Testing hypotheses on the multivariate normal mean vector

I. Testing the multivariate normal mean vector in case of known covariance matrices

- *1-sample case*: Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$ be i.i.d. sample with $n > p$ and $\mathbb{C} > 0$ known. For testing

$$H_0 : \mathbf{m} = \mathbf{m}_0 \quad \text{versus} \quad H_1 : \mathbf{m} \neq \mathbf{m}_0$$

the statistic

$$Z_1 = (\bar{\mathbf{X}} - \mathbf{m}_0)^T \left(\frac{\mathbf{C}}{n}\right)^{-1} (\bar{\mathbf{X}} - \mathbf{m}_0) = n(\bar{\mathbf{X}} - \mathbf{m}_0)^T \mathbf{C}^{-1} (\bar{\mathbf{X}} - \mathbf{m}_0)$$

is used that follows $\chi^2(p)$-distribution under H_0 (generalization of the z-test, sometimes called u-test). This is the immediate consequence of Proposition 3 of Lesson 2 and of the fact, that under H_0: $\bar{\mathbf{X}} - \mathbf{m}_0 \sim \mathcal{N}_p(\mathbf{0}, \frac{1}{n}\mathbf{C})$.

Therefore, we reject H_0 with significance (Type I error) α if $Z_1 \geq \chi^2_\alpha(p)$, where $\chi^2_\alpha(p)$ is the upper α-point ($1-\alpha$ quantile-, or $100(1-\alpha)$ percentile value) of the $\chi^2(p)$-distribution.

- *2-sample case*: Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathcal{N}_p(\mathbf{m}_1, \mathbf{C}_1)$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m \sim \mathcal{N}_p(\mathbf{m}_2, \mathbf{C}_2)$ be i.i.d. samples ($\mathbf{X}_i$ is not necessarily identically distributed with $\mathbf{Y}_j$, but they are independent $\forall i, j$). Suppose that $n, m > p$ and $\mathbf{C}_1 > 0$, $\mathbf{C}_2 > 0$ are known covariance matrices. For testing

$$H_0 : \mathbf{m}_1 = \mathbf{m}_2 \quad \text{versus} \quad H_1 : \mathbf{m}_1 \neq \mathbf{m}_2$$

the statistic

$$Z_2 = (\bar{\mathbf{X}} - \bar{\mathbf{Y}})^T \left(\frac{\mathbf{C}_1}{n} + \frac{\mathbf{C}_2}{m}\right)^{-1} (\bar{\mathbf{X}} - \bar{\mathbf{Y}}) = (\bar{\mathbf{X}} - \bar{\mathbf{Y}})^T \left(\frac{m\mathbf{C}_1 + n\mathbf{C}_2}{nm}\right)^{-1} (\bar{\mathbf{X}} - \bar{\mathbf{Y}})$$

is used that follows $\chi^2(p)$-distribution under H_0.

Therefore, we reject H_0 with significance (Type I error) α if $Z_2 \geq \chi^2_\alpha(p)$, where $\chi^2_\alpha(p)$ is the upper α-point ($1-\alpha$ quantile-, or $100(1-\alpha)$ percentile value) of the $\chi^2(p)$-distribution.

In the special case of $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}$:

$$Z_2 = \frac{nm}{n+m} (\bar{\mathbf{X}} - \bar{\mathbf{Y}})^T \mathbf{C}^{-1} (\bar{\mathbf{X}} - \bar{\mathbf{Y}}) = \frac{nm}{n+m} \cdot D^2(\mathbf{X}, \mathbf{Y}),$$

where D^2 denotes the *Mahalanobis-distance* between the two populations given by the data matrices $\mathbf{X}$ and $\mathbf{Y}$.

This was the generalization of the one- and two-sample, two tail z-test (u-test).

Note that if the sample size(s) is (are) 'large' ($n, m \geq 30p^2$), then even the unknown covariance matrices can be used in the above formulas, since they can be estimated with a 'good' precision.

To extend the Student's t-test to the multivariate situation of unknown covariance matrices (and 'small' sample sizes), we need some definitions.

Definition 1 *Let* $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ *and* $\mathbf{W} \sim \mathcal{W}_p(n, \mathbf{I}_p)$ *be a random vector and a random matrix, independent of each other* ($n > p$)*. Then the random variable*

$$T^2 = n\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X}$$

is said to follow (centered) Hotelling's T^2*-distribution with parameters* n *and* p (n *is also called degree of freedom).*

Note that the Hotelling's T^2 is a generalization of the Student's t-distribution, whereas in the $p = 1$ case $T^2 \equiv (t)^2$.

It is easy to see the following.

Proposition 1 *In the case of* $\mathbf{X} \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$ *and* $\mathbf{W} \sim \mathcal{W}_p(n, \mathbf{C})$*,*

$$T^2 = n(\mathbf{X} - \mathbf{m})^T\mathbf{W}^{-1}(\mathbf{X} - \mathbf{m})$$

also follows the above Hotelling's T^2*-distribution with parameters* n *and* p*.*

Without proof, we state that the Hotelling's T^2-distribution is in fact a Fisher's F-distribution with appropriate parameters.

Theorem 1 $\frac{n-p+1}{p} \cdot \frac{T^2}{n} \sim \mathcal{F}(p, n-p+1)$.

II. Testing the multivariate normal mean vector in case of unknown covariance matrices

- *1-sample case*: Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$ be i.i.d. sample with $n > p$ and $\mathbf{C} > 0$ unknown. For testing

  $$H_0 : \mathbf{m} = \mathbf{m}_0 \quad \text{versus} \quad H_1 : \mathbf{m} \neq \mathbf{m}_0$$

  the statistic

  $$T_1^2 = (n-1)(\bar{\mathbf{X}} - \mathbf{m})^T\hat{\mathbf{C}}^{-1}(\bar{\mathbf{X}} - \mathbf{m}) = n(\bar{\mathbf{X}} - \mathbf{m})^T\hat{\mathbf{C}}^{*-1}(\bar{\mathbf{X}} - \mathbf{m})$$

  is used that under H_0 follows Hotelling's T^2-distribution with parameters $n-1$ and p, where $\hat{\mathbf{C}} = \mathbf{S}/n$ and $\hat{\mathbf{C}}^* = \mathbf{S}/(n-1)$. Hence, under H_0: $F = \frac{n-p}{p}\frac{T_1^2}{n-1} \sim \mathcal{F}(p, n-p)$.

- *2-sample case*: Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathcal{N}_p(\mathbf{m}_1, \mathbf{C})$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m \sim \mathcal{N}_p(\mathbf{m}_2, \mathbf{C})$ 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

  $$H_0 : \mathbf{m}_1 = \mathbf{m}_2 \quad \text{versus} \quad H_1 : \mathbf{m}_1 \neq \mathbf{m}_2$$

  the statistic

  $$T_2^2 = \frac{nm}{n+m}(\bar{\mathbf{X}} - \bar{\mathbf{Y}})^T\hat{\mathbf{C}}^{*-1}(\bar{\mathbf{X}} - \bar{\mathbf{Y}}) = \frac{nm}{n+m}D^2(\mathbf{X}, \mathbf{Y})$$

  is used that under H_0 follows Hotelling's T^2-distribution with parameters $n+m-2$ and p, 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: $F = \frac{n+m-p-1}{p} \frac{T_2^2}{n+m-2} \sim \mathcal{F}(p, n+m-p-1)$.

If the significance of the test is α, we reject H_0 if T_2^2 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_2^2. 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 (see the recommended literature).

To derive the above sampling distributions of the T_1^2 and T_2^2 statistics, we use the known transformations and definition of the Hotelling's T^2-distribution.

EXERCISES

- Prove that the 1-sample Hotelling's T^2-test is a likelihood ratio test.
- **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: $T_2^2 = 22.05$ 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).

- The effects of 3 stimulating pills (A, B, C) are investigated on 20 young men from the point of view of reaction time (sec/100):

$$\bar{X}_A = 21.05 \qquad \bar{X}_B = 21.65 \qquad \bar{X}_C = 28.95$$

$$\mathbf{S_X} = \begin{pmatrix} 45.2 & 43.6 & 32.6 \\ 43.6 & 53.2 & 36.4 \\ 32.6 & 36.4 & 49.4 \end{pmatrix}.$$

Investigate, whether the effects of the 3 pills differ significantly, with 95% confidence. Hint: use a self-control test for the differences $B - A$, $C - B$!

Solution: Because it is a self-control test (the samples are not independent),

$$H_0 : \mathbf{m}_A = \mathbf{m}_B = \mathbf{m}_C$$

is equivalent to

$$H_0 : \mathbf{m}_B - \mathbf{m}_A = 0, \quad \mathbf{m}_C - \mathbf{m}_B = 0.$$

Introducing $\mathbf{Y} = (Y_1, Y_2)^T = (X_B - X_A, X_C - X_B)^T$, $\mathbf{Y} = \mathbf{AX}$, where $\mathbf{X} = (X_A, X_B, X_C)^T$ and

$$\mathbf{A} = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}.$$

Then the above 0-hypothesis is equivalent to

$$H_0 : \mathbb{E}\mathbf{Y} = \mathbf{0}.$$

This can be tested with a one-sample Hotelling T^2 test with $p = 2$ and the transformed multiple of the empirical covariance matrix is

$$\mathbf{S_Y} = \mathbf{AS_X A}^T.$$