UNIVERSITAT DE BARCELONA

Inference about a Mean Vector (results)

Prof. Miquel Salicrú
Prof. Sergi Civit

Inference about a Mean Vector μ

A natural generalization of the squared univariate distance t is the multivariate analog Hotelling's T^2:

Hotelling's T^2

$$T^2 = (\bar{\mathbf{X}} - \boldsymbol{\mu})' \left(\frac{1}{n}\mathbf{S}\right)^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}) = n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})$$

To test

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$$

$$H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

The T^2 statistic can be rewritten as

$$T^2 = \sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)' \left(\frac{\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'}{n-1}\right)^{-1} \sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)$$

Inference about a Mean Vector μ

When the null hypothesis is true, the T^2 statistic can be written as the product of two multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and a Wishart $W_{p,n-1}(\boldsymbol{\Sigma})$.

Relation between T^2 and F

$$T^2 \sim \frac{(n-1)p}{(n-p)} F_{p,n-p}$$

The General Likelihood Ratio Method

Let $\boldsymbol{\theta}$ be the vector of all the unknown parameters that take values in some parameter space $\boldsymbol{\Theta}$ (i.e., $\boldsymbol{\theta} \in \boldsymbol{\Theta}$)

For example, in the p-dimensional multivariate normal case,

$$\boldsymbol{\theta} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_p; \sigma_{11}, \ldots, \sigma_{1p}; \sigma_{21}, \ldots, \sigma_{2p}, \ldots, \sigma_{p1}, \ldots, \sigma_{pp}]$$

Also let $L(\boldsymbol{\theta})$ be the likelihood function obtained by evaluating the joint density of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ at their observed values $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$.

A likelihood ratio test of $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ is rejected in favour of $H_0 : \boldsymbol{\theta} \notin \boldsymbol{\Theta}_0$ if

$$\Lambda = \frac{\max\limits_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} L(\boldsymbol{\theta})}{\max\limits_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta})} < c$$

For a relatively large sample size n, under the null hypothesis,

$$-2\ln(\Lambda) = -2\ln \frac{\max_{\boldsymbol{\theta}\in\Theta_0} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})} \sim \chi^2_{\nu-\nu_0}$$

Paired Comparisons

Let x_{lij} be the value of the i^{th} variable taken from the j^{th} observation of the l^{th} group.

For $g = 2$ groups, create p new variables D_{ij}:

$$D_{lij} = X_{1ij} - X_{2ij} \quad i = 1, \ldots, p \quad j = 1, \ldots, n$$

$$\mathbf{D}_j = \begin{bmatrix} D_{1j} \\ D_{2j} \\ \vdots \\ D_{pj} \end{bmatrix}$$

Assuming that,

$$E(\mathbf{D}_j) = \delta \qquad \text{cov}(\mathbf{D}_j) = \Sigma_{\mathbf{D}}$$

Paired Comparisons

If the $\mathbf{D}_1, \ldots, \mathbf{D}_n$ are independent random vectors, then

$$T^2 = (\bar{\mathbf{D}} - \boldsymbol{\delta})' \left(\frac{1}{n}\mathbf{S}_D\right)^{-1} (\bar{\mathbf{D}} - \boldsymbol{\delta})$$

$$\bar{\mathbf{D}}_j = \frac{1}{n}\sum_{j=1}^{n} \mathbf{D}_j \qquad \mathbf{S}_D = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{D}_i - \bar{\mathbf{D}})(\mathbf{D}_i - \bar{\mathbf{D}})'$$

and we know that

$$T^2 \sim \frac{(n-1p)}{n-p} F_{p,n-p}$$

Hypothesis tests for the mean difference vector δ

Let $\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n$ be he observed difference vectors from a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_d)$ distribution.

The hypothesis testing,

$$H_0 : \boldsymbol{\delta} = 0$$

$$H_1 : \boldsymbol{\delta} \neq 0$$

will be rejected at a level of significance α, if

$$T^2 = n\bar{\mathbf{d}}'\mathbf{S}_d^{-1}\bar{\mathbf{d}} \sim \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$$

Comparing mean vectors from two independent populations

Now to test the hypothesis

$$H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \delta$$

we consider the squared distance from the sample estimate $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$ from the hypothesized difference δ_0

$$E(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$$

Independence of the samples implies,

$$\text{Cov}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) = \frac{1}{n_1}\boldsymbol{\Sigma}_1 + \frac{1}{n_2}\boldsymbol{\Sigma}_2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{\Sigma}$$

Comparing mean vectors from two independent populations

$$\hat{\boldsymbol{\Sigma}} = \mathbf{S}_{\text{pooled}}$$

the estimator of the covariance is,

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{\text{pooled}}$$

as a result

$$T^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))' \left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{\text{pooled}}\right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))$$
$$\sim \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p, n_1+n_2-p-1}(\alpha)$$

Example

When the covariance structures are not equal (i.e.,$\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$), any measure of distance (such as T^2) will depend on the unknowns $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ when at least one of the sample sizes n_1 and n_2 is small relative to p. However, if both sample sizes n_1 and n_2 are large relative to p, we can avoid the complexities due to unequal covariance matrices when making inferences about the difference between the mean vectors $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

Under such conditions we have that

$$T^2 = (\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2 - (\mu_1 - \mu_2))' \left(\left(\frac{1}{n_1} \mathrm{S}_1 + \frac{1}{n_2} \mathrm{S}_2 \right) \right)^{-1} (\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2 - (\mu_1 - \mu_2)) \sim \chi_p^2(\alpha)$$