Análise de Variância

M. Zevallos

IMECC-UNICAMP

Motivação

- Anteriormente discutimos o problema de comparar médias de duas populações usando uma amostra aleatória da cada população.
- Nesta aula discutiremos o problema mais geral de comparar as médias de várias (duas ou mais) populações usando amostras independentes.
- Esta comparação será realizada utilizando a técnica estatística de **Análise de Variância** (ANOVA).
- A Análise de Variância foi proposta por R.A. Fisher nos anos 20 de século 19 e teve um impacto enorme na aplicação da Estatística em vários campos do conhecimento.

Exemplo 1

- Fonte: Box, Hunter and Hunter (2005)
- Interessa comparar quatro tipos diferentes de dietas: A, B, C e D, em termos do tempo de coagulação.
- Para isso foi realizado um experimento no qual foram 24 animais foram alocados aleatóriamente nas dietas; 6 animais para cada dieta.
- O problema é o seguinte: *existe evidência de que o tempo médio de coagulação é o mesmo para os quatro tipos de dietas?*

Exemplo 1

Tabela 1: Tempos de coagulação

	A	**B**	**C**	**D**
	62	63	68	56
	60	67	66	62
	63	71	71	60
	59	64	67	61
	63	65	68	63
	59	66	68	64
$\bar{y}_i$	61	66	68	61

Exemplo 1

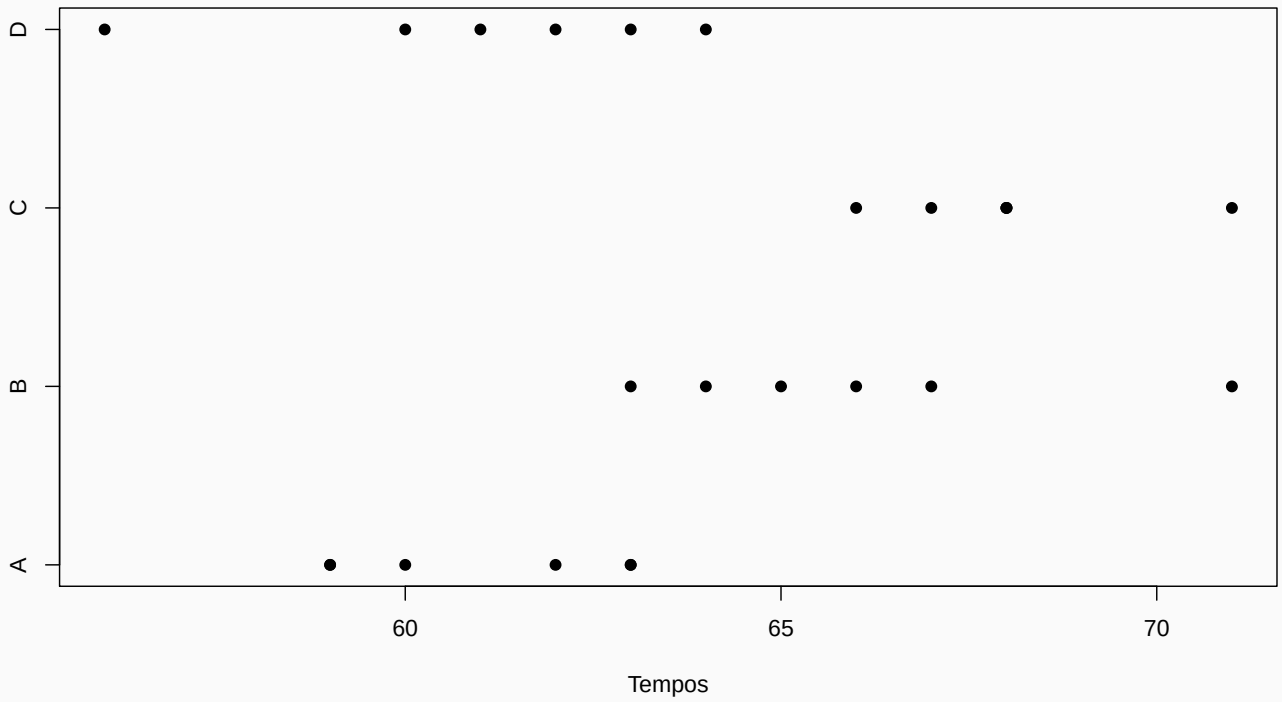

Figura 1: Comparação de quatro dietas em termos de tempos de coagulação

Problema em termos estatísticos

- Temos m populações (grupos ou tratamentos).
- De cada população é extraída uma amostra aleatória de tamanho n.
- Supomos $y_{ij} \sim N(\mu_i, \sigma^2)$ para $i = 1, \ldots, m \quad j = 1, \ldots, n$

Interessa testar

$$H_0 \quad : \quad \mu_1 = \cdots = \mu_m \tag{1}$$

$$H_1 \quad : \quad \mu_i \neq \mu_j \quad \text{para pelo menos um par } i,j \tag{2}$$

Construção do teste

- A idéia é propor um teste baseado em dois estimadores diferentes da variância σ^2.
- Grandes discrepâncias entre esses estimadores constituem evidência contra H_0

Construção do teste

1. Estimador de σ^2 baseado na variabilidade **dentro** de cada grupo
   - No i-ésimo grupo um estimador de σ^2 é

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2 \tag{3}$$

   - Então uma média destes estimadores constitui também um estimador de σ^2

$$\hat{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{m} s_i^2 \tag{4}$$

   - Com os dados do Exemplo 1: $\hat{\sigma}^2 = 5.6$

Exemplo 1

- $m = 4$ grupos, com $n = 6$ observações em cada grupo.
- Para $i \in \{1, 2, 3, 4\}$, $\quad s_i^2 = \frac{1}{5}\sum_{j=1}^{6}(y_{ij} - \bar{y}_i)^2$
- No Grupo A temos as obs:

$$62 \quad 60 \quad 63 \quad 59 \quad 63 \quad 59$$

- Então, $\bar{y}_1 = 61$

$$\begin{aligned} s_1^2 &= \frac{1}{5}\sum_{j=1}^{6}(y_{1j} - \bar{y}_1)^2 \\ &= \frac{1}{5}[(1)^2 + (-1)^2 + (2)^2 + (-2)^2 + (2)^2 + (-2)^2] \\ &= 3.6 \end{aligned}$$

- De forma similar encontramos $s_2^2 = 8, \quad s_2^2 = 2.8 \quad s_4^2 = 8$
- Portanto,

$$\hat{\sigma}^2 = \frac{1}{4}\sum_{i=1}^{4} s_i^2 = \frac{1}{4}[3.6^2 + 8^2 + 2.8^2 + 8^2] = 5.6$$

Construção do teste

2. Estimador de σ^2 baseado na variabilidade **entre** grupos
   - Como $Var(\bar{y}_i) = \sigma^2/n$ então $\sigma^2 = nVar(\bar{y}_i)$
   - Com base em $\bar{y}_1, \ldots, \bar{y}_m$,

$$\widehat{Var}(\bar{y}_i) = \frac{1}{m-1}\sum_{i=1}^{m}(\bar{y}_i - \bar{y})^2 \tag{5}$$

     onde $\bar{y}$ é a media de todas as observações y_{ij}
   - Então um estimador de σ^2 é $n\widehat{Var}(\bar{y}_i)$, ou seja

$$\tilde{\sigma}^2 = \frac{n}{m-1}\sum_{i=1}^{m}(\bar{y}_i - \bar{y})^2 \tag{6}$$

   - Com os dados do Exemplo 1: $\tilde{\sigma}^2 = 76$

Exemplo 1

- $m = 4$ grupos, com $n = 6$ observações em cada grupo.
- $\bar{y}_1 = 61 \quad \bar{y}_2 = 66 \quad \bar{y}_3 = 68 \quad \bar{y}_4 = 61 \quad$ e $\bar{y} = 64$
- Então,

$$\begin{aligned} \tilde{\sigma}^2 &= \frac{6}{3}\sum_{i=1}^{4}(\bar{y}_i - \bar{y})^2 \\ &= \frac{6}{3}[(61-64)^2 + (66-64)^2 + (68-64)^2 + (61-64)^2] \\ &= 76 \end{aligned}$$

Construção do teste

3. Comparação dos estimadores
   - A diferença entre $\hat{\sigma}^2 = 5.6$ e $\tilde{\sigma}^2 = 76$ é resultado do acaso?
   - Pode-se demonstrar que

$$E(\hat{\sigma}^2) = \sigma^2 \tag{7}$$

$$E(\tilde{\sigma}^2) = \sigma^2 + \frac{n}{m-1}\sum_{i=1}^{m}(\mu_i - \mu)^2, \quad \mu = \frac{1}{m}\sum_{i=1}^{m}\mu_i \tag{8}$$

   e sob H_0

$$E(\tilde{\sigma}^2) = \sigma^2 \tag{9}$$

   - Portanto, grandes diferenças entre $\hat{\sigma}^2$ e $\tilde{\sigma}^2$ fornecem evidência contra H_0.

Construção do teste

4. Teste
   - Pode-se demonstrar que sob H_0

$$F = \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \sim F(m-1, m(n-1)) \tag{10}$$

   - Com os dados do Exemplo 1:

$$F_c = \frac{76}{5.6} = 13.57$$

     O valor-P é inferior a 0.00001. Portanto, concluímos que os tempos médios de coagulação são diferentes.

Tabela de Análise de Variância de um fator

- Usualmente os resultados são reportados na forma da *Tabela de Análise de Variância*
- Essa tabela tem como elementos, para cada fonte de variabilidade (**total**, **dentro** e **entre**):
  - Somas de Quadrados (SQ)
  - Graus de liberdade (gl)
  - Quadrados Medios, iguais a SQ/gl

Tabela de Análise de Variância de um fator

- Decomposição da Soma de Quadrados Total (SQ_T),

$$SQ_T = \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2 \tag{11}$$

em termos da Soma de Quadrados de Tratamento (SQ_{trat}) e a Soma de Quadrados de Resíduos (dentro de tratamentos, SQ_E)

$$SQ_{trat} = n\sum_{i=1}^{m}(\bar{y}_i - \bar{y})^2 \tag{12}$$

$$SQ_E = \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2 \tag{13}$$

- Pode-se demonstrar que

$$SQ_T = SQ_{trat} + SQ_E \tag{14}$$

Tabela de Análise de Variância de um fator

- Quadrado Médio de Tratamento (QM_{trat})

$$QM_{trat} = \frac{SQ_{trat}}{m-1} \quad (15)$$

$$= \tilde{\sigma}^2 \quad (16)$$

- Quadrado Médio de Resíduo (QM_E)

$$QM_E = \frac{SQ_E}{m(n-1)} \quad (17)$$

$$= \hat{\sigma}^2 \quad (18)$$

Tabela de Análise de Variância de um fator

Fonte de variabilidade	Soma de quadrados	Graus de liberdade	Quadrado médio	F
Entre tratamentos	SQ_{trat}	$m-1$	QM_{trat}	QM_{trat}/QM_E
Dentre tratamentos	SQ_E	$m(n-1)$	QM_E	
Total	SQ_T	$mn-1$		

Tabela 2: Tabela ANOVA

Exemplo 1: Tabela ANOVA

Com os dados do Exemplo 1

Fonte de variabilidade	Soma de quadrados	Graus de liberdade	Quadrado médio	F
Entre tratamentos	228	3	76	13,57
Dentre tratamentos	112	20	5,6	
Total	340	23		

Tabela 3: Tabela ANOVA

Análise de Variância de um fator no caso desbalanceado

Que acontece quando o número de observações em cada grupo é diferente?

- Conceitualmente a ANOVA é a mesma.
- Seja n_i o número de observações no i-ésimo grupo
- O Quadrado Médio de Tratamento é

$$QM_{trat} = \frac{\sum_{i=1}^{m} n_i(\bar{y}_i - \bar{y})^2}{m-1} \tag{19}$$

- O Quadrado Médio de Resíduo é

$$QM_E = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_{ij} - \bar{y}_i)^2}{\sum_{i=1}^{m} n_i - m} \tag{20}$$

- $F = QM_{trat}/QM_E$ tem distribuição F com $m-1$ e $\sum_{i=1}^{m} n_i - m$ graus de liberdade.

Análise de Variância de um fator

Sejam

- m tratamentos (grupos).
- amostra aleatória de n observações em cada grupo.
- $y_{ij} \sim N(\mu_i, \sigma^2)$ para $i = 1, \ldots, m \quad j = 1, \ldots, n$

Equivalentemente, podemos considerar que as observações foram geradas pelo modelo

$$y_{ij} \quad = \quad \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim NID(0, \sigma^2) \tag{21}$$

ou, supondo que $\mu_i = \mu + \tau_i$ com $\sum_{i=1}^{m} \tau_i = 0$,

$$y_{ij} \quad = \quad \mu + \tau_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim NID(0, \sigma^2) \tag{22}$$

Então interessa testar $H_0 : \tau_1 = \ldots = \tau_m = 0$

Análise de Variância de um fator

Acerca do desempenho do método em relação à falha nas suposições do modelo

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim NID(0, \sigma^2) \tag{23}$$

- ε_{ij} têm distribuição Normal. Quando a distribuição não é muito não-normal, o teste F é aproximadamente válido para amostras moderadas e grandes.
- Variância constante $\sigma^2 = Var(\varepsilon_{ij})$. Se temos igual número de observações em cada grupo, mesmo na presença de grupos com variâncias (não muito) diferentes, o teste F não é muito afetado.
- ε_{ij} independentes. A suposição mais importante. A distribuição da estatística QM_{trat}/QM_E depende fortemente desta suposição.

Exemplo 2

- Fonte: Rice (1995)
- Interessa comparar as medições de um composto quimico fornecidas por 7 laboratórios.
- Em cada laboratório são realizadas 10 medidas.
- Estatísticas

	Lab1	**Lab2**	**Lab3**	**Lab4**	**Lab5**	**Lab6**	**Lab7**
$\bar{y}$	4.062	3.997	4.003	3.920	3.957	3.955	3.998
s	0.033	0.090	0.023	0.033	0.057	0.067	0.085

Exemplo 2

	Lab1	**Lab2**	**Lab3**	**Lab4**	**Lab5**	**Lab6**	**Lab7**
	4.13	3.86	4	3.88	4.02	4.02	4
	4.07	3.85	4.02	3.88	3.95	3.86	4.02
	4.04	4.08	4.01	3.91	4.02	3.96	4.03
	4.07	4.11	4.01	3.95	3.89	3.97	4.04
	4.05	4.08	4.04	3.92	3.91	4	4.1
	4.04	4.01	3.99	3.97	4.01	3.82	3.81
	4.02	4.02	4.03	3.92	3.89	3.98	3.91
	4.06	4.04	3.97	3.9	3.89	3.99	3.96
	4.1	3.97	3.98	3.97	3.99	4.02	4.05
	4.04	3.95	3.98	3.9	4	3.93	4.06
$\bar{y}$	4.062	3.997	4.003	3.920	3.957	3.955	3.998
s	0.033	0.090	0.023	0.033	0.057	0.067	0.085

Exemplo 2

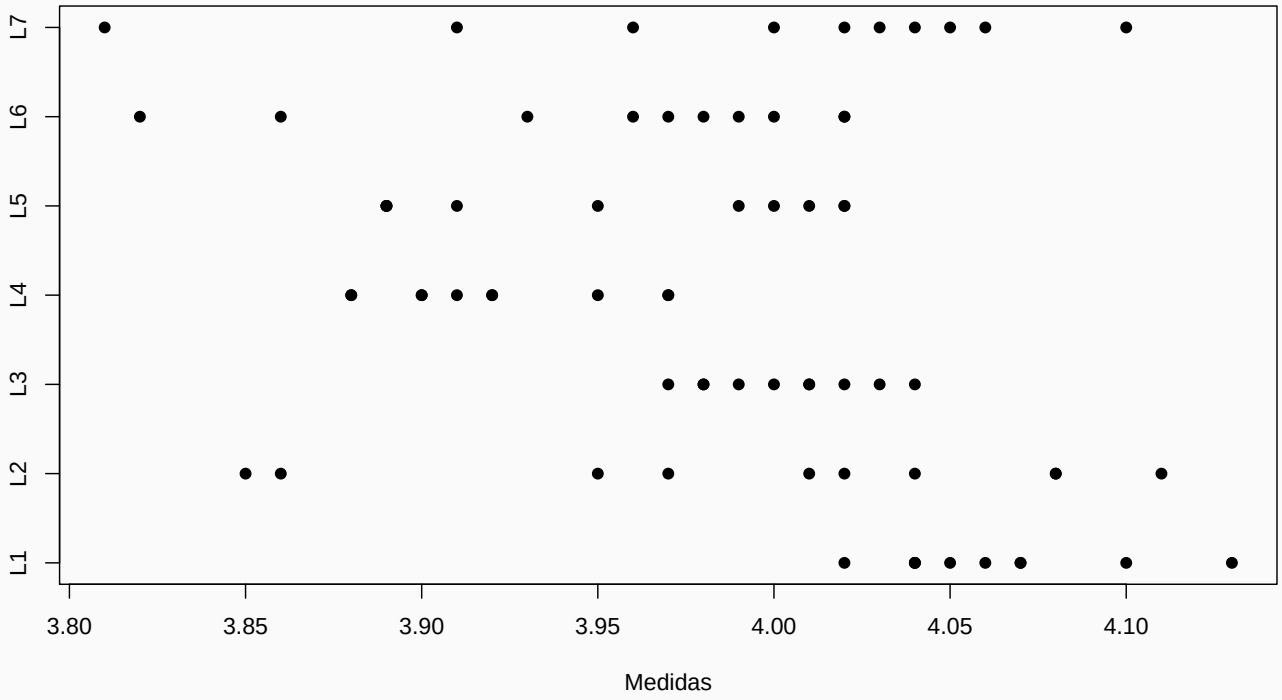

Figura 2: Comparação de medidas em sete laboratórios

Tabela ANOVA do Exemplo 2

Fonte de variabilidade	Soma de quadrados	Graus de liberdade	Quadrado médio	F
Entre tratamentos	0.125	6	0.021	5.66
Dentre tratamentos	0.231	63	0.0037	
Total	0.356	69		

Tabela 4: Tabela ANOVA

Valor-p inferior a 0.001. Portanto, concluímos que as medidas fornecidas pelos laboratórios são diferentes.

Comparações Múltiplas

- Um problema frequente consiste em averiguar quais pares de tratamentos são diferentes. Assim, no caso do Exemplo 2 gostariamos de saber especificamente quais pares (ou grupos) de Labs fornecem resultados diferentes.
- Abordagem *naive*: comparar todos os pares de tratamentos através de testes-t. O problema em fazer isso é que embora cada comparação individual tenha probabilidade de cometer erro tipo I igual a α, o conjunto de todas as comparações consideradas simultaneamente não tem probabilidade de cometer erro tipo I igual a α

Comparações Múltiplas

Vamos supor que temos k testes (por exemplo, comparacoes dois a dois) Sejam

- α_i a probabilidade de cometer erro tipo I no i-ésimo teste para H_{0i}
- α a probabilidade de cometer erro tipo I no conjunto de testes

$$\begin{aligned} \alpha &= P[\text{pelo menos uma } H_{0i} \text{ é rejeitada}] && (24) \\ &= 1 - P[\text{nenhuma } H_{0i} \text{ é rejeitada}] && (25) \end{aligned}$$

Supondo que os testes são independentes, a $P[\text{nenhuma } H_{0i} \text{ é rejeitada}]$ é igual a

$$P[H_{01} \text{ não é rejeitada}]P[H_{02} \text{ não é rejeitada}] \ldots P[H_{0k} \text{ não é rejeitada}] \quad (26)$$

$$\alpha = 1 - \prod_{i=1}^{k}(1 - \alpha_i) \quad (27)$$

Comparações Múltiplas

Por exemplo, com $\alpha_i = 0.05$

- Se $k = 10$ então $\alpha = 0.401$
- Se $k = 100$ então $\alpha = 0.994$

ou seja, com alta probabilidade encontraremos pelo menos uma diferença significativa mesmo que todas as hipóteses nulas sejam verdadeiras.

Comparações Múltiplas: método de Bonferroni

Vamos supor que temos k testes (por exemplo, comparacoes dois a dois)
Sejam

- R_i o evento tal que H_{0i} é rejeitado
- α_i a probabilidade de cometer erro tipo I no i-ésimo teste para H_{0i}
- α a probabilidade de cometer erro tipo I no conjunto de testes

$$\begin{aligned}
\alpha &= P[\text{pelo menos uma } H_{0i} \text{ é rejeitada}] && (28)\\
&= P[R_1 \text{ ou } R_2 \text{ ou } \ldots \text{ ou } R_k] && (29)\\
&\leq P[R_1] + P[R_2] + \ldots + P[R_k] && (30)\\
&= \sum_{i=1}^{k} \alpha_i && (31)
\end{aligned}$$

Portanto, basta escolher

$$\alpha_i = \alpha/k \qquad (32)$$

Comparações Múltiplas: método de Bonferroni

- Se cada hipótese nula é testada ao nível α/k então o nível de significância global é menor ou igual α
- Se k intervalos com confiança $(1 - \alpha/k)$ são construidos, então o conjunto dos k intervalos tem confiança $(1 - \alpha)$

Comparações Múltiplas: método de Bonferroni

Aplicação do método de Bonferroni para a comparação de m grupos com n observações em cada grupo.

Intervalos de confiança simultâneos $(1-\alpha)\%$

$$(\bar{y}_i - \bar{y}_j) \pm \textit{margem} \tag{33}$$

onde

$$\textit{margem} = \textit{ep} \times \textit{vc}, \tag{34}$$

- vc é o quantil $(\alpha/2)/k$ da t-Student com $mn - m$ graus de liberdade com $k = \binom{m}{2}$
- $ep = s_p\sqrt{\frac{2}{n}}$ onde

$$s_p^2 = \frac{(n-1)s_1^2 + \ldots + (n-1)s_m^2}{mn - m} \tag{35}$$

Exemplo 2

Intervalos de confiança simultâneos 95%

Então, $m = 7$, $n = 10$, $\alpha = 0.05$, $k = \binom{7}{2} = 21$

- $vc = -3.1661$ é o quantil $0.025/21$ da t-Student com 63 graus de liberdade.
- $ep = s_p\sqrt{\frac{2}{10}} = 0.0271$ onde

$$s_p^2 = \frac{(10-1)s_1^2 + \ldots + (10-1)s_7^2}{63} = 0.003673 \tag{36}$$

- Portanto, $margem = -0.086$ e

$$(\bar{y}_i - \bar{y}_j) \pm 0.086 \tag{37}$$

Exemplo 2

Labs ordenados por média na ordem decrescente

Lab	Média
1	4.062
3	4.003
7	3.998
2	3.997
5	3.957
6	3.955
4	3.920

$$(\bar{y}_i - \bar{y}_j) \pm 0.086$$

- Lab1 vs Lab5

  $4.062 - 3.957 = 0.105 > 0.086$

- Lab3 vs Lab4

  $4.003 - 3.920 = 0.083 < 0.086$

- Portanto, a média do Lab1 é maior do que as médias dos Lab 4, 5 e 6.

Referências

- Box, G.E.P., Hunter, J.S. and Hunter, W.G.H. (2005). *Statistics for Experimenters: Design, Innovation and Discovery*. Second edition. Wiley.
- Rice, J.A. (1995). *Mathematical Statistics and Data Analysis*. Second edition. Duxbury press.