Inferência para a distribuição normal multivariada: Parte 2

Prof. Caio Azevedo

8 de setembro de 2009

- Considere duas populações (grupos) independentes das quais retiramos duas a.a.'s 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 $= 1,2,\ldots,n_i$ (indivíduo). Notação: X_{ijk} observação referente à variável k do indivíduo j do grupo i.
- Resultando na seguinte matriz de dados ($n = n_1 + n_2$):

$$\mathbf{X}_{(n\times p)} = \begin{bmatrix} X_{111} & X_{112} & \cdots & X_{11p} \\ X_{121} & X_{122} & \cdots & X_{12p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n_1 1} & X_{1n_1 2} & \cdots & X_{1n_1 p} \\ --- & --- & --- & --- \\ X_{211} & X_{212} & \cdots & X_{21p} \\ X_{221} & X_{222} & \cdots & X_{22p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{2n_2 1} & X_{2n_2 2} & \cdots & X_{2n_2 p} \end{bmatrix}$$

- 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).
- Temos que $\mathbf{Y} = \overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 \sim N_p\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \boldsymbol{\Sigma}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$
- Estatística do teste: $T^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} \left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 - \boldsymbol{\Delta}\right)' \widehat{\boldsymbol{\Sigma}}^{-1} \left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 - \boldsymbol{\Delta}\right)$.
- $\widehat{\boldsymbol{\Sigma}}$: estimador conveniente de $\boldsymbol{\Sigma}$.

- O estimador de máxima verossimilhança (e.m.v) de $\boldsymbol{\Sigma}_i, i = 1, 2$ é
  $\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i\right)\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i\right)' = \frac{n_i - 1}{n_i}\mathbf{S}_i^2$, em que
  $\overline{\mathbf{X}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{X}_{ij}$.
- Sob a suposição de que $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ o e.m.v de $\boldsymbol{\Sigma}$ é dado por:
  $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n_1+n_2}\left[(n_1 - 1)\,\mathbf{S}_1^2 + (n_2 - 1)\,\mathbf{S}_2^2\right] = \frac{n_1+n_2-2}{n_1+n_2}\mathbf{S}_P^2$, em que (exercício):

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

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

- Logo: $(n_1 + n_2 - 2)\,\mathbf{S}_P^2 \sim W_p(n_1 + n_2 - 2, \boldsymbol{\Sigma})$.
- Portanto: $T^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 - \boldsymbol{\Delta})' \left(\mathbf{S}_P^2\right)^{-1} (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 - \boldsymbol{\Delta})$.
- Sob H_0, $F = \left[\frac{n_1+n_2-p-1}{(n_1+n_2-2)p}\right] T^2 \sim F_{(p,n_1+n_2-p-1)}$.
- Nível descritivo: $p = P(F > f_{calc} | \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\Delta})$.
- Função do poder do Teste:
  $1 - \beta = P(F > f_c | \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\Delta}, \alpha)$, sob $H_1, F \approx \chi^2_p(\gamma), \gamma = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 - \boldsymbol{\Delta})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 - \boldsymbol{\Delta})$ para n suficientemente grande (Teorema de Slutsky).
- Poder do teste estimado:
  $\widehat{1 - \beta} = P(\widehat{F} > f_c | \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\Delta}, \alpha)$, em que $\widehat{F} \approx \chi^2_p(\widehat{\gamma}), \widehat{\gamma} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2 - \Delta)' \left(\mathbf{s}_P^2\right)^{-1} (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2 - \Delta)$, f_c é o valor crítico, para n suficientemente grande e $\mathbf{s}_P^2 = \frac{1}{n_1+n_2-2} \left[(n_1 - 1)\,\mathbf{s}_1^2 + (n_2 - 1)\,\mathbf{s}_2^2\right]$.

- Aplicação: conjunto de dados da iris (flor): Foram medidas quatro variáveis de três grupos de iris. Variáveis: comprimento e lárgura da pétala e comprimento e largura da sépala (partes da flor). Grupos: iris setosa, iris versicolor e iris virginica.
- Objetivo: caracterizar os diferentes tipos de iris de acordo com as características levantadas.
- Disponível no pacote R através do comando *iris*.
- Variáveis: comprimento e largura da sépala.
- Grupos: iris setosa e iris versicolor.
- Objetivo específico: Testar se $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ vs $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ $\Delta = \mathbf{0}$.

- Utilização da função desenvolvida na linguagem R: *teste.mu1mu2.Homocedast*().
- Entrada de dados:
  - m.X.completa: matriz de dados $n \times p$.
  - v.grupos: vetor contendo os grupos $n \times 1$, v.grupos $= [\mathbf{1}_{(n_1 \times 1)} \; ; \mathbf{2}_{(n_2 \times 1)}]'$.
  - alpha: nível de significância.
  - Utilizando comando *teste.mu1mu2.Homocedast*() no ambiente R, obtemos:
    - Estatística do Teste: 498.5481.
    - Nível descritivo: 0.
    - Estimativa do poder do teste (aproximado): 0.9980378.
- Descobrir onde estão as diferenças: testes sobre combinações lineares, intervalos de confiança simultâneos, modelo linear multivariado.

- Estendível 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).

- Supondo uma única população, podemos estar interessados em testar $H_0 : \mathbf{\Sigma} = \mathbf{\Sigma}_0$, vs $H_1 : \mathbf{\Sigma} \neq \mathbf{\Sigma}_0$, em que $\mathbf{\Sigma}_{0(p\times p)}$ é uma matriz conhecida.
- Se $\mathbf{\Sigma}_0 = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p^2 \end{bmatrix}$.
- Se $\mathbf{\Sigma}_0 = \begin{bmatrix} \sigma^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{12} & \sigma^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \dots & \sigma^2 \end{bmatrix}$.
- Se $\mathbf{\Sigma}_0 = \sigma^2 \mathbf{I}_{(p\times p)}$.
- Solução: Teste da razão de verossimilhanças (exercício).

- A suposição de homecedasticidade é requerida por algumas metodologias de análise multivariada: MANOVA, Análise de discriminante.
- Suponha agora G grupos independentes, tais que $\mathbf{X}_{ij} \overset{ind.}{\sim} N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, i =1,...,G.
- Queremos testar se $H_0 : \Sigma_1 = ... = \Sigma_G$ vs H_1 : pelo menos uma diferença.
- A estatística do t.r.v é tal que (exercício):

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

$$\mathbf{S}_P^2 = \frac{1}{\sum_{i=1}^{G}(n_i-1)} \left[\sum_{i=1}^{G} (n_i-1)\mathbf{S}_i^2 \right]$$

$$\mathbf{S}_i = \frac{1}{n_i-1} \sum_{j=1}^{n_i} \left(\overline{\mathbf{X}}_i - \mathbf{X}_{ij}\right)\left(\overline{\mathbf{X}}_i - \mathbf{X}_{ij}\right)'$$

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

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

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, $Q_B \approx \chi^2_{(\nu)}$.

- Dados da iris: mesmas variáveis anteriormente escolhidas e os três grupos.
- Utilização da função *bartlet.teste.Igual.MCov*() implementada no pacote R.
- Entrada de dados:
  - m.X.completa: matriz de dados $n \times p$.
  - v.grupos: vetor contendo os grupos $n \times 1$, v.grupos $= [\mathbf{1}_{(n_1 \times 1)} ; \mathbf{2}_{(n_2 \times 1)}; \mathbf{3}_{(n_3 \times 1)}]'$.
  - G: número de grupos.
  - v.n: vetor com o número de observações em cada grupo $v.n = (50, 50, 50)'$.

- Utilizando comando *bartlet.teste.Igual.MCov*() no ambiente R, obtemos:
  - Estatística do Teste: 35.65459
  - nível descritivo: 3.21716e-06
  - Matrizes de Covariâncias por grupo:

grupo	
1	0.12424898 0.09921633
1	0.09921633 0.14368980
2	0.26643265 0.08518367
2	0.08518367 0.09846939
3	0.40434286 0.09376327
3	0.09376327 0.10400408