MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Semestre/2022

Já vimos ☺

MAE5776

Matriz de Dados: $Y_{n\times p} = \left(Y_{ij}\right) \in \Re^{n\times p}; \quad Y_{ip\times 1} \overset{\text{iid}}{\sim} \left(\mu_{p\times 1}; \Sigma_{p\times p}\right), i = 1,\ldots,n$

Estatísticas descritivas multivariadas: $\bar{Y}_{p\times 1}, S_{p\times p}, R_{p\times p}, S^{-1}_{p\times p} \qquad D_{n\times n} = \left(d^2_{ij}\right), d^2_{Pij}, d^2_{Mij}$

Regiões (elipsóides) de Concentração de Observações:

$$R\left(Y_i\right) = \left(Y_i \in \Re^p;\ d^2_M(Y_i;\mu) = \left(Y_i - \bar{Y}\right)' S_u^{-1}\left(Y_i - \bar{Y}\right) \le c^2; c^2 = \chi^2_p(\alpha)\right)$$

Matriz de Dados Aleatórios: $Y_{n\times p} \in \Re^{n\times p}; \quad Y_i \overset{\text{iid}}{\sim} N_p\left(\mu_{p\times 1}; \Sigma_{p\times p}\right), i = 1,\ldots,n$

- Caso de **Uma única População**: $H_0: \mu = \mu_0$ Teste T^2 de Hotelling

$$T^2 = n\left(\bar{Y} - \mu\right)' S_u^{-1}\left(\bar{Y} - \mu\right) \le c^2; \quad c^2 = \frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha)$$

- Caso de **Duas Populações**: $H_0: \mu_1 = \mu_2 \Leftrightarrow H_0: \mu_D = 0$ Teste T^2 de Hotelling

$$T^2 = n\left(\bar{D} - \mu_D\right)' S_D^{-1}\left(\bar{D} - \mu_D\right) \le c^2; \quad c^2 = \frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha)$$

Dados dependentes

$$T^2 = \left(\bar{D} - \mu_D\right)'\left(S_c\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)^{-1}\left(\bar{D} - \mu_D\right) \le c^2; \quad c^2 = \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{(p;n_1+n_2-p-1)}(\alpha)$$

Dados independentes e Homecedásticos, $\Sigma_1 = \Sigma_2$

Teste da Igualdade de Matrizes de Covariância

Comparação de Vetores de Médias - Amostras Independentes e Homocedasticidade

$$Y_{1 n_1 \times p}; Y_{1i} \overset{iid}{\sim} N_p(\mu_1; \Sigma_1) \quad Y_{2 n_2 \times p}; Y_{2i} \overset{iid}{\sim} N_p(\mu_2; \Sigma_2)$$

$$\Rightarrow H_0 : \mu_1 - \mu_2 = 0 \quad ; \Sigma_1 = \Sigma_2 = \Sigma$$

$$\begin{cases} S_{ug} = \dfrac{1}{n_g - 1} \displaystyle\sum_{i=1}^{n_1} \left(Y_{gi} - \bar{Y}_g\right)\left(Y_{gi} - \bar{Y}_g\right)' ; g = 1,2 \\ S_c = \dfrac{(n_1 - 1)S_{u1} + (n_2 - 1)S_{u2}}{n_1 + n_2 - 2}; \end{cases}$$

É uma Hipótese condicional.
Logo a homocedasticidade
deve ser verificada.

- Teste M de Box: $\Rightarrow H_0 : \Sigma_g = \Sigma; \;\; \mu_g \in \Re^p$

É a estatística da Razão de Verossimilhanças (sob N_p)
Para p=1 o teste de Box equivale ao teste de Bartlet.

$$M = (1 - c)\left\{\left[\sum_{g=1}^{G}\left(n_g - 1\right)\right]\ln\left|S_c\right| - \sum_{g=1}^{G}\left[\left(n_g - 1\right)\ln\left|S_{ug}\right|\right]\right\} \sim \chi^2_{\frac{1}{2}p(p+1)(G-1)}$$

$$c = \left[\sum_{g=1}^{G}\frac{1}{\left(n_g - 1\right)}\right]\left[\frac{2p^2 + 3p - 1}{6(p+1)(G-1)}\right]$$

Critério “prático” de heterocedasticidade sugerido em Johnson and Wichern 1992:

$$\sigma_{gjj} = 4\sigma_{g'jj}$$

Inferência sobre Vetores de Médias de Duas Populações

Caso Multivariado - Amostras Independentes – Heterocedasticidade

Teoria Assintótica

$$Y_{1_{n_1\times p}}; Y_{1i} \overset{iid}{\sim} N_p(\mu_1;\Sigma_1); \qquad Y_{2_{n_2\times p}}; Y_{2i} \overset{iid}{\sim} N_p(\mu_2;\Sigma_2)$$

$$\overline{D}_{p\times 1} = \overline{Y}_1 - \overline{Y}_2 \sim N_p\left(\mu_D = \delta = \mu_1 - \mu_2; \Sigma_{\overline{D}} = \left(\frac{\Sigma_1}{n_1} + \frac{\Sigma_1}{n_2}\right)\right)$$

$$\Rightarrow H_0 : \mu_D = \delta_0; \quad \Sigma_g \in \Re^{p\times p}, g = 1,2$$

Hipótese condicional, sob heterocedasticidade (resultados assintóticos).

$$T^2 = \left(\overline{D} - \delta_0\right)'\left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1}\left(\overline{D} - \delta_0\right) \overset{\substack{n_1 - p \to \infty \\ n_2 - p \to \infty}}{\sim} \chi^2_p$$

Pesquise:

- Método de Welch para aproximações dos graus de liberdade da estatística de teste sob heterocedasticidade (Kaiser e Bowden, 1983)
- Algoritmo de Behrens-Fisher: Testar a igualdade dos vetores de médias “SEM” suposições sobre as matrizes de covariâncias

Inferência: Vetores de Médias de Duas Populações
Correções para Múltiplos Testes

$$\underbrace{Y_{1i} \overset{iid}{\sim} N_p\left(\mu_1;\Sigma_1\right)} \qquad \underbrace{Y_{2i} \overset{iid}{\sim} N_p\left(\mu_2;\Sigma_2\right)}$$

Dados Dependentes*1: $n_1 = n_2 \quad \Rightarrow \quad D_i, \quad i = 1,\ldots,n \qquad \bar{D} \quad S_{\bar{D}} = S_D \frac{1}{n}$

Dados Independentes*2: $n_1 \Rightarrow \quad \bar{Y}_1 \quad S_1 \qquad n_2 \Rightarrow \quad \bar{Y}_2 \quad S_2 \qquad \bar{D} \quad S_{\bar{D}} = S_c\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$

Intervalos de Confiança Simultâneos (para combinações lineares das p variáveis)

$$\Rightarrow ICS\ \left(\mu_{Dj}\right) a \left(1-\alpha\right)\times 100\% = \left(\bar{D}_j \mp \sqrt{\frac{\nu_1}{\nu_2}F_{\nu_1,\nu_2}(\alpha)}\ \sqrt{S_{\bar{D}\,jj}}\right)$$

$*1: \nu_1 = (n-1)\,p, \quad \nu_2 = (n-p)$

$*2: \nu_1 = (n_1+n_2-2)\,p, \quad \nu_2 = (n_1+n_2-p-1)$

Intervalos de Confiança de Bonferroni (correção para múltiplos testes)

$$\Rightarrow ICB\left(\mu_{D\,j}\right) a \left(1-\alpha\right)\times 100\% = \left(\bar{Y}_{1j}-\bar{Y}_{2j}\right) \pm t_\nu(\alpha/2p)\sqrt{S_{\bar{D}\,jj}}$$

$*1: \nu = (n-1)$

$*2: \nu\left(= n_1+n_2-2\right)$

Inferência sobre Vetores de Médias Comparações Múltiplas e o Problema de Múliplos Testes

$$Y_{1i} \overset{iid}{\sim} N_p(\mu_1; \Sigma_1) \quad \rightarrow n_1$$

$$Y_{2i} \overset{iid}{\sim} N_p(\mu_2; \Sigma_2) \quad \rightarrow n_2$$

$$\Rightarrow \begin{cases} H_0: \mu_{1\,p\times 1} = \mu_{2\,p\times 1} \\ H_1: \mu_{1\,p\times 1} \neq \mu_{2\,p\times 1} \end{cases}$$

Estatística T^2: Teste simultâneo para as p variáveis a um nível de significância α. Ao rejeitar H_0, há diferença entre os grupos para pelo menos uma variável ou alguma combinação linear delas!

Exemplo: Dados Iris
Comparar Virginica x Setosa
Estatística T^2 é significante
Neste caso p=4 Variáveis.
Para qual variável os grupos são diferentes?

$$\begin{cases} H_{01}: \mu_{1g=1} = \mu_{1g=2} & \Rightarrow \alpha_1 \\ H_{02}: \mu_{2g=1} = \mu_{2g=2} & \Rightarrow \alpha_2 \\ H_{03}: \mu_{3g=1} = \mu_{3g=2} & \Rightarrow \alpha_3 \\ H_{04}: \mu_{4g=1} = \mu_{4g=2} & \Rightarrow \alpha_4 \end{cases}$$

α_c?
Valor-p ajustado?

Para testar as 4 hipóteses, qual é o nível de significância global?

Em Estudos envolvendo Comparações Múltiplas é necessário obter um nível de significância corrigido para a Família de Comparações de interesse.

Inferência sobre Vetores de Médias Comparações Múltiplas e o Problema de Múliplos Testes

$$H_{0k} \quad \times \quad H_{1k} \quad k = 1, \dots, K$$

$$P(\text{pelo menos uma Rej H}_0) = 1 - P(\text{nenhuma Rej H}_0)$$

$$= 1 - \prod_{l=1}^{K} P(p_l > \alpha) = 1 - (1-\alpha)^K \quad \overset{K \to \infty}{\approx} \quad 1$$

```
> B<-10000
> K<-500
> set.seed(1299)
minpval <- replicate(B,
min(runif(K,0,1))<0.01)
table(minpval)
table(minpval)[2]/B
```

```
> table(minpval)
minpval
FALSE  TRUE
   78  9922
> table(minpval)[2]/B
  TRUE
0.9922
```

Quando o número de testes (K), realizados simultaneamente, cresce, é alta a probabilidade de obter resultados significantes ao acaso (Falsos Positivos)

Critérios de Correção do nível de significância individual para garantir um nível de significância coletivo igual a α (em geral, 5%): **Bonferroni, Holm, FDR**

Inferência sobre Vetores de Médias Comparações Múltiplas e o Problema de Múliplos Testes

Rejeitar H_{0j} se:	Método
$p_{(k)} < \alpha / K$	Correção de Bonferroni para múltiplos testes
$p_{(j)} < \alpha /(K - j + 1)$	Correção de Holm (Controle "forte" da taxa de erro para os múltiplos testes)
$p_{(j)} < j\alpha / K$	Correção FDR (Taxa de Falsa Descoberta): Benjamini-Hochberg Controle menos conservador da taxa de erro para os múltiplos testes)

K: número total de testes α: nível de significância global fixado

$p_{(j)}$: nível descritivo (p-valor) ordenado, $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(K)}$

Oficina no R
Comparação de Vetores de Médias de Duas Populações

Exemplo 1: Gerar dados de duas populações N_2

Exemplo 2: Analisar os dados de duas espécies de Moscas (Johnson and Wichern, 1992)