Universidade Federal do Paraná
Campus Avançado de Jandaia do Sul
Estatística

Análise de Variância (ANOVA)

Dr. Landir Saviniec

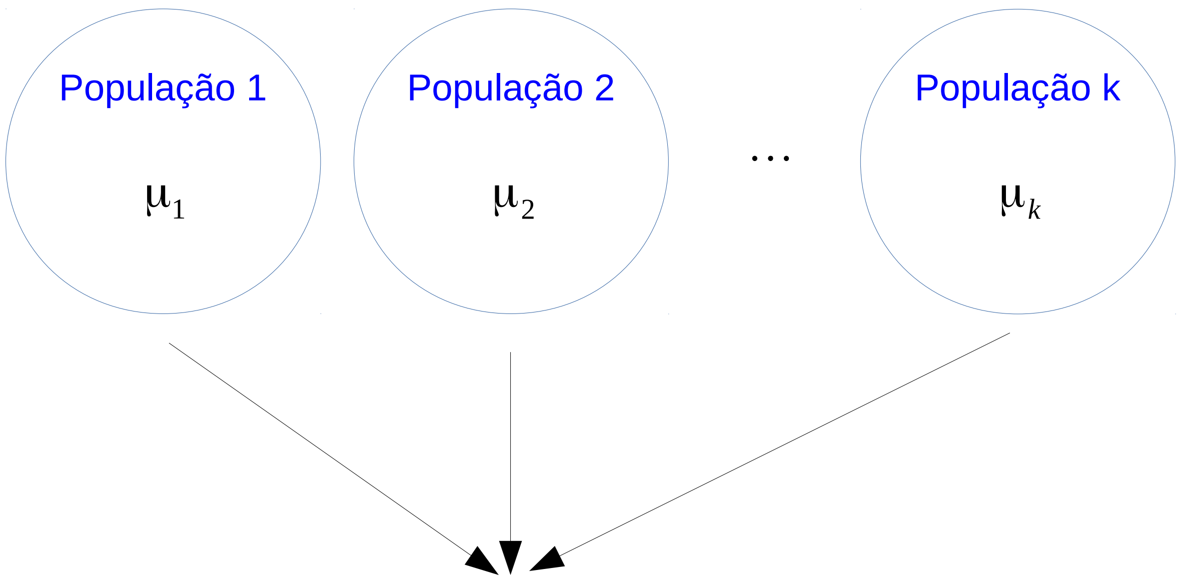
E-mail: landir.saviniec@gmail.com
Homepage: github.com/lansaviniec

Novembro de 2018

ANOVA de um fator:
Cálculo para amostras de tamanhos iguais

Análise de variância

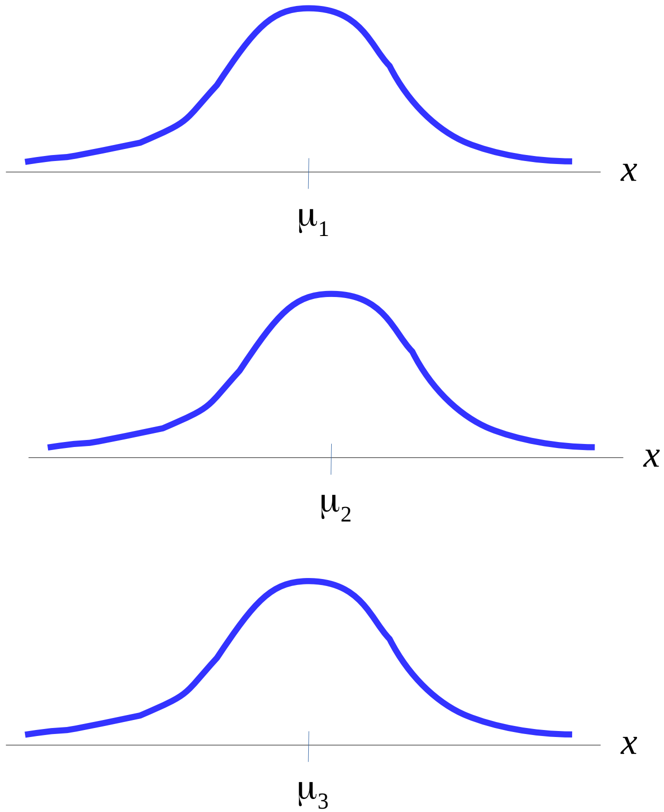
É um **método para teste de igualdade de três ou mais médias populacionais** através da **análise das variâncias amostrais**.



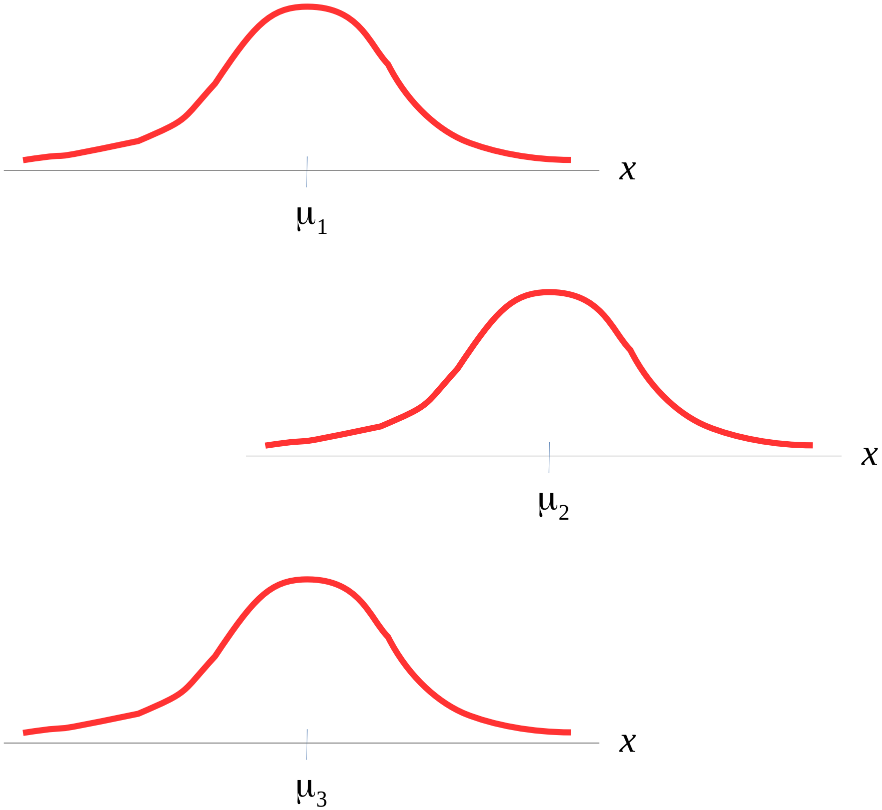
$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_k \qquad (\textit{as médias são todas iguais})$$
$$H_1: \quad \textit{Pelo menos uma das médias é diferente das demais}$$

Exemplo

Médias aproximadamente iguais



Pelo menos uma média difere das demais



Um fator

Usa-se apenas **uma característica para comparar as populações**.

Exemplo: Suponha as populações de pessoas de três países A, B e C. **Uma única característica (fator)** seria considerar apenas a **altura** das pessoas. **Duas características (fatores)** seria considerar **altura e peso** das pessoas.

Objetivo

Testar se todas as médias populacionais são iguais ou se **pelo menos uma difere das demais,** de forma significativa.

Requisitos

1) As populações devem ter distribuições que são aproximadamente normais. O método funciona bem, a não ser que as distribuições se afastem muito da normal.

2) As populações devem ter variâncias aproximadamente iguais.

3) As amostras devem ser amostras aleatórias simples e independentes.

Estatística de teste utilizada

Calculamos a estatística de teste baseada na **distribuição F**:

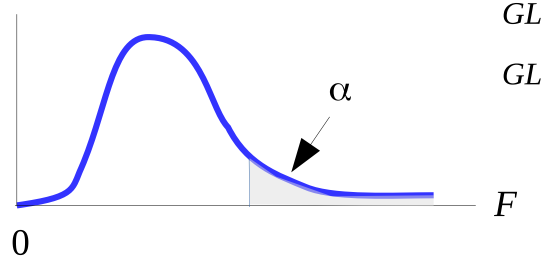
$$F = \frac{variância\ entre\ amostras}{variância\ dentro\ das\ amostras}$$

Distribuição F

Se extrairmos amostras A e B aleatórias de uma mesma população e calcularmos o escore F abaixo:

$$F = \frac{s_A^2}{s_B^2} \quad , \quad com\ s_A^2 \geq s_B^2$$

Os valores de F seguem uma distribuição assimétrica à direita:



$$GL(numerador) = n_A - 1$$

$$GL(denominador) = n_B - 1$$

Exemplo 1: com amostras de mesmo tamanho

Considere a população de dados {3, 4, 5, 6, 7, 8}:

Todas as amostras são da população dada

Amostra 1	Amostra 2	Amostra 3
7	6	4
3	5	7
6	5	6
6	8	7

$n_1=4$ $\quad n_2=4$ $\quad n_3=4$

$\bar{x}_1=5.5$ $\quad \bar{x}_2=6.0$ $\quad \bar{x}_3=6.0$

$s_1^2=3.0$ $\quad s_2^2=2.0$ $\quad s_3^2=2.0$

variância entre amostras :

$s_{\bar{x}}^2=0.0833$

$$n\, s_{\bar{x}}^2=4\left(0.0833\right)=0.3332$$

variância dentro das amostras :

$$s_p^2=\frac{3.0+2.0+2.0}{3}=2.3333$$

Estatística de teste F :

$$F=\frac{n\, s_{\bar{x}}^2}{s_p^2}=\frac{0.3332}{2.3333}=0.1428$$

Uma das amostras não é da população dada

Amostra 1	Amostra 2	Amostra 3
17	6	4
13	5	7
16	5	6
16	8	7

$n_1=4$ $\quad n_2=4$ $\quad n_3=4$

$\bar{x}_1=15.5$ $\quad \bar{x}_2=6.0$ $\quad \bar{x}_3=6.0$

$s_1^2=3.0$ $\quad s_2^2=2.0$ $\quad s_3^2=2.0$

variância entre amostras :

$s_{\bar{x}}^2=30.0833$

$$n\, s_{\bar{x}}^2=4\left(30.0833\right)=120.3332$$

variância dentro das amostras :

$$s_p^2=\frac{3.0+2.0+2.0}{3}=2.3333$$

Estatística de teste F :

$$F=\frac{n\, s_{\bar{x}}^2}{s_p^2}=\frac{120.3332}{2.3333}=51.5721$$

Exemplo 1: com amostras de mesmo tamanho

Considere a população de dados {3, 4, 5, 6, 7, 8}:

Todas as amostras são da população dada

Amostra 1	Amostra 2	Amostra 3
7	6	4
3	5	7
6	5	6
6	8	7

Uma das amostras não é da população dada

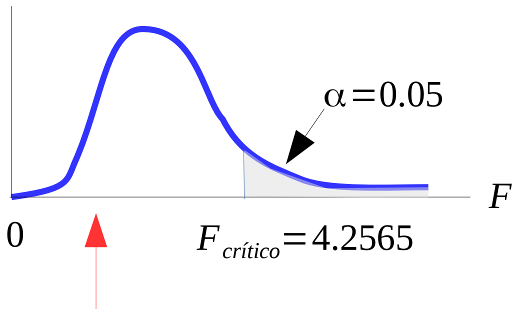
Amostra 1	Amostra 2	Amostra 3
17	6	4
13	5	7
16	5	6
16	8	7

$k = 3$ *quantidade de amostras*

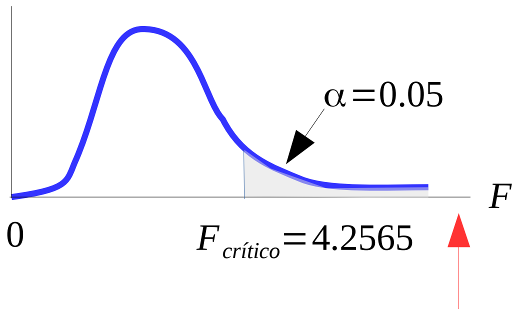
$n = 4$ *tamanho das amostras*

$GL(numerador) = k-1 = 3-1 = 2$

$GL(denominador) = k(n-1) = 3(4-1) = 9$



$F = 0.1428$



$F = 51.5721$

Aceita – se H_0. As amostras são de populações que tem as mesmas médias .

Rejeita – se H_0. Pelo menos uma amostra é de uma população com média diferente das demais .

De onde surgiu a fórmula para calcular a variância entre as amostras?

Pelo teorema do limite central, temos:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad \sigma_{\bar{x}}^2 = \left(\frac{\sigma}{\sqrt{n}}\right)^2 \quad \Rightarrow \quad \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \quad \Rightarrow \quad \sigma^2 = n\,\sigma_{\bar{x}}^2$$

Logo, a **variância entre amostras** ou variâncias das médias amostrais é dada por:

$$variância\ entre\ amostras = n\, s_{\bar{x}}^2$$

Onde:

$$A\ variância\ populacional\ \sigma_{\bar{x}}^2\ e\ estimada\ pela\ variância\ das\ médias\ amostrais\ s_{\bar{x}}^2$$

Exemplo 2

As amostras a seguir apresentam valores de compras de clientes de três lojas A, B e C de uma mesma rede de lojas.

A: 350, 380, 450, 500, 555, 540, 590, 602, 620, 630
B: 520, 405, 540, 580, 610, 650, 555, 590, 645, 615
C: 800, 750, 770, 830, 850, 860, 790, 775, 815, 845

Aplique a ANOVA com nível de significância de 5% para verificar se o **gasto médio por compra é o mesmo nas três lojas**.

$$\alpha = 0.05 \qquad n = 10$$

$$\bar{x}_A = 521.7 \qquad s_A^2 = 9902.23$$

$$\bar{x}_B = 571.0 \qquad s_B^2 = 5221.11$$

$$\bar{x}_C = 808.5 \qquad s_C^2 = 1405.83$$

Passo 1: Formulamos as hipóteses nula e alternativa:

$$H_0: \quad \mu_A = \mu_B = \mu_C \qquad \left(\textit{as médias são todas iguais}\right)$$

$$H_1: \quad \textit{Pelo menos uma das médias é diferente das demais}$$

Exemplo 2

Passo 2: Calculamos a estatística de teste para os dados amostrais:

variância entre amostras :

$$s_{\bar{x}}^2 = 23515.16$$

$$n\, s_{\bar{x}}^2 = 10\left(23515.16\right) = 235151.6$$

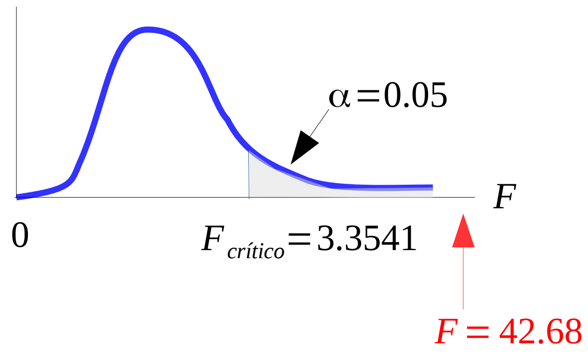
variância dentro das amostras :

$$s_p^2 = \frac{9902.23 + 5221.11 + 1405.83}{3} = 5509.73$$

Estatística de teste F :

$$F = \frac{n\, s_{\bar{x}}^2}{s_p^2} = \frac{235151.6}{5509.73} = 42.68$$

Passo 3: Analisamos os valores críticos na distribuição F:



$k = 3$ *quantidade de amostras*

$n = 10$ *tamanho das amostras*

$GL\left(numerador\right) = k-1 = 3-1 = 2$

$GL\left(denominador\right) = k\left(n-1\right) = 3\left(10-1\right) = 27$

Exemplo 2

Passo 4: Decidimos se aceitamos ou rejeitamos a hipótese nula:

Como a estatística de teste caiu na região de rejeição, rejeitamos a hipótese nula. Logo, há evidências de que em pelo menos uma das lojas o gasto médio por compra é diferente das demais.

ANOVA de um fator:
Cálculo para amostras de tamanhos diferentes

Estatística de teste utilizada

Calculamos a estatística de teste baseada na **distribuição F**:

$$F = \frac{\text{variância entre amostras}}{\text{variância dentro das amostras}} = \frac{\left[\dfrac{\sum n_i (\bar{x}_i - \bar{\bar{x}})^2}{k-1}\right]}{\left[\dfrac{\sum (n_i - 1)\, s_i^2}{\sum (n_i - 1)}\right]}$$

$$GL(\text{numerador}) = k - 1$$
$$GL(\text{denominador}) = \sum (n_i - 1)$$

Obs: neste caso os tamanhos da amostras são levados em consideração, para que amostras maiores tenham mais peso.

Onde:

$\bar{\bar{x}}$: *média de todas os valores amostrais combinados*
k : *número de médias populacionais sendo comparadas*
n_i : *tamanho da i – ésima amostra*
$\bar{x}_i$: *média da i – ésima amostra*
s_i^2 : *variância da i – ésima amostra*

Exemplo 3

Um carro de marca X e mesmo ano de fabricação foi cotado, a preço de mercado (em R$), em três capitais diferentes, obtendo-se as seguintes amostras de preços:

Capital A: 30000, 32100, 29100, 28900, 30500, 33000, 33800
Capital B: 34200, 33100, 32000, 33000, 33800, 34500, 35000, 32800
Capital C: 35200, 34800, 34700, 33900, 35500, 35500

Aplique a ANOVA com nível de significância de 5% para verificar se o preço de mercado deste veículo é o mesmo nas três capitais.

$$k=3 \qquad \alpha=0.05$$
$$\bar{\bar{x}}=33114.29$$
$$n_A=7 \qquad \bar{x}_A=31057.14 \qquad s_A^2=3716190.48$$
$$n_B=8 \qquad \bar{x}_B=33550.00 \qquad s_B^2=994285.71$$
$$n_C=6 \qquad \bar{x}_C=34933.33 \qquad s_C^2=370666.67$$
$$\sum n_i\left(\bar{x}_i-\bar{\bar{x}}\right)^2=29622857.14+1518775.51+19853605.44=50995238.1$$
$$\sum\left(n_i-1\right)s_i^2=22297142.86+6960000.00+1853333.33=31110476.19$$
$$\sum\left(n_i-1\right)=6+7+5=18$$

Exemplo 3

Passo 1: Formulamos as hipóteses nula e alternativa:

$$H_0: \quad \mu_A = \mu_B = \mu_C \qquad \left(\textit{as médias são todas iguais}\right)$$
$$H_1: \quad \textit{Pelo menos uma das médias é diferente das demais}$$

Passo 2: Calculamos a estatística de teste para os dados amostrais:

variância entre amostras :

$$\frac{\sum n_i\left(\bar{x}_i - \bar{\bar{x}}\right)^2}{k-1} = \frac{50995238.1}{3-1} = 25497619.05$$

variância dentro das amostras :

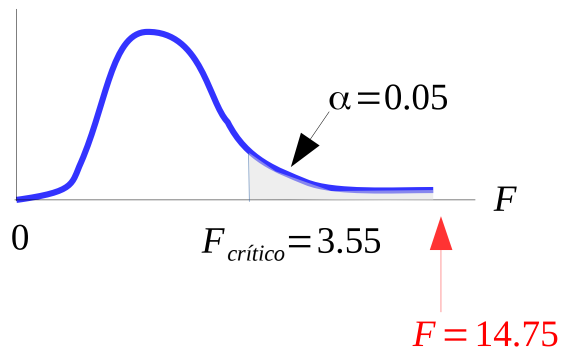
$$\frac{\sum\left(n_i - 1\right)s_i^2}{\sum\left(n_i - 1\right)} = \frac{31110476.19}{18} = 1728359.79$$

Estatística de teste F :

$$F = \frac{\textit{variância entre amostras}}{\textit{variância dentro das amostras}} = \frac{25497619.05}{1728359.79} = 14.75$$

Exemplo 3

Passo 3: Analisamos os valores críticos na distribuição F:



$$GL(numerador) = k - 1 = 3 - 1 = 2$$
$$GL(denominador) = \sum (n_i - 1) = 18$$

Passo 4: Decidimos se aceitamos ou rejeitamos a hipótese nula:

Como a estatística de teste caiu na região de rejeição, rejeitamos a hipótese nula. Logo, há evidências de que em pelo menos uma das três capitais o preço de mercado do veículo X é diferente das demais.

ANOVA de um fator:
Encontrando pares de médias que diferem significativamente

Método de SCHEFFÉ

A ANOVA testa a existência ou não de diferenças significativas entre k>2 médias populacionais. Mas, caso haja diferença, ela não identifica quais médias diferem das demais. Existem vários métodos para isso. Vamos usar o método de SCHEFFÉ.

Scheffé demonstrou que se duas médias μ_i *e* μ_j *diferem significativamente, então:*

$$\left|\bar{x}_i - \bar{x}_j\right| > \Delta\,\alpha$$

onde:

$$\Delta\,\alpha = \sqrt{QMR\left(\frac{1}{n_i}+\frac{1}{n_j}\right)(k-1)\,F_{k-1,N-k}(\alpha)}$$

e:

$$QMR = \frac{\sum (n_i - 1)\, s_i^2}{\sum (n_i - 1)}, \qquad N = \sum n_i$$

Exemplo 4

Determinando os pares de médias que diferem no exemplo 3:

$$N = \sum n_i = 7 + 8 + 6 = 21$$

$$QMR = 1728359.79$$

$$F_{k-1,\,N-k}(\alpha) = F_{2,\,18}(0.05) = 3.55$$

$$\Delta\,\alpha = \sqrt{QMR\left(\frac{1}{n_i} + \frac{1}{n_j}\right)(k-1)\,F_{k-1,\,N-k}(\alpha)}$$

Capitais	$\Delta\,\alpha$	$\lvert \bar{x}_i - \bar{x}_j \rvert$	Conclusão
A e B	1813.00	2492.86	Diferem
A e C	1948.92	3876.19	Diferem
B e C	1891.86	1383.33	Não diferem