

CHAPTER 1

Introduction to multivariate distributions-4

1.1

1.2

1.3

1.4 Testing methods

1.4.1 Tests and confidence regions for μ

Σ known :

As we have seen that $\bar{\mathbf{X}} \sim N(\mu, \frac{1}{n}\Sigma)$, we can deduce that $n(\bar{\mathbf{X}} - \mu)^t \Sigma^{-1}(\bar{\mathbf{X}} - \mu) \sim \chi_p^2$ and therefore the region $\{\mu : n(\bar{\mathbf{X}} - \mu)^t \Sigma^{-1}(\bar{\mathbf{X}} - \mu) \leq \chi_p^2(\alpha)\}$ is a critical region. To test $H_0 : \mu = \mu_0$ against $\mu \neq \mu_0$, the LRT criterion is indeed the quantity $n(\bar{\mathbf{X}} - \mu_0)^t \Sigma^{-1}(\bar{\mathbf{X}} - \mu_0)$. However, it should be noted that this test statistic can behave very poorly for some other alternative hypothesis, and hence needs to be used carefully.

Σ unknown :

The problem becomes more intensive when the population covariance Σ is unknown. The straight-forward generalization of the univariate t-statistics for the multivariate data is the Hotelling's T^2 statistics,

$$T^2 = (n - 1)(\bar{\mathbf{X}} - \mu_0)^t S^{-1}(\bar{\mathbf{X}} - \mu_0)$$

It can be shown that T^2 is a function of the LRT criterion for testing $H_0 : \mu = \mu_0$ against $\mu \neq \mu_0$. To this end, the likelihood is

$$L(\mu, \Sigma) = \frac{1}{|2\pi\Sigma|^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^n (\mathbf{X}_i - \mu)^t \Sigma^{-1} (\mathbf{X}_i - \mu)\right)$$

Under H_0 , the MLE of Σ turns out to be

$$\hat{\Sigma}_0 = \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_i - \mu_0)(\mathbf{X}_i - \mu_0)^t = \hat{\Sigma} + (\bar{\mathbf{X}} - \mu_0)(\bar{\mathbf{X}} - \mu_0)^t$$

In the general unrestricted case, we have the usual MLE-s $\bar{\mathbf{X}}$ and V respectively. Clearly, $L(\mu_0, \hat{\Sigma}_0) = \frac{1}{|2\pi\hat{\Sigma}_0|^{n/2}} \exp(-np/2)$ and $L(\hat{\mu}, \hat{\Sigma}) = \frac{1}{|2\pi\hat{\Sigma}|^{n/2}} \exp(-np/2)$.

Therefore, the likelihood ratio criterion is

$$\begin{aligned} \lambda &= \frac{L(\mu_0, \hat{\Sigma}_0)}{L(\hat{\mu}, \hat{\Sigma})} \\ &= \left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_0|} \right)^{n/2} \\ &= \left(1 + \frac{T^2}{n-1} \right)^{n/2} \end{aligned}$$

(For the last equality, observe that, for any A and \mathbf{b} , we have

$$\begin{vmatrix} A & \mathbf{b} \\ -\mathbf{b}^t & 1 \end{vmatrix} = |A + \mathbf{b}\mathbf{b}^t| = |A|(1 + \mathbf{b}^t A^{-1} \mathbf{b})$$

Therefore, $|\hat{\Sigma}_0| = |\hat{\Sigma} + (\bar{\mathbf{X}} - \mu_0)(\bar{\mathbf{X}} - \mu_0)^t| = |\hat{\Sigma}|(1 + (\bar{\mathbf{X}} - \mu_0)^t \hat{\Sigma}^{-1} (\bar{\mathbf{X}} - \mu_0))$

This quantity T^2 is a pivotal quantity, i.e. its distribution does not depend on μ or Σ under H_0 . It follows the so-called *Hotelling's* T^2 distribution, which can be defined as follows.

Let $Z \sim N_p(0, \Psi)$ and $W \sim W_p(\Psi, m)$, independent to each other. Then the quantity $Z^t W^{-1} Z$ is known to have a T^2 distribution with m d.f.

According to our result, the quantity $T^2 = (n-1)(\bar{\mathbf{X}} - \mu)^t S^{-1} (\bar{\mathbf{X}} - \mu)$ has a T^2 distribution with d.f. $n-1$. It is monotonically related to an F distribution, and can be used to compute its quantiles.

1.4.2 Other testing problems

Two-sample problems :

Let $\mathbf{X}_1, \dots, \mathbf{X}_{N_1}$ be a sample from $N(\mu^{(1)}, \Sigma)$ and $\mathbf{X}_{N_1+1}, \dots, \mathbf{X}_{N_1+N_2}$ be a sample from $N(\mu^{(2)}, \Sigma)$. We need to test $H_0 : \mu^{(1)} = \mu^{(2)}$. Writing $\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2$ as the respective sample means, and $S = \frac{1}{N_1+N_2-2} (\sum_{i=1}^{N_1} (\mathbf{X}_i - \bar{\mathbf{X}}_1)(\mathbf{X}_i - \bar{\mathbf{X}}_1)^t + \sum_{i=N_1+1}^{N_1+N_2} (\mathbf{X}_i - \bar{\mathbf{X}}_2)(\mathbf{X}_i - \bar{\mathbf{X}}_2)^t)$ as the pooled variance, the LRT test statistic turns out to be a function of

$$T^2 = \frac{N_1 N_2}{N_1 + N_2} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^t S^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$$

This T^2 statistic has $N_1 + N_2 - 2$ d.f.

Linear hypothesis :

Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ comes from a normal distribution with parameters μ and Σ . We are interested in testing the hypothesis $H_0 : \mathbf{R}\mu = \mathbf{r}$, i.e. whether the mean vector μ belong to lower dimension hyperplane or not. Without loss of generality, let \mathbf{R} be of full row rank, say q (otherwise, either we can reduce the equation $\mathbf{R}\mu = \mathbf{r}$ to a full row rank hypothesis, or we have an inconsistency and cannot hope to have any solution). The corresponding T^2 statistics turn out to be

$$T^2 = (n - 1)(\mathbf{R}\bar{\mathbf{X}} - \mathbf{r})^t (\mathbf{R}\mathbf{S}\mathbf{R}^t)^{-1} (\mathbf{R}\bar{\mathbf{X}} - \mathbf{r})$$

which has a T^2 distribution with $n - p + q$ d.f. Examples include testing whether the mean vector is symmetric, i.e. $\mu_1 = \dots = \mu_p$, or whether any particular variable has mean zero, and so on.

Behren's Fisher problem :

Consider the two sample problem, except for the two populations having different covariance matrices Σ_1 and Σ_2 . As we seek to test the hypothesis $H_0 : \mu^{(1)} = \mu^{(2)}$, we observe that $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$ have normal distribution with means and covariances $\mu^{(1)}, \mu^{(2)}, \Sigma_1/N_1$ and Σ_2/N_2 respectively, besides being independent. Therefore,

$$\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 \sim N(\mu^{(1)} - \mu^{(2)}, \frac{1}{N_1}\Sigma_1 + \frac{1}{N_2}\Sigma_2)$$

However, $\sum_{i=1}^{N_1} (\mathbf{X}_i - \bar{\mathbf{X}}_1)(\mathbf{X}_i - \bar{\mathbf{X}}_1)^t + \sum_{i=N_1+1}^{N_1+N_2} (\mathbf{X}_i - \bar{\mathbf{X}}_2)(\mathbf{X}_i - \bar{\mathbf{X}}_2)^t$ does not have a Wishart distribution any more since Σ_1 is not same as Σ_2 .

In the simple case $N_1 = N_2 = N$ (say), we can proceed by taking $\mathbf{Y}_i = \mathbf{X}_i - \mathbf{X}_{i+N}$, and use a simple T^2 test using $\bar{\mathbf{Y}}^t S_Y^{-1} \bar{\mathbf{Y}}$, since $\mathbf{Y}_i \sim N(\mu^{(1)} - \mu^{(2)}, \Sigma_1 + \Sigma_2)$. However, this T^2 has only $N - 1$ d.f.

See Anderson, page 119 for the general case, using Scheffe's test.

Union intersection tests :

This technique was devised to test multiple hypotheses simultaneously. To this end, consider a random sample \mathbf{X} from the distribution $N(\mu, I)$, and we wish to test whether $\mu = 0$ or not. For any fixed vector \mathbf{a} , we have $y_a = \mathbf{a}^t \mathbf{X} \sim N(\mathbf{a}^t \mu, \mathbf{a}^t \mathbf{a})$, and the corresponding null hypothesis turns out to be $H_{0a} : y_a \sim N(0, \mathbf{a}^t \mathbf{a})$. In particular, H_0 can be seen as an intersection of all such H_{0a} , where \mathbf{a} varies over all p-vectors.

To test H_{0a} , $z_a = \frac{y_a}{\sqrt{\mathbf{a}^t \mathbf{a}}}$ is the z-score and hence a rejection area would be $R_a = \{z_a : z_a^2 > c^2\}$. Then, the Union Intersection Principle states that the hypothesis H_0 should be accepted if and only if all H_{0a} are accepted, i.e. $R = \bigcup R_a$ will serve as the rejection region for H_0 .

The UIT criteria are important to study since a further analysis can lead us to more detailed idea about which set of hypothesis are still true, and which set fail to hold. In the above example, if H_0 is rejected, one can find out which R_{0a} is responsible for that, and see which hypothesis went wrong.

1.4.3 Tests for Σ

For any specific hypothesis on Σ , when μ is unknown, we need to carry out the MLE for both the constrained and the unconstrained case. Clearly, $\bar{\mathbf{X}}$ serves as the MLE for μ in either scenario and S remains the unconstrained MLE for Σ . To continue, let $\hat{\Sigma}_0$ denote the constrained MLE for Σ under H_0 .

Then, the log-likelihood ratio statistics can be computed as follows :

$$\begin{aligned} l_0^* &= l(\bar{\mathbf{X}}, \hat{\Sigma}_0) = -\frac{n}{2} \log |2\pi \hat{\Sigma}_0| - \frac{n}{2} \text{tr}(\hat{\Sigma}_0^{-1} S) \\ l_1^* &= l(\bar{\mathbf{X}}, S) = -\frac{n}{2} \log |2\pi S| - \frac{np}{2} \end{aligned}$$

Therefore,

$$-2 \log \lambda = 2(l_1^* - l_0^*) = n \text{tr}(\hat{\Sigma}_0^{-1} S) - n \log |\hat{\Sigma}_0^{-1} S| - np$$

Let l_1, \dots, l_p denote the eigenvalues of the matrix $\hat{\Sigma}_0^{-1} S$; a_0 and g_0 being their arithmetic mean and the geometric means respectively. Then, the statistic becomes

$$-2 \log \lambda = na_0 - n \log g_0 - np$$

To go further, we need to study the distributions of eigenvalues for random matrices, and that will be done later in the course. Asymptotically, the above statistics follows a χ^2 distribution, where the degrees of freedom is given by the number of parameters restricted by the hypothesis. Now we consider some special type of hypotheses about Σ that appears frequently.

The hypothesis $\Sigma = \Sigma_0$:

Taking $A = \Sigma_0^{-1} S$ as the corresponding matrix, we need to compute its eigenvalues to use the above formula.

Test for sphericity or other pattern :

Here, the hypothesis $H_0 : \Sigma = k\Sigma_0$ ($\Sigma = kI$ in case of sphericity) is considered for unknown k . The MLE of k turns out to be $\hat{k} = \text{tr}(\Sigma_0^{-1} S)/p$. In this case, the above statistic simplifies to $np \log(a_0/g_0)$ where a_0 and g_0 are the eigenvalues of $\Sigma_0^{-1} S$.

Tests for independence :

let us partition the multivariate observations in two sets of variables, with $p = p_1 + p_2$ components. We partition the mean vector $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and let

$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ be the corresponding partition. The independence between the two sets of variables transpires to $H_0 : \Sigma_{12} = 0$.

Partition S as $\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$. Then, under the null, the estimate of Σ is $\hat{\Sigma}_0 = \begin{pmatrix} S_{11} & 0 \\ 0 & S_{22} \end{pmatrix}$. It turns out that $\text{tr}(\hat{\Sigma}_0^{-1}S)$ is p , and we are left with $|\hat{\Sigma}_0^{-1}S| = \frac{|S|}{|S_{11}||S_{22}|}$ as the test statistics. Again, we need to learn the distribution of the eigenvalues to find the distribution of the test statistic.