BHARATHIDASAN UNIVERSITY

Tiruchirappalli- 620024

Tamil Nadu, India.

Programme: M.Sc. Statistics

Course Title: Multivariate Analysis

Course Code: 23ST12CC

Unit-III

Sampling Distributions in Multivariate Analysis

Dr. T. Jai Sankar
Associate Professor and Head
Department of Statistics

Ms. I. Angel Agnes Mary
Guest Faculty
Department of Statistics

UNIT – III

SAMPLING DISTRIBUTIONS IN MULTIVARIATE ANALYSIS

Hotelling's T^2

Hotelling's T-Squared (Hotelling, 1931) is the multivariate counterpart of the T-test. Multivariate means that you have data for **more than one parameter for each sample**. For example, to compare how well two different sets of students performed in school. Could compare Univariate (e.g. mean test scores) with a t-test. Or, could use Hotelling's T-squared to compare multivariate data (e.g. the multivariate mean of test scores, GPA and class grades).

Hotelling's T-Squared is based on Hotelling's T^2 distribution and forms the basis for various multivariate control charts.

Hotelling's T^2 Distribution

If X is Univariate normal with mean μ and standard deviation σ, then $U = \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma} \sim N(0,1)$, and $V = \frac{1}{\sigma^2}(x_i - \bar{x})^2 = \frac{(n-1)\,s^2}{\sigma^2} \sim \chi^2_{n-1}$, where s^2 is the sample variance from a sample of size n . If U and V are independently distributed, then Student's-t is defined as

$$t = \frac{U}{\sqrt{V/n-1}} = \frac{\sqrt{n}(\bar{x}-\mu)/\sigma}{(n-1)\,s^2/(n-1)\,\sigma^2} = \frac{\sqrt{n}(\bar{x}-\mu)}{s} \sim t_{n-1}\,.$$

The multivariate analogue of Student's − t is Hotelling's T^2.

If x_α (α = 1, 2, . . . , n) is an independent sample of size n from N$_p$(μ, Σ) and, if $\bar{x}$ is the sample mean vector, S the matrix of variance covariance, then the Hotelling's –T^2 is defined by the relation

$$T^2 = n(\bar{x}-\mu)'\,S^{-1}(\bar{x}-\mu)\,.$$

Properties of Hotelling's T^2 Distribution

- **Distribution:** T^2 follows a Hotelling's T-squared distribution, which is a multivariate extension of the F-distribution.
- **Degrees of freedom:** The T^2 distribution has two types of degrees of freedom: p (number of variables) and n-1 (sample size minus one).

- **Mean and variance:** The mean of T^2 is $p(n – 1)/(n – p – 1)$ and the variance is $2p(n – 1)(n – p – 1)/((n – p – 1)^2 (n – p – 3))$.
- **Invariance:** T^2 is invariant under linear transformations of the data.
- **Consistency**: T^2 is a consistent estimator of the population mean vector.
- **Asymptotic distribution:** As n → ∞, T^2 converges in distribution to a chi-squared distribution with p degrees of freedom.

Applications of Hotelling's $\mathbf{T^2}$ Distribution

- **Statistics and Research**
  - **Multivariate hypothesis testing:** T^2 is used to test the significance of differences between sample mean vectors and known population mean vectors.
  - **Discriminant analysis:** T^2 is used to determine the most effective variables for discriminating between groups.
  - **Cluster analysis:** T^2 is used to evaluate the similarity between clusters.

- **Data Science and Machine Learning**
  - **Anomaly detection:** T^2 is used to identify outliers and anomalies in multivariate data.
  - **Feature selection:** T^2 is used to select the most relevant features for model development.
  - **Model validation:** T^2 is used to evaluate the performance of multivariate models.
- **Business and Economics**
  - **Quality control:** $T^2$is used to monitor and control multivariate processes.
  - **Marketing research:** T^2 is used to analyze customer behavior and preferences.
  - **Financial analysis:** T^2 is used to evaluate portfolio performance and risk.

- **Engineering and Computer Science**
  - **Signal processing:** T^2 is used to detect anomalies in multivariate signals.
  - **Image processing:** T^2 is used to analyze and classify images.
  - **Robotics:** T^2 is used to evaluate the performance of robotic systems.

- **Medical and Healthcare**
  - **Disease diagnosis:** T^2 is used to identify biomarkers for disease diagnosis.
  - **Patient monitoring:** T^2 is used to monitor patient health and detect anomalies.
  - **Clinical trials:** T^2 is used to evaluate the efficacy of treatments.

- **Other applications**
  - **Environmental monitoring:** T^2 is used to evaluate water and air quality.
  - **Social sciences:** T^2 is used to analyze social and demographic data.
  - **Sports analytics:** T^2 is used to evaluate team and player performance.

One Sample Hotelling's T^2

As described in One Sample t-Test, the t-test can be used to test the null hypothesis that the population mean of a random variable *x* has a certain value, i.e. H_0: $\mu = \mu_0$. The test statistic is given by

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

The applicable Univariate test of the null hypothesis is based on the fact that $t \sim T(n - 1)$ provided the following assumptions are met:

- The population of *x* has a unique mean: i.e. there are no distinct sub-populations with different means
- The population of *x* has a normal distribution
- The sample is a random sample with each element in the sample taken independently.

The null hypothesis is rejected if $|t| > t_{crit}$, F Distribution, an equivalent test can be made using the test statistic t^2 and noting that $t^2 \sim F(1, n - 1)$.

Now, t^2 can be expressed as follows:

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

where $\bar{x}$ is the sample mean and *S* is the sample standard deviation.

Multivariate case

The population mean of the k × 1 random vector X has a certain value. Here, the null hypothesis is H_0: $\mu = \mu^0$ where μ and μ^0 are vectors.

Since the null hypothesis is true when $\mu_i = \mu_i^0$ for all i, $1 \leq i \leq k$, one way to carry out this test is to perform *k* separate univariate t-tests (or the equivalent *F* tests). The null hypothesis is then rejected if any one of these *k* univariate tests rejects its null hypothesis.

Experiment-wise Error Rate

If use a given value of α for all k tests, then the probability of the multivariate null hypothesis being rejected is much higher than α. For this reason, use a correction factor, usually, the Dunn/Sidak or Bonferroni correction factor, as described in Planned Comparisons. Thus, use a significance level of $1 - (1-\alpha)^{1/k}$ or α/k instead of α for each of the k univariate tests.

This approach is perfectly reasonable when the random variables x_i in X are independent. But when they are not independent then the Dunn/Sidák or Bonferroni correction factors over-correct and the resulting experiment-wise value for α is lower than it needs to be, which results in a test with lower statistical power.

Since it is common to create experiments in which the random variables x_i in X are not independent, it is better to use a different approach. In particular, will use the multivariate test based on Hotelling's T-square test statistic.

T-square statistic

Definition 1: Hotelling's T-square test statistic is

$$T^2 = n(\overline{X} - \mu^0)^T S^{-1} (\overline{X} - \mu^0)$$

where S is the covariance matrix of the sample for X, $\bar{X}$ is the mean of the sample, and where the sample for each random variable x_i in X has n elements.

Note the similarity between the expression for T^2 and the expression for t^2 given above.

Property 1:

$$n(\overline{X} - \mu^0)^T S^{-1} (\overline{X} - \mu^0) \sim \chi^2(k)$$

Corollary 1: For n sufficiently large, $T^2 \sim \chi^2(k)$

For small n, T^2 is not sufficiently accurate and a better estimate is achieved using the following property.

Property 2: Under the null hypothesis

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

If $F > F_{crit}$ then we reject the null hypothesis.

Hotelling’s T-square Test for Two Independent Samples

Univariate case

In the univariate case, we have two independent random variables and want to determine whether the population means of the two random variables are equal, i.e. H_0: Î¼x = Î¼y. To test this hypothesis we create a random sample for each variable. We define the following statistic

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{s\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

where s is the pooled standard deviation defined by

$$s^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x - 1) + (n_y - 1)}$$

The t-statistic defined above has a t distribution with $n_x + n_y - 2$ degrees of freedom, i.e. $t \sim T(n_x+n_y-2)$

Assumptions

- The populations of x and y have unique means and there are no distinct sub-populations with different means
- The populations of x and y have a normal distribution
- The variances of the two populations are equal (homogeneity of variances)
- The samples for x and y are random with each element in the sample taken independently

The null hypothesis is rejected if $|t| > t_{crit}$. Also note that by Property 1 of F Distribution, an equivalent test can be made using the t^2 test statistic is

$t^2 \sim F(1, n_x+n_y-2)$

Also, t^2 can be expressed as follows:

$$t^2 = (\bar{z} - \mu)\left[s^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)\right]^{-1}(\bar{z} - \mu)^T$$

Multivariate case

To test whether the population means of the k Ã – 1 random vectors and Y are equal, i.e. the null hypothesis H_0: Î¼X = Î¼Y.

Definition 1: The Two-sample Hotelling T-square test statistic is

$$T^2 = (\bar{X} - \bar{Y})\left[S\left(\frac{1}{n_x} + \frac{1}{n_y} \right) \right]^{-1} (\bar{X} - \bar{Y})^T$$

where S is the pooled sample covariance matrix of *XÂ* and *Y*, namely

$$S = \frac{(n_x - 1)S_X + (n_y - 1)S_Y}{(n_x - 1) + (n_y - 1)}$$

Note the similarity between the expression for T^2 and the expression for t^2 given above.

Property 1: For n_x and n_y sufficiently large, T^2 ~ Ï ≠ (k)

For small n_xÂ and Ân_y, T^2 is not sufficiently accurate and a better estimate is achieved using the following theorem.

Property 2: Under the null hypothesis

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

If F > F_{crit} then we reject the null hypothesis.

T^2 Statistic as a function of Likelihood Ratio Criterion

Let x_α *(α =1, 2, . . . , n > p)* be a random sample of size n from $N_p(\mu, \Sigma)$. The likelihood function is

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

and the likelihood ratio criterion

$$\lambda = \frac{\max\, L_\omega(\mu, \Sigma)}{\max\, L_\Omega(\mu, \Sigma)} = \frac{\max\, L_0}{\max\, L}$$

In the parameter space Ω, the maximum of L occurs when the parameters μ and Σ are estimated by their maximum likelihood estimators i.e., $\hat{\mu} = \bar{x} \, and \, \hat{\Sigma} = \frac{A}{n}$. In the space ω, we have $\mu = \mu_0 \, and \, \hat{\Sigma} = \frac{1}{n}\sum_\alpha (x_\alpha - \mu_0)\,(x_\alpha - \mu_0)'$, Therefore

$$\max L_\Omega = \frac{1}{(2\pi)^{np/2}\left|\frac{A}{n}\right|^{n/2}} \exp\left[-\frac{1}{2}\sum_\alpha (x_\alpha - \bar{x})'\left(\frac{A}{n}\right)^{-1}(x_\alpha - \bar{x})\right]$$

$$= \frac{n^{np/2}}{(2\pi)^{np/2}|A|^{n/2}} \exp\left[-\frac{1}{2}np\right]$$

Similarly,

$$\max L_\omega = \frac{1}{(2\pi)^{np/2}\left|\frac{1}{n}\sum_\alpha (x_\alpha - \mu_0)\,(x_\alpha - \mu_0)'\right|^{n/2}} \exp\left[-\frac{1}{2} tr\, n\left\{\sum_\alpha (x_\alpha - \mu_0)\,(x_\alpha - \mu_0)'\right\}^{-1}\left\{\sum_\alpha (x_\alpha - \mu_0)'\,(x_\alpha - \mu_0)\right\}\right]$$

$$= \frac{n^{np/2}}{(2\pi)^{np/2}\left|\sum_\alpha (x_\alpha - \mu_0)\,(x_\alpha - \mu_0)'\right|^{n/2}} \exp\left[-\frac{1}{2}np\right]$$

Consider

$$\sum_\alpha (x_\alpha - \mu_0)\,(x_\alpha - \mu_0)' = \sum_\alpha \left[(x_\alpha - \bar{x}) + (\bar{x} - \mu_0)\right]\left[(x_\alpha - \bar{x}) + (\bar{x} - \mu_0)\right]'$$

$$= A + n(\bar{x} - \mu_0)(\bar{x} - \mu_0)'$$

Hence,

$$\max L_\omega = \frac{n^{np/2}}{(2\pi)^{np/2}\left|A + n(\bar{x} - \mu_0)(\bar{x} - \mu_0)'\right|^{n/2}} \exp\left[-\frac{1}{2}np\right]$$

Thus, the likelihood ratio criterion is

$$\lambda = \frac{|A|^{n/2}}{\left|A + n(\bar{x} - \mu_0)(\bar{x} - \mu_0)'\right|^{n/2}}$$

$$\lambda^{2/n} = \frac{|A|}{\begin{vmatrix} 1 & -\sqrt{n}\,(\bar{x} - \mu_0)' \\ \sqrt{n}\,(\bar{x} - \mu_0) & A \end{vmatrix}} = \frac{|A|}{|A|\left|1 + n(\bar{x} - \mu_0)'\,A^{-1}(\bar{x} - \mu_0)\right|}$$

Since $|\Sigma| = |\Sigma_{22}|\,\left|\Sigma_{11} - \Sigma_{12}\,\Sigma_{22}^{-1}\,\Sigma_{21}\right|$

$$= \frac{1}{1 + n(\bar{x} - \mu_0)' A^{-1} (\bar{x} - \mu_0)}$$

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

where, $T^2 = n(\bar{x} - \mu_0)' S^{-1} (\bar{x} - \mu_0) = n(n-1)(\bar{x} - \mu_0)' A^{-1} (\bar{x} - \mu_0)$

The likelihood ratio test is defined by the critical region $\lambda \le \lambda_0$, where, λ_0 is so chosen so as to have level α, i.e., $\Pr[\lambda \le \lambda_0 \mid H_0] = \alpha$.

Thus,

$$\lambda^{2/n} \le \lambda_0^{2/n}, \text{ or } \frac{1}{1 + T^2/(n-1)} \le \lambda_0^{2/n}, \text{or } 1 + \frac{T^2}{n-1} \ge \lambda_0^{-n/n}$$

or $T^2 \ge (n-1)(\lambda_0^{-2/n} - 1) = T_0^2$

$\Rightarrow T^2 \ge T_0^2.$

Therefore, $\Pr\left[T^2 \ge T_0^2 \mid H_0\right] = \alpha.$

Invariance property of T^2 statistic

Let $X \sim N_p(\mu, \Sigma)$, then $T_x^2 = n(\bar{x} - \mu_{0x})' S_x^{-1} (\bar{x} - \mu_{0x})$,

where,

$$S_x = \frac{1}{n-1} \sum_\alpha (x_\alpha - \bar{x})(x_\alpha - \bar{x})' = \frac{1}{n-1}\left(x_1 x_1' + \cdots + x_n x_n' - n\bar{x}\bar{x}'\right)$$

Make a non-singular transformation

$$Y = CX, \Rightarrow y_\alpha = Cx_\alpha$$

Now

$$S_y = \frac{1}{n-1} \sum_\alpha (y_\alpha - \bar{y})(y_\alpha - \bar{y})' = \frac{1}{n-1}\left(y_1 y_1' + \cdots + y_n y_n' - n\bar{y}\bar{y}'\right)$$

$$= \frac{1}{n-1}\left(Cx_1 x_1' C' + \cdots + Cx_n x_n' C' - nC\bar{x}\bar{x}'C'\right)$$

$$= C\left[\frac{1}{n-1}\left(x_1 x_1' + \cdots + x_n x_n' - n\bar{x}\bar{x}'\right)\right]C'$$

$$= CS_x C'.$$

By definition

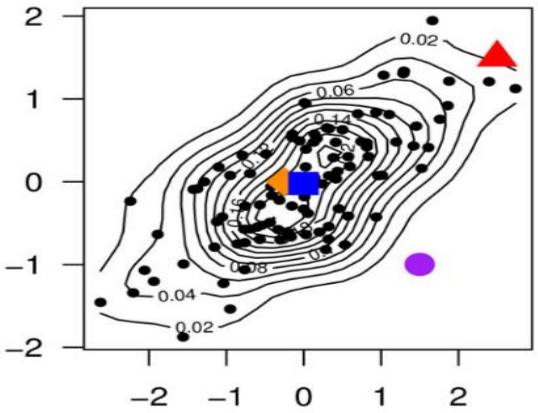
$$T_y^2 = n(\bar{y} - \mu_{0y})' S_y^{-1} (\bar{y} - \mu_{0y}) = n(C\bar{x} - C\mu_{0x})'(CS_x C')^{-1}(C\bar{x} - C\mu_{0x})$$

$$= n(\bar{x} - \mu_{0x})' C' C'^{-1} S_x^{-1} C^{-1} C(\bar{x} - \mu_{0x})$$

$$= n(\bar{x} - \mu_{0x})' S_x^{-1} (\bar{x} - \mu_{0x}) = T_x^2.$$

Mahalanobis D^2 statistic

The Mahalanobis distance (MD) is the distance between two points in multivariate space. In a regular Euclidean space, variables (e.g. x, y, z) are represented by axes drawn at right angles to each other; The distance between any two points can be measured with a ruler. For uncorrelated variables, the Euclidean distance equals the MD. However, if two or more variables are correlated, the axes are no longer at right angles, and the measurements become impossible with a ruler.



Mahalanobis distance plot example. A contour plot overlaying the scatterplot of 100 random draws from a bivariate normal distribution with mean zero, unit variance, and 50% correlation. The centroid defined by the marginal means is noted by a blue square [1].

Uses

The most common use for the Mahalanobis distance is to find multivariate outliers, which indicates unusual combinations of two or more variables. For example, it's fairly common to find a 6′ tall woman weighing 185 lbs, but it's rare to find a 4′ tall woman who weighs that much.

Formal Definition

The Mahalanobis distance between two objects is

$$D^2\ (Mahalanobis) = [(x_B - x_A)^T\ \Sigma^{-1}\ (x_B - x_A)]^{0.5}$$

Where:

- x_A and x_B is a pair of objects
- Σ is the sample covariance matrix and
- T is the transpose operation.

The formula uses distances from each observation to the central mean:

$$D^2 = [x_i - \bar{x})^T\ \Sigma^{-1}(x_i - \bar{x})]^{0.5}$$

Where:

- x_i = an object vector
- $\bar{x}$ = arithmetic mean vector

Distribution of Mahalanobis D^2 statistic

The quantity $(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)})$ is denoted by Δ^2 and was proposed by Mahalanobis as a measure of the distance between the two populations, $N_p(\mu^{(1)}, \Sigma)$, and $N_p(\mu^{(2)}, \Sigma)$. If the parameters are replaced by their unbiased estimates, is denoted by D^2, which is given by

$D^2=(\bar{x}^{(1)}-\bar{x}^{(2)})'S^{-1}(\bar{x}^{(1)}-\bar{x}^{(2)})$ and is known as Mahalanobis's D^2,

where

$$S=\frac{(n_1-1)S^{(1)}+(n_2-1)S^{(2)}}{n_1+n_2-2}=\frac{1}{n_1+n_2-2}\sum_{i=1}^{2}\sum_{\alpha=1}^{n_i}\left(x_\alpha^{(i)}-\bar{x}^{(i)}\right)\left(x_\alpha^{(i)}-\bar{x}^{(i)}\right)'$$

where

$\bar{x}^{(i)}=\frac{1}{n_i}\sum_{\alpha=1}^{n_i}x_\alpha^{(i)}$, i = 1, 2.

It is obvious that $T^2=\frac{n_1n_2}{n_1+n_2}D^2$

i.e. two-sample T^2 and D^2 are almost the same, except for the constant $k^2=\frac{n_1n_2}{n_1+n_2}$.

Let, $Y=k(\bar{x}^{(1)}-\bar{x}^{(2)})$, then expected value of Y is $E(Y)=k(\mu^{(1)}-\mu^{(2)})=\delta$ and the variance covariance matrix of Y is

$$\Sigma_Y=k^2E\left[(\bar{x}^{(1)}-\bar{x}^{(2)})-(\mu^{(1)}-\mu^{(2)})\right]\left[(\bar{x}^{(1)}-\bar{x}^{(2)})-(\mu^{(1)}-\mu^{(2)})\right]'$$

$=k^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\Sigma=\Sigma$, because $\left(\frac{n_1+n_2}{n_1n_2}\right)=\frac{1}{k^2}$.

Therefore,

$Y\sim N_p(\delta, \Sigma)$, then $k^2D^2 = Y'S^{-1}Y$

Since Σ is positive definite matrix there exist a nonsingular matrix C, such that

$C\Sigma C'=I \Longrightarrow CC'=\Sigma^{-1}$.

Define,

$Y^* = CY, S^* = CSC', and\ \delta^* = C\delta,$ then

$k^2D^2 = Y^{*'}S^{*-1}Y^*$, and expected value of Y^* is

$$\Sigma_{Y^*}=CE\left[Y-E(Y)\right]E\left[Y-E(Y)\right]'C'=C\Sigma C'=I.$$

Thus,

$$Y^*\sim N_p(\delta^*,I)\Rightarrow Y^{*'}Y^*\sim\chi_p^2(\delta^{*'},\delta^*)$$

Where,

$$\delta^{*'}\delta^*=\delta'C'C\delta=\delta'\Sigma^{-1}\delta=\lambda^2$$

Let, $(n_1+n_2-2)S = \sum_{\alpha=1}^{n_1+n_2-2} Z_\alpha Z_\delta{}'$, where, $Z_\alpha \sim N_p(0, \Sigma)$

$$\Rightarrow (n_1+n_2-2)S^* = \sum_{\alpha=1}^{n_1+n_2-2} (CZ_\alpha)(CZ_\delta)'$$, where, $CZ_\alpha \sim N_p(0, I)$

Therefore,

$$k^2D^2 = Y^{*'} S^{*-1} Y^* = (n_1+n_2-2)\frac{\chi_p^2(\lambda^2)}{\chi^2_{n_1+n_2-2-(p-1)}},$$

$$\Rightarrow \frac{n_1n_2}{n_1+n_2}\frac{n_1+n_2-p-1}{(n_1+n_2-2)p}D^2 = \frac{\chi_p^2(\lambda^2)/p}{\chi^2_{n_1+n_2-2-(p-1)}/n_1+n_2-p-1} \sim F_{p,n_1+n_2-p-1}(\lambda^2).$$

If $\mu^{(1)} = \mu^{(2)}$, then the F – Distribution is central.

Behrens-Fisher problem

Let $x_\alpha^{(i)}$ (α =1, 2, . . . , n_i ; i =1, 2) be random sample from $N_p(\mu^{(i)}, \Sigma_i)$. Hypothesis of interest is $H_0 : \mu^{(1)} = \mu^{(2)}$. The mean $\bar{x}^{(1)}$ of the first sample is normally distributed with expected value

$E(\bar{x}^{(1)}) = \mu^{(1)}$, and covariance matrix

$$E(\bar{x}^{(1)}-\mu^{(1)})(\bar{x}^{(1)}-\mu^{(1)})' = \frac{1}{n}\Sigma_1, i.e., \bar{x}^{(1)} \sim N_p(\mu^{(1)}, \Sigma_1/n_1).$$

Similarly, $\bar{x}^{(2)} \sim N_p(\mu^{(2)}, \Sigma_2/n_2)$.

Thus, $(\bar{x}^{(1)} - \bar{x}^{(2)}) \sim N_p\left[\mu^{(1)}-\mu^{(2)}, \left(\frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2\right)\right].$

if $n_1 = n_2 = n$.

Let $y_\alpha = x_\alpha^{(1)} - x_\alpha^{(2)}$, (assuming the numbering of the observations in the two samples is independent of the observations themselves), then $y \sim N_p(0, \Sigma_1 + \Sigma_2)$ under H_0.

$$\bar{y} = \frac{1}{n}\sum_{\alpha=1}^{n} y_\alpha = (\bar{x}^{(1)} - \bar{x}^{(2)}) \sim N_p\left[0, \left(\frac{\Sigma_1+\Sigma_2}{n}\right)\right] \Rightarrow \sqrt{n}\bar{y} \sim N_p(0, \Sigma_1+\Sigma_2).$$

Let,

$$S_y = \frac{1}{n-1}\sum_{\alpha=1}^{n-1}(y_\alpha - \bar{y})(y_\alpha - \bar{y})' \text{ or}$$

$$(n-1)S_y = \sum_{\alpha=1}^{n-1} Z_\alpha Z_\alpha{}'$$, where $Z_\alpha \sim N_{p-1}(0, \Sigma_1 + \Sigma_2)$.

Thus, by definition, $T^2 = n\bar{y}' S_y^{-1}\bar{y}$ has T^2 - distribution with $(n-1)$ degrees of freedom.

The critical region is $T^2 \geq \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$.

If $n_1 \neq n_2$, and $n_1 < n_2$.

Define, $y_\alpha = x_\alpha^{(1)} - \sqrt{\frac{n_1}{n_2}} x_\alpha^{(2)} + \frac{1}{\sqrt{n_1 n_2}} \sum_{\beta=1}^{n_1} x_\beta^{(2)} - \frac{1}{n^2} \sum_{\gamma=1}^{n_2} x_\gamma^{(2)}, \alpha = 1, 2, \cdots, n_1,$ then

$$Ey_\alpha = \mu^{(1)} - \sqrt{\frac{n_1}{n_2}} \mu^{(2)} + \frac{1}{\sqrt{n_1 n_2}} \sum_{\beta=1}^{n_1} \mu^{(2)} - \frac{1}{n_2} \sum_{\gamma=1}^{n_2} \mu^{(2)}$$

$$= \mu^{(1)} - \sqrt{\frac{n_1}{n_2}} \mu^{(2)} + \sqrt{\frac{n_1}{n_2}} \mu^{(2)} - \mu^{(2)} = \mu^{(1)} - \mu^{(2)}$$

The covariance matrix of y_α and y_β is

$$E[y_\alpha - E(y_\alpha)][y_\beta - E(y_\beta)]' = E\left[\left[x_\alpha^{(1)} - \mu^{(1)}\right] - \sqrt{\frac{n_1}{n_2}} (x_\alpha^{(2)} - \mu^{(2)}) + \frac{1}{\sqrt{n_1 n_2}} \sum_{\gamma=1}^{n_1} (x_\gamma^{(2)} - \mu^{(2)}) - \frac{1}{n_2} \sum_{\gamma=1}^{n_2} (x_\gamma^{(2)} - \mu^{(2)}) \right]$$

$$\left[\left[x_\beta^{(1)} - \mu^{(1)}\right]' - \sqrt{\frac{n_1}{n_2}} (x_\beta^{(2)} - \mu^{(2)})' + \frac{1}{\sqrt{n_1 n_2}} \sum_{\gamma=1}^{n_1} (x_\gamma^{(2)} - \mu^{(2)})' - \frac{1}{n_2} \sum_{\gamma=1}^{n_2} (x_\gamma^{(2)} - \mu^{(2)})' \right]$$

$$= \Sigma_1 + \frac{n_1}{n_2} \Sigma_2 + \frac{1}{n_1 n_2} n_1 \Sigma_2 + \frac{1}{n_2^2} n_2 \Sigma_2 - 2\sqrt{\frac{n_1}{n_2}} \frac{1}{\sqrt{n_1 n_2}} \Sigma_2 + 2\sqrt{\frac{n_1}{n_2}} \frac{1}{n_2} \Sigma_2 - 2 \frac{1}{\sqrt{n_1 n_2}} \frac{1}{n_2} n_1 \Sigma_2.$$

$$= \Sigma_1 + \Sigma_2 \left(\frac{n_1}{n_2} + \frac{1}{n_2} + \frac{1}{n_2} - \frac{2}{n_2} + \frac{2}{n_2} \sqrt{\frac{n_1}{n_2}} - \frac{2}{n_2} \sqrt{\frac{n_1}{n_2}} \right) = \Sigma_1 + \frac{n_1}{n_2} \Sigma_2 .$$

Hence, $y_\alpha \sim N_p \left(0, \Sigma_1 + \frac{n_1}{n_2} \Sigma_2 \right)$ under H$_0$, then

$$\bar{y} = \frac{1}{n_1} \sum_{\alpha=1}^{n_1} y_\alpha \sim N_p \left(0, \frac{1}{n_1} \left(\Sigma_1 + \frac{n_1}{n_2} \Sigma_2 \right) \right) \Rightarrow \sqrt{n_1} \bar{y} \sim N_p \left(0, \left(\Sigma_1 + \frac{n_1}{n_2} \Sigma_2 \right) \right)$$

Consider $(n_1 - 1)S = \sum_{\alpha=1}^{n_1} (y_\alpha - \bar{y})(y_\alpha - \bar{y})' = \sum_{\alpha=1}^{n_1 - 1} Z_\alpha Z_\alpha{}'$, with Z$_\alpha$ ~ $N_p(0,\ \Sigma_1 + \frac{n_1}{n_2} \Sigma_2)$.

Therefore, by definition

$T^2 = n_1 \bar{y}' S^{-1} \bar{y}.$ This statistic has T^2 - distribution with (n$_1$ −1) degree of freedom.

The critical region of size α is

$$T^2 \geq \frac{(n_1 - 1)p}{n_1 - p} F_{p, n_1 - p}(\alpha).$$

Comparison between Mahalanobis Distance and Euclidean Distance

Aspect	Mahalanobis Distance	Euclidean Distance
Definition and Formula	Measures dissimilarity while considering the covariance structure of the data. It is calculated using the mean vector, covariance matrix, and data point vector.	Measures the straight-line distance between two data points in a multidimensional space. It is calculated as the square root of the sum of squared differences along each dimension.
Sensitivity to Data Distribution	Assumes that the data follows a multivariate normal distribution.	Assumes no specific data distribution; it is applicable to a wide range of data types and distributions.
Robustness to Scaling	Scale-invariant; it is not affected by the scaling of variables.	Sensitive to outliers, extreme values can significantly affect distance calculations.
Handling Correlated Variables	Suitable for datasets with correlated variables; considers variable correlations in the covariance matrix.	Treats variables independently; does not account for correlations between variables.
Dimensionality	Becomes less effective with high-dimensional data due to increased computational complexity and potential data sparsity.	Generally applicable to high-dimensional data, although interpretation can become challenging as dimensions increase.
Outlier Sensitivity	May be less sensitive to outliers due to covariance structure consideration.	It may be less sensitive to outliers due to covariance structure consideration.
Customization of Thresholds	Customizable thresholds can be set to identify outliers or anomalies, providing flexibility.	Thresholds are typically not customized, and outliers are identified based on distance magnitude alone.
Applications	Widely used in various fields, including finance, healthcare, quality control, and image recognition, where correlations between variables are important.	Commonly applied in geometric and spatial analysis, machine learning, and data clustering tasks when correlations between variables are less critical.

SCL PROBLEMS

1. Test for One Sample Hotelling's T^2

A shoe company evaluates new shoe models based on five criteria: style *(x_1)*, comfort *(x_2)*, stability *(x_3)*, cushioning *(x_4)* and durability *(x_5)* with each of the first four criteria evaluated on a scale of 1 to 10 and the durability criteria evaluated on the scale of 1 to 20. Goals for each criterion expected from new products 7, 8, 5, 7 and 9 respectively. Calculate one sample Hotelling's T^2 for the following data:

Subjects	***Style***	***Comfort***	***Stability***	***Cushion***	***Durability***	***Subjects***	***Style***	***Comfort***	***Stability***	***Cushion***	***Durability***
1	6	8	3	5	19	***14***	4	9	10	2	16
2	6	7	3	4	9	***15***	2	9	4	10	14
3	5	7	1	4	16	***16***	7	5	8	6	15
4	10	9	8	4	4	***17***	4	8	8	2	16
5	7	9	7	6	9	***18***	5	10	9	3	11
6	6	6	3	9	17	***19***	7	7	3	7	12
7	5	8	6	7	6	***20***	1	5	2	7	17
8	3	7	3	6	16	***21***	5	6	7	7	20
9	8	8	9	3	8	***22***	4	3	1	2	15
10	8	6	5	3	13	***23***	7	9	6	6	9
11	5	9	5	4	17	***24***	4	5	2	4	12
12	8	8	2	3	5	***25***	8	9	5	7	18
13	5	8	7	5	8						

Procedure

- To calculate the sample mean.
- To calculate the sample covariance.

$$\sigma_{ij} = \sigma_i^2 = \frac{\sum X_i^2}{n} - \overline{X}_i^2$$

$$\sigma_{ij} = \frac{\sum X_i X_j}{n} - \overline{X}_i \overline{X}_j$$

- To calculate T^2

$$T^2 = n(\overline{X} - \mu)^T \, S^{-1} \, (\overline{X} - \mu) \; \sim \chi^2(k)$$

  - where S is the covariance matrix of the sample for X,
  - $\overline{X}$ is the mean of the sample

- Under the null hypothesis

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

If $F_{cal} > F_{crit}$ then we reject the null hypothesis, otherwise we accept it.

Calculation

Hypothesis

H_0: There is no significant difference between the sample means in the five categories and the goals (i.e. population means).

H_1: There is a significant difference between the sample means in the five categories and the goals (i.e. population means).

Calculation of Mean vector and Variance covariance Matrix

Subjects	x_1	x_2	x_3	x_4	x_5	x_1^2	x_2^2	x_3^2	x_4^2	x_5^2	x_1x_2	x_1x_3	x_1x_4	x_1x_5	x_2x_3	x_2x_4	x_2x_5	x_3x_4	x_3x_5	x_4x_5
1	6	8	3	5	19	36	64	9	25	361	48	18	30	114	24	40	152	15	57	95
2	6	7	3	4	9	36	49	9	16	81	42	18	24	54	21	28	63	12	27	36
3	5	7	1	4	16	25	49	1	16	256	35	5	20	80	7	28	112	4	16	64
4	10	9	8	4	4	100	81	64	16	16	90	80	40	40	72	36	36	32	32	16
5	7	9	7	6	9	49	81	49	36	81	63	49	42	63	63	54	81	42	63	54
6	6	6	3	9	17	36	36	9	81	289	36	18	54	102	18	54	102	27	51	153
7	5	8	6	7	6	25	64	36	49	36	40	30	35	30	48	56	48	42	36	42
8	3	7	3	6	16	9	49	9	36	256	21	9	18	48	21	42	112	18	48	96
9	8	8	9	3	8	64	64	81	9	64	64	72	24	64	72	24	64	27	72	24
10	8	6	5	3	13	64	36	25	9	169	48	40	24	104	30	18	78	15	65	39
11	5	9	5	4	17	25	81	25	16	289	45	25	20	85	45	36	153	20	85	68
12	8	8	2	3	5	64	64	4	9	25	64	16	24	40	16	24	40	6	10	15
13	5	8	7	5	8	25	64	49	25	64	40	35	25	40	56	40	64	35	56	40
14	4	9	10	2	16	16	81	100	4	256	36	40	8	64	90	18	144	20	160	32
15	2	9	4	10	14	4	81	16	100	196	18	8	20	28	36	90	126	40	56	140
16	7	5	8	6	15	49	25	64	36	225	35	56	42	105	40	30	75	48	120	90
17	4	8	8	2	16	16	64	64	4	256	32	32	8	64	64	16	128	16	128	32
18	5	10	9	3	11	25	100	81	9	121	50	45	15	55	90	30	110	27	99	33
19	7	7	3	7	12	49	49	9	49	144	49	21	49	84	21	49	84	21	36	84
20	1	5	2	7	17	1	25	4	49	289	5	2	7	17	10	35	85	14	34	119
21	5	6	7	7	20	25	36	49	49	400	30	35	35	100	42	42	120	49	140	140
22	4	3	1	2	15	16	9	1	4	225	12	4	8	60	3	6	45	2	15	30
23	7	9	6	6	9	49	81	36	36	81	63	42	42	63	54	54	81	36	54	54
24	4	5	2	4	12	16	25	4	16	144	20	8	16	48	10	20	60	8	24	48
25	8	9	5	7	18	64	81	25	49	324	72	40	56	144	45	63	162	35	90	126
Total	**140**	**185**	**127**	**126**	**322**	**888**	**1439**	**823**	**748**	**4648**	**1058**	**748**	**686**	**1696**	**998**	**933**	**2325**	**611**	**1574**	**1670**

Mean Vector

Style $\bar{X}_1 = \dfrac{140}{25} = 5.6$ Comfort $\bar{X}_2 = \dfrac{185}{25} = 7.4$

Stability $\bar{X}_3 = \dfrac{127}{25} = 5.08$ Cushion $\bar{X}_4 = \dfrac{126}{25} = 5.04$

Durability $\bar{X}_5 = \dfrac{322}{25} = 12.88$

$$\text{Mean Vector} = \begin{bmatrix} 5.6 \\ 7.4 \\ 5.08 \\ 5.04 \\ 12.88 \end{bmatrix}$$

Covariance Matrix

$$\sigma_{11} = \sigma_1^2 = \frac{888}{25} - (5.6)^2 = 4.16$$

$$\sigma_{22} = \sigma_2^2 = \frac{1439}{25} - (7.4)^2 = 2.8$$

$$\sigma_{33} = \sigma_3^2 = \frac{823}{25} - (5.08)^2 = 7.11$$

$$\sigma_{44} = \sigma_4^2 = \frac{748}{25} - (5.04)^2 = 4.52$$

$$\sigma_{55} = \sigma_5^2 = \frac{4648}{25} - (12.88)^2 = 20.03$$

$$\sigma_{12} = \frac{1058}{25} - (5.6 \times 7.4) = 0.88$$

$$\sigma_{13} = \frac{748}{25} - (5.6 \times 5.08) = 1.472$$

$$\sigma_{14} = \frac{686}{25} - (5.6 \times 5.04) = -0.784$$

$$\sigma_{15} = \frac{1696}{25} - (5.6 \times 12.88) = -4.288$$

$$\sigma_{23} = \frac{998}{25} - (7.4 \times 5.08) = 2.328$$

$$\sigma_{24} = \frac{933}{25} - (7.4 \times 5.04) = 0.024$$

$$\sigma_{25} = \frac{2325}{25} - (7.4 \times 12.88) = -2.312$$

$$\sigma_{34} = \frac{611}{25} - (5.08 \times 5.04) = -1.1632$$

$$\sigma_{35} = \frac{1574}{25} - (5.08 \times 12.88) = -2.4704$$

$$\sigma_{45} = \frac{1670}{25} - (5.04 \times 12.88) = 1.8848$$

$$\text{Covariance Matrix} = \begin{bmatrix} 4.16 & 0.88 & 1.472 & -0.784 & -4.288 \\ 0.88 & 2.8 & 2.328 & 0.024 & -2.312 \\ 1.472 & 2.328 & 7.1136 & -1.1632 & -2.4704 \\ -0.784 & 0.024 & -1.1632 & 4.5184 & 1.8848 \\ -4.288 & -2.312 & -2.4704 & 1.8848 & 20.026 \end{bmatrix}$$

here,

n = 25, k = 5

$$T^2 = n(\overline{X} - \mu)^T S^{-1} (\overline{X} - \mu)$$

$$(\overline{X} - \mu) = \begin{bmatrix} 5.6 \\ 7.4 \\ 5.08 \\ 5.04 \\ 12.88 \end{bmatrix} - \begin{bmatrix} 7 \\ 8 \\ 5 \\ 7 \\ 9 \end{bmatrix} = \begin{bmatrix} -1.4 \\ -0.6 \\ 0.08 \\ -1.96 \\ 3.88 \end{bmatrix}$$

$$b = S^{-1} (\overline{X} - \mu)$$

$$\begin{bmatrix} 4.16 & 0.88 & 1.472 & -0.784 & -4.288 \\ 0.88 & 2.8 & 2.328 & 0.024 & -2.312 \\ 1.472 & 2.328 & 7.1136 & -1.1632 & -2.4704 \\ -0.784 & 0.024 & -1.1632 & 4.5184 & 1.8848 \\ -4.288 & -2.312 & -2.4704 & 1.8848 & 20.026 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} -1.4 \\ -0.6 \\ 0.08 \\ -1.96 \\ 3.88 \end{bmatrix}$$

$$\begin{array}{c} R_1 / 4.16 \\ R_2 - (0.88 \times R_1) \\ R_3 - (1.472 \times R_1) \\ R_4 - (0.784 \times R_1) \\ R_5 - (4.288 \times R_1) \end{array} \begin{bmatrix} 1 & 0.212 & 0.354 & -0.188 & -1.031 \\ 0 & 2.614 & 2.017 & 0.1898 & -1.405 \\ 0 & 2.016 & 6.593 & -0.886 & -0.953 \\ 0 & 0.1898 & -0.886 & 4.371 & 1.077 \\ 0 & -1.405 & -0.890 & 1.077 & 15.606 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} -0.337 \\ -0.304 \\ 0.575 \\ -2.224 \\ 2.437 \end{bmatrix}$$

$$\begin{array}{c} R_1 - (0.212 \times R_2) \\ R_2 / 2.614 \\ R_3 - (2.016 \times R_2) \\ R_4 - (0.1898 \times R_2) \\ R_5 + (1.405 \times R_2) \end{array} \begin{bmatrix} 1 & 0 & 1.191 & -0.204 & -0.917 \\ 0 & 1 & 0.772 & 0.073 & -0.538 \\ 0 & 0 & 5.037 & -1.032 & 0.131 \\ 0 & 0 & -1.032 & 4.357 & 1.179 \\ 0 & 0 & 0.194 & 1.179 & 14.851 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} -0.312 \\ -0.116 \\ 0.810 \\ -2.202 \\ 2.274 \end{bmatrix}$$

$$\begin{array}{c} R_1 - (1.191 \times R_3) \\ R_2 - (0.772 \times R_3) \\ R_3 / 5.037 \\ R_4 + (1.032 \times R_3) \\ R_5 - (0.194 \times R_3) \end{array} \begin{bmatrix} 1 & 0 & 0 & -0.165 & -0.922 \\ 0 & 1 & 0 & 0.231 & -0.558 \\ 0 & 0 & 1 & -0.205 & 0.026 \\ 0 & 0 & 0 & 4.145 & 1.206 \\ 0 & 0 & 0 & 1.218 & 14.846 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} -0.343 \\ -0.240 \\ 0.161 \\ -2.036 \\ 2.242 \end{bmatrix}$$

$$\begin{array}{c} R_1+(0.165\times R_4) \\ R_2-(0.231\times R_4) \\ R_3+(0.205\times R_4) \\ R_4/4.145 \\ R_5-(1.218\times R_4) \end{array}\begin{bmatrix} 1 & 0 & 0 & 0 & -0.874 \\ 0 & 1 & 0 & 0 & -0.625 \\ 0 & 0 & 1 & 0 & 0.086 \\ 0 & 0 & 0 & 1 & 0.291 \\ 0 & 0 & 0 & 0 & 14.492 \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix}=\begin{bmatrix} -0.424 \\ -0.127 \\ 0.060 \\ -0.491 \\ 2.841 \end{bmatrix}$$

$$\begin{array}{c} R_1+(0.874\times R_5) \\ R_2+(0.625\times R_5) \\ R_3-(0.086\times R_5) \\ R_4-(0.291\times R_5) \\ R_5/14.492 \end{array}\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix}=\begin{bmatrix} -0.252 \\ -0.005 \\ 0.043 \\ -0.548 \\ 0.196 \end{bmatrix}$$

$b_1=-0.252$, $b_2=-0.005$, $b_3=0.043$, $b_4=-0.548$ and $b_5=0.196$

$$T^2=25\begin{bmatrix} -1.4 & -0.6 & 0.08 & -1.96 & 3.88 \end{bmatrix}\begin{bmatrix} -0.252 \\ -0.005 \\ 0.043 \\ -0.548 \\ 0.196 \end{bmatrix}$$

= 25× [-1.4×(-0.252) + (-0.6)×(-0.005) + 0.08×0.043 + (-1.96)×(-0.548) + 3.88×(0.196)]

=25×[0.353+0.003+0.003+1.074+0.760] = 25×2.193

T^2 = 54.825

Under the null hypothesis,

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

$$F=\frac{25-5}{5(25-1)}\times 54.825 \sim F(5,20)$$

$$F_{cal}=9.138$$

Table Value

Degrees of freedom = *F(k, n-k)*

= *F(5, 20)*

F_{crit} = 2.71

Conclusion

Since, the calculated value is greater than the critical table value (*9.138* > *2.71*), we reject the null hypothesis. Hence, we conclude that there is a significant difference between the sample means in the five categories and the stated goals.

2. Test for Two Independent Samples Hotelling's T^2

A certain type of tropical disease is characterized by fever, low blood pressure, and body aches. A pharmaceutical company was working on a new drug to treat this type of disease and wanted to determine whether the drug was effective. They took a random sample of 20 people with this type of disease and 18 with a placebo. Apply two sample Hotelling's T^2 method and to test whether the drug is effective at reducing these three symptoms for the following data:

	Drug			**Placebo**		
ID	**Fever (X_1)**	**Blood Pressure (X_2)**	**Body Aches (X_3)**	**Fever (Y_1)**	**Blood Pressure (Y_2)**	**Body Aches (Y_3)**
1	38.4	73	18	40.9	54	14
2	36.8	85	14	39.5	75	18
3	40.0	58	20	39.4	57	24
4	39.8	80	20	38.2	71	24
5	38.6	68	25	39.7	65	22
6	39.1	52	27	38.9	49	30
7	38.9	79	26	38.6	58	25
8	36.8	100	8	39.9	52	17
9	40.4	64	21	41.3	62	18
10	39.4	53	22	38.1	57	20
11	38.0	70	15	39.6	78	19
12	38.6	75	14	37.1	92	15
13	40.1	48	28	39.5	63	13
14	38.1	57	22	40.3	52	25
15	37.2	78	16	41.5	46	27
16	39.5	65	18	39.3	56	14
17	37.3	77	13	37.6	86	16
18	39.1	67	16	40.6	48	21
19	39.9	52	10			
20	37.8	68	13			

Procedure

- To calculate the sample mean.
- To calculate the sample covariance.

$$\sigma_{ij} = \sigma_i^2 = \frac{\sum X_i^2}{n} - \overline{X}_i^2$$

$$\sigma_{ij} = \frac{\sum X_i X_j}{n} - \overline{X}_i \overline{X}_j$$

- To calculate T^2

$$T^2 = (\overline{X} - \overline{Y})\left[S\left(\frac{1}{n_x} + \frac{1}{n_y}\right)\right]^{-1}(\overline{X} - \overline{Y})^T$$

where, S is the pooled sample covariance matrix of *XÂ* and *Y*, namely

$$S = \frac{(n_x - 1)S_X + (n_y - 1)S_Y}{(n_x - 1) + (n_y - 1)}$$

- Under the null hypothesis

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

If $F_{cal} > F_{crit}$ then we reject the null hypothesis, otherwise we accept it.

Calculation

Hypothesis

H_0: There is no significant difference between the mean vectors for the drug and placebo.

H_1: There is significant difference between the mean vectors for the drug and placebo.

Mean and Covariance for Drug

ID	X_1	X_2	X_3	X_1^2	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3
1	38.4	73	18	1474.56	5329	324	2803.2	691.2	1314
2	36.8	85	14	1354.24	7225	196	3128	515.2	1190
3	40.0	58	20	1600	3364	400	2320	800	1160
4	39.8	80	20	1584.04	6400	400	3184	796	1600
5	38.6	68	25	1489.96	4624	625	2624.8	965	1700
6	39.1	52	27	1528.81	2704	729	2033.2	1055.7	1404
7	38.9	79	26	1513.21	6241	676	3073.1	1011.4	2054
8	36.8	100	8	1354.24	10000	64	3680	294.4	800
9	40.4	64	21	1632.16	4096	441	2585.6	848.4	1344
10	39.4	53	22	1552.36	2809	484	2088.2	866.8	1166
11	38.0	70	15	1444	4900	225	2660	570	1050
12	38.6	75	14	1489.96	5625	196	2895	540.4	1050
13	40.1	48	28	1608.01	2304	784	1924.8	1122.8	1344
14	38.1	57	22	1451.61	3249	484	2171.7	838.2	1254
15	37.2	78	16	1383.84	6084	256	2901.6	595.2	1248
16	39.5	65	18	1560.25	4225	324	2567.5	711	1170
17	37.3	77	13	1391.29	5929	169	2872.1	484.9	1001
18	39.1	67	16	1528.81	4489	256	2619.7	625.6	1072
19	39.9	52	10	1592.01	2704	100	2074.8	399	520
20	37.8	68	13	1428.84	4624	169	2570.4	491.4	884
Total	**773.8**	**1369**	**366**	**29962.2**	**96925**	**7302**	**52777.7**	**14222.6**	**24325**

Mean Vector

Fever $\overline{X}_1 = \dfrac{773.8}{20} = 38.69$ Blood pressure $\overline{X}_2 = \dfrac{1369}{20} = 68.45$

Body aches $\overline{X}_3 = \dfrac{366}{20} = 18.3$

$$\text{Mean Vector} = \begin{bmatrix} 38.69 \\ 68.45 \\ 18.3 \end{bmatrix}$$

Covariance Matrix

$$\sigma_{11} = \sigma_1^2 = \frac{29962.2}{20} - (38.69)^2 = 1.19$$

$$\sigma_{22} = \sigma_2^2 = \frac{96925}{20} - (68.45)^2 = 160.85$$

$$\sigma_{33} = \sigma_3^2 = \frac{7302}{20} - (18.3)^2 = 30.21$$

$$\sigma_{12} = \frac{52777.7}{20} - (38.69 \times 68.45) = -9.45$$

$$\sigma_{13} = \frac{14222.6}{20} - (38.69 \times 18.3) = 3.10$$

$$\sigma_{23} = \frac{24325}{20} - (68.45 \times 18.3) = -36.39$$

$$\text{Covariance Matrix} = \begin{bmatrix} 1.19 & -9.45 & 3.10 \\ -9.45 & 160.85 & -36.39 \\ 3.10 & -36.39 & 30.21 \end{bmatrix}$$

Mean and Covariance for Placebo

ID	Y_1	Y_2	Y_3	Y_1^2	Y_2^2	Y_3^2	Y_1Y_2	Y_1Y_3	Y_2Y_3
1	40.9	54	14	1672.81	2916	196	2208.6	572.6	756
2	39.5	75	18	1560.25	5625	324	2962.5	711	1350
3	39.4	57	24	1552.36	3249	576	2245.8	945.6	1368
4	38.2	71	24	1459.24	5041	576	2712.2	916.8	1704
5	39.7	65	22	1576.09	4225	484	2580.5	873.4	1430
6	38.9	49	30	1513.21	2401	900	1906.1	1167	1470
7	38.6	58	25	1489.96	3364	625	2238.8	965	1450
8	39.9	52	17	1592.01	2704	289	2074.8	678.3	884
9	41.3	62	18	1705.69	3844	324	2560.6	743.4	1116
10	38.1	57	20	1451.61	3249	400	2171.7	762	1140
11	39.6	78	19	1568.16	6084	361	3088.8	752.4	1482
12	37.1	92	15	1376.41	8464	225	3413.2	556.5	1380
13	39.5	63	13	1560.25	3969	169	2488.5	513.5	819
14	40.3	52	25	1624.09	2704	625	2095.6	1007.5	1300
15	41.5	46	27	1722.25	2116	729	1909	1120.5	1242
16	39.3	56	14	1544.49	3136	196	2200.8	550.2	784
17	37.6	86	16	1413.76	7396	256	3233.6	601.6	1376
18	40.6	48	21	1648.36	2304	441	1948.8	852.6	1008
Total	***710***	***1121***	***362***	***28031***	***72791***	***7696***	***44039.9***	***14289.9***	***22059***

Mean Vector

Fever $\bar{Y}_1 = \frac{710}{18} = 39.44$ Blood pressure $\bar{Y}_2 = \frac{1121}{18} = 62.28$

Body aches $\bar{Y}_3 = \frac{362}{18} = 20.11$

$$\text{Mean Vector} = \begin{bmatrix} 39.44 \\ 62.28 \\ 20.11 \end{bmatrix}$$

Covariance Matrix

$$\sigma_{11} = \sigma_1^2 = \frac{28031}{18} - (39.44)^2 = 1.41$$

$$\sigma_{22} = \sigma_2^2 = \frac{72791}{18} - (62.28)^2 = 165.42$$

$$\sigma_{33} = \sigma_3^2 = \frac{7969}{18} - (20.11)^2 = 23.10$$

$$\sigma_{12} = \frac{44039.9}{18} - (39.44 \times 62.28) = -9.85$$

$$\sigma_{13} = \frac{14289.9}{18} - (39.44 \times 20.11) = 0.61$$

$$\sigma_{23} = \frac{22059}{18} - (62.28 \times 20.11) = -26.98$$

$$\text{Covariance Matrix} = \begin{bmatrix} 1.41 & -9.85 & 0.61 \\ -9.85 & 165.42 & -26.98 \\ 0.61 & -26.98 & 23.10 \end{bmatrix}$$

Pooled covariance Matrix

$$S = \frac{(n_x - 1)S_X + (n_y - 1)S_Y}{(n_x - 1) + (n_y - 1)}$$

$$\sigma_{11} = \sigma_1^2 = \frac{(19\times 1.19) + (17\times 1.41)}{19+17} = 1.30$$

$$\sigma_{22} = \sigma_2^2 = \frac{(19\times 160.85) + (17\times 165.42)}{19+17} = 163.01$$

$$\sigma_{33} = \sigma_3^2 = \frac{(19\times 30.21) + (17\times 23.10)}{19+17} = 26.85$$

$$\sigma_{12} = \frac{(19\times(-9.45)) + (17\times(-9.85))}{19+17} = -9.64$$

$$\sigma_{13} = \frac{(19\times 3.10) + (17\times 0.61)}{19+17} = 1.93$$

$$\sigma_{23} = \frac{(19\times(-36.39)) + (17\times(-26.98))}{19+17} = -31.94$$

Pooled Covariance Matrix *(S)* = $\begin{bmatrix} 1.30 & -9.63 & 1.93 \\ -9.63 & 163.01 & -31.94 \\ 1.93 & -31.94 & 26.85 \end{bmatrix}$

here, n = 38, k = 3

$$T^2 = (\overline{X} - \overline{Y})\left[S\left(\frac{1}{n_x} + \frac{1}{n_y}\right)\right]^{-1}(\overline{X} - \overline{Y})^T$$

$$(\overline{X} - \overline{Y}) = \begin{bmatrix} 38.69 \\ 68.45 \\ 18.30 \end{bmatrix} - \begin{bmatrix} 39.44 \\ 62.28 \\ 20.11 \end{bmatrix} = \begin{bmatrix} -0.75 \\ 6.17 \\ -1.81 \end{bmatrix}$$

$$\left[S\left(\frac{1}{n_x} + \frac{1}{n_y}\right)\right] = \begin{bmatrix} 1.30 & -9.63 & 1.93 \\ -9.63 & 163.01 & -31.94 \\ 1.93 & -31.94 & 26.85 \end{bmatrix}\left[\frac{1}{20} + \frac{1}{18}\right]$$

$$= \begin{bmatrix} 1.30 & -9.64 & 1.93 \\ -9.64 & 163.01 & -31.94 \\ 1.93 & -31.94 & 26.85 \end{bmatrix}[0.11]$$

$$= \begin{bmatrix} 0.14 & -1.06 & 0.21 \\ -1.06 & 17.93 & -3.51 \\ 0.21 & -3.51 & 2.95 \end{bmatrix}$$

$$b = \left[S\left(\frac{1}{n_x} + \frac{1}{n_y}\right)\right]^{-1}(\overline{X} - \overline{Y})$$

$$\begin{bmatrix} 0.14 & -1.06 & 0.21 \\ -1.06 & 17.93 & -3.51 \\ 0.21 & -3.51 & 2.95 \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} -0.75 \\ 6.17 \\ -1.81 \end{bmatrix}$$

$$\begin{array}{r} R_1/0.14 \\ R_2+(1.06\times R_1) \\ R_3-(0.21\times R_1) \end{array}\begin{bmatrix} 1 & -7.57 & 1.50 \\ 0 & 9.90 & -1.92 \\ 0 & -1.92 & 2.64 \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} -5.36 \\ 0.49 \\ -0.69 \end{bmatrix}$$

$$\begin{array}{r} R_1+(7.57\times R_2) \\ R_2/9.90 \\ R_3+(1.92\times R_2) \end{array}\begin{bmatrix} 1 & 0 & 0.03 \\ 0 & 1 & -0.19 \\ 0 & 0 & 2.26 \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} -4.98 \\ 0.05 \\ -0.59 \end{bmatrix}$$

$$\begin{array}{r} R_1-(0.03\times R_3) \\ R_2+(0.19\times R_3) \\ R_3/2.26 \end{array}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} -4.97 \\ 0.00 \\ -0.26 \end{bmatrix}$$

$b_1 = -4.97$, $b_2 = 0$ and $b_3 = -0.26$

$$T^2 = 38\begin{bmatrix} -0.75 & 6.17 & -1.81 \end{bmatrix}\begin{bmatrix} -4.97 \\ 0.00 \\ -0.26 \end{bmatrix}$$

= [-0.75×(-4.97) + 6.17×0 + (-1.81)×(-0.26)]

= 3.73 + 0 + 0.47

$T^2 = 4.20$

Under the null hypothesis,

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

$$F = \frac{38-3}{3(38-1)}\times 4.20 \sim F(3, 35)$$

$$F_{cal} = 1.32$$

Table Value

Degrees of freedom = *F(k, n-k)*

= *F(3, 35)*

$F_{crit} = 2.87$

Conclusion

Since, the calculated value is less than the critical table value (*1.32 < 2.87*), we do not reject the null hypothesis. Hence, we conclude that there is no significant difference between the mean vectors for the drug and placebo.

3. Test for Paired Sample Hotelling's T^2

A shoe company evaluates two new model shoes, **Model 1** based on five criteria: style *(x_1)*, comfort *(x_2)*, stability *(x_3)*, cushioning *(x_4)* and durability *(x_5)* and **Model 2** based on five criteria: style *(y_1)*, comfort *(y_2)*, stability *(y_3)*, cushioning *(y_4)* and durability *(y_5)* with each of the first four criteria evaluated on a scale of 1 to 10 and the durability criteria evaluated on the scale of 1 to 20. Calculate the paired sample Hotelling's T^2 for the following data:

	Model 1					**Model 2**				
Subjects	**Style**	**Comfort**	**Stability**	**Cushion**	**Durability**	**Style**	**Comfort**	**Stability**	**Cushion**	**Durability**
1	6	8	3	5	19	8	6	5	6	10
2	6	7	3	4	9	8	6	3	6	4
3	5	7	1	4	16	7	5	6	4	17
4	10	9	8	4	4	9	8	6	3	4
5	7	9	7	6	9	8	5	6	8	11
6	6	6	3	9	17	8	7	4	4	13
7	5	8	6	7	6	7	3	6	3	8
8	3	7	3	6	16	6	6	5	8	14
9	8	8	9	3	8	6	9	7	5	12
10	8	6	5	3	13	7	5	9	6	11
11	5	9	5	4	17	7	5	4	6	15
12	8	8	2	3	5	5	7	4	4	6
13	5	8	7	5	8	6	4	6	4	12
14	4	9	10	2	16	8	7	8	5	12
15	2	9	4	10	14	5	6	5	7	12
16	7	5	8	6	15	10	5	7	6	6
17	4	8	8	2	16	9	6	9	5	11
18	5	10	9	3	11	8	7	10	5	5
19	7	7	3	7	12	6	2	5	3	8
20	1	5	2	7	17	5	7	5	5	8
21	5	6	7	7	20	8	4	8	8	10
22	4	3	1	2	15	3	2	4	4	15
23	7	9	6	6	9	8	6	3	6	12
24	4	5	2	4	12	5	4	6	5	9
25	8	9	5	7	18	6	3	4	8	8

Procedure

- To calculate the difference between model 1 and model 2.
- To calculate the sample mean.
- To calculate the sample covariance.

$$\sigma_{ij} = \sigma_i^2 = \frac{\sum X_i^2}{n} - \bar{X}_i^2 \quad \text{and} \quad \sigma_{ij} = \frac{\sum X_i X_j}{n} - \bar{X}_i \bar{X}_j$$

- To calculate T^2

$$T^2 = n(\bar{X} - \bar{Y})^T S^{-1} (\bar{X} - \bar{Y}) \sim \chi^2(k)$$

  - where S is the covariance matrix of the sample for X,
  - $\bar{X}$ is the mean of the sample

- Under the null hypothesis

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

If $F > F_{crit}$ then we reject the null hypothesis, otherwise we accept it.

Calculation

Hypothesis

H_0: There is no significant difference between the two shoe models.

H_1: There is a significant difference between the two shoe models.

Calculation of difference between model 1 and model 2

	Model 1					*Model 2*					*Z = Model 1 - Model 2*				
Subjects	x_1	x_2	x_3	x_4	x_5	y_1	y_2	y_3	y_4	y_5	z_1	z_2	z_3	z_4	z_5
1	6	8	3	5	19	8	6	5	6	10	-2	2	-2	-1	9
2	6	7	3	4	9	8	6	3	6	4	-2	1	0	-2	5
3	5	7	1	4	16	7	5	6	4	17	-2	2	-5	0	-1
4	10	9	8	4	4	9	8	6	3	4	1	1	2	1	0
5	7	9	7	6	9	8	5	6	8	11	-1	4	1	-2	-2
6	6	6	3	9	17	8	7	4	4	13	-2	-1	-1	5	4
7	5	8	6	7	6	7	3	6	3	8	-2	5	0	4	-2
8	3	7	3	6	16	6	6	5	8	14	-3	1	-2	-2	2
9	8	8	9	3	8	6	9	7	5	12	2	-1	2	-2	-4
10	8	6	5	3	13	7	5	9	6	11	1	1	-4	-3	2
11	5	9	5	4	17	7	5	4	6	15	-2	4	1	-2	2
12	8	8	2	3	5	5	7	4	4	6	3	1	-2	-1	-1
13	5	8	7	5	8	6	4	6	4	12	-1	4	1	1	-4
14	4	9	10	2	16	8	7	8	5	12	-4	2	2	-3	4
15	2	9	4	10	14	5	6	5	7	12	-3	3	-1	3	2
16	7	5	8	6	15	10	5	7	6	6	-3	0	1	0	9
17	4	8	8	2	16	9	6	9	5	11	-5	2	-1	-3	5
18	5	10	9	3	11	8	7	10	5	5	-3	3	-1	-2	6
19	7	7	3	7	12	6	2	5	3	8	1	5	-2	4	4
20	1	5	2	7	17	5	7	5	5	8	-4	-2	-3	2	9
21	5	6	7	7	20	8	4	8	8	10	-3	2	-1	-1	10
22	4	3	1	2	15	3	2	4	4	15	1	1	-3	-2	0
23	7	9	6	6	9	8	6	3	6	12	-1	3	3	0	-3
24	4	5	2	4	12	5	4	6	5	9	-1	1	-4	-1	3
25	8	9	5	7	18	6	3	4	8	8	2	6	1	-1	10

Calculation of Mean vector and Variance covariance Matrix

Subjects	z_1	z_2	z_3	z_4	z_5	z_1^2	z_2^2	z_3^2	z_4^2	z_5^2	z_1z_2	z_1z_3	z_1z_4	z_1z_5	z_2z_3	z_2z_4	z_2z_5	z_3z_4	z_3z_5	z_4z_5
1	-2	2	-2	-1	9	4	4	4	1	81	-4	4	2	-18	-4	-2	18	2	-18	-9
2	-2	1	0	-2	5	4	1	0	4	25	-2	0	4	-10	0	-2	5	0	0	-10
3	-2	2	-5	0	-1	4	4	25	0	1	-4	10	0	2	-10	0	-2	0	5	0
4	1	1	2	1	0	1	1	4	1	0	1	2	1	0	2	1	0	2	0	0
5	-1	4	1	-2	-2	1	16	1	4	4	-4	-1	2	2	4	-8	-8	-2	-2	4
6	-2	-1	-1	5	4	4	1	1	25	16	2	2	-10	-8	1	-5	-4	-5	-4	20
7	-2	5	0	4	-2	4	25	0	16	4	-10	0	-8	4	0	20	-10	0	0	-8
8	-3	1	-2	-2	2	9	1	4	4	4	-3	6	6	-6	-2	-2	2	4	-4	-4
9	2	-1	2	-2	-4	4	1	4	4	16	-2	4	-4	-8	-2	2	4	-4	-8	8
10	1	1	-4	-3	2	1	1	16	9	4	1	-4	-3	2	-4	-3	2	12	-8	-6
11	-2	4	1	-2	2	4	16	1	4	4	-8	-2	4	-4	4	-8	8	-2	2	-4
12	3	1	-2	-1	-1	9	1	4	1	1	3	-6	-3	-3	-2	-1	-1	2	2	1
13	-1	4	1	1	-4	1	16	1	1	16	-4	-1	-1	4	4	4	-16	1	-4	-4
14	-4	2	2	-3	4	16	4	4	9	16	-8	-8	12	-16	4	-6	8	-6	8	-12
15	-3	3	-1	3	2	9	9	1	9	4	-9	3	-9	-6	-3	9	6	-3	-2	6
16	-3	0	1	0	9	9	0	1	0	81	0	-3	0	-27	0	0	0	0	9	0
17	-5	2	-1	-3	5	25	4	1	9	25	-10	5	15	-25	-2	-6	10	3	-5	-15
18	-3	3	-1	-2	6	9	9	1	4	36	-9	3	6	-18	-3	-6	18	2	-6	-12
19	1	5	-2	4	4	1	25	4	16	16	5	-2	4	4	-10	20	20	-8	-8	16
20	-4	-2	-3	2	9	16	4	9	4	81	8	12	-8	-36	6	-4	-18	-6	-27	18
21	-3	2	-1	-1	10	9	4	1	1	100	-6	3	3	-30	-2	-2	20	1	-10	-10
22	1	1	-3	-2	0	1	1	9	4	0	1	-3	-2	0	-3	-2	0	6	0	0
23	-1	3	3	0	-3	1	9	9	0	9	-3	-3	0	3	9	0	-9	0	-9	0
24	-1	1	-4	-1	3	1	1	16	1	9	-1	4	1	-3	-4	-1	3	4	-12	-3
25	2	6	1	-1	10	4	36	1	1	100	12	2	-2	20	6	-6	60	-1	10	-10
Total	***-33***	***50***	***-18***	***-8***	***69***	***151***	***194***	***122***	***132***	***653***	***-54***	***27***	***10***	***-177***	***-11***	***-8***	***116***	***2***	***-91***	***-34***

Mean Vector

Style $\bar{z}_1 = \frac{-33}{25} = -1.32$ Comfort $\bar{z}_2 = \frac{50}{25} = 2$

Stability $\bar{z}_3 = \frac{-18}{25} = -0.72$ Cushion $\bar{z}_4 = \frac{-8}{25} = -0.32$

Durability $\bar{z}_5 = \frac{69}{25} = 2.76$

$$\text{Mean Vector} = \begin{bmatrix} -1.32 \\ 2.00 \\ -0.72 \\ -0.32 \\ 2.76 \end{bmatrix}$$

Covariance Matrix

$$\sigma_{11} = \sigma_1^2 = \frac{151}{25} - (-1.32)^2 = 4.30$$

$$\sigma_{22} = \sigma_2^2 = \frac{194}{25} - (2)^2 = 3.76$$

$$\sigma_{33} = \sigma_3^2 = \frac{122}{25} - (-0.72)^2 = 4.36$$

$$\sigma_{44} = \sigma_4^2 = \frac{132}{25} - (-0.32)^2 = 5.18$$

$$\sigma_{55} = \sigma_5^2 = \frac{653}{25} - (2.76)^2 = 18.50$$

$$\sigma_{12} = \frac{-54}{25} - (-1.32 \times 2) = 0.48$$

$$\sigma_{13} = \frac{27}{25} - (-1.32 \times (-0.72)) = 0.13$$

$$\sigma_{14} = \frac{10}{25} - (-1.32 \times (-0.32)) = -0.02$$

$$\sigma_{15} = \frac{-177}{25} - (-1.32 \times (2.76)) = -3.44$$

$$\sigma_{23} = \frac{-11}{25} - (2 \times (-0.72)) = 1.00$$

$$\sigma_{24} = \frac{-8}{25} - (2 \times (-0.32)) = 0.32$$

$$\sigma_{25} = \frac{116}{25} - (2 \times 2.76) = -0.88$$

$$\sigma_{34} = \frac{2}{25} - (-0.72 \times (-0.32)) = -0.15$$

$$\sigma_{35} = \frac{-91}{25} - (-0.72 \times 2.76) = -1.65$$

$$\sigma_{45} = \frac{-34}{25} - (-0.32 \times 2.76) = -0.48$$

$$\text{Covariance Matrix} = \begin{bmatrix} 4.30 & 0.48 & 0.13 & -0.02 & -3.44 \\ 0.48 & 3.76 & 1.00 & 0.32 & -0.88 \\ 0.13 & 1.00 & 4.36 & -0.15 & -1.65 \\ -0.02 & 0.32 & -0.15 & 5.18 & -0.48 \\ -3.44 & -0.88 & -1.65 & -0.48 & 18.50 \end{bmatrix}$$

here,

n = 25, k = 5

$$T^2 = n\,\bar{z}^T\, S^{-1}\,\bar{z}$$

$$(\bar{x} - \bar{y}) = \bar{z} = \begin{bmatrix} -1.32 \\ 2.00 \\ -0.72 \\ -0.32 \\ 2.76 \end{bmatrix}$$

$$b = S^{-1}\,(\overline{X} - \overline{Y})$$

$$\begin{bmatrix} 4.30 & 0.48 & 0.13 & -0.02 & -3.44 \\ 0.48 & 3.76 & 1.00 & 0.32 & -0.88 \\ 0.13 & 1.00 & 4.36 & -0.15 & -1.65 \\ -0.02 & 0.32 & -0.15 & 5.18 & -0.48 \\ -3.44 & -0.88 & -1.65 & -0.48 & 18.50 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} -1.32 \\ 2.00 \\ -0.72 \\ -0.32 \\ 2.76 \end{bmatrix}$$

$$\begin{array}{r} R_1 / 4.30 \\ R_2 - (0.48 \times R_1) \\ R_3 - (0.13 \times R_1) \\ R_4 - (-0.02 \times R_1) \\ R_5 - (-3.44 \times R_1) \end{array} \begin{bmatrix} 1 & 0.11 & 0.03 & -0.01 & -0.80 \\ 0 & 3.71 & 0.99 & 0.32 & -0.50 \\ 0 & 0.99 & 4.36 & -0.15 & -1.55 \\ 0 & 0.32 & -0.15 & 5.18 & -0.49 \\ 0 & -0.50 & -1.55 & -0.49 & 15.75 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} -0.31 \\ 2.15 \\ -0.68 \\ -0.33 \\ 1.70 \end{bmatrix}$$

$$\begin{array}{r} R_1 - (0.11 \times R_2) \\ R_2 / 3.71 \\ R_3 - (0.99 \times R_2) \\ R_4 - (0.32 \times R_2) \\ R_5 + (0.50 \times R_2) \end{array} \begin{bmatrix} 1 & 0 & 0 & -0.01 & -0.78 \\ 0 & 1 & 0.27 & 0.09 & -0.13 \\ 0 & 0 & 4.09 & -0.24 & -1.42 \\ 0 & 0 & -0.23 & 5.15 & -0.45 \\ 0 & 0 & -1.42 & -0.45 & 15.69 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} -0.37 \\ 0.58 \\ -1.25 \\ -0.51 \\ 1.99 \end{bmatrix}$$

$$\begin{array}{r} R_1 - (0 \times R_3) \\ R_2 - (0.27 \times R_3) \\ R_3 / 4.09 \\ R_4 + (0.23 \times R_3) \\ R_5 + (1.42 \times R_3) \end{array} \begin{bmatrix} 1 & 0 & 0 & -0.01 & -0.78 \\ 0 & 1 & 0 & 0.10 & -0.04 \\ 0 & 0 & 1 & -0.06 & -0.35 \\ 0 & 0 & 0 & 5.14 & -0.53 \\ 0 & 0 & 0 & -0.53 & 15.19 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} -0.37 \\ 0.66 \\ -0.31 \\ -0.58 \\ 1.56 \end{bmatrix}$$

$$\begin{array}{r} R_1 + (0.01 \times R_4) \\ R_2 - (0.10 \times R_4) \\ R_3 + (0.06 \times R_4) \\ R_4 / 5.14 \\ R_5 + (0.53 \times R_4) \end{array} \begin{bmatrix} 1 & 0 & 0 & 0 & -0.79 \\ 0 & 1 & 0 & 0 & -0.03 \\ 0 & 0 & 1 & 0 & -0.35 \\ 0 & 0 & 0 & 1 & -0.10 \\ 0 & 0 & 0 & 0 & 15.14 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} -0.37 \\ 0.67 \\ -0.31 \\ -0.11 \\ 1.50 \end{bmatrix}$$

$$\begin{array}{r} R_1+(0.79\times R_5)\\ R_2+(0.03\times R_5)\\ R_3+(0.35\times R_5)\\ R_4+(0.10\times R_5)\\ R_5/15.14 \end{array}\begin{bmatrix}1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{bmatrix}\begin{bmatrix}b_1\\b_2\\b_3\\b_4\\b_5\end{bmatrix}=\begin{bmatrix}-0.29\\0.72\\-0.28\\-0.10\\0.10\end{bmatrix}$$

$b_1=-0.29$, $b_2=-0.72$, $b_3=-0.28$, $b_4=-0.10$ and $b_5=0.10$

$$T^2=25\begin{bmatrix}-1.32 & 2.00 & -0.72 & -0.32 & 2.76\end{bmatrix}\begin{bmatrix}-0.29\\0.72\\-0.28\\-0.10\\0.10\end{bmatrix}$$

= 25× [-1.32×(-0.29) + 2)×(0.72) + (-0.72)×(-0.28) + (-0.32)×(-0.10) + 2.76×(0.10)]

=25×[0.39+1.44+0.20+0.03+0.27] = 25×2.33

$T^2 = 58.36$

Under the null hypothesis,

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

$$F=\frac{25-5}{5(25-1)}\times 58.36 \sim F(5,20)$$

$$F_{cal}=9.73$$

Table Value

Degrees of freedom = *F(k, n-k)*
= *F(5, 20)*

$F_{crit} = 2.71$

Conclusion

Since, the calculated value is greater than the critical table value (*9.73* > *2.71*), we reject the null hypothesis. Hence, we conclude that there is a significant difference between the two shoe models.

4. Test for Mahalanobis D^2 Statistic

Calculate Mahalanobis distance for the following data:

Months (2021)	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
Stocks (X_1)	5.00%	1.90%	9.80%	-2.00%	2.10%	2.50%	1.60%	6.20%	5.30%	-0.40%
Bonds (X_2)	-0.30%	1.20%	1.30%	0.30%	0.00%	2.10%	0.90%	1.40%	0.60%	0.70%

Procedure

- To calculating Mean vector **(**μ**).**
- To finding Covariance Matrix (Σ).

$$\sigma_{ij} = \sigma_i^2 = \frac{\sum X_i^2}{n} - \overline{X}_i^2 \quad \text{and} \quad \sigma_{ij} = \frac{\sum X_i X_j}{n} - \overline{X}_i \overline{X}_j$$

- To calculate Mahalanobis Distance Components

$$D^2 = [x_i - \bar{x})^T \Sigma^{-1}(x_i - \bar{x})]^{0.5}$$

where, x_i = an object vector

$\bar{x}$ = arithmetic mean vector

Σ = variance covariance matrix

- Mahalanobis Distance **(**D_M**)** calculation

Calculation

Calculate Mean Vector and Variance-covariance Matrix

X_1	X_1	X_1^2	X_2^2	X_1X_2
5	-0.3	25	0.09	-1.5
1.9	1.2	3.61	1.44	2.28
9.8	1.3	96.04	1.69	12.74
-2	0.3	4	0.09	-0.6
2.1	0	4.41	0	0
2.5	2.1	6.25	4.41	5.25
1.6	0.9	2.56	0.81	1.44
6.2	1.4	38.44	1.96	8.68
5.3	0.6	28.09	0.36	3.18
-0.4	0.7	0.16	0.49	-0.28
32	**8.2**	**208.56**	**11.34**	**31.19**

Mean Vector

$$\text{Stocks } \overline{X}_1 = \frac{32\%}{10} = 3.2\% \qquad \text{Bonds } \overline{X}_2 = \frac{8.2\%}{10} = 0.82\% \qquad \text{Mean Vector} = \begin{bmatrix} 3.2 \\ 0.82 \end{bmatrix}$$

Covariance Matrix

$$\sigma_{11} = \sigma_1^2 = \frac{208.56\%}{10} - (3.2\%)^2 = 10.616$$

$$\sigma_{22} = \sigma_2^2 = \frac{11.34}{10} - (0.82)^2 = 0.4616$$

$$\sigma_{12} = \frac{31.19}{10} - (3.2 \times 0.82) = 0.495$$

$$\text{Covariance Matrix } (\Sigma) = \begin{bmatrix} 10.616 & 0.495 \\ 0.495 & 0.4616 \end{bmatrix}$$

Inverse of Covariance

$$\Sigma^{-1} = \frac{1}{|\Sigma|}(adj\,\Sigma) = \begin{vmatrix} 10.616 & 0.495 \\ 0.495 & 0.4616 \end{vmatrix} = |4.900 - 0.245| = |4.655|$$

$$adj\,\Sigma = \begin{bmatrix} 0.4616 & -0.495 \\ -0.495 & 10.616 \end{bmatrix}^T$$

$$adj\,\Sigma = \begin{bmatrix} 0.4616 & -0.495 \\ -0.495 & 10.616 \end{bmatrix}$$

$$\Sigma^{-1} = \frac{1}{4.655}\begin{bmatrix} 0.4616 & -0.495 \\ -0.495 & 10.616 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} 0.0992 & -0.1063 \\ -0.1063 & 2.2806 \end{bmatrix}$$

Mahalanobis Distance Components

$$D^2 = [x_i - \bar{x})^T \Sigma^{-1}(x_i - \bar{x})]^{0.5}$$

$$= \left[\begin{bmatrix} 1.8 & -1.3 & 6.6 & -5.2 & -1.1 & -0.7 & -1.6 & 3.0 & 2.1 & -3.6 \\ -1.12 & 0.38 & 0.48 & -0.52 & -0.82 & 1.28 & 0.08 & 0.58 & -0.22 & -0.12 \end{bmatrix}^T \begin{bmatrix} 0.0992 & -0.1063 \\ -0.1063 & 2.2806 \end{bmatrix} \begin{bmatrix} 1.8 & -1.3 & 6.6 & -5.2 & -1.1 & -0.7 & -1.6 & 3.0 & 2.1 & -3.6 \\ -1.12 & 0.38 & 0.48 & -0.52 & -0.82 & 1.28 & 0.08 & 0.58 & -0.22 & -0.12 \end{bmatrix} \right]^{0.5}$$

c_{11} = 1.8 × 1.8 + (-1.3) × (-1.3) + 6.6 × 6.6 + (-5.2) × (-5.2) + (-1.1) × (-1.1) + (-0.7) × (-0.7) + (-1.6) × (-1.6) + 3 × 3 + 2.1 × 2.1 + (-3.6) × (-3.6) = **106.16**

c_{12} = 1.8 × (-1.12) + (-1.3) × 0.38 + 6.6 × 0.48 + (-5.2) × (-0.52) + (-1.1) × (-0.82) + (-0.7) × 1.28 + (-1.6) × 0.08 + 3 × 0.58 + 2.1 × (-0.22) + (-3.6) × (-0.12) = **4.95**

c_{21} = -1.12 × 1.8 + 0.38 × (-1.3) + 0.48 × 6.6 + (-0.52) × (-5.2) + (-0.82) × (-1.1) + 1.28 × (-0.7) + 0.08 × (-1.6) + 0.58 × 3 + (-0.22) × 2.1 + (-0.12) × (-3.6) = **4.95**

c_{22} = -1.12 × (-1.12) + 0.38 × 0.38 + 0.48 × 0.48 + (-0.52) × (-0.52) + (-0.82) × (-0.82) + 1.28 × 1.28 + 0.08 × 0.08 + 0.58 × 0.58 + (-0.22) × (-0.22) + (-0.12) × (-0.12) = **4.616**

$$= \left[\begin{bmatrix} 106.16 & 4.95 \\ 4.95 & 4.616 \end{bmatrix} \times \begin{bmatrix} 0.0992 & -0.1063 \\ -0.1063 & 2.2806 \end{bmatrix} \right]^{0.5}$$

$$= \begin{bmatrix} 10.0049 & 0.0042 \\ 0.0004 & 10.0011 \end{bmatrix}^{0.5}$$

$$D^2 = \begin{bmatrix} 3.163 & 0.065 \\ 0.02 & 3.162 \end{bmatrix}$$

Result

The Mahalanobis distance is a positive value that quantifies are similar between a Stocks (X_1) and Bonds (X_2), and the mean of the data set. Hence, we conclude that two variables are correlated.