MULTIVARIATE STATISTICS, Lesson 6.
Continuation: testing the mean vector in multivariate normal model

- *Definition*: 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).

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

- *Proposition*: In 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.

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

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

1. *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, $F = \frac{n-p}{p}\frac{T_1^2}{n-1} \sim \mathcal{F}(p, n-p)$.

2. *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

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

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

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

   is the *pooled variance*. Hence, $\hat{\mathbf{C}}^*$ is unbiased estimator of $\mathbf{C}$. Further, D^2 denotes the Mahalanobis-distance. Consequently, $F = \frac{n+m-p-1}{p}\frac{T_2^2}{n+m-2} \sim \mathcal{F}(p, n+m-p-1)$.