ANOVA

Análise de Variâncias

ANOVA : ANalysis Of VAriance

A ANOVA é usada para determinar se existem diferenças entre três ou mais médias populacionais.

1) ANOVA de 1 via: examina o efeito de um fator (variável categórica) em uma variável quantitativa.

2) ANOVA de 2 vias: examina os efeitos de dois fatores (variáveis categóricas), que podem ou não “interagir”, em uma variável quantitativa.

COMPARANDO 3 MÉDIAS (OU MAIS)

Comparando médias

Comparamos três ou mais médias entre 1 ou 2 grupos:

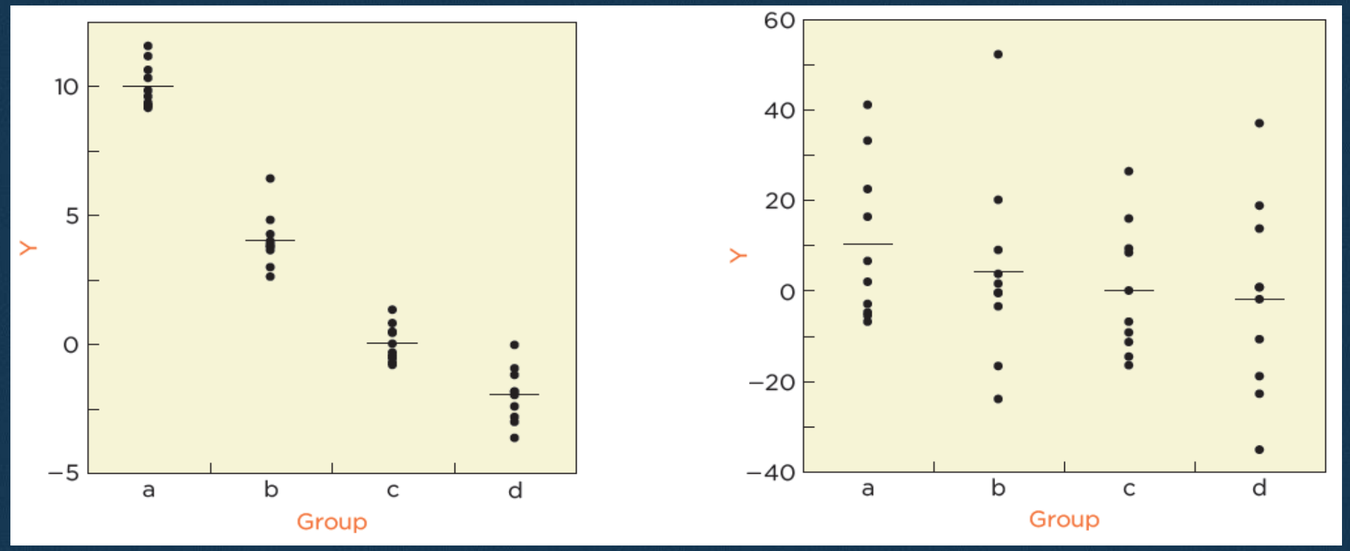
Há diferença de salários entre os cargos de analista de dados, cientista de dados e engenheiro de dados em uma empresa?

Teste ANOVA de 1 via

Há diferença de salários entre os cargos de analista de dados, cientista de dados e engenheiro de dados em uma empresa por nível de escolaridade?

Teste ANOVA de 2 vias

Comparando médias



TESTE DE HIPÓTESES: ANOVA

Teste de Hipóteses

As hipóteses para a ANOVA são:

Ho: μ1 = μ2 = ... = μc (todas as médias são iguais)
Ha: Nem todas as médias da população são iguais

Metodologia

A Análise de Variância compara dois tipos de variações calculadas a partir das médias da amostra e das variâncias da amostra.

Variação entre grupos: variabilidade entre as médias das amostras

Variação dentro dos grupos: variabilidade dentro de cada amostra

Examinamos as proporções dessas estimativas:

- Se estiverem próximas, Ho não é rejeitado
- Se forem estatisticamente diferentes, Ho é rejeitado

Exemplo

Queremos testar se fertilizantes diferentes afetam a produção de milho.

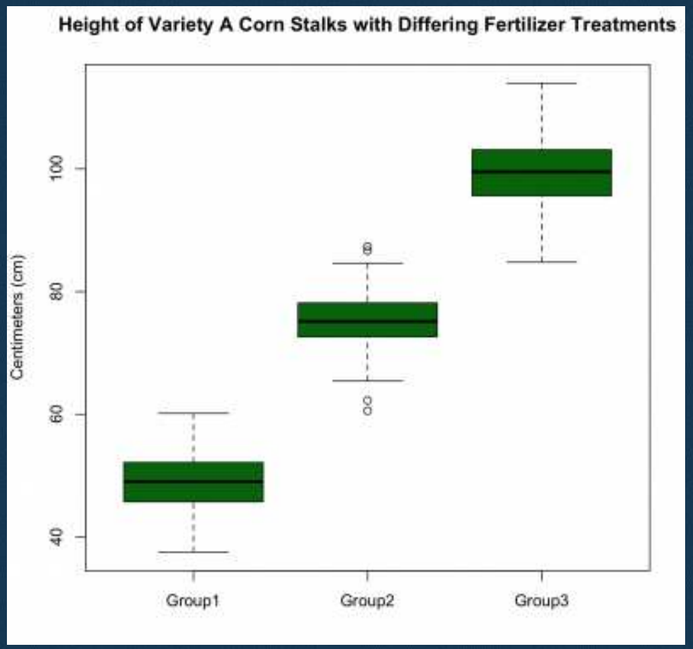
Os campos de milho são divididos em três grupos e um fertilizante diferente é aplicado a cada grupo.

Em cada grupo, medimos o tamanho do caule do milho para comparar a variação dentre dos grupos e entre os grupos.

Exemplo 1

A variação do crescimento do caule do milho entre os grupos é maior do que a variação dentro de cada grupo.

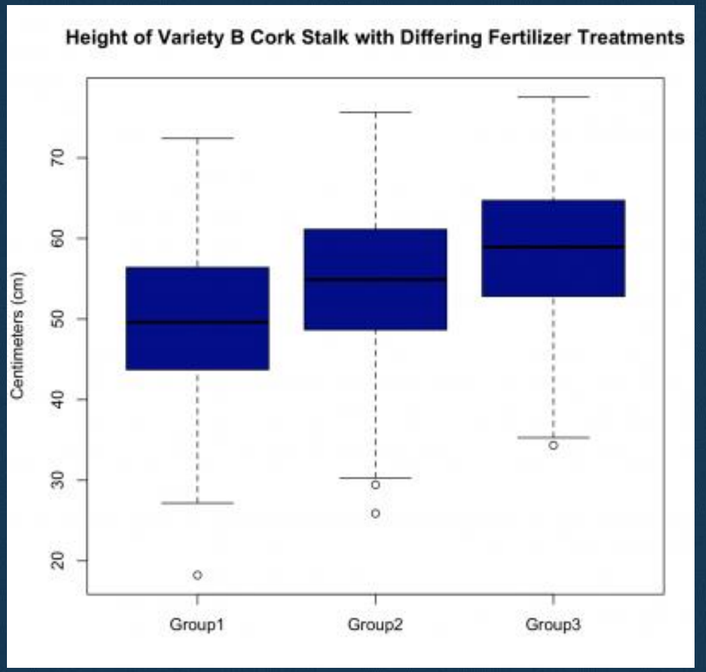
É mais provável que as médias sejam diferentes (rejeitando Ho).



Exemplo 2

A variação dentro de cada grupo é relativamente grande, enquanto não há muita variação entre os grupos.

É mais provável que as médias sejam similares (não rejeitando Ho).



Suposições

1. Independência das Observações

2. Normalidade dos Resíduos: testes estatísticos e QQ plot.

3. Homoscedasticidade: testes estatísticos e gráfico de dispersão dos resíduos.

ANOVA DE 1 VIA

Teste de Hipóteses

Passo 1: Defina as hipóteses

Passo 2: Escolha o nível de significância α

Passo 3: Calcule o p-valor

Passo 4: Conclua e interprete o resultado

Tabela ANOVA

Medimos a variabilidade entre os grupos comparando com a variabilidade dentro dos grupos.

Fonte	Graus de liberdade	Soma dos quadrados (SQ)	Quadrados médios (QM)	F
Entre grupos	k-1	SQE	QME	QME/QMD
Dentro dos grupos	N-k	SQD	QMD	
Total	N-1	SQE+SQD		

Exemplo

A gerente de um banco está buscando a melhor maneira de contratar funcionários para sua agência. Ela sabe que precisa de mais caixas às sextas-feiras do que nos outros dias, mas está tentando descobrir se a necessidade de caixas é constante durante o resto da semana. Ela coleta o número de transações a cada dia (segunda, terça, quarta e quinta-feira) durante dois meses.

Pergunta: O número médio de transações difere por dia?

Dados

Segunda	Terça	Quarta	Quinta
276	243	288	254
323	279	292	279
298	301	310	241
256	285	267	227
277	274	243	278
309	243	293	276
312	228	255	256
265	298	273	262
311	255		
$n_1=9$	$n_2=9$	$n_3=8$	$n_4=8$

Variação Entre Grupos

Para medir a variabilidade entre os tratamentos, comparamos as médias da amostra com a média total.

	Segunda	Terça	Quarta	Quinta
Média	291,89	267,33	277,63	259,13
Variância	569,11	681,25	491,98	348,70
Observações	$n_1=9$	$n_2=9$	$n_3=8$	$n_4=8$

A média total é a média de todas as 34 transações: 274,32

Variação Entre Grupos

Fonte	Graus de liberdade	Soma dos quadrados (SQ)	Quadrados médios (QM)	F
Entre grupos	k-1	SQE	QME	QME/QMD
Dentro dos grupos	N-k	SQD	QMD	
Total	N-1	SQE+SQD		

Chamamos de variação explicada a soma dos quadrados entre os grupos (SQE):

$$SQE = \sum_{i=1}^{c} n_i(\bar{x}_i - \bar{\bar{x}})^2$$

Calculamos o quadrado médio entre os grupos (QME), que é a SQE dividido pelos graus de liberdade k-1:

$$QME = \frac{SQE}{k-1}$$

Dados

Fonte	Graus de liberdade	Soma dos quadrados (SQ)	Quadrados médios (QM)	F
Entre grupos	k-1	SQE	QME	QME/QMD
Dentro dos grupos	N-k	SQD	QMD	
Total	N-1	SQE+SQD		

	Segunda	Terça	Quarta	Quinta
Média	291,89	267,33	277,63	259,13
Variância	569,11	681,25	491,98	348,70
Observações	n_1=9	n_2=9	n_3=8	n_4=8

SQE = $9(291{,}89 - 274{,}32)^2 + 9(267{,}33 - 274{,}32)^2 + 8(277{,}63 - 274{,}32)^2 + 8(259{,}13 - 274{,}32)^2 = 5151{,}8$

k-1 = 4-1 = 3

A variação entre os grupos (QME) é calculada por: $\frac{SQE}{k-1}$ = 1717,3

Variação Dentro dos Grupos

Fonte	Graus de liberdade	Soma dos quadrados (SQ)	Quadrados médios (QM)	F
Entre grupos	k-1	SQE	QME	QME/QMD
Dentro dos grupos	N-k	SQD	QMD	
Total	N-1	SQE+SQD		

Chamamos de variação inexplicada a soma dos quadrados dentro dos grupos (SQD):

$$SQD = \sum_{i=1}^{c} (n_i - 1) {s_i}^2$$

Calculamos o quadrado médio dentro dos grupos (QMD), dividindo a SQD pelos graus de liberdade:

$$QMD = \frac{SQD}{N - k}$$

Dados

Fonte	Graus de liberdade	Soma dos quadrados (SQ)	Quadrados médios (QM)	F
Entre grupos	k-1	SQE	QME	QME/QMD
Dentro dos grupos	N-k	SQD	QMD	
Total	N-1	SQE+SQD		

	Segunda	Terça	Quarta	Quinta
Média	291,89	267,33	277,63	259,13
Variância	569,11	681,25	491,98	348,70
Observações	n_1=9	n_2=9	n_3=8	n_4=8

SQD =$8(569{,}11) + 8(681{,}25) + 7(491{,}98) + 7(348{,}70)$ = 15887,64

N-k = 34-4 = 30

A variação entre os grupos (QMD) é calculada por: $\frac{SQD}{N-k}$= 529,6

Estatística F

Fonte	Graus de liberdade	Soma dos quadrados (SQ)	Quadrados médios (QM)	F
Entre grupos	k-1	SQE	QME	QME/QMD
Dentro dos grupos	N-k	SQD	QMD	
Total	N-1	SQE+SQD		

Variância entre grupos QME = 1717,3

Variância dentro dos grupos QMD = 529,6

A estatística do teste é

$$F = \frac{QME}{QMD} = \frac{1717,3}{529,6} = 3,24$$

Estatística F

Fonte	Graus de liberdade	Soma dos quadrados (SQ)	Quadrados médios (QM)	F
Entre grupos	k-1	SQE	QME	QME/QMD
Dentro dos grupos	N-k	SQD	QMD	
Total	N-1	SQE+SQD		

SQE+SQD é a variação total.

SQE é a variação explicada, esta variação é "explicada" pelo fato de que as amostras podem vir de populações com médias diferentes.

SQD é a variação inexplicada, porque é uma variação aleatória da amostragem.

Quanto maior for a variação explicada em relação à variação inexplicada, mais provável será que as médias da população sejam diferentes.

Resultado

Fonte	Graus de liberdade	Soma dos quadrados (SQ)	Quadrados médios (QM)	F
Entre grupos	k-1	SQE	QME	QME/QMD
Dentro dos grupos	N-k	SQD	QMD	
Total	N-1	SQE+SQD		

ANOVA					
Fonte da variação	SQ	gl	MQ	F	p-valor
Entre grupos	5151,802	3	1717,267	3,242648	0,035735
Dentro dos grupos	15887,64	30	529,588		
Total	21039,44	33			

Teste de Hipóteses

Passo 1:

- Ho: μSeg = μTer = μQua = μQui
- Ha: Nem todas as médias são iguais

Passo 2: Defina α = 0,05

Passo 3: O p-valor é 0,036

0,036 < 0,05: Rejeito Ho

Passo 4: O número médio de transações difere por dia da semana, ou pelo menos há um dia diferente dos outros.

TESTE POST-HOC

Métodos de Comparação

Quando o teste da ANOVA de 1 via encontra diferenças significativas entre as médias da população (rejeita Ho), é natural nos perguntarmos quais médias diferem.

O teste ANOVA não responde a essa pergunta.

Se a hipótese nula for rejeitada, a próxima etapa é realizar uma análise post-hoc para ver quais médias diferem.

Para isso, as diferenças de pares são investigadas usando intervalos de confiança.

Métodos de Comparação

- Teste LSD de Fisher
- Teste de Bonferroni
- Teste HSD de Tukey

Teste LSD de Fisher

O teste LSD de Fisher compara pares de médias sem ajuste para múltiplas comparações.

É usado para identificar diferenças específicas entre pares de grupos.

Problema: não ajusta para múltiplas comparações, o que aumenta a probabilidade de cometer um erro do tipo I (rejeitar a hipótese nula quando ela é verdadeira).

Teste de Bonferroni

O método de Bonferroni faz uma correção do nível de significância α para reduzir a chance do erro tipo I.

Isso faz com que o intervalo seja um pouco mais amplo do que o teste de Fisher.

Mais conservador, tem um maior poder do teste quando são poucos grupos.

Teste HSD de Tukey

O método de Diferença Honestamente Significativa (HSD) de Tukey também reduz a chance do erro tipo I.

Isso faz com que o intervalo seja um pouco mais amplo do que o teste de Fisher.

Ele compara todas as possíveis combinações de pares de médias de grupos e é ideal quando temos um número igual de observações em cada grupo.

Exemplo

Intervalo de confiança de 95% para testar se há diferença no número médio de transações às segundas em relação às terças-feiras.

Fisher: [2,41; 46,71]

Bonferroni: [-6,09; 55,21]

Tukey: [-4,94; 54,06]

ANOVA DE 2 VIAS

Análise de Variâncias

A ANOVA é usada para determinar se existem diferenças entre três ou mais médias populacionais.

1) ANOVA de 1 via: examina o efeito de um fator (variável categórica) em uma variável quantitativa.

2) ANOVA de 2 vias: examina os efeitos de dois fatores (variáveis categóricas), que podem ou não “interagir”, em uma variável quantitativa.

Exemplos

(1) Há diferença de renda com base na área de trabalho e nível de escolaridade?

A renda é a variável numérica que está sendo testada.

Área de trabalho e nível escolar são as duas variáveis categóricas (fatores ou tratamento).

(2) Há diferença na produção entre operadores de máquina com base no método de treinamento e experiência?

A produção é a variável numérica que está sendo testada.

Método de treinamento e experiência são as duas variáveis categóricas.

Exemplo

Uma estudante está tentando decidir qual carreira seguir em três áreas. Ela entrevista quatro pessoas em cada área e pergunta seus salários (em US$1.000/ano).

Field of Employment (Factor *A*)		
Educational Services	**Financial Services**	**Medical Services**
18	25	26
35	45	43
46	58	62
75	90	110
$\bar{x}_{education} = 43.50$	$\bar{x}_{financial} = 54.50$	$\bar{x}_{medical} = 60.25$

Exemplo

Neste caso, um teste ANOVA de uma via indica que os salários médios nas áreas não são significativamente diferentes (p-valor = 0,73)

ANOVA					
Fonte da variação	SQ	gl	MQ	F	p-valor
Entre grupos	579,5	2	289,75	0,3299	0,7273
Dentro dos grupos	7902,75	9	878,08		
Total	8482,25	11			

Verificando outro fator

Se for considerado o nível de escolaridade dos trabalhadores, a história muda. Está claro que o nível de escolaridade também impacta os salários.

Education Level (Factor *B*)	Field of Employment (Factor *A*)			
	Educational Services	Financial Services	Medical Services	Factor *B* Means
High School	18	25	26	$\overline{x}_{high\ school} = 23.00$
Bachelor's	35	45	43	$\overline{x}_{bachelor's} = 41.00$
Master's	46	58	62	$\overline{x}_{master's} = 53.3333$
Ph.D.	75	90	110	$\overline{x}_{Ph.D.} = 91.6667$
Factor *A* Means	$\overline{x}_{education} = 43.50$	$\overline{x}_{financial} = 54.50$	$\overline{x}_{medical} = 60.25$	$\overline{\overline{x}} = 52.75$

Layout da ANOVA de 2 vias

Source of Variation	SS	df	MS	F	p-value
Rows	SSB	$r-1$	$MSB=\frac{SSB}{r-1}$	$F_{(df_1,df_2)}=\frac{MSB}{MSE}$	$P\left(F_{(df_1,df_2)}\geq\frac{MSB}{MSE}\right)$
Columns	SSA	$c-1$	$MSA=\frac{SSA}{c-1}$	$F_{(df_1,df_2)}=\frac{MSA}{MSE}$	$P\left(F_{(df_1,df_2)}\geq\frac{MSA}{MSE}\right)$
Error	SSE	$n_T-c-r+1$	$MSE=\frac{SSE}{n_T-c-r+1}$		
Total	SST	n_T-1			

Observe três fontes de variação:

- Variabilidade da linha (devido ao Fator B);
- Variabilidade da coluna (devido ao Fator A);
- Variabilidade devido ao acaso ou SSE

Teste de Hipóteses

Para as médias nas linhas:
Ho: As médias da linhas são as mesmas
Ha: A média de pelo menos uma linha é diferente

Para as médias das colunas:
Ho: As médias das colunas são as mesmas
Ha: A média de pelo menos uma coluna é diferente

Resultado

ANOVA					
Fonte da variação	SQ	gl	MQ	F	p-valor
Linhas	7632,917	3	2544,306	56,57505	8,6E-05
Colunas	579,5	2	289,75	6,442866	0,03206
Erro	269,8333	6	44,97222		
Total	8482,25	11			

- Linhas: O salário difere por nível educacional, conforme indicado pelo p-valor.
- Colunas: Quando levamos em conta o nível educacional, os salários médios diferem por área.

ANOVA COM INTERAÇÃO

ANOVA de 2 vias com interação

Interação significa que o efeito de um fator depende do nível do outro fator.

Por exemplo, talvez a educação tenha impacto sobre os salários no setor financeiro, mas não nos esportes profissionais.

As duas categorias, setor de emprego e educação, interagem de maneira diferente dependendo do setor.

Na ANOVA de 2 vias com interação, particionamos a variabilidade total do conjunto de dados em quatro componentes:
SSA, SSB, SSAB e SSE.

Exemplo

Cada célula representa o salário médio de 3 pessoas que se enquadram na categoria específica:

Education Level (Factor *B*)	Field of Employment (Factor *A*)			
	Educational Services	Financial Services	Medical Services	Factor *B* Means
High School	22.3333	25.6667	25.00	24.3333
Bachelor's	33.00	46.00	43.3333	40.7778
Master's	47.6667	54.6667	59.3333	53.8889
Ph.D.	77.00	92.3333	98.3333	89.2222
Factor *A* Means	45.00	54.6667	56.50	$\overline{\overline{x}} = 52.0556$

Tabela ANOVA com interação

Source of Variation	*SS*	*df*	*MS*	*F*
Rows	SSB	r-1	MSB	$F_{(df1,df2)} = \frac{MSB}{MSE}$
Columns	SSA	c-1	MSA	$F_{(df1,df2)} = \frac{MSA}{MSE}$
Interaction	SSAB	(r-1)(c-1)	MSAB	$F_{(df1,df2)} = \frac{MSAB}{MSE}$
Error	SSE	rc(w-1)	MSE	
Total	SST			

Hipóteses

Para um teste ANOVA de dois fatores com interação, existem três conjuntos de hipóteses:

(1) Interação:
Ho: Não há interação
Ha: Há interação

(2) Médias das Linhas
Ho: As médias da linha são as mesmas
Ha: pelo menos uma média de linha é diferente

Hipóteses

(3) Médias das Colunas

Ho: As médias da coluna são as mesmas

Ha: A média de pelo menos uma coluna é diferente

Observação: Se houver interação (Ho for rejeitada), não testamos as médias das linhas e colunas.

Resultado

Interação: p-valor = 0,002. Ao nível de significância de 5%, podemos concluir que existem evidências de um efeito de interação entre a área de trabalho e o nível de educação.

Esse resultado implica que o salário médio de quem buscou níveis educacionais superiores é maior em algumas áreas do que em outras.

Source of Variation	SS	df	MS	F	p-value
Sample (Rows)	20523.8889	3	6841.2963	658.521	3.58E-23
Columns	916.2222	2	458.1111	44.096	9.18E-09
Interaction	318.4444	6	53.0741	5.109	0.002
Within (Error)	249.3333	24	10.3889		
Total	22007.8889	35			

Importante

Por causa da interação, as diferenças entre os níveis de educação não são as mesmos para todas as áreas.

Esse resultado dificulta a interpretação dos efeitos principais, já que as diferenças de um fator não são constantes em outro fator.

Se a interação não for significativa, podemos proceder focando nos efeitos principais: testando se as médias das linhas ou das colunas diferem.

Se a interação for significativa (como neste caso),
podemos usar a análise de regressão.