SME0822 Análise Multivariada e Aprendizado Não-Supervisionado

Aula 5a: **Testes de Hipóteses sobre a média**

Prof. Cibele Russo

cibele@icmc.usp.br

http://www.icmc.usp.br/~cibele

Baseado em Johnson, R. A., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. Prentice Hall.

Testes de hipóteses para a média

Seja $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ uma amostra aleatória de uma distribuição normal p-variada com vetor de médias $\underset{\sim}{\mu}$ e matriz de variâncias e covariâncias Σ.
Sejam $\overline{\underset{\sim}{X}}$ e S o vetor de médias amostrais e a matriz de variâncias e covariâncias amostrais.
Queremos avaliar se

$$\begin{array}{l} H_0 : \underset{\sim}{\mu} = \underset{\sim}{\mu}_0 \text{ contra} \\ H_1 : \underset{\sim}{\mu} \neq \underset{\sim}{\mu}_0, \end{array} .$$

Testes de hipóteses para a média

Relembramos o resultado anterior

Resultado

1. $\underset{\sim}{\overline{X}} \sim N_p\left(\underset{\sim}{\mu}, \frac{\Sigma}{n}\right).$
2. $(n-1)S \sim Wishart(n-1).$
3. $\underset{\sim}{\overline{X}}$ *e* S *são independentes.*

Além disso, sob H_0,

$$T^2 = \sqrt{n}(\underset{\sim}{\overline{X}} - \underset{\sim}{\mu}_0)^\top \left(\frac{(n-1)S}{n-1}^{-1}\right) \sqrt{n}(\underset{\sim}{\overline{X}} - \underset{\sim}{\mu}_0) \sim \frac{(n-1)p}{n-p} F_{p,n-p}$$

Testes de hipóteses para a média

A quantidade

$$T^2 = n(\underset{\sim}{\overline{X}} - \underset{\sim}{\mu}_0)^\top S^{-1}(\underset{\sim}{\overline{X}} - \underset{\sim}{\mu}_0) \sim \frac{(n-1)p}{n-p} F_{p,n-p}$$

é conhecida como a **Estatística T^2 de Hotelling**.

Testes de hipóteses para a média

Assim, rejeitamos H_0 a um nível de significância α se

$$T^2_{obs} = n(\overline{\underset{\sim}{X}} - \underset{\sim}{\mu}_0)^\top S^{-1}(\overline{\underset{\sim}{X}} - \underset{\sim}{\mu}_0) > \frac{(n-1)p}{n-p}\ q_{F_{p,n-p,\alpha}}$$

em que $q_{F_{p,n-p,\alpha}}$ é o quantil $\alpha-$superior de uma distribuição $F_{p,n-p}$.

Região de Confiança

Seja $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ uma amostra aleatória de uma distribuição $N_p(\underset{\sim}{\mu}, \Sigma)$. Uma **região com** $100(1-\alpha)\%$ **de confiança** para $\underset{\sim}{\mu}$ é dada pelo elipsóide determinado pelos valores de $\underset{\sim}{\mu}^\star$ que satisfazem

$$n(\overline{\underset{\sim}{x}} - \underset{\sim}{\mu}^\star)^\top S^{-1}(\overline{\underset{\sim}{x}} - \underset{\sim}{\mu}^\star) \leq \frac{(n-1)p}{n-p} F_{p,n-p,\alpha} = c$$

em que $\overline{\underset{\sim}{x}}$ e S são, respectivamente, a média e a matriz de variâncias e covariâncias amostrais observadas.

Assim, para verificar se $\underset{\sim}{\mu}_0$ está dentro da região de confiança, verificamos se

$$n(\overline{\underset{\sim}{x}} - \underset{\sim}{\mu}_0)^\top S^{-1}(\overline{\underset{\sim}{x}} - \underset{\sim}{\mu}_0) \leq \frac{(n-1)p}{n-p} F_{p,n-p,\alpha}$$

Se for verdadeiro, $\underset{\sim}{\mu}_0$ está dentro da região de confiança.

Região de Confiança

Seja $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ uma amostra aleatória de uma distribuição $N_p(\underset{\sim}{\mu}, \Sigma)$.
Uma **região com** $100(1-\alpha)\%$ **de confiança** para $\underset{\sim}{\mu}$ é dada pelo elipsóide determinado pelos valores de $\underset{\sim}{\mu}^\star$ que satisfazem

$$n(\overline{\underset{\sim}{x}} - \underset{\sim}{\mu}^\star)^\top S^{-1}(\overline{\underset{\sim}{x}} - \underset{\sim}{\mu}^\star) \leq \frac{(n-1)p}{n-p} F_{p,n-p,\alpha} = c$$

em que $\overline{\underset{\sim}{x}}$ e S são, respectivamente, a média e a matriz de variâncias e covariâncias amostrais observadas.
Assim, para verificar se $\underset{\sim}{\mu}_0$ está dentro da região de confiança, verificamos se

$$n(\overline{\underset{\sim}{x}} - \underset{\sim}{\mu}_0)^\top S^{-1}(\overline{\underset{\sim}{x}} - \underset{\sim}{\mu}_0) \leq \frac{(n-1)p}{n-p} F_{p,n-p,\alpha}$$

Se for verdadeiro, $\underset{\sim}{\mu}_0$ está dentro da região de confiança.

Comparação de médias multivariadas

Considere o problema de avaliar a igualdade de médias multidimensionais de diferentes populações.

Para isso, vamos considerar três casos básicos:

1. Comparações pareadas ou medidas repetidas
2. Comparação de médias em duas populações independentes
3. Comparação de médias em mais que duas populações independentes

Comparação de médias multivariadas

Considere o problema de avaliar a igualdade de médias multidimensionais de diferentes populações.

Para isso, vamos considerar três casos básicos:

1. Comparações pareadas ou medidas repetidas
2. Comparação de médias em duas populações independentes
3. Comparação de médias em mais que duas populações independentes

Comparação de médias multivariadas

Considere o problema de avaliar a igualdade de médias multidimensionais de diferentes populações.

Para isso, vamos considerar três casos básicos:

1. Comparações pareadas ou medidas repetidas
2. Comparação de médias em duas populações independentes
3. Comparação de médias em mais que duas populações independentes

Comparação de médias multivariadas

Considere o problema de avaliar a igualdade de médias multidimensionais de diferentes populações.

Para isso, vamos considerar três casos básicos:

1. Comparações pareadas ou medidas repetidas
2. Comparação de médias em duas populações independentes
3. Comparação de médias em mais que duas populações independentes

Comparações pareadas

Sejam

- $\underset{\sim}{X}_{11}, \ldots, \underset{\sim}{X}_{1n}$ vetores aleatórios $p \times 1$ referentes a uma população normal multivariada antes de um tratamento com $\mathrm{E}(\underset{\sim}{X}_{1j}) = \underset{\sim}{\mu}_1$ para $j = 1, \ldots, n$,
- $\underset{\sim}{X}_{21}, \ldots, \underset{\sim}{X}_{2n}$ vetores aleatórios $p \times 1$ referentes a uma população normal multivariada após de um tratamento com $\mathrm{E}(\underset{\sim}{X}_{2j}) = \underset{\sim}{\mu}_2$ para $j = 1, \ldots, n$,

sendo que $\underset{\sim}{X}_{11}, \ldots, \underset{\sim}{X}_{1n}$ e $\underset{\sim}{X}_{21}, \ldots, \underset{\sim}{X}_{2n}$ são amostras aleatórias de uma mesma população em diferentes situações, em que $\underset{\sim}{X}_{1j}$ e $\underset{\sim}{X}_{2j}$ são correlacionadas (por exemplo, vetores aleatórios de medições antes e após um tratamento).

Comparações pareadas

Sejam

- $\underset{\sim}{X}_{11}, \ldots, \underset{\sim}{X}_{1n}$ vetores aleatórios $p \times 1$ referentes a uma população normal multivariada antes de um tratamento com $\mathrm{E}(\underset{\sim}{X}_{1j}) = \underset{\sim}{\mu}_1$ para $j = 1, \ldots, n$,
- $\underset{\sim}{X}_{21}, \ldots, \underset{\sim}{X}_{2n}$ vetores aleatórios $p \times 1$ referentes a uma população normal multivariada após de um tratamento com $\mathrm{E}(\underset{\sim}{X}_{2j}) = \underset{\sim}{\mu}_2$ para $j = 1, \ldots, n$,

sendo que $\underset{\sim}{X}_{11}, \ldots, \underset{\sim}{X}_{1n}$ e $\underset{\sim}{X}_{21}, \ldots, \underset{\sim}{X}_{2n}$ são amostras aleatórias de uma mesma população em diferentes situações, em que $\underset{\sim}{X}_{1j}$ e $\underset{\sim}{X}_{2j}$ são correlacionadas (por exemplo, vetores aleatórios de medições antes e após um tratamento).

Comparações pareadas

Sejam

- $\underset{\sim}{X}_{11}, \ldots, \underset{\sim}{X}_{1n}$ vetores aleatórios $p \times 1$ referentes a uma população normal multivariada antes de um tratamento com $\mathrm{E}(\underset{\sim}{X}_{1j}) = \underset{\sim}{\mu}_1$ para $j = 1, \ldots, n$,
- $\underset{\sim}{X}_{21}, \ldots, \underset{\sim}{X}_{2n}$ vetores aleatórios $p \times 1$ referentes a uma população normal multivariada após de um tratamento com $\mathrm{E}(\underset{\sim}{X}_{2j}) = \underset{\sim}{\mu}_2$ para $j = 1, \ldots, n$,

sendo que $\underset{\sim}{X}_{11}, \ldots, \underset{\sim}{X}_{1n}$ e $\underset{\sim}{X}_{21}, \ldots, \underset{\sim}{X}_{2n}$ são amostras aleatórias de uma mesma população em diferentes situações, em que $\underset{\sim}{X}_{1j}$ e $\underset{\sim}{X}_{2j}$ são correlacionadas (por exemplo, vetores aleatórios de medições antes e após um tratamento).

Comparações pareadas

Sejam $\underset{\sim}{\mu}_1$ e $\underset{\sim}{\mu}_2$ os vetores de médias em situações 1 e 2, respectivamente. Deseja-se testar se não há diferença entre as situações 1 e 2 para verificar, por exemplo, que o tratamento não produz nenhum efeito, ou seja, se $\underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2$.
Para avaliar as hipóteses

$$H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 \text{ contra}$$
$$H_1 : \underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2$$

vamos considerar as diferenças

$$\underset{\sim}{D}_j = \underset{\sim}{X}_{1j} - \underset{\sim}{X}_{2j}.$$

Assim, $\underset{\sim}{D}_1, \ldots, \underset{\sim}{D}_n$ são i.i.d e $\underset{\sim}{D}_j \sim N(\underset{\sim}{\mu}_D, \Sigma_D)$.

Comparações pareadas

Então, avaliamos se

$$H_0 : \underset{\sim}{\mu}_D = \underset{\sim}{0} \text{ contra}$$
$$H_1 : \underset{\sim}{\mu}_D \neq \underset{\sim}{0}$$

com a estatística T^2 de Hotelling:

$$T^2 = n(\bar{\underset{\sim}{D}} - \underset{\sim}{0})^\top S_D^{-1}(\bar{\underset{\sim}{D}} - \underset{\sim}{0}) \overset{sob\ H_0}{\sim} \frac{(n-1)p}{n-p} F_{p,n-p},$$

em que $\bar{\underset{\sim}{D}}$ e S_D são o vetor de médias e a matriz de variâncias e covariâncias amostrais de $\underset{\sim}{D}$.

Comparações pareadas

Um teste análogo poderia ser desenvolvido para avaliar

$$H_0 : \underset{\sim}{\mu}_D = \underset{\sim}{\mu}_{D0} \text{ contra}$$
$$H_1 : \underset{\sim}{\mu}_D \neq \underset{\sim}{\mu}_{D0}$$

com a estatística T^2 de Hotelling:

$$T^2 = n(\bar{\underset{\sim}{D}} - \underset{\sim}{\mu}_{D0})^\top S_D^{-1}(\bar{\underset{\sim}{D}} - \underset{\sim}{\mu}_{D0}) \overset{sob\ H_0}{\sim} \frac{(n-1)p}{n-p} F_{p,n-p},$$

em que $\bar{\underset{\sim}{D}}$ e S_D são o vetor de médias e a matriz de variâncias e covariâncias amostrais de $\underset{\sim}{D}$.

Comparações pareadas

A região de confiança, com nível de confiança $100(1-\alpha)\%$ nesse caso seria

$$\{\underset{\sim}{\mu}^{\star}_{D}; n(\bar{\underset{\sim}{d}} - \underset{\sim}{\mu}^{\star}_{D})^{\top} S_D^{-1} (\bar{\underset{\sim}{d}} - \underset{\sim}{\mu}^{\star}_{D}) \leq \frac{(n-1)p}{n-p} q_{F_{p,n-p,\alpha}}\}$$

em que $\bar{\underset{\sim}{D}}$ e S_D são o vetor de médias e a matriz de variâncias e covariâncias amostrais de $\underset{\sim}{D}$.

Comparação de médias de duas populações independentes

Sejam

- $\underset{\sim}{X}_{11}, \ldots, \underset{\sim}{X}_{1n_1}$ vetores aleatórios $p \times 1$ referentes a uma população com $E(\underset{\sim}{X}_{1j}) = \underset{\sim}{\mu}_1$ para $j = 1, \ldots, n_1$,
- $\underset{\sim}{X}_{21}, \ldots, \underset{\sim}{X}_{2n_2}$ vetores aleatórios $p \times 1$ referentes a uma população com $E(\underset{\sim}{X}_{2j}) = \underset{\sim}{\mu}_2$ para $j = 1, \ldots, n_2$,

supondo que a população 1 é independente da população 2.

Comparações pareadas

Temos

	Média amostral	Matriz de covariâncias amostrais
População 1	$\overline{\underset{\sim}{X}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} \underset{\sim}{X}_{1j}$	$S_1 = \frac{1}{n_1}\sum_{j=1}^{n_1}(\underset{\sim}{X}_{1j} - \overline{\underset{\sim}{X}}_1)(\underset{\sim}{X}_{1j} - \overline{\underset{\sim}{X}}_1)^\top$
População 2	$\overline{\underset{\sim}{X}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \underset{\sim}{X}_{2j}$	$S_2 = \frac{1}{n_2}\sum_{j=1}^{n_2}(\underset{\sim}{X}_{2j} - \overline{\underset{\sim}{X}}_2)(\underset{\sim}{X}_{2j} - \overline{\underset{\sim}{X}}_2)^\top$

Comparação de médias de duas populações independentes

Suposições adicionais

1. Ambas as populações têm distribuição normal multivariada
2. $\Sigma_1 = \Sigma_2 = \Sigma$

Deseja-se avaliar as hipóteses

$$H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 \text{ contra}$$
$$H_1 : \underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2$$

Comparação de médias de duas populações independentes

Primeiramente, consideramos um estimador para Σ, por exemplo

$$S = S_{pooled} = \frac{\sum_{j=1}^{n_1}(\underset{\sim}{X}_{1j} - \overline{\underset{\sim}{X}}_1)(\underset{\sim}{X}_{1j} - \overline{\underset{\sim}{X}}_1)^\top + \sum_{j=1}^{n_2}(\underset{\sim}{X}_{2j} - \overline{\underset{\sim}{X}}_2)(\underset{\sim}{X}_{2j} - \overline{\underset{\sim}{X}}_2)^\top}{n_1 + n_2 - 2}$$

ou seja

$$S_{pooled} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

Comparação de médias de duas populações independentes

Então, reescrevemos as hipóteses de interesse na forma mais geral

$$H_0 : \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 = \underset{\sim}{\delta}_0 \text{ contra}$$
$$H_1 : \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 \neq \underset{\sim}{\delta}_0.$$

e rejeitamos H_0 ao nível de significância α se

$$T^2_{obs} = (\overline{\underset{\sim}{x}}_1 - \overline{\underset{\sim}{x}}_2 - \underset{\sim}{\delta}_0)\left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S\right]^{-1} (\overline{\underset{\sim}{x}}_1 - \overline{\underset{\sim}{x}}_2 - \underset{\sim}{\delta}_0) > c^2$$

com $c^2 = \dfrac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} q_{F_{p, n_1+n_2-p-1, \alpha}}$.

Comparação de médias de duas populações independentes

Note que

- $\mathrm{E}(\overline{\underset{\sim}{X}}_1 - \overline{\underset{\sim}{X}}_2) = \mathrm{E}(\overline{\underset{\sim}{X}}_1) - \mathrm{E}(\overline{\underset{\sim}{X}}_2) = \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$
- $\mathrm{Var}(\overline{\underset{\sim}{X}}_1 - \overline{\underset{\sim}{X}}_2) = \mathrm{Var}(\overline{\underset{\sim}{X}}_1) + \mathrm{Var}(\overline{\underset{\sim}{X}}_2)$ pois $\overline{\underset{\sim}{X}}_1$ é independente de $\overline{\underset{\sim}{X}}_2$

Logo, como as populações originais tem distribuição normal multivariada,

$$\overline{\underset{\sim}{X}}_1 - \overline{\underset{\sim}{X}}_2 \sim N(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma)$$

Comparação de médias de duas populações independentes

Temos:

$(n_1 - 1)S_1 \sim Wishart(n_1 - 1, \Sigma)$

$(n_2 - 1)S_2 \sim Wishart(n_2 - 1, \Sigma)$

Como as observações da população 1 são independentes das da população 2, S_1 é independente de S_2. Uma propriedade garante que $(n_1 - 1)S_1 + (n_2 - 1)S_2 \sim Wishart(n_1 + n_2 - 2, \Sigma)$.

Comparação de médias em duas populações independentes

Então

$$T^2 = \left[\overline{\underset{\sim}{X}}_1 - \overline{\underset{\sim}{X}}_2 - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)\right]^\top \frac{S_{pooled}^{-1}}{n_1 + n_2 - 2} \left[\overline{\underset{\sim}{X}}_1 - \overline{\underset{\sim}{X}}_2 - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)\right],$$

o que é novamente o produto de uma v.a. normal multivariada pela inversa de uma v.a. Wishart dividida pelos seus g.l. e uma normal multivariada.

Comparação de médias em duas populações independentes

Se as hipóteses de interesse são

$$H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 \text{ contra } \; H_1 : \underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2$$

então a estatística se simplifica, sob H_0, em

$$T^2 = (\overline{\underset{\sim}{X}}_1 - \overline{\underset{\sim}{X}}_2)^\top \frac{S_{pooled}^{-1}}{n_1 + n_2 - 2} (\overline{\underset{\sim}{X}}_1 - \overline{\underset{\sim}{X}}_2) \sim \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1}.$$