Análise de Variância - Parte I

Rodrigo Sant'Ana[1]

[1] Universidade do Vale do Itajaí - UNIVALI
Centro de Ciências Tecnológicas, da Terra e do Mar - CTTMar
Curso de Engenharia Ambiental e Sanitária - EAS
Curso de Engenharia Ambiental - EA
Laboratório do Grupo de Estudos Pesqueiros - GEP

Março, 2015

Sumário

UNIVALI

Análise de Variância

Análise de Variância - ANOVA

Trata-se de um método estatístico que permite realizar comparações simultâneas entre três ou mais médias, ou seja, permite testas **hipóteses** sobre médias de diferentes populações.

Análise de Variância

Análise de Variância - ANOVA

Trata-se de um método estatístico que permite realizar comparações simultâneas entre três ou mais médias, ou seja, permite testas **hipóteses** sobre médias de diferentes populações.

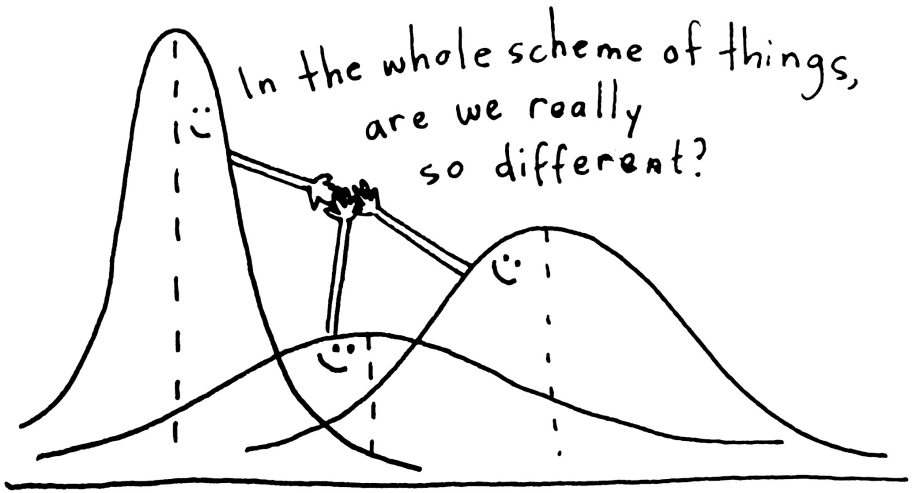
Importante

Na ANOVA o termo **Variância** se refere ao método utilizado e **não** à estatística que está sendo testada.

Estamos testando médias...

Ideia básica

A Análise de Variância (ANOVA) se estrutura no particionamento da variabilidade contrariando o próprio nome. Neste sentido, este método não está preocupado em analisar variância, mas sim, a variabilidade nas médias ou no em torno delas.



- $H_0 : \mu_1 = \mu_2 = ... = \mu_k$
- H_1: pelo menos uma média é distinta das demais.

Por que aprender um novo método chamado ANOVA, quando simplesmente poderíamos conduzir uma série de testes T encadeados??

Por que aprender um novo método chamado ANOVA, quando simplesmente poderíamos conduzir uma série de testes *T* encadeados??

ANOVA

$$\mu_1 = \mu_2 = \mu_3$$

Um único teste com 95% de confiança

ou seja,

nível de significância $\alpha = 0.05$

(% Confiança)^c	$1 - (1 - \alpha)^c$

"Boa ideia"

Teste *T*

$$\mu_1 = \mu_2, \mu_1 = \mu_3 \text{ e } \mu_2 = \mu_3$$

Três testes, cada um com 95% de confiança

ou seja,

cada teste com um $\alpha = 0.05$

(% Confiança)^c	$1 - (1 - \alpha)^c$

"Péssima ideia"

Quando mais de um teste T é ajustado de forma encadeada, cada um com seu respectivo nível de significância, a probabilidade de termos um erro sobre a decisão é aumentada exponencialmente (**Erro do Tipo I**).

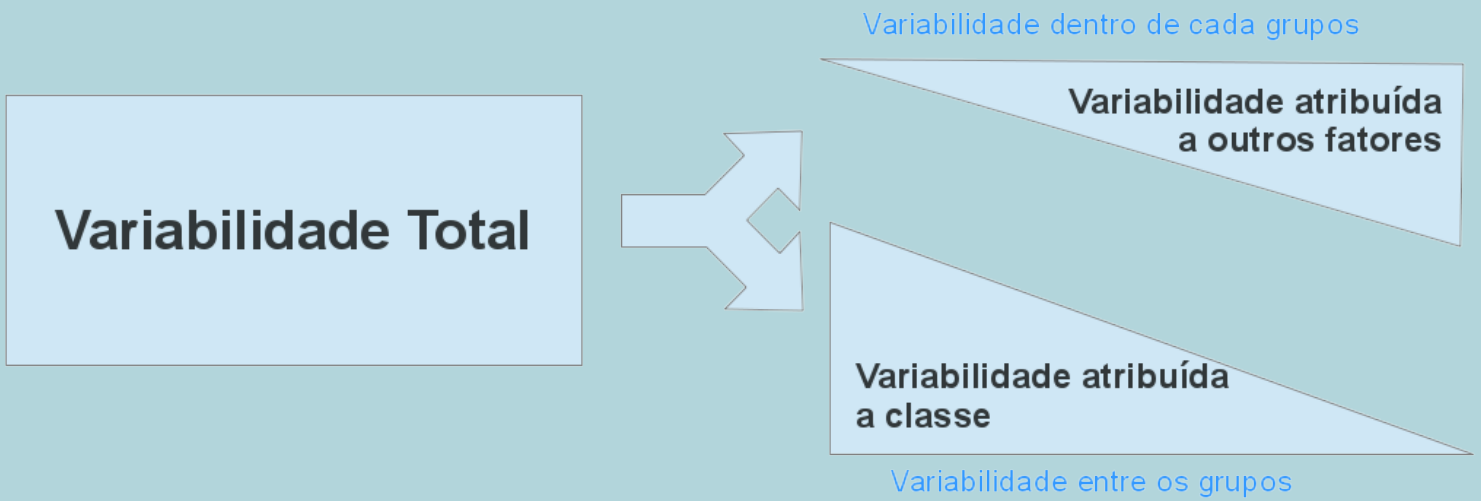
		Decisão	
		Falha em rejeitar H_0	rejeita H_0
Verdade	H_0 é verdadeira	OK	**Erro do Tipo I**
	H_1 é verdadeira	**Erro do Tipo II**	OK

"We (almost) never know if H_0 ou H_1 is true, but we need to consider all possibilities"

Dr. Mine Çetinkaya-Rundel
Duke University

UNIVALI

Particionamento da variabilidade



Pressupostos do método

I) Todas as observações devem ser independentes uma das outras e devem ser selecionadas aleatoriamente da população que representam;

II) As populações de onde foram retiradas as amostras devem ter distribuições aproximadamente normais;

III) As variâncias devem ser aproximadamente as mesmas entre as diferentes populações (homocedasticidade das variâncias).

UNIVALI

Independência - considerações sobre o experimento

Consiste em pressupor que os erros são *variáveis aleatórias independentes*. Mas o que significa isto???

Independência - considerações sobre o experimento

Consiste em pressupor que os erros são *variáveis aleatórias independentes*. Mas o que significa isto???

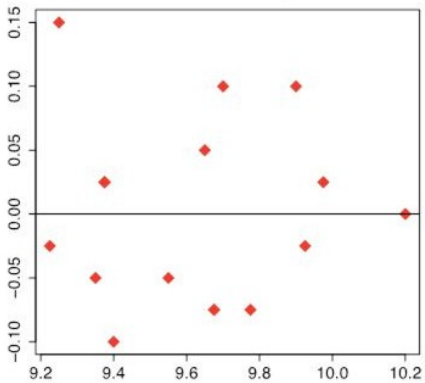
Consideremos um experimento com voluntários...

É razoável assumir “Independência”

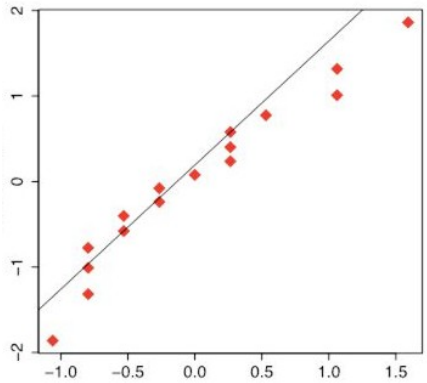
É razoável assumir “Dependência”

Independência - análise de resíduos

A análise gráfica dos resíduos é extremamente útil, porém não pode associar um nível de probabilidade à conclusão de que os erros não são independentes.



É razoável assumir “Independência”



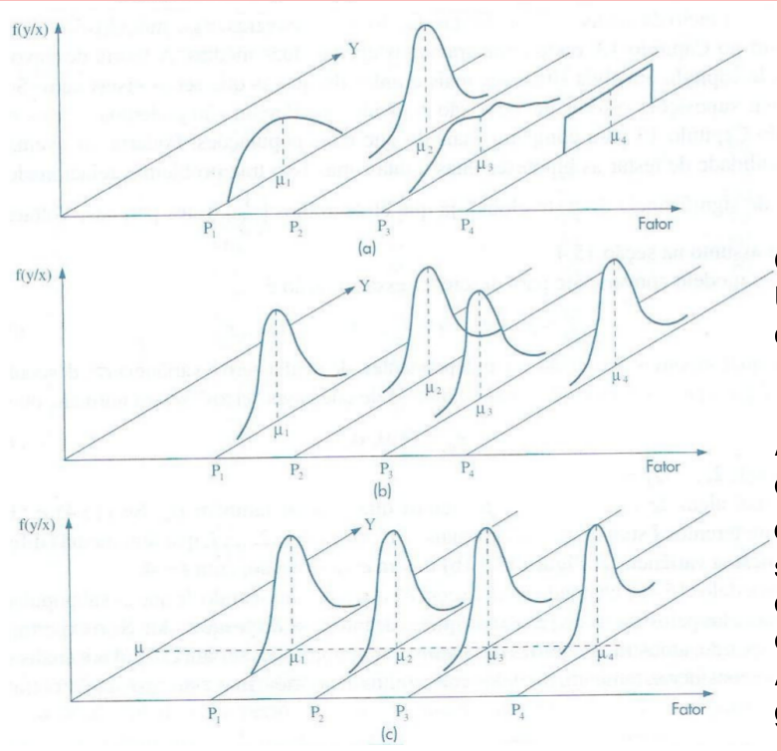
É razoável assumir “Dependência”

“Variáveis ou dados coletados em sequência - no tempo ou no espaço - geralmente têm correlação. Em outras palavras, medidas feitas na mesma unidade ou em unidades agrupadas, estão, muito frequentemente, correlacionadas”

Sonia Vieira, *Análise de Variância (ANOVA)*

UNIVALI

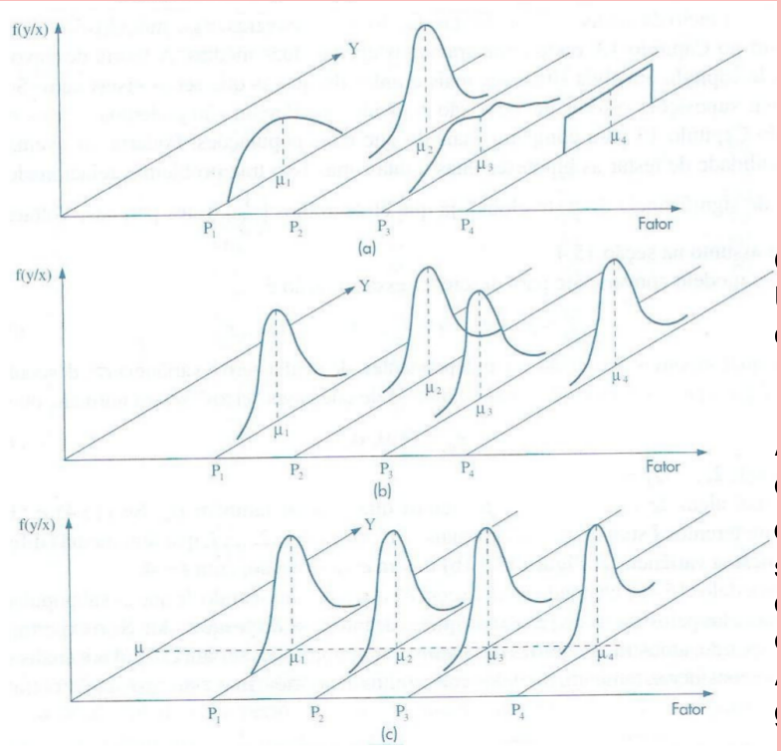
Igualdade das variâncias



Regra prática - “Cuidado ao usar!!!”

- Maior variância não exceda em três vezes a menor (Dean & Voss, 1999);
- Maior variância não exceda em quatro vezes a menor (Box, 2000).

Igualdade das variâncias



Regra prática - “Cuidado ao usar!!!”

- Maior variância não exceda em três vezes a menor (Dean & Voss, 1999);
- Maior variância não exceda em quatro vezes a menor (Box, 2000).

Alternativa, testar igualdade de variâncias - “Embora nenhum deles tenha ampla recomendação!!!”

- teste de Cochran;
- teste de Hartley;
- teste de Bartlett;
- teste de Levene.

Variáveis obtidas por processos de contagem

Em geral, variáveis discretas não possuem variância constante, muito menos, distribuição aproximadamente normal. No entanto são utilizadas comumente em análises de variância.

Para estes casos, recomenda-se extrair a **raiz quadrada**, assim, esta nova variável (variável transformada), em geral, terá variância constante.

Este tipo de transformação é bastante eficaz pois reduz a heterocedasticidade das variâncias.

Normalidade

Em geral, apesar de um pressuposto, este é o que menos impacta a análise de variância como um todo. A menos que, a distribuição dos erros (análise dos resíduos) mostre uma curtose positiva e assimetria.

Nestes casos, as transgressões à pressuposição de normalidade irão afetar o nível de significância do teste.

Em outras palavras, o pesquisador/analista de dados pensa que o nível de significância (α) do teste é 0,05 ou 5%, mas na realidade está trabalhando com um nível de significância de 7% ou 8%.

De todo modo, o teste *F* é bastante robusto, pequenas transgressões à pressuposição de normalidade são usuais e não afetam substancialmente os resultados da análise. No entanto, a hipótese de normalidade dos erros pode ser testada:

- teste χ^2;
- teste de Kolmogorov-Smirnov;
- teste de Shapiro-Wilks.

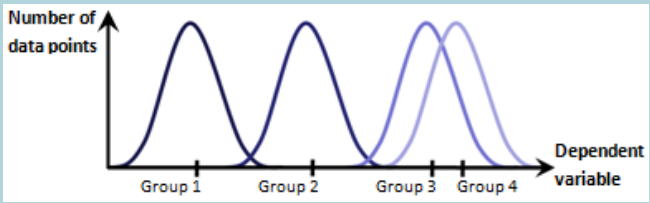
UNIVALI

ANOVA com um fator

Suponhamos que estamos interessados em uma variável Y_{ij}, e que poderíamos classificar os dados de acordo com uma característica i.

Em uma ANOVA simples, estes dados são observados por um modelo linear simples:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, ..., k \quad \text{e} \quad j = 1, ..., n$$



- H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
- H_1: pelo menos uma média é diferente.

Assim, as estimativas para as médias populacionais (parâmetros) de cada grupo são as médias amostrais (estatísticas) para cada respectivo grupo.

$$\mu_1 = \bar{y}_1 \quad \mu_2 = \bar{y}_2 \quad \mu_3 = \bar{y}_3 \quad \mu_4 = \bar{y}_4$$

UNIVALI

Tabela da ANOVA

A análise de variância consiste em preencher esta tabela, calculando cada uma da quantidades abaixo apresentadas.

Fonte de variação	G.L.	SQ	MQ	F
Entre fatores (grupos)	k-1	SQEnt	MQEnt	$\frac{MQEnt}{MQDen}$
Dentro fatoras (grupos)	n-k	SQDen	MQDen	
Total	n-1	SQT		

Passo I - Calcular a SQT

$$SQT = \sum_{i}^{k} \sum_{j}^{n} (y_{ij} - \bar{y})^2$$

Passo I - Calcular a SQT

$$SQT = \sum_{i}^{k} \sum_{j}^{n} (y_{ij} - \bar{y})^2$$

Passo II - Calcular a SQEnt

$$SQEnt = \sum n_i (\bar{y}_i - \bar{y})^2$$

Passo I - Calcular a SQT

$$SQT = \sum_{i}^{k} \sum_{j}^{n} (y_{ij} - \bar{y})^2$$

Passo II - Calcular a SQEnt

$$SQEnt = \sum n_i (\bar{y}_i - \bar{y})^2$$

Passo III - Calcular a SQDen

$$SQDen = SQT - SQEnt$$

Passo IV - Calcular os Graus de Liberdade

$$G.L._{total} = n - 1$$

$$G.L._{Entre} = k - 1$$

$$G.L._{Dentro} = n - k$$

Passo IV - Calcular os Graus de Liberdade

$$G.L._{total} = n - 1$$

$$G.L._{Entre} = k - 1$$

$$G.L._{Dentro} = n - k$$

Passo V - Calcular os QMEnt e QMDen

$$MQEnt = \frac{SQEnt}{G.L._{entre}}$$

$$MQDen = \frac{SQDen}{G.L._{dentro}}$$

Passo VI - Calcular a estatística do teste

$$F = \frac{MQEnt}{MQDen}$$

Passo VI - Calcular a estatística do teste

$$F = \frac{MQEnt}{MQDen}$$

Passo VII - Calcular o coeficiente de explicação da variação total

$$r^2 = \frac{SQEnt}{SQT}$$

Passo VI - Calcular a estatística do teste

$$F = \frac{MQEnt}{MQDen}$$

Passo VII - Calcular o coeficiente de explicação da variação total

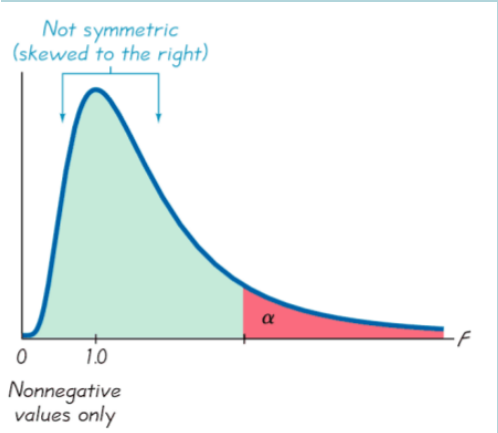
$$r^2 = \frac{SQEnt}{SQT}$$

Passo VIII - Comparar a estatística do teste com o valor crítico

Identificar na tabela da distribuição F o valor crítico para comparação e decisão sobre as hipóteses testadas.

- Nível de significância do teste (α);
- G.L.$_{entre}$
- G.L.$_{dentro}$

Decisão/Inferência



$$F = \frac{Varinciaentrefatores}{Varinciadentrodecadafator}$$

Assim, se o $F_{calculado}$ cair dentro da área α de "Rejeição", temos evidências suficientes para refutar a H_0 em favorecimento da H_1.

Caso contrário, não temos evidências suficientes e falhamos em rejeitar a H_0.