two populations. In such a situation one can proceed by taking the prior probabilities to be equal and minimize the expected cost. Therefore the regions R_1 and R_2 can be defined as given below:

$$R_1 : p_1(x)/p_2(x) \geq [C(1|2)/[C(2|1)]$$

$$R_2 : p_1(x)/p_2(x) < [C(1|2)]/[C(2|1)] \quad (15)$$

If the misclassification cost is also indeterminate, it is taken to be unity. The classification rule for such a case would reduce to

$$R_1 : p_1(x)/p_2(x) \geq 1$$

$$R_2 : p_1(x)/p_2(x) < 1 \quad (16)$$

Another approach would be to use the property of admissibility of Bayes procedure to get the minimax rule. First of all let us prove the admissibility of Bayes procedure. Let $R = (R_1, R_2)$ be a Bayes procedure for a given prior probabilities (q_1, q_2). Is there a procedure $R^* = (R_1^*, R_2^*)$ such that $P(1|2,R^*) \leq P(1|2,R)$ and $P(2|1,R^*) \leq P(2|1,R)$ with at least one inequality? If it is so the R is inadmissible. However, R is Bayes procedure

$$q_1\, P(2|1,R) + q_2\, P(1|2,R) \leq q_1\, P(2|1,R^*) + q_2\, P(1|2,R^*) \quad (17)$$

This inequality can be rewritten as

$$q_1\, [P(2|1,R) - P(2|1,R^*)] \leq q_2\, [P(1|2,R^*) - P(1|2,R)] \quad (18)$$

Suppose that $0 < q_1 < 1$. Then if $P(1|2,R^*) < P(1|2,R)$, the right hand side of Eqn.(18) is less than zero and therefore $P(2|1,R) < P(2|1,R^*)$. Similarly, if $P(2|1,R^*) < P(2|1,R)$, the left hand side of Eqn.(18) is greater than zero and therefore

$P(1|2,R) < P(1|2,R^*)$. Thus R^* is not better than R and R is admissible rule. If $q_1 = 0$, it may be noted from Eqn.(18) that $0 \le [P(1|2,R^*) - P(1|2,R)]$ and from Eqn.(14) we see that Bayes procedure R_1 consists of those points for which $p_2(x) = 0$. Therefore $P(1|2,R) = 0$ and R^* will be better than R only if $P(1|2,R^*) = 0$. But, if $P(1|2,R) = 0, P(2|1,R) = P_r\{p_2(x) > 0 | \pi_1\} = 1$. Similarly, if $P(1|2,R^*) = 0, P(2|1,R^*) = P_r\{p_2(x) > 0 | \pi_1\} = 1$ and hence R^* is not better than R. We may now conclude that if $\Pr\{p_2(x) = 0 | \pi_1\} = 0$ and $\Pr\{p_1(x) = 0 | \pi_1\} = 0$ then every Bayes procedure is admissible.

Now we will prove that every admissible procedure is a Bayes procedure. We assume that

$$\Pr\left(\frac{p_1(x)}{p_2(x)} = k \mid \pi_i\right) = 0 \quad i = 1, 2; 0 \le k < \infty \tag{19}$$

Then for any q_1, the Bayes procedure is unique. Moreover, the cdf of $p_1(x)/p_2(x)$ for π_1 and π_2 is continuous. Let R be an admissible procedure. Then there exists a k such that

$$P(2|1,R) = \Pr\left(\frac{p_1(x)}{p_2(x)} \le k \mid \pi_1\right) = P(2|1,R) = P(2|1,R^*) \tag{20}$$

where R^* is the Bayes Procedure corresponding to $q_2/q_1 = k$ i.e. $q_1 = 1/(1+k)$. Since R is admissible, $P(1|2,R) \le P(1|2,R^*)$. However, as discussed above R^* is admissible, $P(1|2,R) \ge P(1|2,R^*)$ and hence $P(1|2,R) = P(1|2,R^*)$. Therefore R is also a Bayes Procedure; by uniqueness of the Bayes procedures R is same as R^*. You may note that if Eqn.(19) holds, every admissible procedure is a Bayes procedure and the class of Bayes procedures is minimal complete.

Let us now consider the minimax procedure. Let R be the Bayes procedure corresponding to q_1 and denote $P(i|j,R) = P(i|j,q_1)$. $P(i|j,q_1)$ is a continuous function of q_1. $P(2|1,q_1)$ varies from 1 to 0 and $P(1|2,q_1)$ varies from 0 to 1. Thus there is a value of q_1, say q_1^*, such that $P(2|1,q_1^*) = P(1/2,q_1^*)$ and the corresponding Bayes procedure R would be the minimax procedure. Because if there were another R^* such that $\max\{P(1|2,R^*), P(2|1,R^*)\} \le P(2|1,q_1^*) = P(1|2,q_1^*)$, this would contradict the fact that every Bayes procedure is admissible.

Example 1: A researcher has enough data available to estimate the density function $p_1(x)$ and $p_2(x)$ associated with the populations π_1 and π_2 respectively. Suppose $C(2|1) = 5$ unit and $C(1|2) = 10$ units. If it is known a priory that about 20% of the observations belongs to π_2 then the prior probabilities is $q_1 = 0.8$ and $q_2 = 0.2$.

Now with given values of cost of misclassification and prior probabilities we can make region R_1 and R_2 by the help of Eqn.(14) as

$$\begin{aligned} R_1 : p_1(x)/p_2(x) &\ge [C(1|2)q_2]/[C(2|1)q_1] \\ &= (10/5)(0.2/0.8) \\ &= 0.5 \end{aligned}$$

$$\begin{aligned} R_2 : p_1(x)/p_2(x) &< [C(1|2)q_2]/[C(2|1)q_1] \\ &= (10/5)(0.2/0.8) \\ &= 0.5 \end{aligned}$$

Suppose the probability function evaluated at new observation x_0 give $p_1(x_0) = 0.3$ and $p_2(x_0) = 0.4$. Now the problem arises whether we classify the new observation as from population π_1 or from π_2. To find the solution we calculate the ratio $p_1(x_0)/p_2(x_0)$ as given below:

$$\frac{p_1(x_0)}{p_2(x_0)} = \frac{0.3}{0.4}$$
$$= 0.75$$

which is greater than 0.5. It implies that our observation satisfy the condition of R_1 as

$$R_1 : p_1(x_0) = 0.75 > [C(1|2)q_2]/[C(2|1)q_1] = 0.50$$

Thus, we conclude that observation $x_0 \in R_1$ and classify it as belonging to π_1.

Now try the following exercise..

E1) Consider two populations π_1 and π_2 having density functions $p_1(x)$ and $p_2(x)$ respectively. The cost of assigning items as π_2 given that π_1 is true is 100 and the cost of assigning items as π_1 given that π_2 is true is 50. It is also know that 40% of all objects (for which the measurements x can be recorded) belong to π_1.

i) Find the prior probabilities.

ii) Write the cost of misclassification.

iii) Derive the classification regions R_1 and R_2.

iv) If the measurements recorded on a new item yield the density values $p_1(x) = 0.3$ and $p_2(x) = 0.5$. Assign this item to population π_1 or population π_2.

So far, we discussed the procedure of classification in general when probability distribution is known. In the following section, we shall discuss the same problem the populations have known when multivariate normal distributions.

23.4 PROCEDURE OF CLASSIFICATION INTO ONE OF TWO KNOWN MULTIVARIATE NORMAL POPULATIONS

Now we shall use the general procedure discussed in the previous section for the case of two multivariate normal populations with equal covariance matrices. Let two population be π_1 and π_2 with same variance covariance matrix Σ and mean vectors $\mu^{(1)}$ and $\mu^{(2)}$ respectively so that

$$\pi_1 \approx N_p(\mu^{(1)}, \Sigma)$$

$$\pi_2 \approx N_p(\mu^{(2)'}, \Sigma)$$

The density functions of populations π_1 and π_2 are given by

$$p_1(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp[-\frac{1}{2}(x - \mu^{(1)})'\Sigma^{-1}(x - \mu^{(1)})] \quad (21)$$

$$p_2(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp[-\frac{1}{2}(x-\mu^{(2)})'\Sigma^{-1}(x-\mu^{(2)})] \tag{22}$$

respectively.

If R_1 be the region of classification for the population π_1, then we know that

$$\frac{p_1(x)}{p_2(x)} \geq K \text{ where } K = [C(1|2)\,q_2]/[C(2|1)\,q_1]. \tag{23}$$

Now, from Eqn. (21) and Eqn. (22), we have

$$\frac{p_1(x)}{p_2(x)} = \frac{\exp[-\frac{1}{2}(x-\mu^{(1)})'\Sigma^{-1}(x-\mu^{(1)})]}{\exp[-\frac{1}{2}(x-\mu^{(2)})'\Sigma^{-1}(x-\mu^{(2)})]}$$

$$= \exp[-\frac{1}{2}(x-\mu^{(1)})'\Sigma^{-1}(x-\mu^{(1)}) + \frac{1}{2}(x-\mu^{(2)})'\Sigma^{-1}(x-\mu^{(2)})]$$

Taking log on both sides of Eqn. (23) and after substituting the value of $p_1(x)/p_2(x)$ from above, we get

$$x'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) - \frac{1}{2}(\mu^{(1)}+\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) \geq \log_e K \tag{24}$$

The first term is well known discriminant function. It is a linear function of the components of the observation vector. The classification rule can now be defined as; classify the observation x into π_1 or π_2 according as it lies in the region R_1 or in R_2 respectively where R_1 and R_2 are defined below:

$$R_1 : x'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) - \frac{1}{2}(\mu^{(1)}+\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) \geq \log_e K \tag{25}$$

and

$$R_2 : x'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) - \frac{1}{2}(\mu^{(1)}+\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) < \log_e K \tag{26}$$

where

$$K = [C(1|2)q_2]/[C(2|1)q_1]$$

Now we consider two cases, one when priori probabilities are known and another with unknown priori probabilities.

Case 1: When a Priori Probabilities are Known

Let q_1, q_2 are known and assume for sake of simplicity that the two are equal. The costs of misclassifications are also assumed to be equal. Then we can take $K = 1$ and hence $\log_e K = 0$. In this situation the region R_1 and R_2 simply to

$$R_1 : x'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) \geq \frac{1}{2}(\mu^{(1)}+\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) \tag{27}$$

and

$$R_2 : x'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) < \frac{1}{2}(\mu^{(1)}+\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) \tag{28}$$

Putting

$$\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) = d_{p\times 1}$$

we get

$$\left.\begin{aligned} R_1: \ x'd \geq \frac{1}{2}(\mu^{(1)} + \mu^{(2)})' d \\ R_2: \ x'd < \frac{1}{2}(\mu^{(1)} + \mu^{(2)})' d \end{aligned}\right\} \tag{29}$$

Case 2: When a Priori Probabilities are Unknown

If we do not know prior probabilities, then K and hence $\log_e K$ is unknown and hence it is to be treated as an unknown constant i.e. $\log K = C$ (say). Now, if we denote

$$U = x' \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right) - \frac{1}{2}\left(\mu^{(1)} + \mu^{(2)}\right)' \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right) \tag{30}$$

Then the regions of classification are

$$R_1 : U \geq C \tag{31}$$

$$R_2 : U < C \tag{32}$$

Let us suppose that X is distributed as $N(\mu^{(1)}, \Sigma)$, then U is distributed normally with mean

$$E_1(U) = E\left[x' \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right) - \frac{1}{2}\left(\mu^{(1)} + \mu^{(2)}\right)' \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)\right]$$

$$= \frac{1}{2}\left[\left(\mu^{(1)} - \mu^{(2)}\right)' \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)\right]$$

$$= \frac{1}{2}\alpha$$

where

$$\alpha = \left(\mu^{(1)} - \mu^{(2)}\right)' \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$$

is measure of distance between $N\left(\mu^{(1)}, \Sigma\right)$ and $N\left(\mu^{(2)}, \Sigma\right)$ also known as Mahalanobis D^2 distance.

Now

$$V_1(U) = E_1\left[u - E(u)\right]^2$$

$$= E_1\left[X'\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right) - \frac{1}{2}\left(\mu^{(1)} + \mu^{(2)}\right)' \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right) - \frac{1}{2}\left(\mu^{(1)} - \mu^{(2)}\right)' \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)\right]^2$$

$$= \left(\mu^{(1)} - \mu^{(2)}\right)' \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$$

$$= \alpha$$

Thus U is distributed according to $N(\alpha/2, \alpha)$ if x is distributed according to $N\left(\mu^{(1)}, \Sigma\right)$. Now if x is distributed according to $N\left(\mu^{(2)}, \Sigma\right)$

$$E_2(U) = E_2\left[X'\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right) - \frac{1}{2}\left(\mu^{(1)} - \mu^{(2)}\right)' \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)\right]$$

$$= \mu^{(2)'} \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right) - \frac{1}{2}\left(\mu^{(1)} + \mu^{(2)}\right)' \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$$

$$= \frac{1}{2}\left(\mu^{(2)} - \mu^{(1)}\right)' \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$$

$$= -\frac{1}{2}\alpha \tag{33}$$

The variance is the same as when x distributed according to $N(\mu^{(1)}, \Sigma)$ (because it depends only on the second order moments of x). So

$$V_2(U) = \alpha$$

Hence the region of classification is obtained in terms of U assuming the two populations are

$$\pi_1 : U \approx N\left(\frac{1}{2}\alpha, \alpha\right) \tag{34}$$

$$\pi_2 : U \approx N\left(-\frac{1}{2}\alpha, \alpha\right) \tag{35}$$

Now, we shall discuss the classification into one of two multivariate normal populations when the parameters are unknown and therefore must be estimated.

Let us consider a sample from each of the two normal populations and we use this information in classifying another observation as coming from one of the two populations.

Suppose that we have a sample $x_1^{(1)}, \ldots, x_{N_1}^{(1)}$ from $N(\mu^{(1)}, \Sigma)$ and a sample $x_1^{(2)}, \ldots, x_{N_2}^{(2)}$ from $N(\mu^{(2)}, \Sigma)$. On the basis of this information we wish to classify the observation x as coming from π_1 or π_2. We know that our best estimate of $\mu^{(1)}$ is $\overline{X}^{(1)}$, of $\mu^{(2)}$ it is $\overline{X}^{(2)}$ and that of Σ is S which are defined below:

$$\overline{X}^{(1)} = \frac{1}{N_1}\sum_{\alpha=1}^{N_1} x_\alpha^{(1)}$$

$$\overline{X}^{(2)} = \frac{1}{N_2}\sum_{\alpha=1}^{N_2} x_\alpha^{(2)}$$

and

$$S = \frac{1}{N_1 + N_2 - 2}\left\{\sum_{\alpha=1}^{N_1}\left(x_\alpha^{(1)} - \overline{X}^{(1)}\right)\left(x_\alpha^{(1)} - \overline{X}^{(1)}\right)' + \sum_{\alpha=1}^{N_2}\left(x_\alpha^{(2)} - \overline{X}^{(2)}\right)\left(x_\alpha^{(2)} - \overline{X}^{(2)}\right)'\right\}. \tag{36}$$

We substitute these estimates for the parameters in Eqn.(30) to obtain

$$W(x) = x'S^{-1}(\overline{X}^{(1)} - \overline{X}^{(2)}) - \frac{1}{2}(\overline{X}^{(1)} + \overline{X}^{(2)})'S^{-1}(\overline{X}^{(1)} - \overline{X}^{(2)}) \tag{37}$$

The first term of Eqn. (37) is the estimated discriminate function based on two samples. It is the linear function that has greatest variance between samples relative to the variance within samples. W(x) defined in Eqn. (37) is used in the criterion of classification. When the populations are known, we can argue the classification criterion is the best in the sense that its use minimizes the expected loss in the case of known apriori probabilities and generates the class of admissible procedures when a priori probabilities are not known. We cannot justify the use of statistic Eqn. (37) in the same way. However, it seems intuitively reasonable that should give good results.

Suppose we have a sample $x_1, \ldots, x_N$ which may be from either π_1 or π_2, but not known from which and we wish to classify them as a whole in one of them. Then we define S by

$$(N_1 + N_2 + N - 3)S = \sum_{\alpha=1}^{N_1}\left(x_\alpha^{(1)} - \overline{X}^{(1)}\right)\left(x_\alpha^{(1)} - \overline{X}^{(1)}\right)' + \sum_{\alpha=1}^{N_2}\left(x_\alpha^{(2)} - \overline{X}^{(2)}\right)\left(x_\alpha^{(2)} - \overline{X}^{(2)}\right)' +$$

$$\sum_{\alpha=1}^{N}(x_\alpha - \bar{X})(x_\alpha - \bar{X})' \quad (38)$$

where

$$\bar{X} = \frac{1}{N}\sum_{\alpha-1}^{N} x_\alpha \quad (39)$$

Then the criterion of classification uses statistic

$$\left[\bar{X} - \frac{1}{2}\left(\bar{X}^{(1)} + \bar{X}^{(2)}\right)\right]' S^{-1}\left(\bar{X}^{(1)} - \bar{X}^{(2)}\right) \quad (40)$$

The larger the N is, the smaller are the probabilities of misclassification.

Example 2: Suppose for detecting potential hemophilia A carriers, blood samples were assayed for two groups of women and measurements on two variables X_1 and X_2 were recorded, which are

$$X_1 = \log(\text{AHF activity})$$
$$X_2 = \log(\text{AHF - like activity})$$

Here "AHF" means antihemophilic factor. A sample of $n_1 = 30$ was selected from population π_1 of women who did not carry hemophilia gene. Another sample of $n_2 = 22$ women were selected from population π_2 of known hemophilia A carriers. On the basis of pair of observations $(X_1, X_2)'$ it is given that

$$\mu^{(1)} = \begin{bmatrix} -0.0065 \\ -0.0390 \end{bmatrix} \qquad \mu^{(2)} = \begin{bmatrix} -0.2483 \\ 0.0262 \end{bmatrix}$$

and

$$\Sigma^{-1} = \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}$$

Now measurements of AHF activity and AHF-like antigen on a women who may be a hemophilia A carrier gave $X_1 = -0.210$ and $X_2 = 0.044$. Our problem now is to classify this woman in one of the populations π_1 or π_2.

Solution: Case 1: Prior probabilities and costs are equal

Let us consider the situation of equal cost and same prior, the discriminant function defined in Eqn. (29) is

$$x'd = [X_1 \ X_2]\begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}\begin{bmatrix} 0.2418 \\ 0.0652 \end{bmatrix}$$
$$= 37.61X_1 - 28.92X_2 \quad (41)$$

Right hand side of Eqn.(28) is

$$(1/2)\left\{[37.61 - 28.92]\begin{bmatrix} -0.0065 \\ -0.0390 \end{bmatrix} + [37.61 - 28.92]\begin{bmatrix} -0.2483 \\ 0.0262 \end{bmatrix}\right\}$$
$$= (0.88 - 10.10)/2$$
$$= 4.61$$

So the classification rule is now allocate (x_1, x_2) to π_1 if

$$R_1 : 37.61X_1 - 28.92X_2 \geq -4.6 \quad (42)$$

And to π_2 if

$$R_2 : 37.61X_1 - 28.92X_2 < -4.6 \quad (43)$$

Since in our case $x_1 = -0.210$ and $x_2 = -0.044$ so putting these values in Eqn.(41) we get

$$x'd = -0.62 \quad (44)$$

Comparing Eqn.(44) with Eqn.(42) and Eqn.(43) we classify the woman as of population π_2. That is, she is carrier of hemophilia A.

Case 2: Prior probabilities are known

Suppose now that prior probabilities of group membership are known assuming that cost of classification is equal. Let the genetic chance of being a hemophilia A carrier is 0.25, then we can take $q_1 = 0.75$ and $q_2 = 0.25$ with $C(1|2) = C(2|1)$ and using classification statistics given in Eqn. (30).

$$U = x'\Sigma^{-1}\left[\mu^{(1)} - \mu^{(2)}\right] - \frac{1}{2}\left(\mu^{(1)} + \mu^{(2)}\right)\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$$

we have

$$U = -6.62 - (-4.61)$$
$$= -2.01$$

To find value of C we evaluate

$$C = \log k = \log(C(1|2)q_2]/[C(2|1)q_1) = \log(0.25/0.75) = -1.10$$

Since $U < C$.

So according to classification rule given in Eqn.(32) we classify the woman as π_2.

Now try the following exercises.

E2) Consider the data given in Example 2 and check whether $X_1 = 0.04$ and $X_2 = 0.08$ is the one from the population π_1 or π_2.

E3) Suppose that $n_1 = 10$ and $n_2 = 15$ observations are made on two random variables X_1 and X_2. Given is $\mu^{(1)} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $\mu^{(2)} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ and

$$\Sigma^{-1} = \begin{bmatrix} 0.15 & 0.05 \\ 0.05 & 0.35 \end{bmatrix}.$$

i) Considering equal costs and equal prior probabilities, check whether the observation [0, 1] belongs to population π_1 or π_2.

ii) Considering $q_1 = 0.6$ and $q_2 = 0.4$ with $C(1/2) = C(2/1)$, check whether the observation [0, 1] belongs to π_1 or π_2.

In the following section, we deal with the classification into one of several populations.

23.5 CLASSIFICATION INTO ONE OF SEVERAL POPULATIONS

Let us now consider the problem of classifying an observation into more than two populations. We shall extend the consideration of the previous sections to the case of more than two populations. Let $\pi_1, \ldots, \pi_m$ be m populations with density functions $p_1(x), \ldots, p_m(x)$ respectively. We partition p-dimensional observations space into regions $R_1, \ldots, R_m$. If an observation x falls into R_i we shall say that it comes from π_i. Let the cost of misclassifying an observation from π_i as coming from π_j be $C(j/i)$ $i \neq j = 1, \cdots, m$. The probability of this misclassification is

$$P(j \mid i, R) = \int_{R_j} p_i(x)dx. \tag{45}$$

Suppose we have a priori probabilities of the m populations, as $q_1, \ldots, q_m$ respectively. Then the expected loss is

$$\sum_{i=1}^{m} q_i \left\{ \sum_{\substack{j=1 \\ j \neq i}}^{m} C(j|i) P(j|i, R) \right\} \tag{46}$$

We should like to choose $R_1, \ldots, R_m$ to make this minimum.
Since we have a priori probabilities for the populations, we can define the conditional probability of an observation coming from a population given the observation vector x. The conditional probability of the observation coming from π_i is

$$\frac{q_i p_i(x)}{\sum_{k=1}^{m} q_k p_k(x)} \tag{47}$$

If we classify the observation as from π_j, the expected loss is

$$\sum_{\substack{i=1 \\ i \neq j}}^{m} \frac{q_i p_i(x)}{\sum_{k=1}^{m} q_k p_k(x)} C(j \mid i). \tag{48}$$

We minimize the expected loss at this point if we choose j so as to minimize Eqn.(48); that is, we calculate

$$\sum_{\substack{i=1 \\ i \neq j}}^{m} q_i p_i(x) C(j \mid i) \tag{49}$$

for $j = 1, 2, \cdots, m$ and select that j that gives the minimum of all. (If two or more different indices give the minimum, it is irrelevant which index is selected.) This procedure assigns the point x to one of the R_j. Following this procedure for each x, we define our regions $R_1, \ldots, R_m$. The classification procedure, then, is to classify an observation as coming from π_j if it falls in R_j.

Bayes Classification Procedure: If q_i is the a priori probability of drawing an observation from population π_i with density $p_i(x)$ $(i = 1, \ldots, m)$ and the cost of misclassifying an observation from π_i as from π_j is $C(j \mid i)$, then the regions of classification, $R_1,\ldots,R_m$ that minimize the expected cost are defined by assigning x to R_k if

$$\sum_{\substack{i=1\\i\neq k}}^{m} q_i p_i(x)C(k\mid i) < \sum_{\substack{i=1\\i\neq j}}^{m} q_i p_i(x)C(j\mid i) \quad \text{for all } j=1,\cdots,m,\ j\neq k. \tag{50}$$

[If Eqn. (50) holds for all $j(j \neq k)$, except for h indices and the inequality is replaced by equality for those indices, then this point can be assigned to any of the $h+1\,\pi's$]. If the probability of equality between right-hand and left hand sides of Eqn.(50) is zero for each k and j, then the minimizing procedure is unique except for sets of probability zero.

We now verify this result. Let

$$h_j(x) = \sum_{\substack{i=1\\i\neq j}}^{m} q_i p_i(x) C(j\mid i). \tag{51}$$

Then the expected loss of a procedure R is

$$\sum_{j=1}^{m} \int_{R_1} h_j(x)dx = \int h(x)dx, \tag{52}$$

where $h(x) = h_j(x)$ for x in R_j. For the procedure described above

$h(x)$ is $h^*(x) \min_i h_i(x)$. Thus, the difference between the expected loss for any procedure R and for R^* is

$$\int\left[h(x)-h^*(x)\right]dx = \sum_j \int_{R_j} \left[h_j(x) - \min_i h_i(x)dx\right] \tag{53}$$

Equality can hold only if $h_j(x) = \min_i h_i(x)$ for x in R_i (except for sets of probability zero).

Let us see how this method applies when $C(j\mid i)=1$ for all i and $j(i\neq j)$. Then in R_k

$$\sum_{\substack{i=1\\i\neq k}}^{m} q_i p_i(x) < \sum_{\substack{i=1\\i\neq j}}^{m} q_i p_i(x)\ (j\neq k) \tag{54}$$

Subtracting $\sum_{\substack{i=1\\i\neq k,j}}^{m} q_i p_i(x)$ from both sides of Eqn.(54), we obtain

$$q_j p_j(x) < q_k p_k(x)(j\neq k) \tag{55}$$

In this case the point x is in R_k if k is the index for which $q_j\, p_j(x)$ for $j = 1, 2,\cdots,m$ is a maximum; that is, π_k is the most probable population.

Now suppose that we do not have a priori probabilities. Then we cannot define an unconditional expected loss for a classification procedure. However, we can define an expected loss on the condition that the observation comes from a given population. The conditional expected loss of the observation is from π_i is

$$\sum_{\substack{j=1\\j\neq i}}^{m} C(j\mid i)P(j\mid i,R) = r(i,R). \tag{56}$$

A procedure R is at least as good as R^* if $r(i,R) \le r(i,R^*), i = 1,\ldots,m$; R is better if at least one inequality is strict. R is admissible if there is no procedure R^* that is better. A class of procedures is complete if for every procedure R outside the class there is a procedure R^* in the class that is better.

Now let us show that a Bayes procedure is admissible. Let R be a Bayes procedure; let R^* be another procedure. Since R is Bayes

$$\sum_{i=1}^{m} q_i\, r(i,R) \le \sum_{i=1}^{m} q_i\, r(i,R^*). \tag{57}$$

Suppose $r(i,R) \le r(i,R), i = 2,\ldots,m$, and $q_1 > 0$. Then

$$q_1\left[r(1,R) - r(1,R^*)\right] \le \sum_{i=2}^{m} q_i\left[r(i,R^*) - r(i,R)\right] \le 0, \tag{58}$$

and $r(1,R) \le r(1,R^*)$. Similarly, if $q_i > 0$ then $r(i,R) \le r(i,R^*)$. Thus R^* cannot be better than R; R is admissible.

In the following section, we shall discuss the classification into one of several multivariate normal population.

23.6 CLASSIFICATION INTO ONE OF SEVERAL MULTIVARIATE NORMAL POPULATIONS

We shall apply the theory described above to the case in which populations have multivariate normal distributions. We assume that the means of populations are different and the covariance matrices are equal. Let $N(\mu^{(i)}, \Sigma)$ be the distribution of π_i $(i = 1,2,\ldots,m)$ having the density given below:

$$p_i(x) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu^{(i)})'\Sigma^{-1}(x-\mu^{(i)})\right] \tag{59}$$

Let us consider that the parameters are known. For general costs with known a priori probabilities, we can form m functions as given in Eqn.(49) and define the region R_j as consisting of the points x such that the j^{th} function is minimum.
Let us now consider that costs of misclassification are same. In such a case, we have seen in Eqn.(55) the observation x is to be classified to the population k of $q_k p_k(x) > q_j p_j(x)$ for every $j(\ne k) = 1.2,\ldots,m$. Let us define a function

$$U_{kj}(x) = \log\frac{p_k(x)}{p_j(x)} = \left[x - \frac{1}{2}\left(\mu^{(k)} + \mu^{(j)}\right)\right]' \Sigma^{-1}\left(\mu^{(k)} - \mu^{(j)}\right) \tag{60}$$

If a priori probabilities are known R_k is defined by those x satisfying

$$R_k : U_{kj}(x) > \log\frac{q_j}{q_k}, \quad j = 1,2,\ldots,m; j \ne k \tag{61}$$

It should be noted that each $U_{kj}(x)$ is the classification function related to the k^{th} and j^{th} population and $U_{kj}(x) = -U_{jk}(x)$.

In case prior probabilities are not known, the region R_k is defined by the inequalities

$$U_{kj} \ge c_k - c_j \quad j = 1,\ldots,m, j \ne k \tag{62}$$

The c_j can be taken to be non-negative. These sets of regions form the class of admissible procedures. Under the minimax procedure these constants are determined so that all $P(i\mid i, R)$ are equal.

Let us now consider the evaluation of the probabilities of correct classification. If X is a random observation, we consider the random variables

$$U_{kj} = \left[X - \frac{1}{2}\left(\mu^{(k)} + \mu^{(j)}\right)\right]' \Sigma^{-1}\left(\mu^{(k)} - \mu^{(j)}\right) \tag{63}$$

Here $U_{kj} = -U_{jk}$. Thus we use m(m–1)/2 classification functions if the means span an (m–1) - dimensional hyperplane. If X is from π_k, then U_{kj} is distributed as $N(\frac{1}{2}\Delta_{kij,}\Delta_{kij})$ where

$$\Delta_{kij} = \frac{1}{2}\left(\mu^{(k)} - \mu^{(j)}\right)\Sigma^{-1}\left(\mu^{(k)} - \mu^{(j)}\right) \tag{64}$$

And covariance of U_{ki} and U_{kj} is

$$\Delta_{kij} = \frac{1}{2}\left(\mu^{(k)} - \mu^{(i)}\right)' \Sigma^{-1}\left(\mu^{(k)} - \mu^{(j)}\right) \tag{65}$$

To determine the constants c_k, we consider the integrals

$$P(k\mid k, R) = \int_{Ck-Cm}^{\infty} \cdots \int_{Ck-Cl}^{\infty} f_k du_{k.1} \ldots du_{k.k-1}\, du_{k.k+1} \ldots du_{k.m} \tag{66}$$

where f_k is the joint density of U_{kj} $(j = 1, 2, \ldots, m)(j \neq k)$.

Example 3: Let us consider a problem of classification among three populations. Suppose that the three populations π_1, π_2 and π_3 represent Brahmin caste , Artisan caste and the Korwa caste of country. The measurements for each individual of a caste are stature (x_1), sitting height (x_2), nasal depth (x_3) and nasal height (x_4). The means of these variables in three populations are given below.

	Brahmin	**Artisan**	**Korwa**
	π_1	π_2	π_3
Stature (x_1)	164.51	160.53	158.17
Sitting height (x_2)	86.43	81.87	81.16
Nasal depth (x_3)	25.49	23.48	21.44
Nasal height (x_4)	51.24	48.62	46.72

The matrix of correlation for all the populations is given below, which can be converted into variance-covariance matrix easily, given standard deviations.

$$\begin{bmatrix} 1.0000 & 0.5849 & 0.1774 & 0.1974 \\ 0.5849 & 1.0000 & 0.2094 & 0.2170 \\ 0.1774 & 0.2094 & 1.0000 & 0.2190 \\ 0.1974 & 0.2170 & 0.2910 & 1.0000 \end{bmatrix}$$

The standard deviations are $\sigma_1 = 5.74$, $\sigma_2 = 3.20$, $\sigma_3 = 1.75$ and $\sigma_4 = 3.50$. Our problem is to decide the space of four variables x_1, x_2, x_3 and x_4 into three regions of classification. Here we can assume that the three populations are normally distributed. The solution of problem consist of to find

(i) a set of regions under the assumption that drawing a new observation from each population is equally probable.

(ii) a set of regions such that the larger probability of misclassification is minimized.

On the basis of computed values of $\Sigma^{-1}\left(\mu^{(1)}-\mu^{(3)}\right)$ and $-\Sigma^{-1}\left(\mu^{(1)}-\mu^{(2)}\right)$ and then $\Sigma^{-1}\left(\mu^{(2)}-\mu^{(3)}\right)=\Sigma^{-1}\left(\mu^{(1)}-\mu^{(3)}\right)-\Sigma^{-1}\left(\mu^{(1)}-\mu^{(2)}\right)$ we calculate all $\frac{1}{2}\left(\mu^{(k)}+\mu^{(j)}\right)'\Sigma^{-1}\left(\mu^{(k)}-\mu^{(j)}\right)$

We obtain discriminant functions Eqn.(60)

$$U_{12}(x) = -0.0708x_1 + 0.4990x_2 + 0.3373x_3 + 0.0887x_4 - 43.13$$
$$U_{13}(x) = 0.0003x_1 + 0.3550x_2 + 1.1063x_3 + 0.1375x_4 - 62.49$$
$$U_{23}(x) = 0.0711x_1 + 0.1440x_2 + 0.7690x_3 + 0.0488x_4 - 19.36$$

Other three discriminant functions can be obtained by the relation

$$U_{kj} = -U_{jk}$$

That is

$$U_{21}(x) = -U_{12}(x), U_{31}(x) = -U_{13}(x) \text{ and } U_{32}(x) = -U_{23}(x)$$

Now if there are equal aprior probabilities then best region of classification can be defined as

$$R_1 : U_{12}(x) \geq 0,\ U_{13}(x) \geq 0$$

$$R_2 : U_{21}(x) \geq 0,\ U_{23}(x) \geq 0$$

$$R_3 : U_{31}(x) \geq 0,\ U_{13}(x) \geq 0$$

For example if we find an individual with measurement x such that $U_{12}(x) \geq 0$, and $U_{13}(x) \geq 0$, we classify it as Brahmin.

To find the probability of misclassification when an individual is drawn from a population π_k we need the mean, variances and covariances of proper pair of U's. These are given in following table

Population of x	U	Means	Standard deviations	Correlation
π_1	U_{12}	1.491	1.727	0.8658
	U_{13}	3.487	2.641	
π_2	U_{21}	1.491	1.727	− 0.3894
	U_{23}	1.031	1.436	
π_3	U_{31}	3.487	2.641	0.7983
	U_{32}	1.031	1.436	

The probabilities of misclassification are then obtained by use of the tables for the bivariate normal distribution. These probabilities are 0.21 for $\bar{\pi}_1$, number is missing for

π_2 and 0.25 for π_3. For example, if measurements are made on a Brahmin, the probability that he is classified as Korwa or Artisan is 0.21.

Try the following exercises.

E4) In a diabetic center the fasting blood sugar levels of three groups of patients are recorded two times, one of before the treatment (x_1) and another after the treatment(x_2).

S.No	Patients of Different Age (in years) Group					
	<40(Π_1)		40-50(Π_2)		50 (Π_3)	
	x_1	x_2	x_1	x_2	x_1	x_2
1	172	174	329	310	210	198
2	222	200	314	303	112	105
3	110	206	228	343	294	328
4	233	218	215	311	213	200
5	366	148	153	215	121	170
6	181	366	179	303	160	176
7	185	205	156	130	369	319
8	144	236	196	167	157	150
9	244	206	253	269	186	160
10	241	217	208	234	136	215

If a new observation vector is x = [180 2107]. Allocate this new object into its population.

Now let us summarize this unit.

23.7 SUMMARY

In this unit, we covered the following point.

1. We have introduced the concepts of classification. However, it is expected that after going through this unit you will be able to have the basic idea about the classification problems.
2. We discussed how to formulate the classification problems along with developing a classification rules. The examples given in different sections are of low dimension; however they will learn the procedure of classification and use it for higher dimension situation too.

23.8 SOLUTIONS/ANSWERS

E1) i) Prior probabilities $q_1 = 0.4$ and $q_2 = 0.6$

ii) $C(2/1) = 100$

$C(1/2) = 50$

iii) $R_1 = \frac{p_1(x)}{p_2(x)} \geq \frac{C(1/2)q_2}{C(2/1)q_1} = \frac{50 \times 0.6}{100 \times 0.4} = 0.75$

$R_2 < 0.75$

iv) Given that $p_1(x) = 0.4$ and $p_2(x) = 0.5$, then the ration

$\frac{p_1(x)}{p_2(x)} = \frac{0.3}{0.5} = 0.6$ which is less than 0.75. It implies that it satisfies the condition of R_2. Therefore, the new item belongs to Π_2.

E2) Here, from Eqn. (41)

$$
\begin{aligned}
x'd &= 37.61\, X_1 - 28.92 X_2 \\
&= 37.61 \times 0.04 - 28.92 \times 0.08 \\
&= 1.4044 - 2.3136 \\
&= -0.9092
\end{aligned}
$$

$$R_1 : 37.61X_1 - 28.92X_2 \geq -4.61$$

and $R_2 = 37.61X_1 - 28.92X_2 < -4.61$

which implies that the given observation belongs to π_2 as it falls in the region R_2.

E3) i)

$$x'd = [X_1 \;\; X_2]\begin{bmatrix} 0.15 & 0.05 \\ 0.05 & 0.35 \end{bmatrix}\begin{bmatrix} 3 \\ 0 \end{bmatrix}$$

$$= 0.45X_1 + 0.15X_2$$

$$R_1 : 0.45X_1 + 0.15X_2 \geq \frac{1}{2}[1 \;\; 2]\begin{bmatrix} 0.15 & 0.05 \\ 0.05 & 0.35 \end{bmatrix}\begin{bmatrix} 3 \\ 0 \end{bmatrix} = 0.375$$

$R_2 : 0.45X_1 + 0.15X_2 < 0.375$

for $X_1 = 0$ and $X_2 = 1$

$0.45X_1 + 0.15X_2 = 0.15$, which falls in the region R_2, therefore observation [0, 1] is from the population π_2.

ii) $q_1 = 0.6,\; q_2 = 0.4$ and $C(2/1) = C(1/2)$

using Eqn. (30)

$$U = [0 \;\; 1]\begin{bmatrix} 0.15 & 0.05 \\ 0.05 & 0.35 \end{bmatrix}\begin{bmatrix} 3 \\ 0 \end{bmatrix} - \frac{1}{2}[1 \;\; 2]\begin{bmatrix} 0.15 & 0.05 \\ 0.05 & 0.35 \end{bmatrix}\begin{bmatrix} 3 \\ 0 \end{bmatrix}$$

$$= -0.225$$

$$C = \log k = \log\frac{C(1/2)q_2}{C(2/1)q_1} = \log\frac{2}{3}$$

$$= -0.4055$$

Since $U \geq C$, therefore the given observation belongs to population π_1.

E4) Here $n_1 = n_2 = n_3 = 10$

$$\overline{X}_1 = \begin{bmatrix} 209.8 \\ 217.6 \end{bmatrix}, \qquad S_1 = \begin{bmatrix} 4439.16 & -1385.88 \\ -1385.88 & 2984.44 \end{bmatrix}$$

$$\overline{X}_2 = \begin{bmatrix} 223.1 \\ 258.5 \end{bmatrix}, \qquad S_2 = \begin{bmatrix} 3278.49 & 2236.05 \\ 2236.05 & 4399.65 \end{bmatrix}$$

$$\overline{X}_3 = \begin{bmatrix} 195.8 \\ 202.1 \end{bmatrix}, \qquad S_3 = \begin{bmatrix} 5921.56 & 4565.12 \\ 4565.12 & 4533.09 \end{bmatrix}$$

$$\overline{X} = [209.57 \quad 226.07], \quad S = \begin{bmatrix} 5051.56 & 2005.66 \\ 2005.66 & 4413.77 \end{bmatrix}$$

Obtain the discriminant functions U_{12}, U_{13}, U_{23}, U_{21}, U_{31} and U_{32} accordingly obtain R_1, R_2 and R_3.

23.9 PRACTICAL ASSIGNMENT

Session 8

1. Write a program in 'C' language to find the classification of a population into one among k populations.

Table 1: STANDARD NORMAL PROBABILITIES

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9667
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

Source: Applied Multivariate Statistical Analysis, Third Edition, Richard A. Johnson, Dean W. Wichern.

Table 2: STUDENT'S t-DISTRIBUTION CRITICAL POINTS

If X has a t-distribution with v degree of freedom, the table given the value of x such that $P(X \leq x) = \alpha$

d.f. v	α							
	0.250	0.100	0.050	0.025	0.010	0.00833	0.00625	0.005
1	1.000	3.078	6.314	12.706	31.821	38.190	50.923	63.657
2	0.816	1.886	2.920	4.303	6.965	7.649	8.860	9.925
3	0.765	1.638	2.353	3.182	4.541	4.857	5.392	5.841
4	0.741	1.533	2.132	2.776	3.747	3.961	4.315	4.604
5	0.727	1.476	2.015	2.571	3.365	3.534	3.810	4.032
6	0.718	1.440	1.943	2.447	3.143	3.287	3.521	3.707
7	0.711	1.415	1.895	2.365	2.998	3.128	3.335	3.499
8	0.706	1.397	1.860	2.306	2.896	3.016	3.206	3.355
9	0.703	1.383	1.833	2.262	2.821	2.933	3.111	3.250
10	0.700	1.372	1.812	2.228	2.764	2.870	3.038	3.169
11	0.697	1.363	1.796	2.201	2.718	2.820	2.981	3.106
12	0.695	1.356	1.782	2.179	2.681	2.779	2.934	3.055
13	0.694	1.350	1.771	2.160	2.650	2.746	2.896	3.012
14	0.692	1.345	1.761	2.145	2.624	2.718	2.864	2.977
15	0.691	1.341	1.753	2.131	2.602	2.694	2.837	2.947
16	0.690	1.337	1.746	2.120	2.583	2.673	2.813	2.921
17	0.689	1.333	1.740	2.110	2.567	2.655	2.793	2.898
18	0.688	1.330	1.734	2.101	2.552	2.639	2.775	2.878
19	0.688	1.328	1.729	2.093	2.539	2.625	2.759	2.861
20	0.687	1.325	1.725	2.086	2.528	2.613	2.744	2.845
21	0.686	1.323	1.721	2.080	2.518	2.601	2.732	2.831
22	0.686	1.321	1.717	2.074	2.508	2.591	2.720	2.819
23	0.685	1.319	1.714	2.069	2.500	2.582	2.710	2.807
24	0.685	1.318	1.711	2.064	2.492	2.574	2.700	2.797
25	0.684	1.316	1.708	2.060	2.485	2.566	2.692	2.787
26	0.684	1.315	1.706	2.056	2.479	2.559	2.684	2.779
27	0.684	1.314	1.703	2.052	2.473	2.552	2.676	2.771
28	0.683	1.313	1.701	2.048	2.467	2.546	2.669	2.763
29	0.683	1.311	1.699	2.045	2.462	2.541	2.663	2.756
30	0.683	1.310	1.697	2.042	2.457	2.536	2.657	2.750
40	0.681	1.303	1.684	2.021	2.423	2.499	2.616	2.704
60	.679	1.296	1.671	2.000	2.390	2.463	2.575	2.660
120	0.677	1.289	1.658	1.980	2.358	2.428	2.536	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.394	2.498	2.576

Source: Applied Multivariate Statistical Analysis, Third Edition, Richard A. Johnson, Dean W. Wichern.

Table 3: χ^2 Distribution

If X has a χ^2 distribution with v degree of freedom this table gives the value of x such that $\Pr(X \le x) = \alpha$

d.f.	α								
v	0.990	0.950	0.900	.500	0.100	0.050	0.025	0.010	0.005
1	0.0002	0.004	0.02	0.45	2.71	3.84	5.02	6.63	7.88
2	0.02	0.10	0.21	1.39	4.61	5.99	7.38	9.21	10.60
3	0.11	0.35	0.58	2.37	6.25	7.81	9.35	11.34	12.84
4	0.30	0.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	0.55	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	0.87	1.64	2.20	5.35	10.64	12.59	14.45	16.81	18.55
7	1.24	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.65	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.95
9	2.09	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.56	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	3.05	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.57	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	4.11	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.66	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	5.23	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.81	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	6.41	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	7.01	9.39	10.86	17.34	25.99	28.87	31.53	34.81	37.16
19	7.63	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	8.26	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.90	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	9.54	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	10.20	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	10.86	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	11.52	14.61	16.47	24.34	34.38	37.65	40.65	44.31	46.93
26	12.20	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	12.88	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.64
28	13.56	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	14.26	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	14.95	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	22.16	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	29.71	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	37.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	45.44	51.74	55.33	69.33	85.53	90.53	95.02	100.43	104.21
80	53.54	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	61.75	69.13	73.29	89.33	107.57	113.15	118.14	124.12	128.30
100	70.06	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17

Source: Applied Multivariate Statistical Analysis, Third Edition, Richard A. Johnson, Dean W. Wichern.

Table 4: F-DISTRIBUTION CRITICAL POINTS (α =.10)

v	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.05	62.26	62.53	62.79
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.17	5.17	5.16	5.15
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.81	2.80	2.78	2.76
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.57	2.56	2.54	2.15
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.27	2.25	2.23	2.21
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.17	2.16	2.13	2.11
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.03	2.01	1.99	1.96
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.93	1.91	1.89	1.86
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.89	1.87	1.85	1.82
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.86	1.84	1.81	1.78
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.83	1.81	1.78	1.75
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.80	1.78	1.75	1.72
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.78	1.76	1.73	1.70
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.76	1.74	1.71	1.68
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.74	1.72	1.69	1.66
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.71	1.69	1.66	1.62
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.68	1.66	1.63	1.59
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.67	1.65	1.61	1.58
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.66	1.64	1.60	1.57
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.65	1.63	1.59	1.56
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.64	1.62	1.58	1.55
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.63	1.61	1.57	1.54
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.50	1.48	1.44	1.40
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24

Source: Applied Multivariate Statistical Analysis, Third Edition, Richard A. Johnson, Dean W. Wichern.

Table 5: F-DISTRIBUTION CRITICAL POINTS (α =.05)

	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60
1	161.5	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	246.0	248.0	249.3	250.1	251.1	252.2
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.48
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.63	8.62	8.59	8.57
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.52	4.50	4.46	4.43
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.83	3.81	3.77	3.74
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.40	3.38	3.34	3.30
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.11	3.08	3.04	3.01
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.89	2.86	2.83	2.79
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.73	2.70	2.66	2.62
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90		2.79	2.72	2.65	2.60	2.57	2.53	2.49
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.50	2.47	2.43	2.38
13	4.67	3.81	3.14	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.41	2.38	2.34	2.30
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.34	2.31	2.27	2.22
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.28	2.25	2.20	2.16
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.23	2.19	2.15	2.11
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.18	2.15	2.10	2.06
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.14	2.11	2.06	2.02
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.07	2.04	1.99	1.95
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.02	1.98	1.94	1.89
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.00	1.96	1.91	1.86
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.97	1.94	1.89	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.94	1.90	1.85	1.80
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.92	1.88	1.84	1.79
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.89	1.85	1.81	1.75
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.88	1.84	1.79	1.74
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.78	1.74	1.69	1.64
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.69	1.65	1.59	1.53
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.60	1.55	1.50	143
∞	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.51	1.46	1.39	1.32

Source: Applied Multivariate Statistical Analysis, Third Edition, Richard A. Johnson, Dean W. Wichern.

Table 6: F-DISTRIBUTION CRITICAL POINTS (α =.01)

	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60
1	4052.	5000.	5403.	5625.	5764.	58.59	5928.	5981.	6023.	6056.	6106.	6157.	6209.	6240.	6261.	6287.	6313.
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.58	26.50	26.41	26.32
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.91	13.84	13.75	13.65
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.45	9.38	9.29	9.20
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.30	7.23	7.14	7.06
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.06	5.99	5.91	5.82
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.26	5.20	5.12	5.03
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.71	4.65	4.57	4.48
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.31	4.25	4.17	4.08
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.01	3.94	3.86	3.78
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.76	3.70	3.62	3.54
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.57	3.51	3.43	3.34
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.41	3.35	3.27	3.18
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.28	3.21	3.13	3.05
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.16	3.10	3.02	2.93
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.07	3.00	2.92	2.83
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.98	2.92	2.84	2.75
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.91	2.84	2.76	2.67
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.84	2.78	2.69	2.61
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.79	2.72	2.64	2.55
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.73	2.67	2.58	2.50
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.69	2.62	2.54	2.45
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.64	2.58	2.49	2.40
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.60	2.54	2.45	2.36
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.57	2.50	2.42	2.33
27	7.68	5.49	4.60	4.11	3.78	3.56	3.93	3.26	3.15	3.06	2.93	2.78	2.63	2.54	2.47	2.38	2.29
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.51	2.44	2.35	2.26
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.48	2.41	2.33	2.23
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.45	2.39	2.30	2.21
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.27	2.20	2.11	2.02
50	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.10	2.03	1.94	1.84
20	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.93	1.86	1.76	1.66
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.78	1.70	1.59	1.47

Source: Applied Multivariate Statistical Analysis, Third Edition, Richard A. Johnson, Dean W. Wichern