Análise de Variância – Aula 09

Statistics for Business and Economics 7 edição, by Paul Newbold , William Carlson , Betty Thorne (cap. ANOVA Cap. 15)

Bussab e Morettin (Análise de Regressão)

Statistics for Economics, Accounting and Business Studies, capítulo 6, Barrow (The x2 and F distributions)

Marislei Nishijima

Análise de variância unidirecional

- Avaliar a diferença entre as médias de três ou mais grupos

Exemplos: Média de produção para a 1st, 2nd, e 3rd mudanças
Milhagem esperada para cinco marcas de pneus

- Pressupostos
  - As populações são normalmente distribuídas
  - Populações têm variâncias iguais
  - As amostras são sorteadas de forma aleatória e independente

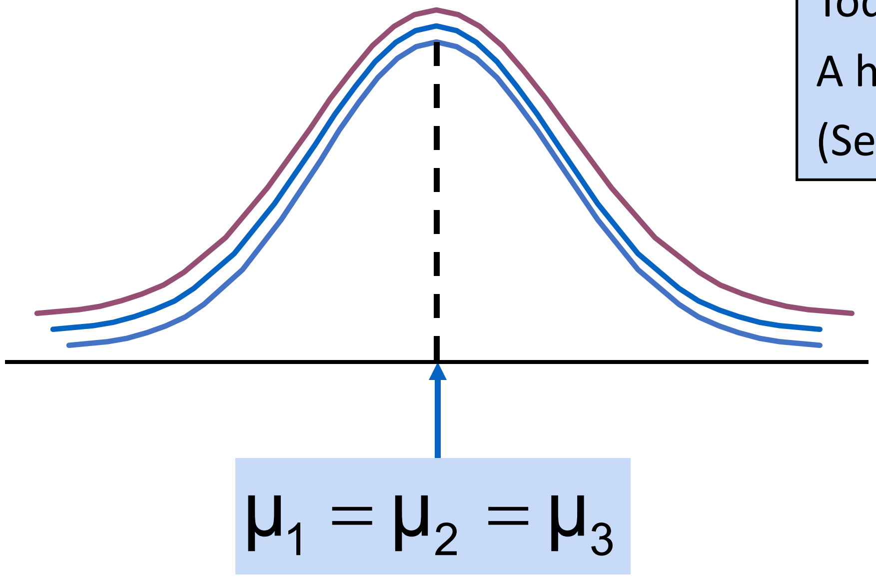
Hipóteses da ANOVA unidirecional

- $H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_K$
  - Todas as médias da população são iguais
  - ou seja, nenhuma variação nas médias entre os grupos
- $H_1 : \mu_i \neq \mu_j \quad \text{para pelo menos um par } i, j$
  - Pelo menos uma média populacional é diferente
  - ou seja, há variação entre os grupos
  - Não significa que todas as médias populacionais sejam diferentes (alguns pares podem ser iguais)

ANOVA unidirecional

$$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_K$$

$$H_1: Nem\ todas\ \mu_i\ são\ iguais$$



Todas as médias são iguais:

A hipótese nula é verdadeira

(Sem variação entre os grupos)

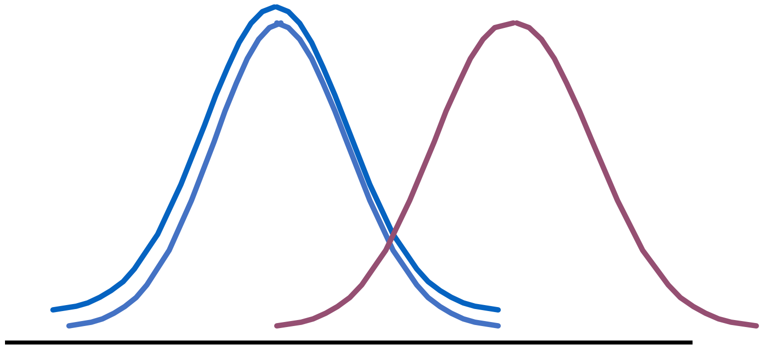
ANOVA unidirecional

(cont.)

$$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_K$$

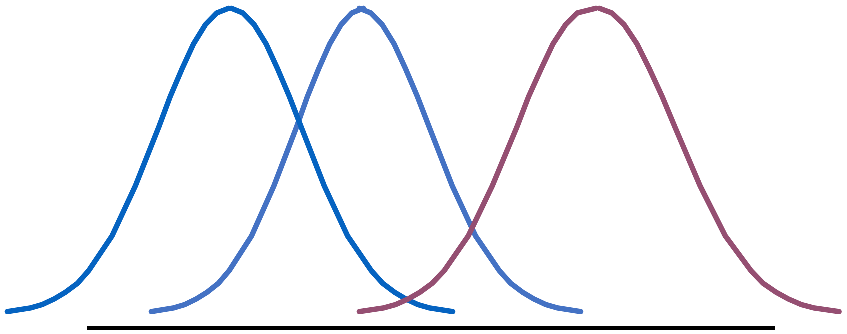
$$H_1: Nem\ todas\ \mu_i\ são\ iguais$$

Pelo menos uma média é diferente:
A hipótese nula NÃO é verdadeira
(Variação está presente entre os grupos)



$\mu_1 = \mu_2 \neq \mu_3$

or

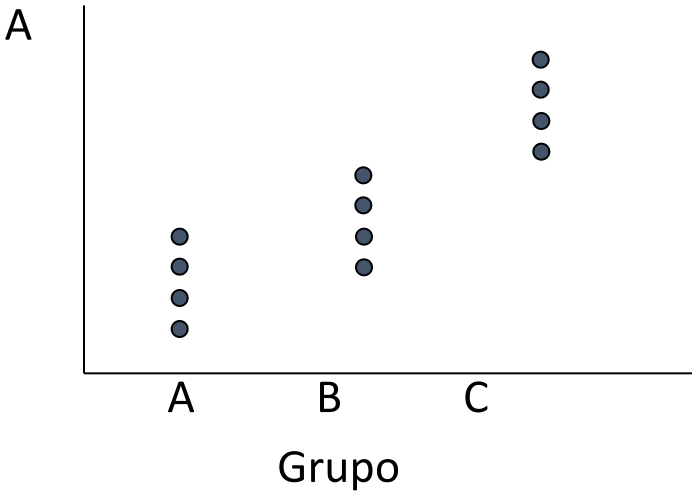


$\mu_1 \neq \mu_2 \neq \mu_3$

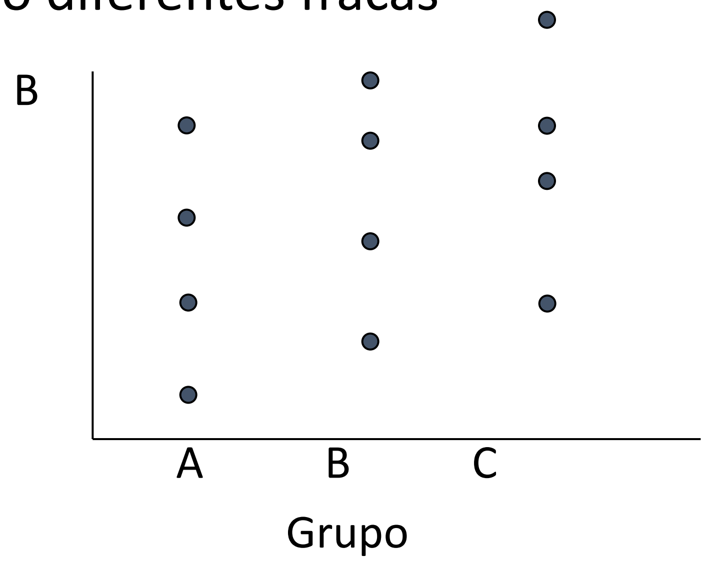
Variabilidade

- A variabilidade dos dados é um fator chave para testar a igualdade das médias

- Em cada caso abaixo, as médias podem parecer diferentes, mas uma grande variação dentro (within) dos grupos em B torna as evidências de que as médias são diferentes fracas



Pequena variação dentro dos grupos



Grande variação dentro dos grupos

Particionando a Variação Total

- A variação total pode ser dividida em duas partes:

$$\boxed{SQT = SQW + SQB}$$

SQT = Soma Total dos Quadrados

Variação total = a dispersão agregada dos valores de dados individuais entre os vários grupos

SQW = Soma dos quadrados dentro dos grupos

Variação dentro do grupo = dispersão que existe entre os valores de dados dentro de um determinado grupo

SQB = Soma dos quadrados entre grupos

Variação entre grupos = dispersão entre as médias da amostra do grupo

Particionando a Variação Total

Soma dos Quadrados Total (SQT)

= **Variação devido à amostragem aleatória (SQW)** + **Variação devido às diferenças entre grupos (SQB)**

Soma dos Quadrados Total

$$\boxed{SQT} = SQW + SQB$$

$$SQT = \sum_{i=1}^{K} \sum_{j=1}^{n_1} (x_{ij} - \bar{x})^2$$

Sendo:

SQT = Total dos Quadrados Total

K = número de grupos (níveis ou tratamento)

n_i = número de observações no grupo i

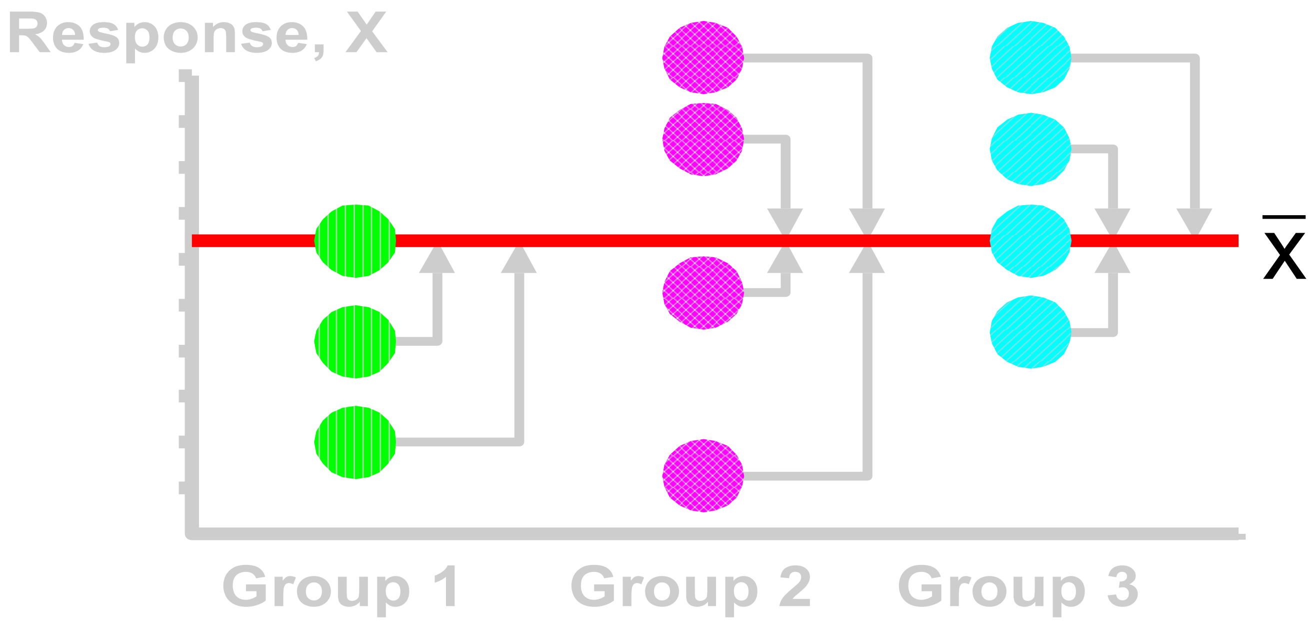
x_{ij} = j^{th} observação do grupo i

$\bar{x}$ = Média geral da amostra

Variação Total

(cont.)

$$SQT = (x_{11} - \bar{x})^2 + (x_{12} - \bar{x})^2 + \ldots + (x_{Kn_K} - \bar{x})^2$$



Variação dentro do grupo

$$SQT = \boxed{SQW} + SQB$$

$$SQW = \sum_{i=1}^{K} \sum_{j=1}^{n_1} (x_{ij} - \overline{x_i})^2$$

Sendo:

SQW = Suma dos quadrados dentro dos grupos

K = número de grupos

n_i = tamanho da amostra do grupo i

$\overline{x}_i$ = media amostral do grupo i

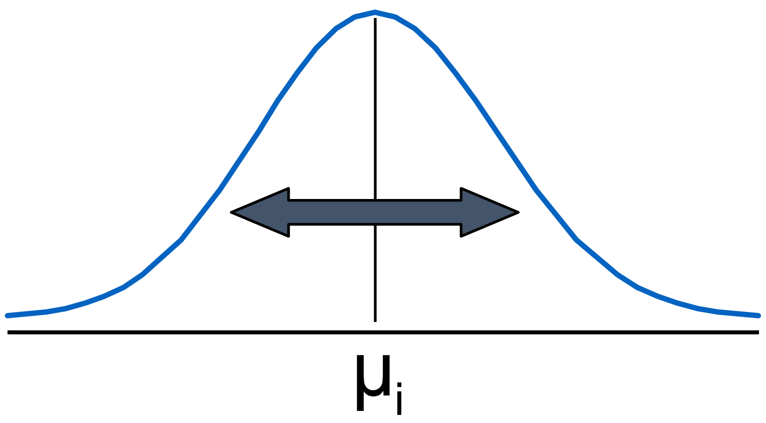
x_{ij} = j^{th} observação no grupo i

Variação dentro do grupo (within)

(cont.)

$$SQW = \sum_{i=1}^{K} \sum_{j=1}^{n_1} (x_{ij} - \overline{x_i})^2$$

Somando a variação dentro de cada grupo e adicionando sobre todos os grupos



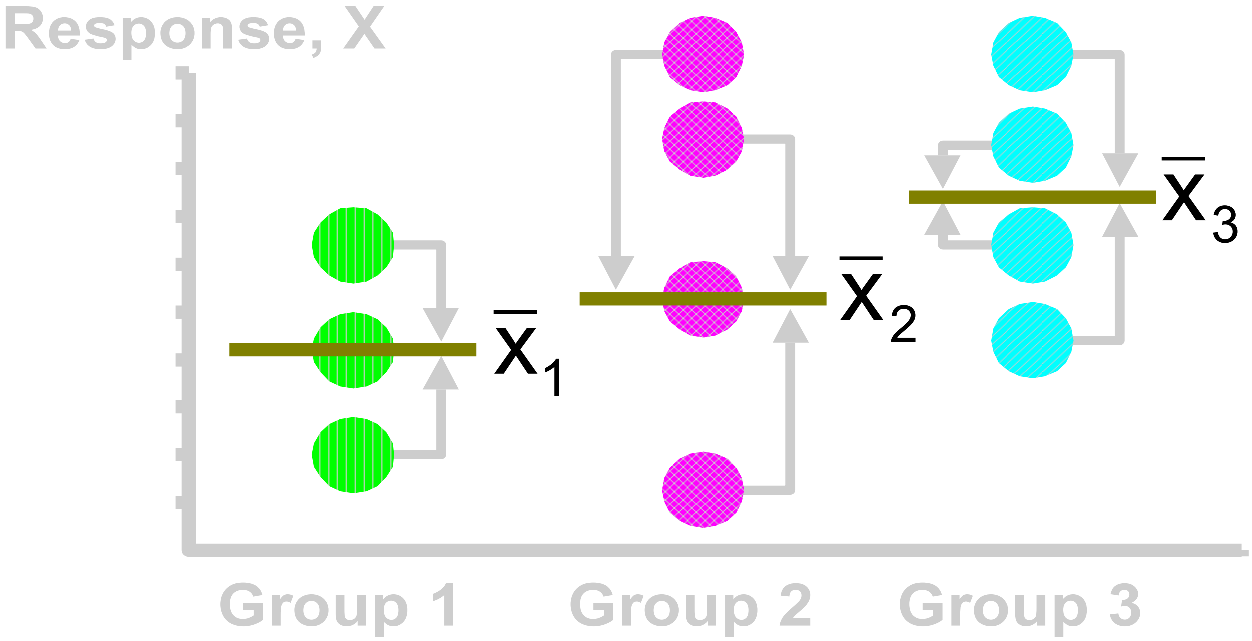
$$MQW = \frac{SQW}{n - K}$$

Média Quadrática Within = SQW/graus de liberdade

Variação dentro do grupo (within)

(cont.)

$$SQW = (x_{11} - \bar{x}_1)^2 + (x_{12} - \bar{x}_1)^2 + \ldots + (x_{Kn_K} - \bar{x}_K)^2$$



Variação Entre Grupos (Between)

$$SQT = SQW + \boxed{SQB}$$

$$SQB = \sum_{i=1}^{K} n_i \, (\bar{x}_i - \bar{x})^2$$

Sendo:

SQB = Soma dos Quadrados entre Grupos

K = número de grupos

n_i = tamanho da amostra do grupo i

$\bar{x}_i$ = média amostral do grupo i

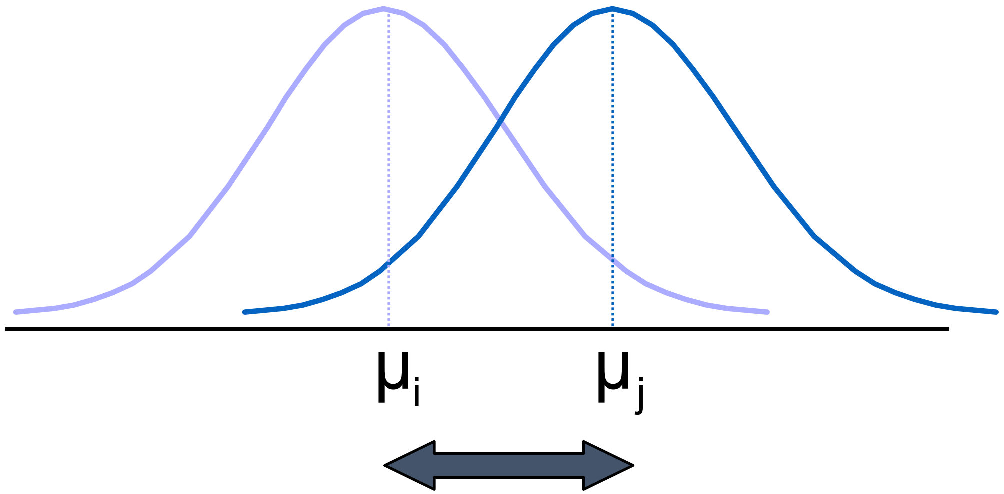
$\bar{x}$ = média geral (média dos valores de todos os dados)

Variação Entre Grupos (Between)

(cont.)

$$SQB = \sum_{i=1}^{K} n_i \, (\bar{x}_i - \bar{x})^2$$

Variação devida às diferenças entre grupos



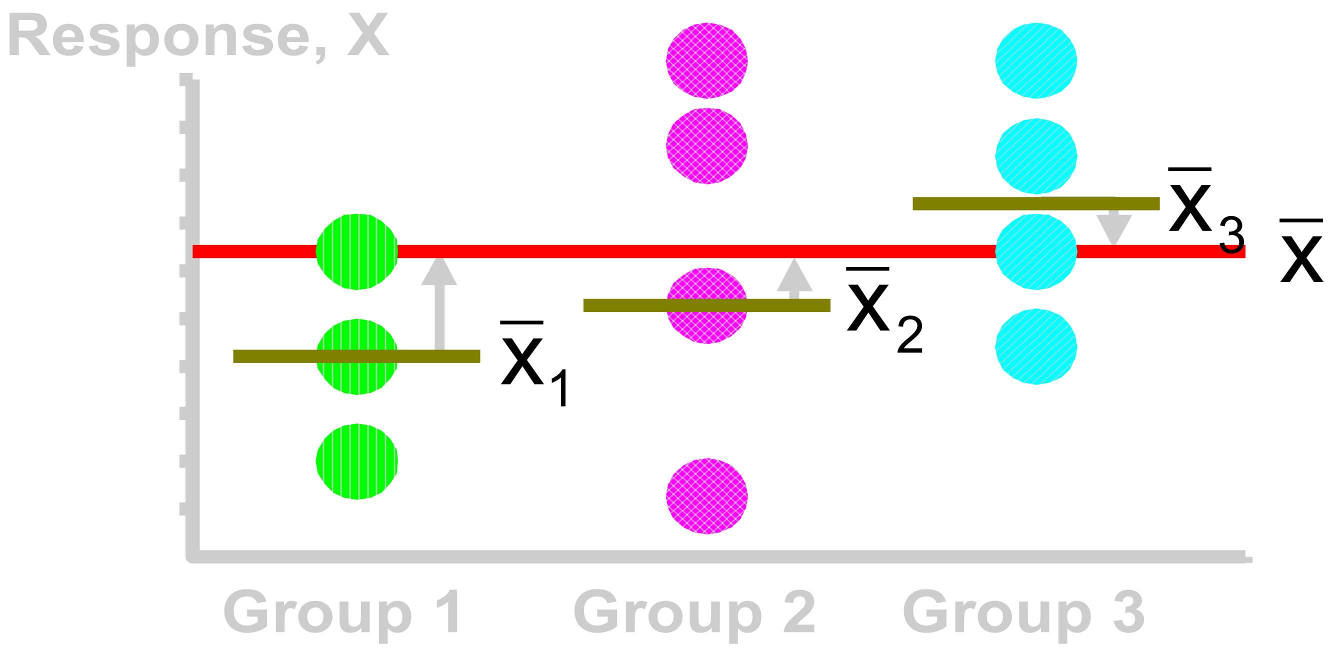
$$MQB = \frac{SQB}{K - 1}$$

Média dos Quadrados Entre Grupos = SQB/graus de liberdade

Variação Entre Grupos (Between)

(cont.)

$$SQB = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \ldots + n_K(\bar{x}_K - \bar{x})^2$$



Obtendo as Médias Quadradas

$$MQT = \frac{SQT}{n-1}$$

$$MQW = \frac{SQW}{n-K}$$

$$MQB = \frac{SQB}{K-1}$$

Tabela ANOVA Unidirecional

Fonte de Variação	SQ	gl	MQ (Variance)	Razão F
Entre Grupos (Between)	SQB	K - 1	MQB = $\frac{SQB}{K - 1}$	F = $\frac{MQB}{MQW}$
Dentro do Grupo (Within)	SQW	n - K	MQW = $\frac{SQW}{n - K}$	
Total	SQT = SQB+SQW	n - 1		

K = número de grupos

n = Tamanho da amostra com todos os grupos

gl = graus de liberdade

ANOVA Um Fator
Estatística de Teste F

H_0: $\mu_1 = \mu_2 = \ldots = \mu_K$

H_1: At least two population means are different

- Estatística de Teste

$$F = \frac{MQB}{MQW}$$

MQB é a média dos Quadrados entre (between) variâncias

MQW é a média dos Quadrados dentro (within) das variâncias

- Graus de Liberdade
  - $gl_1 = K - 1$ (K = número de grupos)
  - $gl_2 = n - K$ (n = Tamanho da amostra com todos os grupos)

Interpretando a Estatística F

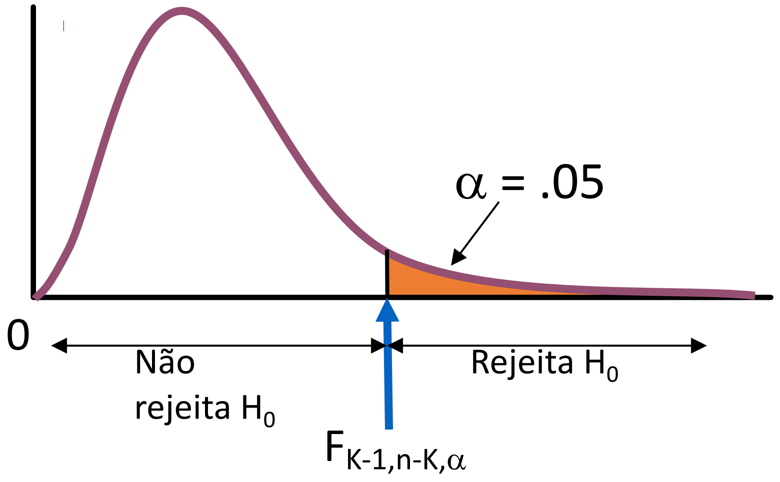
- A Estatística F é a razão da estimativa da variância entre grupos (between) e da estimativa da variância dentro dos grupos (within).

- A taxa sempre deve ser positiva
  - gl_1 = K -1 tipicamente será pequeno
  - gl_2 = n - K tipicamente será grande

Regra de Decisão:

- Rejeita H_0 se

  $F > F_{K-1,n-K,\alpha}$



ANOVA Um Fator
Estatística de Teste F: Exemplo

Você quer saber se três clubes de golfe geram distâncias diferentes. Você, então, seleciona aleatoriamente cinco medidas de testes em uma máquina de condução automática para cada clube. No nível de significância de 0,05, há uma diferença na distância média?

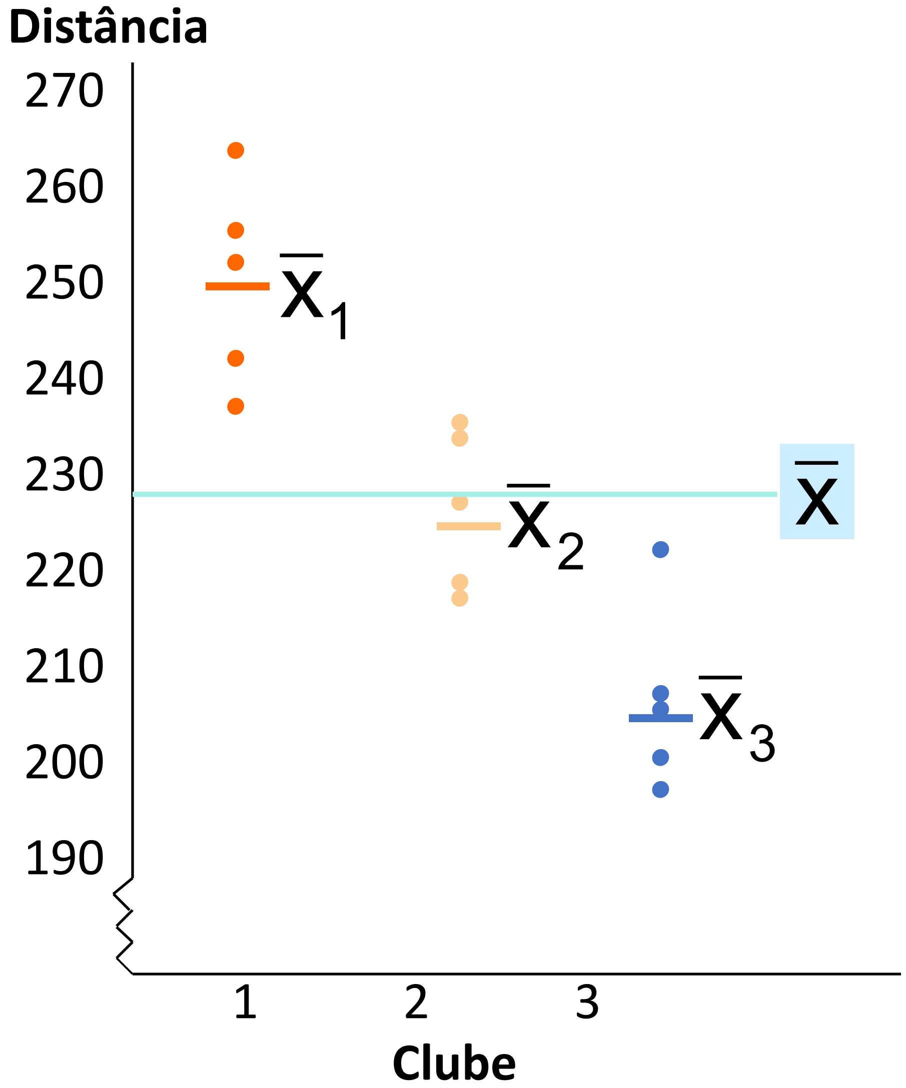
Clube 1	**Clube 2**	**Clube 3**
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

ANOVA Um Fator Exemplo: Diagrama Scatter

Clube 1	Clube 2	Clube 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

$\bar{x}_1 = 249.2$ $\quad \bar{x}_2 = 226.0$ $\quad \bar{x}_3 = 205.8$

$\bar{x} = 227.0$



ANOVA Um Fator Exemplo Contabilizando

Clube 1	**Clube 2**	**Clube 3**
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

$\bar{x}_1 = 249.2$ $n_1 = 5$

$\bar{x}_2 = 226.0$ $n_2 = 5$

$\bar{x}_3 = 205.8$ $n_3 = 5$

$n = 15$

$\bar{x} = 227.0$

$K = 3$

$SQB = 5\,(249.2 - 227)^2 + 5\,(226 - 227)^2 + 5\,(205.8 - 227)^2 = 4716.4$

$SQW = (254 - 249.2)^2 + (263 - 249.2)^2 + \ldots + (204 - 205.8)^2 = 1119.6$

$MQB = 4716.4 / (3\text{-}1) = 2358.2$

$MQW = 1119.6 / (15\text{-}3) = 93.3$

$$F = \frac{2358.2}{93.3} = 25.275$$

ANOVA Um Fator Exemplo - Solução

H_0: $\mu_1 = \mu_2 = \mu_3$

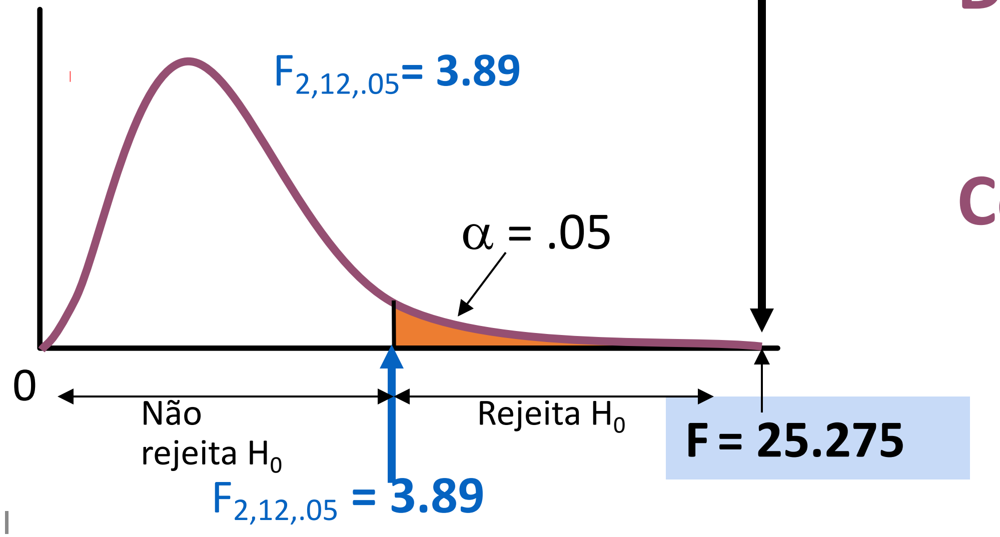
H_1: μ_i pelo menos uma diferente

$\alpha = 0.05$

$gl_1 = 2 \quad gl_2 = 12$

Valor Crítico:

$F_{2,12,.05}$ **= 3.89**



Estatística de Teste:

$$F = \frac{MQB}{MQW} = \frac{2358.2}{93.3} = 25.275$$

Decisão:

Rejeita H_0 à $\alpha = 0.05$

Conclusão:

Há evidência de que pelo menos uma μ_i difere do resto

ANOVA – Fator Único: Saída de Excel

EXCEL: data | data analysis | ANOVA: single factor

SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Club 1	5	1246	249.2	108.2		
Club 2	5	1130	226	77.5		
Club 3	5	1029	205.8	94.2		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	**4716.4**	**2**	**2358.2**	**25.275**	4.99E-05	**3.89**
Within Groups	**1119.6**	**12**	**93.3**			
Total	5836.0	14				

Como fazer o teste em stata e excel

Stata: https://www.youtube.com/watch?v=XEFGGkFRdD4

Excel: https://www.youtube.com/watch?v=0V5scynrVjY

Teste Kruskal-Wallis

- Use quando a hipótese de normalidade para ANOVA unidirecional for violada
- Hipóteses:
  - As amostras são aleatórias e independentes
  - variáveis têm uma distribuição contínua
  - os dados podem ser classificados
  - as populações têm a mesma variabilidade
  - as populações têm a mesma forma

Procedimento para o Teste Kruskal-Wallis

- Obtenha classificações relativas para cada valor
- Em caso de empate, cada um dos valores empatados obtém a classificação média
- Somar as classificações para os dados de cada um dos K grupos
- Calcular a estatística de teste Kruskal-Wallis
- Avalie usando a distribuição qui-quadrado com K - 1 grau de liberdade

Procedimento para o Teste Kruskal-Wallis

(cont.)

- Teste Kruskal-Wallis:

  (quiquadrado com K – 1 graus de liberdade)

$$W = \left[\frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} \right] - 3(n+1)$$

sendo:

n = soma de todas amostras em todos os grupos

K = Número de amostras

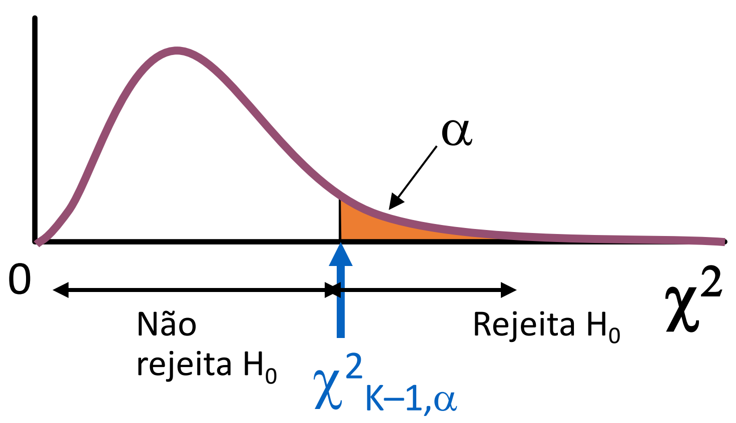
R_i = Soma das classificações no i^{esimo} grupo

n_i = Tamanho do i^{esimo} grupo

Procedimento para o Teste Kruskal-Wallis

(cont.)

- Complete o teste comparando o valor H calculado com um valor crítico χ^2 da distribuição qui-quadrada com K – 1 graus de liberdade



Regra de decisão

- Rejeita H_0 **se** $W > \chi^2_{K-1,\alpha}$
- Caso contrário não rejeita H_0

Kruskal-Wallis Exemplo

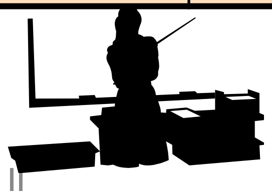
- Diferentes departamentos tem diferentes tamanhos de classes?

Tamanho da classe (Matema, M)	Tamanho da classe (Inglês, E)	Tamanho da classe (Biololgia, B)
23	**55**	**30**
45	**60**	**40**
54	**72**	**18**
78	**45**	**34**
66	**70**	**44**

Kruskal-Wallis Exemplo

- Diferentes departamentos tem diferentes tamanhos de classes?

Tamanho da classe (Matema, M)	Classificação	Tamanho da classe (Inglês, E)	Classificação	Tamanho da classe (Biololgia, B)	Classificação
23	2	55	10	30	3
41	6	60	11	40	5
54	9	72	14	18	1
78	15	45	8	34	4
66	12	70	13	44	7
	$\Sigma = 44$		$\Sigma = 56$		$\Sigma = 20$



Kruskal-Wallis Exemplo

(cont.)

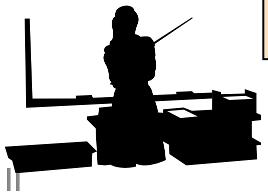
H_0: $\mu_M = \mu_I = \mu_B$

H_1: Nem todas as medias populacionais são iguais

- A estatística W é

$$W = \left[\frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} \right] - 3(n+1)$$

$$= \left[\frac{12}{15(15+1)} \left(\frac{44^2}{5} + \frac{56^2}{5} + \frac{20^2}{5} \right) \right] - 3(15+1) = 6.72$$



Kruskal-Wallis Exemplo

(cont.)

- Comparar W = 6.72 com o valor crítico da distribuição qui-quadrada com 3 – 1 = 2 graus de Liberdade e α = .05:

$$\chi^2_{2,0.05} = 5.991$$

Como W = 6.72 > $\chi^2_{2,0.05} = 5.991$

rejeita H_0

Há evidências suficientes para rejeitar que as médias da população são todas iguais

Análise de Variância Bidirecional

- Examine o efeito de
  - Dois fatores de interesse sobre a variável dependente
    - por exemplo, porcentagem de carbonatação e velocidade da linha no processo de engarrafamento de refrigerantes
  - Interação entre os diferentes níveis desses dois fatores
    - por exemplo: O efeito de um determinado nível de carbonatação depende de qual nível a velocidade da linha é definida?

ANOVA bidirecional

(cont.)

- Pressupostos
  - Populações são normalmente distribuídas
  - Populações tem a mesma variância
  - Amostras aleatórias independentes são retiradas

Projeto de Bloco Randomizado

Dois Fatores de interesse: A e B

K = número de grupos do fator A

H = número de níveis do fator B

Bloco	Grupo			
	1	2	…	K
1	x_{11}	x_{21}	…	x_{K1}
2	x_{12}	x_{22}	…	x_{K2}
.	.	.	.	.
.	.	.	.	.
H	x_{1H}	x_{2H}	…	x_{KH}

Notação Bidirecional

- Considere que x_{ji} denota a observação no j^{esimo} grupo e i^{esimo} bloco
- Suponha que existam K grupos e H blocos, para um total de $n = KH$ observações
- Seja a media geral $\bar{x}$
- Denomine as medias amostrais de grupos por

$$\boxed{\bar{x}_{j\bullet} \quad (j = 1,2,\ldots,K)}$$

- Denomine as médias amostrais dos blocos por

$$\boxed{\bar{x}_{\bullet i} \quad (i = 1,2,\ldots,H)}$$

Partição da Variância Total

- SQT = SQB + SQBb + SQE

Soma dos Quadrados Total (SQT)	=	**Variação devida às diferenças entre grupos (between) (SQB)**
		+
		Variação devida às diferenças entre blocos (SQBb)
		+
Os termos erros são assumidos serem independentes, normalmente distribuídos, e possuirem a mesma variância	→	**Variação devida à aleatoriedade amostral (erro não explicado) (SQE)**

Soma dos Quadrados Bidirecional

- Soma dos Quadrados

	Soma dos Quadrados	Graus de Liberdade
Total	$SQT = \sum_{j=1}^{K}\sum_{i=1}^{H}(x_{ji} - \bar{x})^2$	n – 1
Entre (Between) Grupos	$SQB = H\sum_{j=1}^{K}(\bar{x}_{j\circ} - \bar{x})^2$	K – 1
Entre (Between) Blocos	$SQBb = K\sum_{j=1}^{H}(\bar{x}_{\circ i} - \bar{x})^2$	H – 1
Erro	$SQE = \sum_{j=1}^{K}\sum_{i=1}^{H}(x_{ji}^2 - x_{j\circ}^2 - x_{\circ i}^2 + \bar{x})^2$	(K – 1)(K – 1)

Média dos Quadrados Bidirecionais

- As Médias dos Quadrados são

$$MQT = \frac{SQT}{n-1}$$

$$MQB = \frac{SQB}{K-1}$$

$$MQBb = \frac{SQBb}{H-1}$$

$$MQE = \frac{SQE}{(K-1)(H-1)}$$

ANOVA Bidirecional: A Estatística de Teste F

H_0: As médias dos K grupos da população iguais

Teste F para Grupos

$$F = \frac{MQB}{MQE}$$

Rejeita H_0 se

$F > F_{K-1,(K-1)(H-1),\alpha}$

H_0: As médias dos H blocos da população são iguais

Teste F para Blocos

$$F = \frac{MQBb}{MQE}$$

Rejeita H_0 se

$F > F_{H-1,(K-1)(H-1),\alpha}$

Formato Geral da Tabela Bidirecional

Origem da variação	Soma dos Quadrados	Graus de Liberdade	Médias dos Quadrados	Razão F
Entre (Between) Grupos	SQB	K – 1	$MQB = \frac{SQB}{K-1}$	$\frac{MSB}{MSE}$
Entre (Between) Blocos	SQBb	H – 1	$MQBb = \frac{SQBb}{H-1}$	$\frac{MSBb}{MSE}$
Erro	SQE	(K – 1)(H – 1)	$MQE = \frac{SQE}{(K-1)(H-1)}$	
Total	SQT	n - 1		

Mais de uma observação por célula

- Um desenho bidirectional com mais de uma observação por célula permite uma fonte de variação a mais
- A interação entre grupos e blocos também pode ser identificada
- Seja
  - K = número de grupos
  - H = número de blocos
  - L = número de observações por célula
  - n = KHL = número total de observações

Mais de uma observação por célula

(cont.)

SQT = SQB + SQBb + SQI + SQE

Graus de Liberdade:

SQT
Variação Total

n – 1

SQB. Variação entre grupos (between) — K – 1

SQBb. Variação entre blocos (between) — H – 1

SQI. Variação devida à interação Entre grupos e blocos (between) — (K – 1)(H – 1)

SQE
Variação aleatótia (Erro) — KH(L – 1)

Soma dos Quadrados com Interação

		Graus de Liberdade
Total	$SQT = \sum_{j}^{H}\sum_{i}^{K}\sum_{l}^{L}(x_{jil} - \bar{x})^2$	n - 1
Entre (Between) Grupos	$SQB = HL\sum_{j=1}^{K}(\bar{x}_{j\circ\circ} - \bar{x})^2$	K – 1
Entre (Between) Blocos	$SQBb = KL\sum_{j=1}^{H}(\bar{x}_{\circ i\circ} - \bar{x})^2$	H – 1
Interação	$SQI = L\sum_{j=1}^{K}\sum_{i=1}^{H}(\bar{x}_{ji\circ} - \bar{x}_{j\circ\circ} - \bar{x}_{\circ i\circ} + \bar{x})^2$	(K – 1)(H – 1)
Erro	$SQE = \sum_{i}^{K}\sum_{j}^{H}\sum_{l}^{L}(x_{jil} - \bar{x}_{ji\circ}^2)^2$	KH(L – 1)

Média dos quadrados bidirectional com Interação

- As médias quadradas são

$$MQT = \frac{SQT}{n-1}$$

$$MQB = \frac{SQB}{K-1}$$

$$MQBb = \frac{SQBb}{H-1}$$

$$MSI = \frac{SQI}{(K-1)(H-1)}$$

$$MQE = \frac{SQE}{KH(L-1)}$$

ANOVA bidirecional: A estatística de Teste F

H_0: As médias dos K grupos populacionais são iguais

Teste F para efeito do grupo

$$F = \frac{MSB}{MSE}$$

Rejeita H_0 se

$F > F_{K-1,KH(L-1),\alpha}$

H_0: As médias dos H blocos populacionais são iguais

Teste F para efeito do bloco

$$F = \frac{MSBb}{MSE}$$

Rejeita H_0 se

$F > F_{H-1,KH(L-1),\alpha}$

H_0: a interação de grupos e blocos é igual a zero

Teste F para efeito interação

$$F = \frac{MSI}{MSE}$$

Rejeita H_0 se

$F > F_{(K-1)(H-1),KH(L-1),\alpha}$

Two-Way ANOVA Summary Table

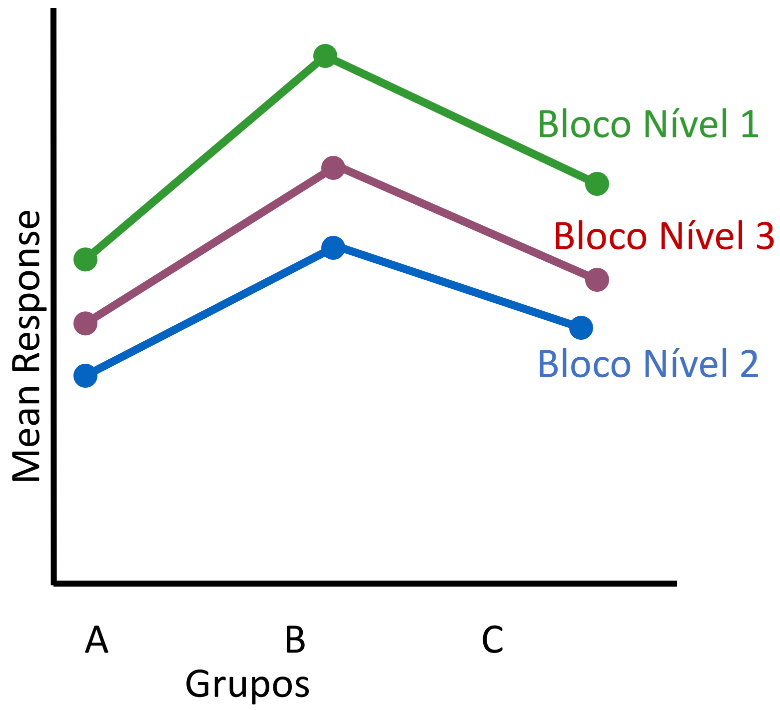
Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Médias dos Quadrados	Razão F
Entre (Between) Grupos	**SQB**	**K – 1**	**MQB** = SQB / (K – 1)	$\frac{\textbf{MQB}}{\textbf{MQE}}$
Entre (Between) Blocos	**SQBb**	**H – 1**	**MQBb** = SQBb / (H – 1)	$\frac{\textbf{MQBb}}{\textbf{MQE}}$
Interação	**SQI**	**(K – 1)(H – 1)**	**MQI** = SSI / (K – 1)(H – 1)	$\frac{\textbf{MQI}}{\textbf{MQE}}$
Erro	**SQE**	**KH(L – 1)**	**MQE** = SSE / KH(L – 1)	
Total	**SQT**	**n – 1**		

Características do teste F da ANOVA bidirectional Features

- Graus de liberdade sempre somam
  - n-1 = KHL-1 = (K-1) + (H-1) + (K-1)(H-1) + KH(L-1)
  - Total = grupos + blocos + interação + erro
- O denominador do Teste F é sempre o mesmo, mas o numerador é diferente
- As somas dos quadrados sempre somam
- SQT = SQB + SQBb + SQI + SQE
- Total = grupos + blocos + interação + erro

Exemplos:
Com Interação vs. Sem Interação

- Sem interação:



- Com Interação: