Testing hypotheses on the multivariate normal mean vector

In clinical tests (see BMDP outputs) usually more than one variables are observed on several patients. The variables are far not independent of each other, but we assume that their expectations, variances, and pairwise covariances (or correlations) characterize their joint distribution. Such a random vector is denoted by $\mathbf{X}$ and we assume that it follows d-dimensional normal distribution with expectation (population mean) vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$. Notation: $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \mathbf{C})$, where the d-dimensional vector $\boldsymbol{\mu}$ contains the expectations of the components of $\mathbf{X}$, and the entries of the $d \times d$ symmetric, positive definite matrix $\mathbf{C}$ are the pairwise covariances c_{ij}'s for $i \neq j$ and c_{ii} is the variance of the ith component of $\mathbf{X}$.

- *1-sample case*: Let $\mathbf{X}_1, \dots, \mathbf{X}_n \sim \mathcal{N}_d(\boldsymbol{\mu}, \mathbf{C})$ be i.i.d. sample with $n > d$ and $\mathbf{C}$ unknown. For testing

  $$H_0 \,:\, \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{versus} \quad H_1 \,:\, \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

  the test statistic

  $$T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^T \hat{\mathbf{C}}^{*-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}_0)$$

  is used that under H_0 follows Hotelling's T^2-distribution with parameters $n-1$ and d, where $\hat{\mathbf{C}}^*$ is the corrected empirical covariance matrix. It can be transformed into F-distribution, and so, under H_0: $F = \frac{n-d}{d} \frac{T^2}{n-1} \sim \mathcal{F}(d, n-d)$. The rejection region with Type I. error probability α is $R = \{F \geq F_\alpha(d, n-d)\}$.

- *2-sample case*: Let $\mathbf{X}_1, \dots, \mathbf{X}_n \sim \mathcal{N}_d(\boldsymbol{\mu}_1, \mathbf{C})$ and $\mathbf{Y}_1, \dots, \mathbf{Y}_m \sim \mathcal{N}_d(\boldsymbol{\mu}_2, \mathbf{C})$ be i.i.d. samples, $\mathbf{X}_i$'s are independent of $\mathbf{Y}_j$'s and they have the same unknown covariance matrix $\mathbf{C}$. For testing

  $$H_0 \,:\, \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{versus} \quad H_1 \,:\, \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

  the test statistic

  $$T^2 = \frac{nm}{n+m} (\bar{\mathbf{X}} - \bar{\mathbf{Y}})^T \hat{\mathbf{C}}^{*-1} (\bar{\mathbf{X}} - \bar{\mathbf{Y}}) = \frac{nm}{n+m} D^2(\mathbf{X}, \mathbf{Y})$$

  is used that under H_0 follows Hotelling's T^2-distribution with parameters $n+m-2$ and d, where $\hat{\mathbf{C}}^* = \mathbf{S}/(n+m-2)$ is the so-called *pooled covariance matrix* and

  $$\mathbf{S} = \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T + \sum_{j=1}^{m} (\mathbf{Y}_j - \bar{\mathbf{Y}})(\mathbf{Y}_j - \bar{\mathbf{Y}})^T.$$

  Hence, $\hat{\mathbf{C}}^*$ is an unbiased estimator of the common $\mathbf{C}$. Further, $D^2(\mathbf{X}, \mathbf{Y})$ denotes the Mahalanobis-distance between the two populations. Consequently, under H_0: $F = \frac{n+m-d-1}{d} \frac{T^2}{n+m-2} \sim \mathcal{F}(d, n+m-d-1)$.
  If the significance of the test is α, we reject H_0 if F exceeds the upper α-point of the above F-distribution. Program packages usually output the smallest α at which H_0 can just be rejected based on T^2. This α^* is sometimes called P-value. For testing the equality of covariance matrices, and acting if they are not equal, further theory and distributions of the multivariate statistics are needed, which exceed the scope of this note.

Example: 49 countries are classified into 2 groups according to their economic policies (Group I. and Group II.). We register 4 macroeconomic (yearly) indicators in each of the countries, for which the averages within the groups are as follows:

	$I.(n = 37)$	$II.(m = 12)$
1.	12.57	8.75
2.	9.57	5.33
3.	11.49	8.50
4.	7.97	4.75

Investigate, whether the two groups differ significantly based on these 4 indicators. Assume that the indicators follow 4-variate normal distribution with the same within-group covariance matrix, for which, the inverse of the pooled covariance matrix is

$$\hat{\mathbf{C}}^{*-1} = \begin{pmatrix} 0.52 & -0.28 & -0.12 & -0.12 \\ -0.28 & 0.38 & -0.08 & -0.02 \\ -0.12 & -0.08 & 0.30 & -0.04 \\ -0.12 & -0.02 & -0.04 & 0.42 \end{pmatrix}.$$

Solution: $T^2 = 22.05$ and $F = \frac{37+12-4-1}{4} \frac{T_2^2}{37+12-2} = 5.16$ follows $\mathcal{F}(4, 44)$ if the null-hypothesis of equality of the macroeconomic indicators holds. Since this F-value is larger than the upper 0.01-point of the corresponding F-distribution, we can reject the null-hypothesis with significance 0.01. It means that the two groups differ significantly with respect to these indicators (it is 0.01 probability that we claim this without any reason).