UNIT 21 INFERENCE-III

Structure

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

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

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

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

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

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

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

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

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

$$\lambda = \frac{\max\limits_{\Sigma} L(\mu_0, \Sigma \mid X)}{\max\limits_{\mu,\Sigma} L(\mu, \Sigma \mid X)}. \tag{6}$$

and,

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

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 $\lambda \le \lambda_\alpha$ is equivalent to $(\bar{X} - \mu_0)'S^{-1}(\bar{X} - \mu_0) \ge C_\alpha$. 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

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

Thus the value of Z_i has the form

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

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

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

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

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 $T^2(p, m) = \dfrac{mp}{m-p+1}F_{p,\,m-p+1}$ 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

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

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

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

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

$= 0.0749$

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

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 $d = [-4.18 \ 1.79 \ 12.86]'$

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

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

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

$$T^2(p, f^*) = d' \; U^{-1} d$$

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

$$= 13.8190.$$

Hence $p = 3 \; f^* = 54$. 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

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

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

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

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

from the given data

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

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

$$= 4.1$$

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,

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

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

$$= 30.78$$

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

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

$$= 13.68$$

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

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

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

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

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

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

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,

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

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

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

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

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

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

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

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

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.