

Inferência para a Distribuição Normal Multivariada: parte 3

Prof. Caio Azevedo

Inferência para duas populações normais multivariadas

- Considere duas populações (grupos) independentes, das quais retiramos duas amostras aleatórias de tamanhos n_1 e n_2 , respectivamente.
- Por suposição, temos que $\mathbf{X}_{ij} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, em que $i = 1, 2$ (grupo) e $j = 1, 2, \dots, n_i$ (indivíduo). Notação: X_{ijk} observação referente à variável k do indivíduo j do grupo i .

Inferência para duas populações normais multivariadas

- Resultando na seguinte matriz de dados ($n = n_1 + n_2$):

$$\mathbf{X}_{(n \times p)} = \begin{bmatrix} X_{111} & X_{112} & \dots & X_{11p} \\ X_{121} & X_{122} & \dots & X_{12p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n_11} & X_{1n_12} & \dots & X_{1n_1p} \\ \text{---} & \text{---} & \text{---} & \text{---} \\ X_{211} & X_{212} & \dots & X_{21p} \\ X_{221} & X_{222} & \dots & X_{22p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{2n_21} & X_{2n_22} & \dots & X_{2n_2p} \end{bmatrix}$$

Teste para a igualdade entre os vetores de médias

- Desejamos testar $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\Delta}$ vs $H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \boldsymbol{\Delta}$, em que $\boldsymbol{\Delta}_{(p \times 1)}$ é um vetor conhecido, considerando que $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ (desconhecida).
- Defina $\bar{\mathbf{X}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{X}_{ij} = \frac{1}{n_i} \left[\sum_{j=1}^{n_i} X_{ij1} \quad \sum_{j=1}^{n_i} X_{ij2} \quad \dots \quad \sum_{j=1}^{n_i} X_{ijp} \right]', i = 1, 2$.
- Temos que $\mathbf{Y} = \bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 \sim N_p \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \boldsymbol{\Sigma} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right)$ (exercício).
- Candidata à estatística do teste:
$$T = \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\Delta})' \hat{\boldsymbol{\Sigma}}^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\Delta}).$$
- Seja $\hat{\boldsymbol{\Sigma}}$ algum estimador conveniente de $\boldsymbol{\Sigma}$.

Teste para a igualdade entre os vetores de médias

- Sob a suposição de que $\Sigma_1 = \Sigma_2 = \Sigma$, um estimador não viciado de Σ é dado por (exercício):

$$\mathbf{S}_P^2 = \frac{1}{n_1 + n_2 - 2} [(n_1 - 1) \mathbf{S}_1^2 + (n_2 - 1) \mathbf{S}_2^2].$$

- Por outro lado, temos que $(n_i - 1) \mathbf{S}_i^2 \stackrel{ind.}{\sim} W_p(n_i - 1, \Sigma)$.
- Resultado: Se $W_i \stackrel{ind.}{\sim} W_p(k_i, \Sigma)$, $i = 1, 2$, então
 $W = W_1 + W_2 \sim W_p(k_1 + k_2, \Sigma)$.
- Logo: $(n_1 + n_2 - 2) \mathbf{S}_P^2 \sim W_p(n_1 + n_2 - 2, \Sigma)$.

Teste para a igualdade entre os vetores de médias

- Portanto:

$T^2 = \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \Delta)' (\mathbf{S}_p^2)^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \Delta)$ segue uma distribuição T^2 de Hotelling.

- Logo, sob H_0 , $F = \left[\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} \right] T^2 \sim F_{(p, n_1 + n_2 - p - 1)}$.

- Defina: $\bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$, $i = 1, 2$ e

$$\mathbf{s}_p^2 = \frac{1}{n_1 + n_2 - 2} [(n_1 - 1) \mathbf{s}_1^2 + (n_2 - 1) \mathbf{s}_2^2].$$

Teste para a igualdade entre os vetores de médias

■ Resumo sobre a estatística F :

- Nível descritivo: $p = P(F > f_{calc} | \boldsymbol{\mu} = \boldsymbol{\mu}_0)$, sob

$H_0, F \sim F_{(p, n_1+n_2-p-1)}$, em que

$$f_{calc} = \left[\frac{n_1+n_2-p-1}{(n_1+n_2-2)p} \right] \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\Delta})' (\mathbf{s}_p^2)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\Delta}).$$

- Função poder: $1 - \beta = P(F > f_c | \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \alpha)$, sob $H_1, F \sim$

$$F_{(p, n_1+n_2-p-1, \delta)}, \delta = \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 - \boldsymbol{\Delta})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 - \boldsymbol{\Delta}),$$

em que f_c é o valor crítico para um dado α (nível de significância).

- Poder do teste estimado: $\tilde{\phi} = \widetilde{1 - \beta} = P(\tilde{F} > f_c | \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \alpha)$, em que

$$\tilde{F} \sim F_{(p, n_1+n_2-p-1, \tilde{\delta})}, \tilde{\delta} =$$

$$\left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\Delta})' (\mathbf{s}_p^2)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\Delta}).$$

Conjunto de dados de Potthoff and Roy

- Aplicação para comparar dois grupos: feminino e masculino.
- Objetivo : Testar se $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ vs $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ ($\Delta = 0_{(4 \times 1)}$).
- Resultados: $f_{calc} = 3,63(0,0203)$, $\tilde{\phi} = \widetilde{1 - \beta} = 0,2408$.

Teste para combinações lineares para diferenças entre dois vetores médias

- Extensível para o caso $H_0 : \mathbf{R}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \boldsymbol{\Delta}$ vs $H_1 : \mathbf{R}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \neq \boldsymbol{\Delta}$ (exercício).
- Se $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$.
 - Teste da razão de verossimilhanças (distribuição assintótica) (exercício).
 - Modelos Lineares Multivariados (na forma vetorial). Veremos adiante.

Teste para uma única matriz de covariâncias

- Supondo uma única população, podemos estar interessados em testar $H_0 : \Sigma = \Sigma_0$, vs $H_1 : \Sigma \neq \Sigma_0$, em que $\Sigma_{0(p \times p)}$ é uma matriz conhecida.

- Dois exemplos:

$$\Sigma_0 = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p^2 \end{bmatrix}; \Sigma_0 = \begin{bmatrix} \sigma^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{12} & \sigma^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \dots & \sigma^2 \end{bmatrix}.$$

- Outra possibilidade: $\Sigma_0 = \sigma^2 \mathbf{I}_{(p \times p)}$.

- Solução: Teste da razão de verossimilhanças (exercício).

Teste de igualdade de matrizes de covariâncias

- A suposição de homocedasticidade é requerida por algumas metodologias de análise multivariada: MANOVA, Análise discriminante, entre outras.
- Suponha agora G grupos independentes, tais que $\mathbf{X}_{ij} \stackrel{ind.}{\sim} N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, \dots, G$ e que queremos testar se $H_0 : \boldsymbol{\Sigma}_1 = \dots = \boldsymbol{\Sigma}_G$ vs H_1 : pelo menos uma diferença.
- A estatística do t.r.v é tal que (exercício):

$$\Lambda \propto \prod_{i=1}^G \left[\frac{|\mathbf{S}_i^2|}{|\mathbf{S}_P^2|} \right]^{(n_i-1)/2}$$

$$\mathbf{S}_P^2 = \frac{1}{\sum_{i=1}^G (n_i - 1)} \left[\sum_{i=1}^G (n_i - 1) \mathbf{S}_i^2 \right]; \mathbf{S}_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\bar{\mathbf{X}}_i - \mathbf{x}_{ij}) (\bar{\mathbf{X}}_i - \mathbf{x}_{ij})'$$

Cont.

- Sob H_0 , $-2 \ln \Lambda \approx \chi^2_{(\nu)}$, para $n_g, g = 1, 2, \dots, G$; suficientemente grandes, em que $\nu = (G - 1)p(p + 1)/2$.
- Correção proposta por Box para melhorar a performance da estatística acima é:

$$\begin{aligned} Q_B &= (1 - u)(-2 \ln \Lambda) = \\ &= (1 - u) \left\{ \left[\sum_{i=1}^G (n_i - 1) \right] \ln |\mathbf{S}_P^2| - \sum_{i=1}^G \left[(n_i - 1) \ln |\mathbf{S}_i^2| \right] \right\} \end{aligned}$$

$$\text{em que } u = \left[\sum_{i=1}^G \frac{1}{n_i - 1} - \frac{1}{\sum_{i=1}^G (n_i - 1)} \right] \left[\frac{2p^2 + 3p - 1}{6(p+1)(g-1)} \right]$$

- Sob H_0 , $Q_B \approx \chi^2_{(\nu)}$, para $n_g, g = 1, 2, \dots, G$; suficientemente grandes.

Aplicação ao conjunto de dados de Potthoff and Roy

■ Resultados: $q_{B(calc)} = 17,33(0,0673)$.

■ Estimativas das matrizes de covariâncias:

grupo	d8	d10	d12	d14
1	4,51	3,35	4,33	4,36
1	3,35	3,62	4,03	4,08
1	4,33	4,03	5,59	5,47
1	4,36	4,08	5,47	5,94
2	6,02	2,29	3,63	1,61
2	2,29	4,56	2,19	2,81
2	3,63	2,19	7,03	3,24
2	1,61	2,81	3,24	4,35