Inferência para a distribuição normal multivariada: Parte 1

Prof. Caio Azevedo

8 de setembro de 2009

- Sejam $\mathbf{X} = (X_1, ..., X_p)' \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, em que

$$\boldsymbol{\mu} = (\mu_1, ..., \mu_p)' \text{ e } \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \ldots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \ldots & \sigma_p^2 \end{bmatrix}.$$

- Considere uma amostra aleatória (a.a.) de tamanho n do vetor $\mathbf{X}$, ou seja, $\mathbf{X}_1, ..., \mathbf{X}_n \overset{i.i.d.}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Concantenando os vetores aleatórias acima, temos:

$$\mathbf{X}_{(nxp)} = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix}$$

- Já vimos como testar as hipóteses $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ (*), com $\boldsymbol{\Sigma}$ conhecida, em que $\boldsymbol{\mu}_0 = (\mu_{01}, ..., \mu_{0p})'$.
- Suponha que queiramos testar as hipóteses $\mathbf{R}\boldsymbol{\mu} = \mathbf{b}$ vs $\mathbf{R}\boldsymbol{\mu} \neq \mathbf{b}$, com $\boldsymbol{\Sigma}$ conhecida, em que $\mathbf{b}$ é um vetor conhecido $(c \times 1)$ e $\mathbf{R}$ é uma matriz conhecida de dimensão $c \times p, c \leq p$, de posto coluna completo.
- Vimos que uma estatítica útil para testar a hipótese (*) é $Q = n\left(\overline{\mathbf{X}} - \boldsymbol{\mu}_0\right)' \boldsymbol{\Sigma}^{-1} \left(\overline{\mathbf{X}} - \boldsymbol{\mu}_0\right)$.
- Utilizar o mesmo raciocínio.

- Sabemos que $\overline{\mathbf{X}} \sim N\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right)$, logo, pelo resultado 2 visto em sala de aula, temos que $\mathbf{Y} = \mathbf{R}\overline{\mathbf{X}} \sim N_c\left(\mathbf{R}\boldsymbol{\mu}, \frac{1}{n}\left(\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}'\right)\right)$, ou seja , $\mathbf{Y} \sim N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$.
- Considere a seguinte estatística (forma quadrática)

$$\begin{aligned} Q_{\mathbf{Y}} &= n\left(\mathbf{Y}-\mathbf{b}\right)'\left(\boldsymbol{\Sigma}^*\right)^{-1}\left(\mathbf{Y}-\mathbf{b}\right) \\ &= n\left(\mathbf{R}\overline{\mathbf{X}}-\mathbf{b}\right)'\left(\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}'\right)^{-1}\left(\mathbf{R}\overline{\mathbf{X}}-\mathbf{b}\right) \end{aligned} \quad (1)$$

- Sob H_0, ou seja, se $\mathbf{R}\boldsymbol{\mu} = \mathbf{b}$, então $Q_{\mathbf{Y}} \sim \chi^2_{(c)}$.
- Lembre-se deque $Q = \left(N_p(\mathbf{0}, \boldsymbol{\Sigma})\right)'\left(\boldsymbol{\Sigma}\right)^{-1}\left(N_p(\mathbf{0}, \boldsymbol{\Sigma})\right) \sim \chi^2_{(p)}$

- Suponha que p = 2, ou seja, $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ e $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$.
- Considere que, para um certo conjunto de dados, rejeitou-se a hipótese de que $\boldsymbol{\mu} = \boldsymbol{\mu}_0$.
- Queremos detectar onde as diferenças residem, levando-se em consideração a estrutura multivariada dos dados.
- Uma abordagem: testar $H_0 : \mathbf{R}\boldsymbol{\mu} = \mathbf{b}$, em duas situações:
  - $\mathbf{R} = [1 \ \ 0]$ e $\mathbf{b} = \mu_{01}$, ou seja testar se $H_0 : \mu_1 = \mu_{01}$ vs $H_1 : \mu_1 \neq \mu_{01}$.
  - $\mathbf{R} = [0 \ \ 1]$ e $\mathbf{b} = \mu_{02}$, ou seja testar se $H_0 : \mu_2 = \mu_{02}$ vs $H_1 : \mu_2 \neq \mu_{02}$.
- Neste caso $Y_i = \dfrac{\left(\overline{X}_i - \mu_{0i}\right)^2}{\sigma_i^2/n} \overset{\text{sob} H_0}{\sim} \chi^2_{(1)}$.

- Suponha que p = 2, em que cada variável representa as observações antes e depois de alguma condição de interesse (exposição à alguma tratamento): ministração de um remédio, exposição à algum sistema de ensino etc. Assim, temos que:

$$\mathbf{X}_{(nx2)} = \begin{bmatrix} \text{antes} & \text{depois} \\ \text{-------} & \text{-------} \\ X_{11} & X_{12} \\ X_{21} & X_{22} \\ \vdots & \vdots \\ X_{n1} & X_{n2} \end{bmatrix}$$

- Desejamos testar se o tratamento surtiu efeitos levando-se em consideração a natureza multivariada dos dados e não considerando a diferença entre as variáveis.

- Considerar $\mathbf{R} = [1 \quad -1]$ e $\mathbf{b} = 0$, ou seja, testaremos se $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$.
- Neste caso, a estatística do teste torna-se $Y = \frac{\left(\overline{X}_1 - \overline{X}_2\right)^2}{(\sigma_1^2 + \sigma_2^2 - 2\sigma_{11})/n} \overset{\text{sob} H_0}{\sim} \chi^2_{(1)}$.
- Note que o teste proposto é útil também para testar hipóteses do tipo $H_0 : \mathbf{R}\boldsymbol{\mu} = \mathbf{b}$ vs $H_1 : \mathbf{R}\mu > \mathbf{b}$ ou $H_1 : \mathbf{R}\mu < \mathbf{b}$, desde que façam sentido tais hipótese.
- No exemplo acima, poderíamos etar interessados em testar $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$, popr exemplo.

- Novamente, desejamos testar as hipóteses $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, com $\boldsymbol{\Sigma}$ desconhecida, em que $\boldsymbol{\mu}_0 = (\mu_{01}, ..., \mu_{0p})'$.
- Já vimos que $\overline{\mathbf{X}} \sim N_p\left(\mu, \frac{1}{n}\boldsymbol{\Sigma}\right)$, $(n-1)\mathbf{S}^2 \sim W_p(n-1, \Sigma)$ e que $\overline{\mathbf{X}} \perp \mathbf{S}^2$, em que $\mathbf{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)'$.
- Por outro lado, se $\mathbf{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, $\mathbf{W} \sim W(k, \boldsymbol{\Sigma})$ e $\mathbf{X} \perp \mathbf{W}$, então $T^2 = \mathbf{X}'\left(\frac{\mathbf{W}}{k}\right)^{-1}\mathbf{X} \sim T^2$ de Hotelling.
- Além disso, temos que $F = \frac{n-p}{(n-1)p}T^2 \sim F_{(p,n-p)}$
- Sob estas suposições e sob H_0 temos que $T^2 = n\left(\overline{\mathbf{X}} - \boldsymbol{\mu}_0\right)'\mathbf{S}^{-2}\left(\overline{\mathbf{X}} - \boldsymbol{\mu}_0\right) \sim T^2 de$ de Hotelling e, consequentemente, temos que $F = \frac{n-p}{(n-1)p}T^2 \sim F_{(p,n-p)}$.

- Pois $T^2 = \underbrace{\left[\sqrt{n}\left(\overline{\mathbf{X}} - \boldsymbol{\mu}_0\right)\right]'}_{N_p(\mathbf{0},\boldsymbol{\Sigma})}[\underbrace{\frac{n-1}{n-1}\mathbf{S}^2}_{\frac{W_p(n-1,\boldsymbol{\Sigma})}{n-1}}]^{-1}\underbrace{\left[\sqrt{n}\left(\overline{\mathbf{X}} - \boldsymbol{\mu}_0\right)\right]}_{N_p(\mathbf{0},\boldsymbol{\Sigma})}$
- Resumo sobre a estatística Q:
  - Nível descritivo: $p = P(Q > q_{calc}|\boldsymbol{\mu} = \boldsymbol{\mu}_0)$, sob H_0, $Q \sim \chi^2_p$
  - Função do poder do Teste: $1 - \beta = P(Q > q_c|\boldsymbol{\mu} = \boldsymbol{\mu}_0, \alpha)$, sob $H_1, Q \sim \chi^2_p(\gamma), \gamma = (\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$.
  - Poder do teste estimado: $\widehat{1-\beta} = P(\widehat{Q} > q_c|\boldsymbol{\mu} = \boldsymbol{\mu}_0, \alpha)$, em que $\widehat{Q} \sim \chi^2_p(\widehat{\gamma}), \widehat{\gamma} = (\overline{\mathbf{x}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)$, q_c é o valor crítico.

- Resumo sobre a estatística $F = \frac{n-p}{(n-1)p}T^2$:
  - Nível descritivo: $p = P(F > f_{calc}|\boldsymbol{\mu} = \boldsymbol{\mu}_0)$, sob $H_0, F \sim F_{(p,n-p)}$
  - Função do poder do Teste: $1 - \beta = P(F > f_c|\boldsymbol{\mu} = \boldsymbol{\mu}_0, \alpha)$, sob $H_1, F \approx \chi^2_p(\gamma), \gamma = (\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$ para n suficientemente grande (Teorema de Slutsky).
  - Poder do teste estimado: $\widehat{1 - \beta} = P(\widehat{F} > f_c|\boldsymbol{\mu} = \boldsymbol{\mu}_0, \alpha)$, em que $\widehat{F} \approx \chi^2_p(\widehat{\gamma}), \widehat{\gamma} = (\overline{\mathbf{x}} - \boldsymbol{\mu}_0)'\left(\mathbf{s}^2\right)^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)$, f_c é o valor crítico, para n suficientemente grande.