AULA 12
Análise de Variância

Ernesto F. L. Amaral

26 de setembro de 2012

Faculdade de Filosofia e Ciências Humanas (FAFICH)
Universidade Federal de Minas Gerais (UFMG)

Fonte:

Triola, Mario F. 2008. "Introdução à estatística". 10 ª ed. Rio de Janeiro: LTC. Capítulo 12 (pp.506-537).

ESQUEMA DA AULA

– ANOVA de um fator.

– ANOVA de dois fatores.

VISÃO GERAL

– Vimos procedimentos para o teste de hipótese de duas médias populacionais serem iguais (capítulo 9).

– Porém, tais testes não se aplicam quando há três ou mais médias envolvidas.

– A **análise de variância (ANOVA)** é um método para se testar a igualdade de três ou mais médias populacionais através da análise das variâncias amostrais.

– Em vez de considerarmos apenas médias amostrais, consideramos quantidades de variação, tamanhos amostrais e natureza da distribuição das médias amostrais.

POR QUE NOVO MÉTODO?

– Por que não podemos testar duas médias amostrais de cada vez?

– **Por que precisamos de novo procedimento**, quando podemos testar igualdade de duas médias (capítulo 9)?

– À medida que aumentamos o **número de testes** de significância individuais, aumentamos o risco de encontrar diferenças por puro acaso (nível de confiança baixo), em vez de diferença real nas médias.

– **Risco de erro tipo I** (encontrar diferença em um dos pares quando tal diferença não existe) é muito alto.

– A análise de variância **evita rejeitar hipótese nula verdadeira**, com uso de teste de igualdade de várias médias.

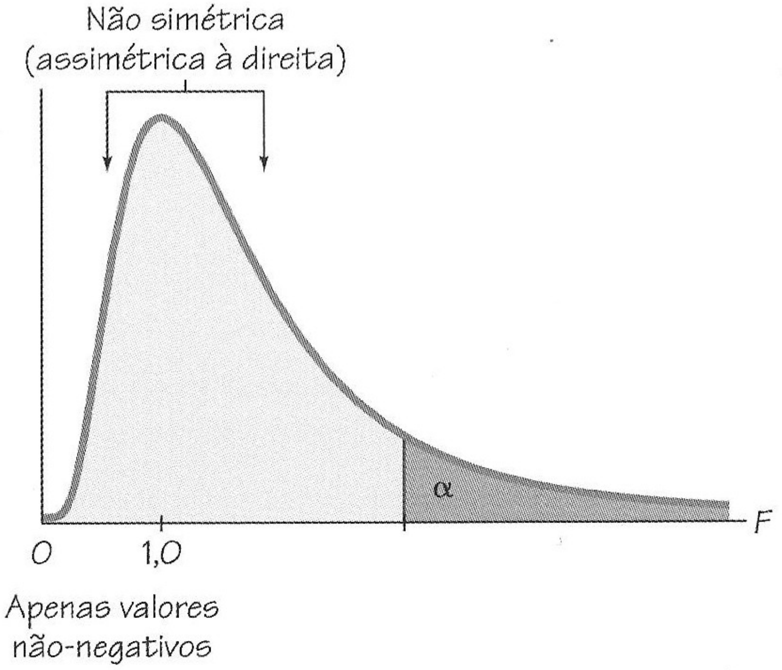
DISTRIBUIÇÃO *F*

– Os métodos de ANOVA requerem a distribuição *F*:

  – Assimétrica à direita.

  – Valores de *F* podem ser 0 ou positivos, mas não podem ser negativos.

  – Há uma distribuição *F* diferente para cada par de graus de liberdade para numerador e denominador.



COMPARAÇÃO DE VARIÂNCIAS

– A **análise de variância** se baseia na comparação de duas estimativas diferentes da variância comum de duas populações diferentes:

  – Variância entre amostras.

  – Variância dentro das amostras.

– O termo de **um fator** é usado porque os dados amostrais são separados em grupos por uma característica (fator).

– A análise de variância de **dois fatores** permite comparar populações separadas em categorias usando duas características (fatores).

– Se o **valor *P*** for pequeno (menor que 0,05), rejeite igualdade das médias. Caso contrário, deixe de rejeitar a igualdade das médias.

ANOVA DE UM FATOR

ANOVA DE UM FATOR

– O método da análise de variância de um fator é usado para testes de hipóteses de que três ou mais médias populacionais são iguais (H_0: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$).

– Estratégia de estudo:

  – Pequeno valor P ($\leq$0,05) leva à rejeição da hipótese nula de médias iguais. Grande valor P deixa de rejeitar H_0.

  – Entenda a natureza dos valores SQ (soma dos quadrados) e dos MQ (média quadrática), além de seus papéis no cálculo de teste F.

– A análise de variância de fator único usa uma propriedade para categorizar as populações.

– Essa propriedade (característica, tratamento, fator) permite distinguir diferentes populações umas das outras.

REQUISITOS

– Populações têm **distribuições** que são aproximadamente **normais** (método funciona bem se população não tem distribuição muito afastada da normal).

– Populações têm a **mesma variância $\boldsymbol{\sigma}^2$** ou desvio padrão σ (método é eficiente se variâncias não diferirem por grandes quantidades).

– **Amostras aleatórias simples**.

– **Amostras independentes** umas das outras (não são emparelhadas).

– Diferentes amostras são de populações que são categorizadas de apenas uma maneira (**um fator**).

PROCEDIMENTOS

– Procedimentos para teste de H_0: $\mu_1 = \mu_2 = \mu_3 = ...$
  – Use programa estatístico para obter resultados.
  – Identifique o valor *P*.
  – Forme conclusão com base nestes critérios:
    – Se valor $P \leq \alpha$, rejeite hipótese nula de médias iguais e conclua que pelo menos uma das médias populacionais é diferente das demais.
    – Se valor $P > \alpha$, deixe de rejeitar hipótese nula de médias iguais.

– Ao concluirmos que há evidência para rejeitar afirmativa de médias populacionais iguais, não dizemos que qualquer média particular seja diferente das demais.

EXEMPLO

– Testar hipótese nula de que médias populacionais do índice tradicional-secular (tradrat5) são iguais para todas categorias de educação (x025r).

education level (recoded)	Summary of traditional/secular rational values Mean
lower	.05828157
middle	.23798683
upper	.49092825
Total	.23695413

```
oneway tradrat5 x025r
```

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	1892.90935	2	946.454674	1220.50	0.0000
Within groups	58511.7713	75454	.775462816		
Total	60404.6807	75456	.800528529		

Bartlett's test for equal variances: chi2(2) = 912.3005 Prob>chi2 = 0.000

– Valor *P*<0,05: há evidência suficiente para apoiar afirmativa de que as três médias populacionais não são todas iguais.

FUNDAMENTOS

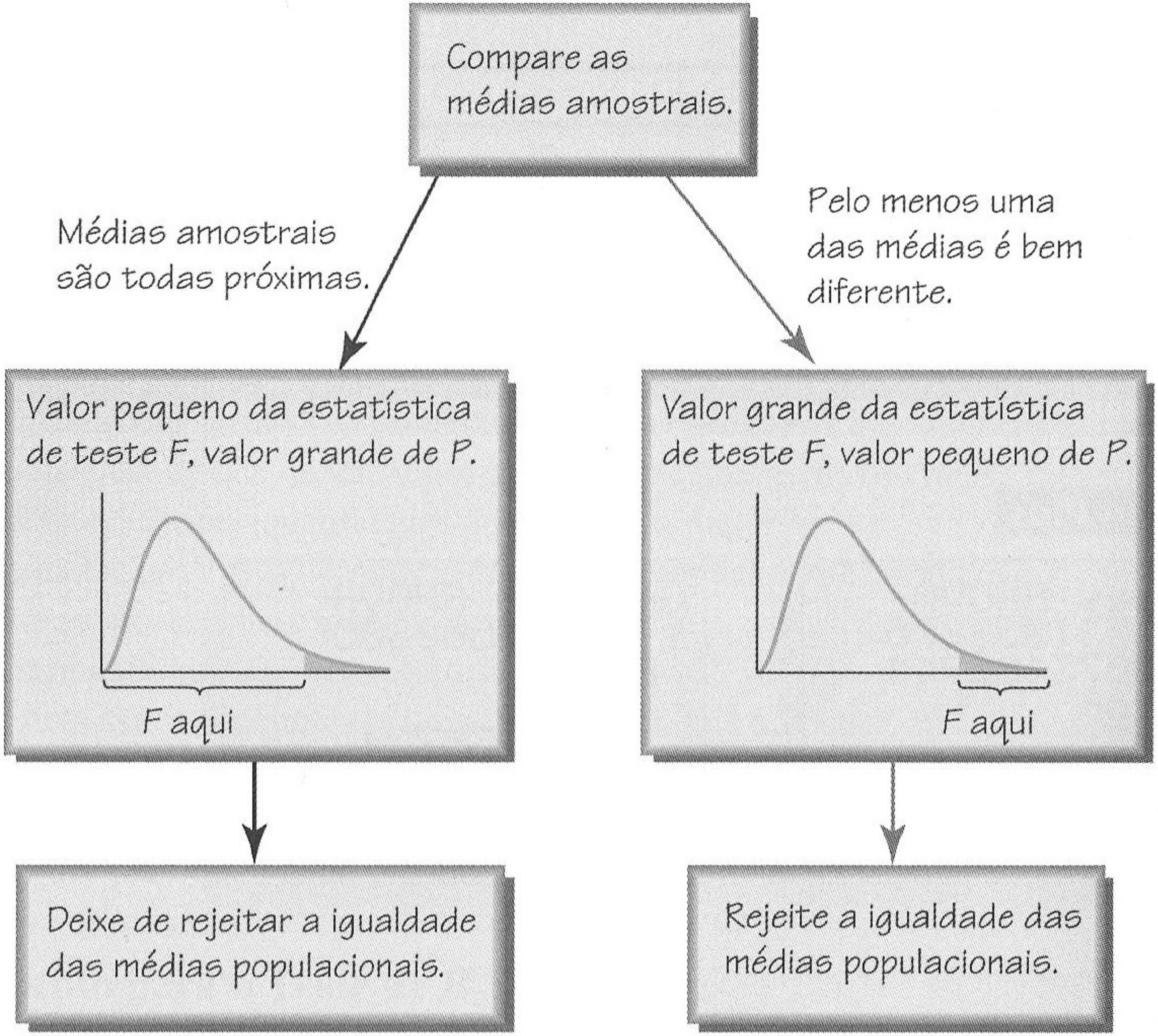
– Com a suposição de que as populações tenham a mesma variância, a estatística de teste *F* é a razão de duas estimativas de σ^2:

– **Variação entre amostras** (com base na variação entre médias amostrais).

– **Variação dentro das amostras** (com base nas variâncias amostrais).

– Estatística de teste *F* significativamente grande é evidência contra médias populacionais iguais.

RELAÇÃO ENTRE ESTATÍSTICA *F* E VALOR *P*



ESTATÍSTICA DE TESTE PARA ANOVA DE UM FATOR

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

– **Numerador** da estatística de teste *F* mede variação entre médias amostrais.

– Estimativa da variância no **denominador** depende apenas das variâncias amostrais e não é afetada pelas diferenças entre as médias amostrais.

– Médias próximas (variância pequena no numerador) causam teste *F* pequeno (não rejeitamos H_0).

– Se valor de *F* for grande, rejeitamos H_0 de médias iguais.

TAMANHOS AMOSTRAIS IGUAIS A *n*

- **Primeiro:**
  - Calcule a variância **entre** amostras: $ns_{\bar{x}}^2$
  - Variância das médias amostrais: $s_{\bar{x}}^2$
  - Tamanho de cada uma das amostras: *n*
  - Ou seja, as médias amostrais são consideradas como um conjunto de valores para calcular sua variância.
- **Segundo:**
  - Calcule a variância **dentro** das amostras (variância combinada obtida pelo cálculo da média das variâncias amostrais): $s_p^2 = (s_1^2 + s_2^2 + s_n^2)/n$
- **Terceiro:**
  - Calcule estatística de teste *F*: $ns_{\bar{x}}^2/s_p^2$
  - Graus de liberdade do numerador: $k - 1$
  - Graus de liberdade do denominador: $k(n - 1)$
  - Sendo *k* (nº de amostras) e *n* (tamanho amostral)

EFEITO DE UMA MÉDIA SOBRE A ESTATÍSTICA *F*

A **somar 10** B

A

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

$n_1 = 4$, $\bar{x}_1 = 5{,}5$, $s^2_1 = 3{,}0$

$n_2 = 4$, $\bar{x}_2 = 6{,}0$, $s^2_2 = 2{,}0$

$n_3 = 4$, $\bar{x}_3 = 6{,}0$, $s^2_3 = 2{,}0$

Variância entre amostras: $ns^2_{\bar{x}} = 4\ (0{,}0833) = 0{,}3332$

Variância dentro das amostras: $s^2_p = \dfrac{3{,}0 + 2{,}0 + 2{,}0}{3} = 2{,}3333$

Estatística de teste *F*: $F = \dfrac{ns^2_{\bar{x}}}{s^2_p} = \dfrac{0{,}3332}{2{,}3333} = \mathbf{0{,}1428}$

Valor *P* (encontrado pelo Excel): Valor $P = 0{,}8688$

B

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

$n_1 = 4$, $\bar{x}_1 = \mathbf{15{,}5}$, $s^2_1 = 3{,}0$

$n_2 = 4$, $\bar{x}_2 = 6{,}0$, $s^2_2 = 2{,}0$

$n_3 = 4$, $\bar{x}_3 = 6{,}0$, $s^2_3 = 2{,}0$

$ns^2_{\bar{x}} = 4\ (30{,}0833) = 120{,}3332$

$s^2_p = \dfrac{3{,}0 + 2{,}0 + 2{,}0}{3} = 2{,}3333$

$F = \dfrac{ns^2_{\bar{x}}}{s^2_p} = \dfrac{120{,}3332}{2{,}3333} = \mathbf{51{,}5721}$

Valor $P = 0{,}0000118$

TAMANHOS AMOSTRAIS DIFERENTES

– Os cálculos se tornam complicados quando os tamanhos amostrais não são os mesmos.

– Também é calculada a estatística *F* que é a razão de duas estimativas diferentes da variância populacional comum (σ^2) e envolvem medidas ponderadas pelos tamanhos amostrais:

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

k = número de médias populacionais sendo comparadas

n_i = número de valores na *i*-ésima amostra

s_i^2 = variância dos valores na *i*-ésima amostra

$\bar{\bar{x}}$ = média de todos valores amostrais combinados

$\bar{x}_i$ = média dos valores na *i*-ésima amostra

COMPONENTES DO MÉTODO DE ANOVA

– **SQ(total)**, ou soma dos quadrados total, é uma medida da variação total em todos dados amostrais combinados:

$$SQ(total) = \sum (x - \bar{\bar{x}})^2$$

– **SQ(tratamento)**, ou SQ(fator) ou SQ(entre grupos) ou SQ(entre amostras), é uma medida da variância entre médias amostrais:

$$SQ(tratamento) = \sum n_i(\bar{x}_i - \bar{\bar{x}})^2$$

– **SQ(erro)**, ou SQ(dentro dos grupos) ou SQ(dentro das amostras), é uma soma de quadrados que representa a variação que se supõe comum a todas populações:

$$SQ(erro) = \sum (n_i - 1)s_i^2$$

COMPONENTES NO MÉTODO DE ANOVA (cont.)

SQ(total) = SQ(tratamento) + SQ(erro)

– Sendo *N*, o número total de valores em todas amostras combinadas, temos:

– **MQ(tratamento)** é uma média quadrática para tratamento:

$$MQ(tratamento) = \frac{SQ(tratamento)}{k-1}$$

– **MQ(erro)** é uma média quadrática para o erro:

$$MQ(erro) = \frac{SQ(erro)}{N-k}$$

– **MQ(total)** é uma média quadrática para a variação total:

$$MQ(total) = \frac{SQ(total)}{N-1}$$

ESTATÍSTICA DE TESTE

– Considerando a hipótese nula como:

$$H_0: \mu_1 = \mu_2 = ... = \mu_k$$

– A estatística de teste para ANOVA com tamanhos amostrais desiguais é dada por:

$$F = \frac{MQ(tratamento)}{MQ(erro)}$$

– Possui distribuição F, com graus de liberdade dados por:

  – Graus de liberdade do numerador = $k - 1$

  – Graus de liberdade do denominador = $N - k$

– **Numerador** é afetado pelas diferenças **entre** médias amostrais.

– **Denominador** depende das variâncias amostrais que medem variação **dentro** dos tratamentos.

IDENTIFICANDO MÉDIAS QUE SÃO DIFERENTES

– Testamos se médias populacionais são diferentes, mas não sabemos se uma média particular é diferente das demais.

– Há **procedimentos informais** para identificar as médias específicas que são diferentes:

  – Construir **diagramas de caixa** com mesma escala.

  – Estimar **intervalos de confiança** e compará-los.

– **Procedimentos formais**:

  – **Testes de amplitude**: identificar subconjuntos de médias que não são diferentes umas das outras.

  – **Testes de comparações múltiplas**: usam pares de médias, mas ajustam o problema de ter nível de confiança que diminui à medida que aumenta número de testes individuais.

TESTE DE COMPARAÇÃO MÚLTIPLA DE BONFERRONI

– Não há consenso sobre qual teste é o melhor.

– O Teste de Bonferroni mostra que as médias do índice tradicional-secular são todas diferentes entre si.

```
. oneway tradrat5 x025r, bonferroni

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      1892.90935      2   946.454674   1220.50     0.0000
 Within groups      58511.7713  75454   .775462816
------------------------------------------------------------------------
    Total           60404.6807  75456   .800528529

Bartlett's test for equal variances:  chi2(2) = 912.3005  Prob>chi2 = 0.000

                Comparison of traditional/secular rational values
                           by education level (recoded)
                                  (Bonferroni)
Row Mean-|
Col Mean |      lower     middle
---------+----------------------
  middle |    .179705
         |      0.000
         |
   upper |    .432647    .252941
         |      0.000      0.000
```

ANOVA DE DOIS FATORES

ANOVA DE DOIS FATORES

– O método da **análise da variância de dois fatores** é usado com dados divididos em categorias de acordo com dois fatores.

– **Primeiro**, testamos em relação a uma interação entre os dois fatores.

– **Depois**, testamos para determinar: (1) se o fator linha tem algum efeito; e (2) se o fator coluna tem algum efeito.

– O **ponto central** é que há uma interação entre dois fatores se o efeito de um dos fatores muda para diferentes categorias do outro fator.

REQUISITOS

– Para cada célula, os valores amostrais provêm de uma população com distribuição que é aproximadamente normal.

– Populações têm mesma variância σ^2 (ou desvio padrão σ).

– Amostras aleatórias simples.

– Amostras são independentes umas das outras.

– Valores amostrais são categorizados de duas maneiras.

– Todas células têm mesmo número de valores amostrais (planejamento balanceado).

PROCEDIMENTOS

– **Efeito da interação:** comece testando a hipótese nula de que não há qualquer interação entre os dois fatores:

$$F = \text{MQ(interação)} / \text{MQ(erro)}$$

– Se $P > 0{,}05$, não há evidência de que média da variável de interesse seja afetada por interação entre os dois fatores.

– **Efeitos de linha/coluna:**

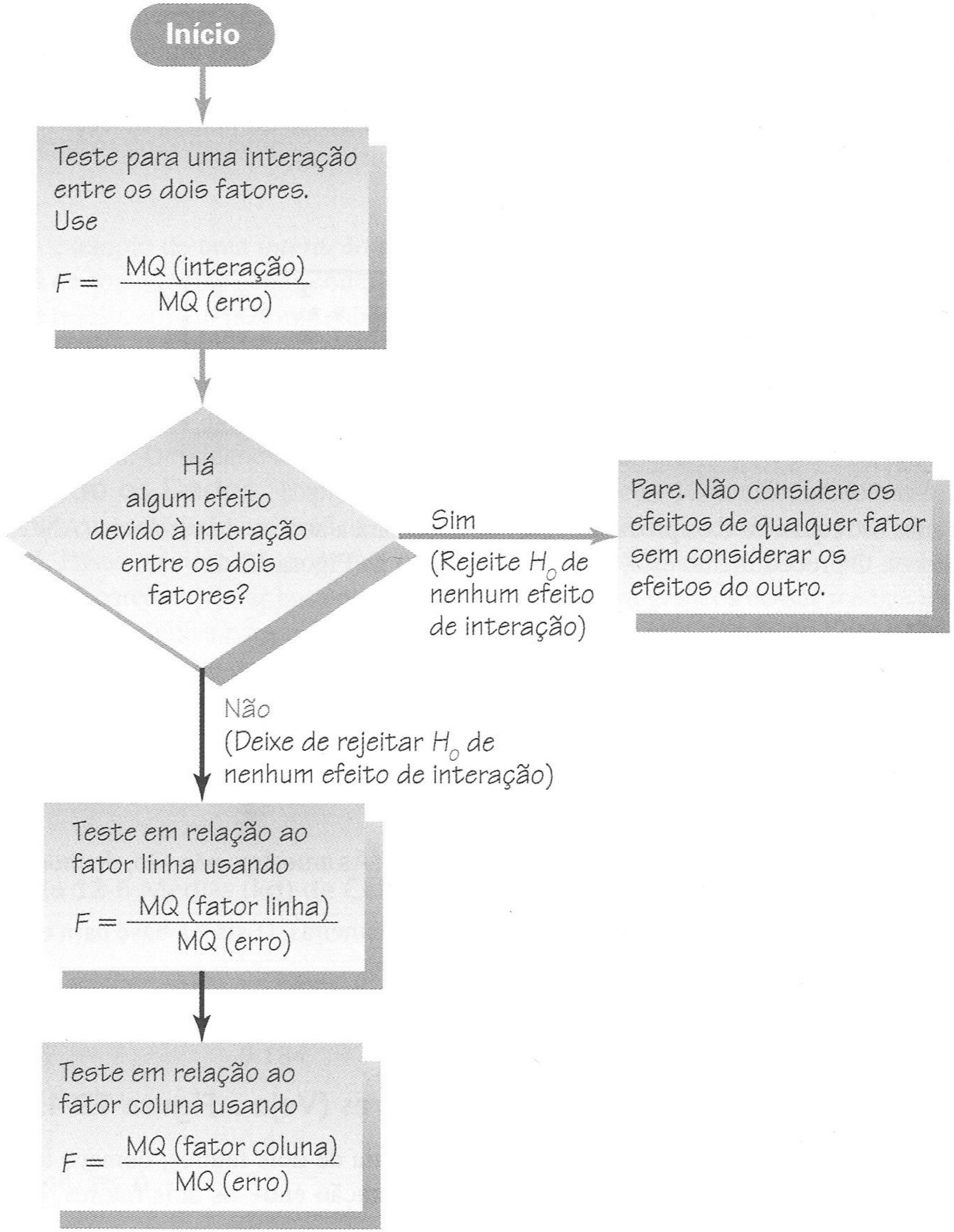
– Se rejeitamos hipótese nula de nenhuma interação entre fatores, não devemos prosseguir com os testes adicionais.

– Se deixamos de rejeitar a hipótese nula de nenhuma interação, devemos testar:

– H_0: não há qualquer efeito do fator linha.

– H_0: não há qualquer efeito do fator coluna.

DIAGRAMA DE PROCEDIMENTOS



EXEMPLO

– A tabela abaixo mostra as médias do índice tradicional-secular por categorias de educação e sexo:

```
. tab x001 x025r, sum(tradrat5) mean
```

Means of traditional/secular rational values

sex	education level (recoded) lower	middle	upper	Total
male	.0715126	.21245551	.44831954	.22501274
female	.04687808	.26288998	.53839587	.24906808
Total	.05870758	.23789516	.49090413	.23709384

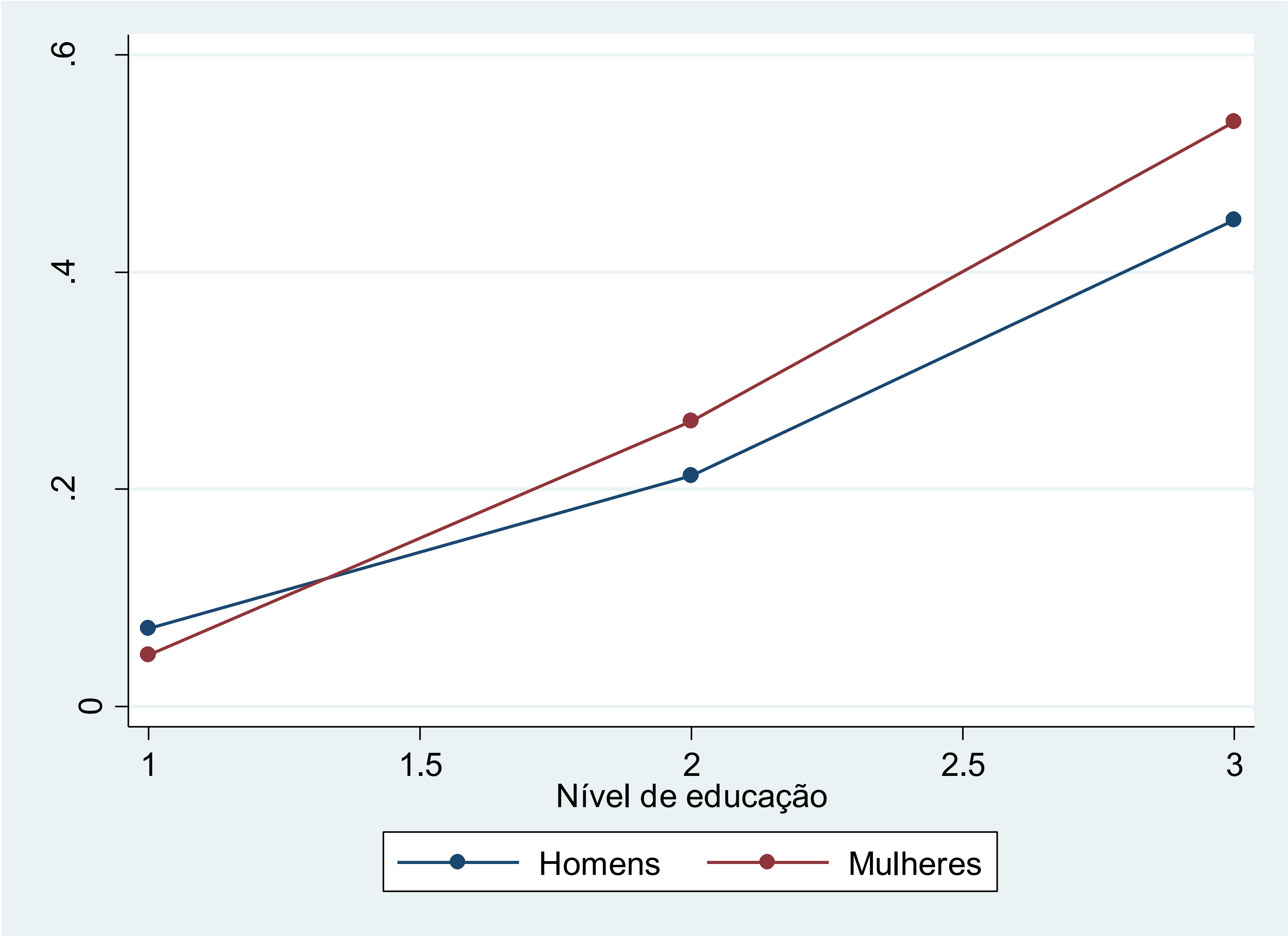
– Dados não são balanceados:

```
. tab x001 x025r
```

sex	education level (recoded) lower	middle	upper	Total
male	11,799	16,688	9,049	37,536
female	12,772	16,985	8,114	37,871
Total	24,571	33,673	17,163	75,407

GRÁFICO DO EXEMPLO

– Índice tradicional-secular por educação e sexo:



INTERPRETANDO ANOVA DE DOIS FATORES

– Resultado sugere que o efeito interação é significativo (probabilidade de rejeitar hipótese nula é pequena).

– As médias do índice tradicional-secular são afetadas por uma interação entre nível educacional e sexo.

```
                      Number of obs =   75407     R-squared     =  0.0323
                      Root MSE      = .880252     Adj R-squared =  0.0322

        Source |  Partial SS    df       MS           F     Prob > F
    -----------+----------------------------------------------------
         Model |  1947.39018     5  389.478036     502.65     0.0000
               |
         x025r |  1900.15933     2  950.079666    1226.16     0.0000
          x001 |  26.0420484     1  26.0420484      33.61     0.0000
    x025r*x001 |  36.7932638     2  18.3966319      23.74     0.0000
               |
      Residual |   58424.004 75401  .774843888
    -----------+----------------------------------------------------
         Total |  60371.3942 75406  .800617911
```