Unidade 5. Análise de dados de classificação simples e dupla

5.1. Introdução e modelos estatísticos

5.2. Parâmetros do modelo de classificação simples e inferências sobre esses parâmetros

5.3. Parâmetros do modelo de classificação dupla e inferências sobre esses parâmetros

5.4. Discriminação da variação de tratamento: testes de comparações múltiplas: teste DMS de Fisher e teste de Tukey

5.5. Uso de programa estatístico para processamento das análises

Análise de dados de classificação simples e dupla

Esta unidade tratará de situações de pesquisa em que as variáveis preditoras são, necessariamente, do tipo fator, ou seja, são variáveis categorizadas que possibilitam o agrupamento das unidades de observação.

Embora uma pesquisa possa envolver várias variáveis do tipo fator, nesta unidade trataremos apenas de pesquisas que compreendem um ou dois fatores.

Um fator de interesse $\rightarrow$ modelo de classificação simples

Dois fatores de interesse $\rightarrow$ modelo de classificação dupla

Modelos de classificação simples

Experimentos com um único fator

Experimento com um único fator

Exemplo: Uma pesquisa foi realizada para estudar a resistência à compressão do concreto.
A hipótese do pesquisador é que a resistência do concreto varia de acordo com a técnica de mistura utilizada.
Para verificar sua hipótese, produziu corpos de prova utilizando quatro diferentes técnicas de mistura e avaliou a resistência desses corpos.
Foram produzidos 16 corpos de prova, quatro de cada técnica de mistura, e a ordem de avaliação desses corpos foi atribuída por sorteio.

Experimento com um único fator

Exemplo: Uma pesquisa foi realizada para estudar a resistência à compressão do concreto.
A hipótese do pesquisador é que a resistência do concreto varia de acordo com a técnica de mistura utilizada.
Para verificar sua hipótese, produziu corpos de prova utilizando quatro diferentes técnicas de mistura e avaliou a resistência desses corpos.
Foram produzidos 16 corpos de prova, quatro de cada técnica de mistura, e a ordem de avaliação desses corpos foi atribuída por sorteio.

Técnica de mistura	Resistência à compressão (psi)			
A	3129	3000	2865	2890
B	3200	3300	2975	3150
C	2800	2900	2985	3050
D	2600	2700	2600	2765

Experimento com um único fator

Exemplo: Uma pesquisa foi realizada para estudar a resistência à compressão do concreto.
A hipótese do pesquisador é que a resistência do concreto varia de acordo com a técnica de mistura utilizada.
Para verificar sua hipótese, produziu corpos de prova utilizando quatro diferentes técnicas de mistura e avaliou a resistência desses corpos.
Foram produzidos 16 corpos de prova, quatro de cada técnica de mistura, e a ordem de avaliação desses corpos foi atribuída por sorteio.

Fator de tratamento: técnica de mistura (4 níveis: A, B, C e D)

Variável resposta (y): resistência à compressão do concreto

Unidade de pesquisa: corpo de prova

Objetivo: estudar o efeito do fator técnica de mistura sobre o comportamento da variável resposta

Experimento com um único fator

Exemplo: Uma pesquisa foi realizada para estudar a resistência à compressão do concreto.
A hipótese do pesquisador é que a resistência do concreto varia de acordo com a técnica de mistura utilizada.
Para verificar sua hipótese, produziu corpos de prova utilizando quatro diferentes técnicas de mistura e avaliou a resistência desses corpos.
Foram produzidos 16 corpos de prova, quatro de cada técnica de mistura, e a ordem de avaliação desses corpos foi atribuída por sorteio.

Técnica de mistura	Resistência à compressão (psi)				Média
A	3129	3000	2865	2890	2971,00
B	3200	3300	2975	3150	3156,25
C	2800	2900	2985	3050	2933,75
D	2600	2700	2600	2765	2666,25
				Média geral	2931,81

Estrutura dos dados

Os experimentos de classificação simples são aqueles que apresentam apenas um fator de classificação (ou de agrupamento) das unidades de observação.

Técnica de mistura	Repetição				Média observada	Média populacional
	1	2	3	4		
A	y_{A1}	y_{A2}	y_{A3}	y_{A4}	$\bar{y}_A$	μ_A
B	y_{B1}	y_{B2}	y_{B3}	y_{B4}	$\bar{y}_B$	μ_B
C	y_{C1}	y_{C2}	y_{C3}	y_{C4}	$\bar{y}_C$	μ_C
D	y_{D1}	y_{D2}	y_{D3}	y_{D4}	$\bar{y}_D$	μ_D
					$\bar{y}$	μ

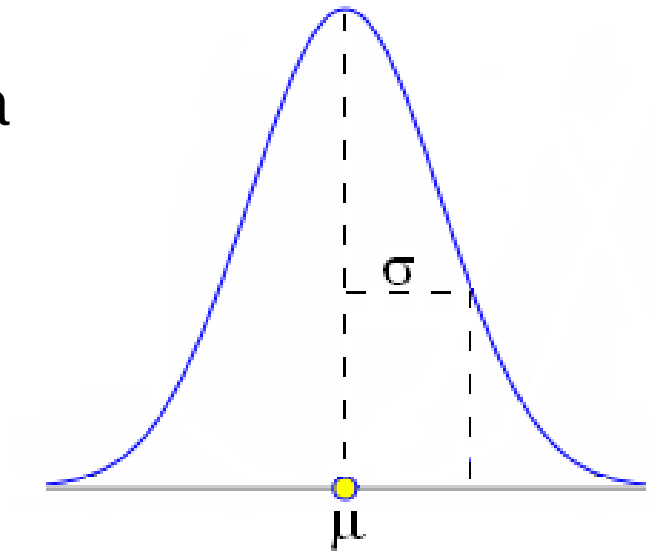
Fator de tratamento, com 4 níveis : Técnica de mistura = {A, B, C, D}

Número de repetições é constante para todos os níveis.

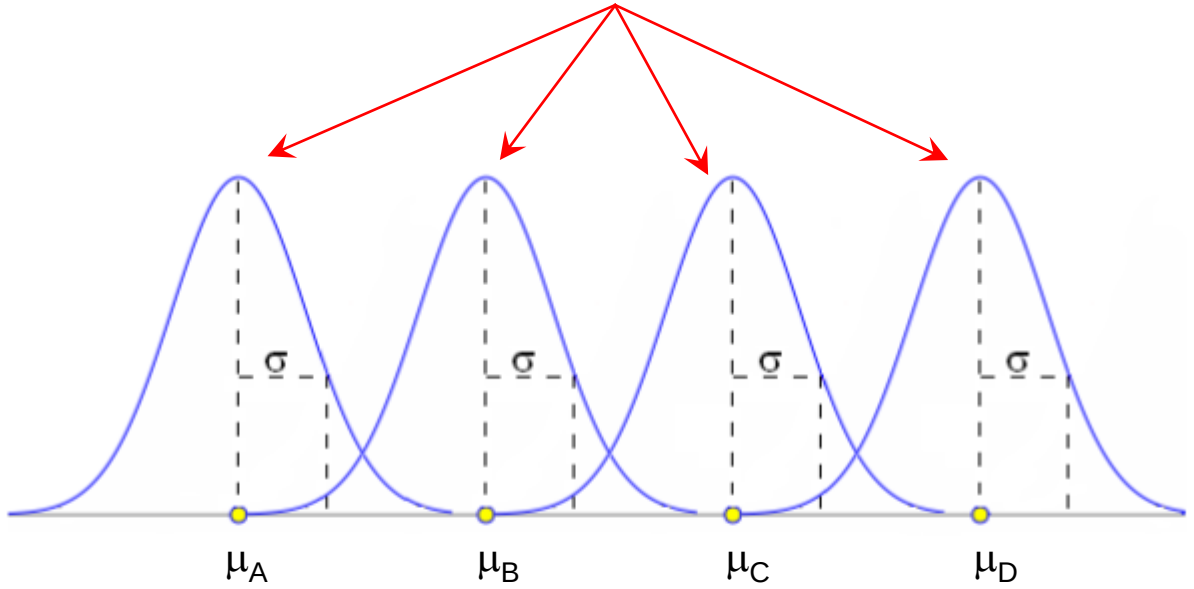
População estatística

y: variável resposta

$y \sim N(\mu, \sigma)$



Tratamentos distintos geram populações distintas?



Estrutura dos dados (generalização)

Os experimentos de classificação simples são aqueles que apresentam apenas um fator de classificação (ou de agrupamento) das unidades de observação.

Fator A (i)	Repetição ou observação (j)				Média observada	Média populacional
	1	2	...	r		
A_1	y_{11}	y_{12}	...	y_{1r}	$\bar{y}_1$	μ_1
A_2	y_{21}	y_{22}	...	y_{2r}	$\bar{y}_2$	μ_2
...	...	...	y_{ij}	...	...	...
A_{n_a}	y_{n_a1}	y_{n_a2}	...	y_{n_ar}	$\bar{y}_{n_a}$	μ_{n_a}
					$\bar{y}$	μ

Fator A, com n_a níveis: $\{A_1, A_2, ..., A_{n_a}\}$

i = nível do Fator A, sendo i=1,2, ..., n_a

j = número da repetição do nível, sendo j=1, 2, ..., r

Modelos estatísticos

Modelo de médias:

$$y_{ij} = \mu_i + e_{ij},$$

onde:

μ_i é a média do nível i do fator (parâmetro)
e_{ij} é o erro aleatório presente na repetição j do nível i do fator

Modelo de efeitos:

$$y_{ij} = \mu + \tau_i + e_{ij},$$

onde:

μ é a média sem efeito ou efeito constante (parâmetro)
τ_i é o efeito do nível i do fator (parâmetro)
e_{ij} é o erro aleatório presente na repetição j do nível i do fator

Pressuposições

O modelo estatístico se completa com as seguintes **pressuposições** referentes aos termos da equação:

1. Os erros (e_{ij}) são aleatórios, têm média zero e variância constante, ou seja, $E(e_{ij}) = 0$ e $V(e_{ij}) = \sigma^2$.

2. Os erros (e_{ij}) têm distribuição normal.

3. Os erros (e_{ij}) são independentes.

Estimação dos parâmetros dos modelos

Parâmetro Estimador

Modelo de médias

$$y_{ij} = \mu_i + e_{ij}$$

$$\mu_i \longrightarrow \hat{\mu}_i = \bar{y}_i = \frac{\sum_j y_{ij}}{r}$$

Fator A (i)	Repetição ou observação (j)				Média observada	Média populacional
	1	2	...	r		
A_1	y_{11}	y_{12}	...	y_{1r}	$\bar{y}_1$	μ_1
A_2	y_{21}	y_{22}	...	y_{2r}	$\bar{y}_2$	μ_2
...	...	...	y_{ij}	...	...	...
A_{n_a}	y_{n_a1}	y_{n_a2}	...	y_{n_ar}	$\bar{y}_{n_a}$	μ_{n_a}
					$\bar{y}$	μ

Parâmetro Estimador

Modelo de médias

$$y_{ij} = \mu_i + e_{ij}$$

$$\mu_i \longrightarrow \hat{\mu}_i = \bar{y}_i = \frac{\sum_j y_{ij}}{r}$$

Modelo de efeitos

$$y_{ij} = \mu + \tau_i + e_{ij}$$

$$\mu \longrightarrow \hat{\mu} = \bar{y} = \frac{\sum_{ij} y_{ij}}{n}$$

Fator A (i)	Repetição ou observação (j)				Média observada	Média populacional
	1	2	...	r		
A_1	y_{11}	y_{12}	...	y_{1r}	$\bar{y}_1$	μ_1
A_2	y_{21}	y_{22}	...	y_{2r}	$\bar{y}_2$	μ_2
...	...	...	y_{ij}	...	...	...
A_{n_a}	y_{n_a1}	y_{n_a2}	...	y_{n_ar}	$\bar{y}_{n_a}$	μ_{n_a}
					$\bar{y}$	μ

Parâmetro

Estimador

Modelo de médias

$$y_{ij} = \mu_i + e_{ij}$$

$$\mu_i \longrightarrow \hat{\mu}_i = \bar{y}_i = \frac{\sum_j y_{ij}}{r}$$

Modelo de efeitos

$$y_{ij} = \mu + \tau_i + e_{ij}$$

$$\mu \longrightarrow \hat{\mu} = \bar{y} = \frac{\sum_{ij} y_{ij}}{n}$$

$$\tau_i = \mu_i - \mu \longrightarrow \hat{\tau}_i = \bar{y}_i - \bar{y}$$

Erro aleatório

$$y_{ij} = \mu_i + e_{ij}$$

$$e_{ij} = y_{ij} - \mu_i$$

Estimativa do erro (resíduo)

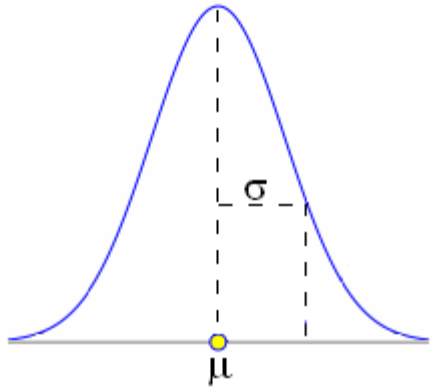
$$\hat{e}_{ij} = y_{ij} - \hat{\mu}_i$$

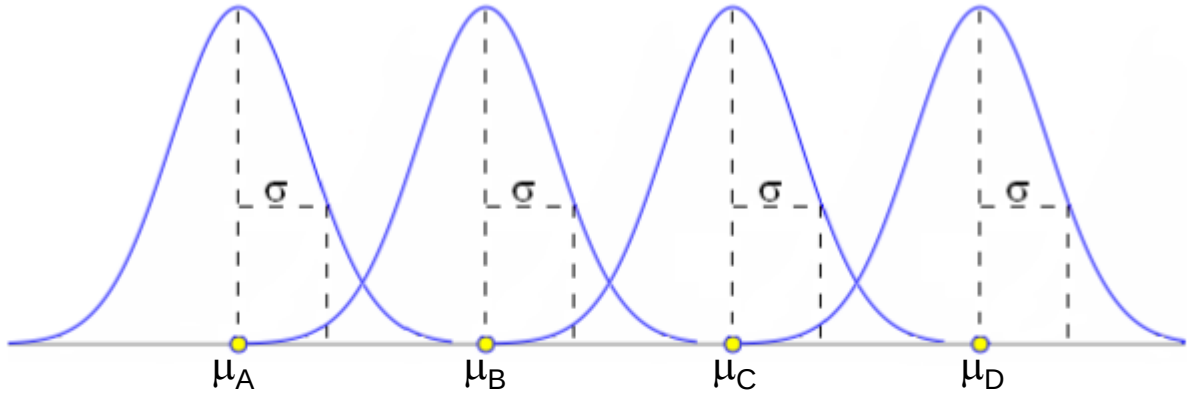
$$\hat{e}_{ij} = y_{ij} - \bar{y}_i$$

Populações estatísticas

y: variável resposta

$y \sim N(\mu, \sigma)$





Testar hipóteses sobre os parâmetros dos modelos

Hipóteses de interesse

$$y_{ij} = \mu_i + e_{ij}$$

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_{n_a} = \mu \\ H_1 : \mu_i \neq \mu \end{cases}$$

$$y_{ij} = \mu + \tau_i + e_{ij}$$

$$\begin{cases} H_0 : \tau_1 = \tau_2 = \ldots = \tau_{n_a} = 0 \\ H_1 : \tau_i \neq 0 \end{cases}$$

A hipótese pode ser testada pela análise da variância

Análise da variância

A análise da variância decompõe a variação total das observações, representada pelos desvios $(y_{ij} - \bar{y})$, em duas partes:

- → a variação provocada pelo efeito do fator de tratamento, representada pelos desvios $(\bar{y}_i - \bar{y})$;
- → a variação aleatória, representada pelos desvios $(y_{ij} - \bar{y}_i)$.

Técnica de mistura	Resistência à compressão (psi)				Média
A	3129	3000	2865	2890	2971,00
B	3200	3300	2975	3150	3156,25
C	2800	2900	2985	3050	2933,75
D	2600	2700	2600	2765	2666,25
				Média geral	2931,81

Análise da variância

A análise da variância decompõe a variação total das observações, representada pelos desvios$(y_{ij} - \bar{y})$, em duas partes:

→ a variação provocada pelo efeito do fator de tratamento, representada pelos desvios $(\bar{y}_i - \bar{y})$;

→ a variação aleatória, representada pelos desvios $(y_{ij} - \bar{y}_i)$.

Assim, a variação de cada observação pode ser representada pela seguinte expressão:

$$(y_{ij} - \bar{y}) = (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

E a variação total das observações pode ser representada por:

$$\sum_{ij}(y_{ij} - \bar{y})^2 = \sum_{ij}\left[(\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)\right]^2$$

$$\sum_{ij}(y_{ij} - \bar{y})^2 = \sum_{i} r(\bar{y}_i - \bar{y})^2 + \sum_{ij}(y_{ij} - \bar{y}_i)^2$$

$$SQ_{Total} = SQ_A + SQ_{Resíduo}$$

Tabela da análise da variância

Fonte de variação	GL (ν)	SQ	QM (S^2)	F
Fator A (Trat.)	$\nu_A = n_a - 1$	$\sum_i r(\bar{y}_i - \bar{y})^2$	$S^2_A = \frac{SQ_A}{\nu_A}$	$\frac{S^2_A}{S^2}$
Resíduo	$\nu = n - n_a$	$\sum_{ij} (y_{ij} - \bar{y}_i)^2$	$S^2 = \frac{SQ}{\nu}$	-
Total	$\nu_{Total} = n - 1$	$\sum_{ij} (y_{ij} - \bar{y})^2$	-	-

$S^2 = \frac{SQ_{Res}}{\nu} \rightarrow$ Variância do resíduo $\rightarrow$ estima a variação aleatória (σ^2)

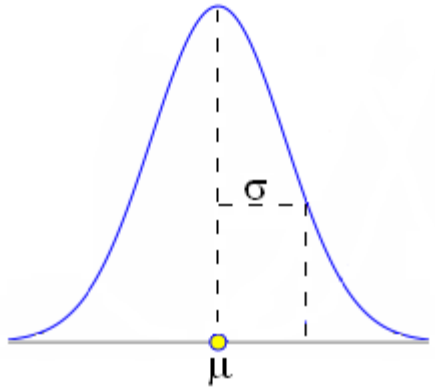
$S^2_A = \frac{SQ_A}{\nu_A} \rightarrow$ Variância do fator A $\rightarrow$ estima a variação do tratamento (σ^2_A) que é composta pela variação aleatória mais o efeito do fator de tratamento ($\sigma^2 + \tau_A$).

Ao estimarmos o efeito do tratamento, não podemos isolá-lo da variação aleatória, pois ele está totalmente confundido com esses efeitos aleatórios.

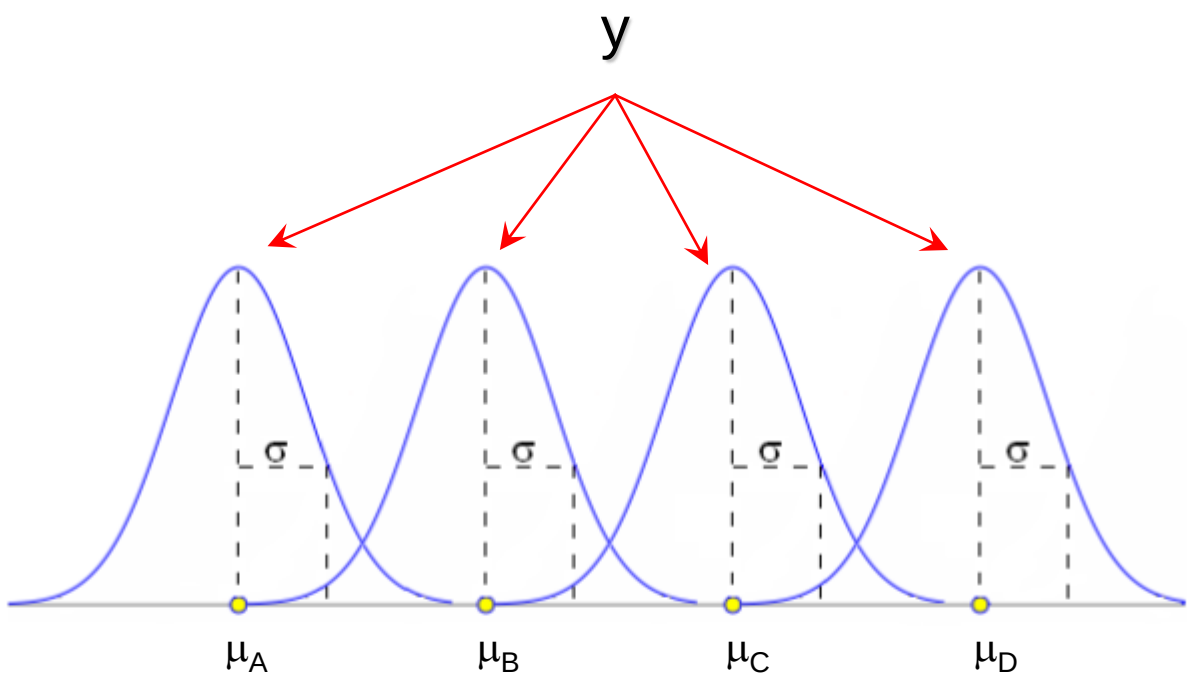
Populações estatísticas

y: variável resposta

$y \sim N(\mu, \sigma)$



Tratamentos distintos geram populações distintas?



Princípio da inferência experimental

Comparar a variação aleatória com o efeito do fator de tratamento confundido com a variação aleatória

Hipótese estatística

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_{n_a} = \mu \\ H_1 : \mu_i \neq \mu \end{cases} \Rightarrow \begin{cases} H_0 : \sigma_A^2 = \sigma^2 \\ H_1 : \sigma_A^2 > \sigma^2 \end{cases}$$

É possível verificar se o efeito do tratamento é significativo, obtendo-se o valor da estatística F, razão entre a variância do fator de tratamento e a variância do resíduo.

Estatística do teste: $F = \frac{S_A^2}{S^2} \sim F(\nu_A, \nu)$

Sob H_0 verdadeira, as duas variâncias estimam o mesmo parâmetro, devendo apresentar valores não muito diferentes.

Devemos esperar, portanto, que a razão entre as variâncias seja um valor não muito maior que 1.

Tabela da análise da variância

Fonte de variação	GL (ν)	SQ	QM (S^2)	F
Fator A (Trat.)	$\nu_A = n_a - 1$	$\sum_i r(\bar{y}_i - \bar{y})^2$	$S^2_A = \frac{SQ_A}{\nu_A}$	$\frac{S^2_A}{S^2}$
Resíduo	$\nu = n - n_a$	$\sum_{ij}(y_{ij} - \bar{y}_i)^2$	$S^2 = \frac{SQ}{\nu}$	-
Total	$\nu_{Total} = n - 1$	$\sum_{ij}(y_{ij} - \bar{y})^2$	-	-

Hipótese estatística:

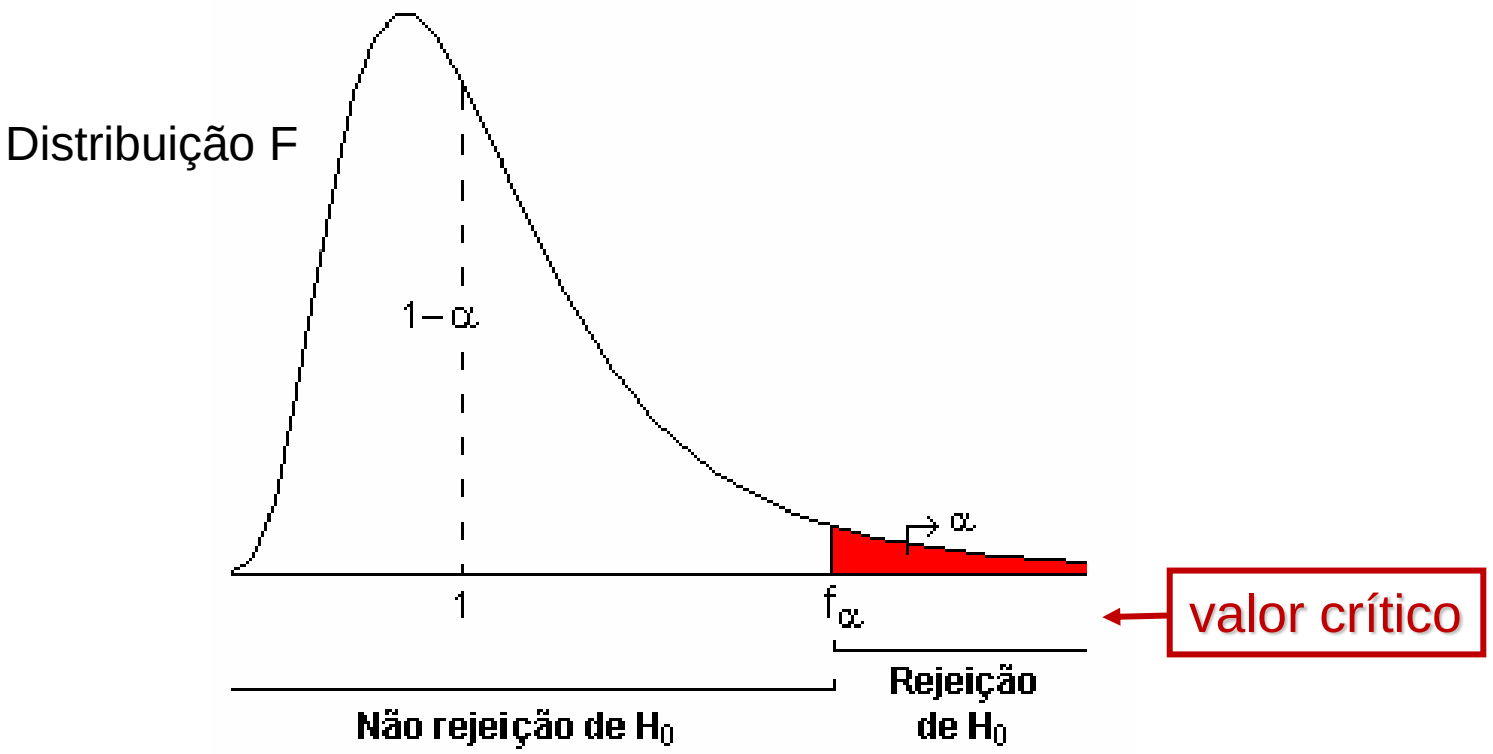
$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_{n_a} = \mu \\ H_1 : \mu_i \neq \mu \end{cases}$$

Estatística do teste:

$$F = \frac{s^2_A}{s^2} \sim F(\nu_A, \nu)$$

Como tomar a decisão a respeito de H_0?

Se H_0 é verdadeira, devemos esperar que o valor da estatística F seja próximo de 1.



Critério de decisão

- ⇨ Se $f > f_\alpha$, rejeitamos H_0 ⇒ f é atípico
- ⇨ Se $f < f_\alpha$, não temos motivos para rejeitar H_0 ⇒ f é típico

Tabela da análise da variância

Fonte de variação	GL (ν)	SQ	QM (S^2)	F
Fator A (Trat.)				
Resíduo				
Total				

Hipótese estatística:

$$\begin{cases} H_0 : \mu_A = \mu_B = \mu_C = \mu_D \\ H_1 : \mu_i \neq \mu_{i'} \end{cases}$$

$$SQ_{Total} = \sum_{ij} (y_{ij} - \bar{y})^2$$

$$SQ_A = \sum_{i} r(\bar{y}_i - \bar{y})^2$$

Experimento com um único fator

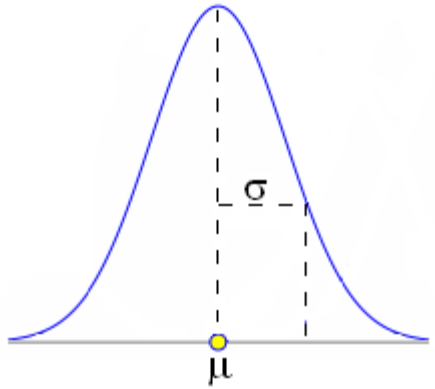
Exemplo: Uma pesquisa foi realizada para estudar a resistência à compressão do concreto.
A hipótese do pesquisador é que a resistência do concreto varia de acordo com a técnica de mistura utilizada.
Para verificar sua hipótese, produziu corpos de prova utilizando quatro diferentes técnicas de mistura e avaliou a resistência desses corpos.
Foram produzidos 16 corpos de prova, quatro de cada técnica de mistura, e a ordem de avaliação desses corpos foi atribuída por sorteio.

Técnica de mistura	Resistência à compressão (psi)				Média
A	3129	3000	2865	2890	2971,00
B	3200	3300	2975	3150	3156,25
C	2800	2900	2985	3050	2933,75
D	2600	2700	2600	2765	2666,25
				Média geral	2931,81

Populações estatísticas

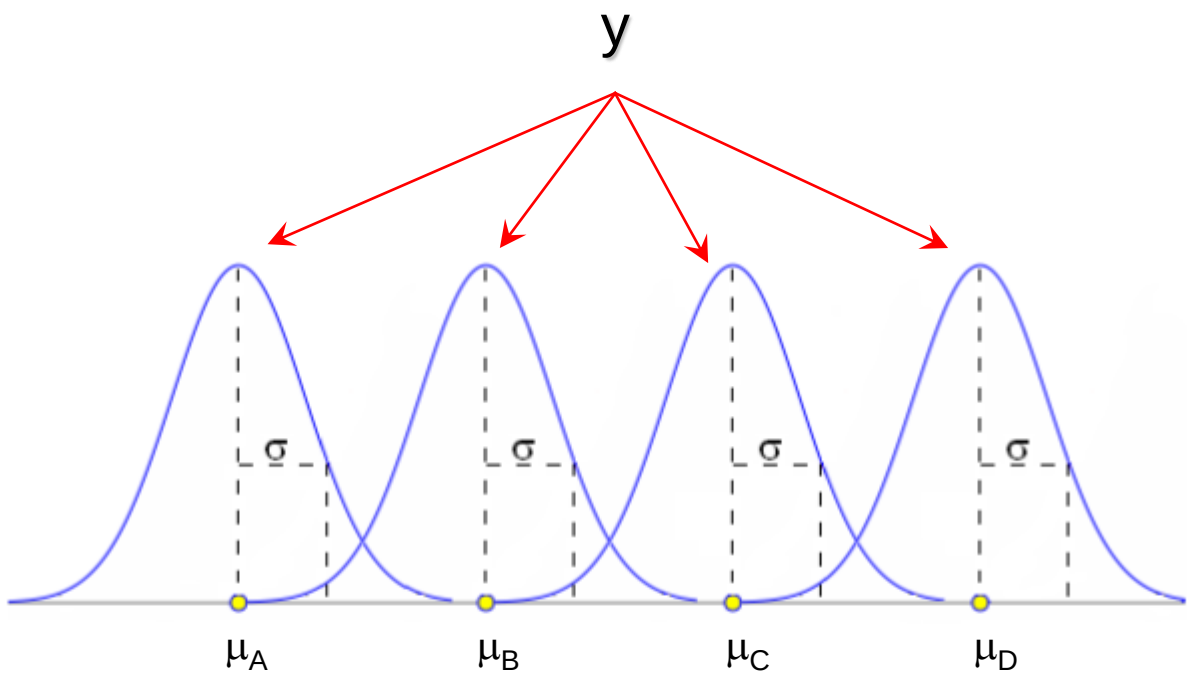
y: variável resposta

$y \sim N(\mu, \sigma)$



$$\begin{cases} H_0 : \mu_A = \mu_B = \mu_C = \mu_D = \mu \\ H_1 : \mu_i \neq \mu \end{cases}$$

Tratamentos distintos geram populações distintas?



No exemplo: Análise da variância

Fonte de variação	GL (ν)	SQ	QM (S^2)	F
Fator A (Trat.)	3	489740,19	163246,73	12,73
Resíduo	12	153908,25	12825,69	-
Total	15	643648,44	-	-

Hipóteses estatísticas:

$$\begin{cases} H_0 : \mu_A = \mu_B = \mu_C = \mu_D = \mu \\ H_1 : \mu_i \neq \mu \end{cases}$$

Decisão e conclusão

$$f_{\alpha(3;12)} = 3{,}49$$

$$f = 12{,}73 > f_{\alpha(3,12)} = 3{,}49 \quad \leftarrow \text{Rejeitamos } H_0$$

Concluímos, ao nível de 5% de significância, que existe um efeito significativo do método de mistura sobre a resistência do concreto à compressão, ou seja pelo menos duas médias diferem entre si.

Coeficiente de variação (CV)

- O coeficiente de variação, denotado por CV, é o desvio padrão do resíduo expresso como uma percentagem da média geral

- É dado pela razão entre o desvio padrão do resíduo e a média geral multiplicada por 100:

$$CV = \frac{\sqrt{s^2}}{\bar{y}} 100$$

O coeficiente de variação também é utilizado para medir a qualidade de um experimento. Pode-se dizer que quanto menor o CV melhor é a qualidade do experimento.

No exemplo: Análise da variância

Fonte de variação	**GL**	**SQ**	**QM**	**F**	**fα**
Técnica	3	489740,19	163246,7	12,73	> 3,49
Resíduo	12	153908,25	12825,69	-	-
Total	15	643648,44	-	-	-

Técnica de mistura	Resistência à compressão (psi)				Média
A	3129	3000	2865	2890	2971,00
B	3200	3300	2975	3150	3156,25
C	2800	2900	2985	3050	2933,75
D	2600	2700	2600	2765	2666,25
				Média geral	2931,81

$$CV = \frac{\sqrt{s^2}}{\bar{y}} 100$$

No exemplo: Análise da variância

Fonte de variação	GL	SQ	QM	F	fα
Técnica	3	489740,19	163246,7	12,73	> 3,49
Resíduo	12	153908,25	12825,69	-	-
Total	15	643648,44	-	-	-

$$\begin{cases} H_0 : \mu_A = \mu_B = \mu_C = \mu_D = \mu \\ H_1 : \mu_i \neq \mu \end{cases}$$

← Rejeitamos H_0

Tabela de médias

Técnica de mistura	Resistência média
A	2971,00
B	3156,25
C	2933,75
D	2666,25

Quais médias diferem entre si?

Bibliografia consultada

SILVA, J.G.C. da **Estatística experimental: análise estatística de experimentos.** Pelotas, RS: Instituto de Física e Matemática, Universidade Federal de Pelotas, 2000. 318p.

Sistema Galileu de Educação Estatística. Disponível em: http://www.galileu.esalq.usp.br

VIEIRA, S. **Estatística Experimental**. 2 ed. São Paulo: Atlas, 1999. 185p.