

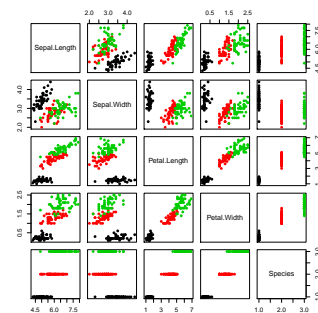
Análise de Variância (ANOVA)

A Regressão Linear visa modelar uma variável resposta numérica (quantitativa), à custa de uma ou mais variáveis preditoras, **igualmente numéricas**.

Mas uma variável resposta numérica pode depender de uma ou mais variáveis **qualitativas (categóricas)**, ou seja, de um ou mais **factores**.

Em tais situações pode ser útil uma **Análise de Variância (ANOVA)**, metodologia estatística desenvolvida nos anos 30 na Estação Experimental Agrícola de Rothamstead (Reino Unido), por **R.A. Fisher**.

Exemplo: os lírios por espécie



As medições das pétalas variam muito entre as espécies dos lírios. As medições das sépalas nem por isso.

A ANOVA como caso particular do Modelo Linear

Embora a Análise de Variância tenha surgido como método autónomo, quer a Análise de Variância, quer a Regressão Linear, são particularizações do **Modelo Linear**.

Introduzir a ANOVA através das suas semelhanças com a Regressão Linear permite aproveitar boa parte da teoria estudada até aqui.

Terminologia:

Variável resposta Y : uma variável **numérica** (quantitativa), que se pretende estudar e modelar.

Factor: uma variável preditora **categórica** (qualitativa);

Níveis do factor: “valores” (distintas categorias) do factor, ou seja, diferentes situações experimentais onde se farão observações de Y .

A ANOVA a um Factor

Começamos por analisar o mais simples de todos os modelos ANOVA: a **ANOVA a um Factor** (totalmente casualizado).

Consideramos que a **variável resposta (numérica) Y** depende de um **único factor**, com k níveis. Efectuamos observações de Y nas k diferentes situações experimentais.

Admite-se que os valores de Y poderão variar por corresponderem a níveis diferentes do factor, ou ainda devido a flutuação aleatória.

As n observações

Para estudar os efeitos dum factor, com k níveis, sobre uma variável resposta Y , admitimos que temos n observações independentes de Y , sendo n_i ($i = 1, \dots, k$) correspondentes ao nível i do factor. Logo,

$$n_1 + n_2 + \dots + n_k = n.$$

Embora fosse possível continuar a indexar as n observações de Y com um único índice, variando de 1 a n (como se fez na Regressão), é preferível utilizar **dois índices para indexar as observações de Y** :

- um para indicar o **nível do factor a que a observação corresponde**;
- outro para **distinguir cada observação dentro de um dado nível**.

As n observações (cont.)

Em geral, Y_{ij} indica a j -ésima observação no i -ésimo nível do factor, com $i = 1, \dots, k$ e $j = 1, \dots, n_i$.

No caso de **igual número de observações em cada nível**,

$$n_1 = n_2 = n_3 = \dots = n_k \quad (= n_c),$$

diz-se que estamos perante um **delineamento equilibrado**.

Os delineamentos equilibrados são **aconselháveis**, por várias razões.

A modelação de Y

A natureza mais pobre da nossa variável preditora estará associada a um modelo mais simples do que na regressão.

Em geral, admitimos que o valor esperado (médio) de Y pode diferir em cada uma das k situações (níveis do factor) em que é observado.

Uma primeira formulação do modelo pode assim ser dada pela equação de base:

$$E[Y_{ij}] = \mu_i .$$

A modelação de Y (cont.)

Para poder enquadrar a ANOVA na teoria do Modelo Linear já estudada, é conveniente re-escrever as médias de nível na forma:

$$E[Y_{ij}] = \mu_i = \mu + \alpha_i .$$

O parâmetro μ é comum a todas as observações, enquanto os parâmetros α_i são específicos para cada nível (i) do factor. Cada α_i é designado o efeito do nível i.

Admite-se que Y_{ij} oscila aleatoriamente em torno do seu valor médio:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} ,$$

com $E[\varepsilon_{ij}] = 0$.

O modelo ANOVA como um Modelo Linear

A equação de base do modelo ANOVA a um factor pode ser escrito na forma vectorial/matricial, como no modelo de regressão linear. Seja

\mathbf{Y} o vector n-dimensional com a totalidade das observações da variável resposta. Admite-se que as n_1 primeiras correspondem ao nível 1 do factor, as n_2 seguintes ao nível 2, e assim de seguida.

$\mathbf{1}_n$ o vector de n uns, já considerado na regressão.

\mathcal{I}_i a variável indicatriz de pertença ao nível i do factor. Para cada observação, esta variável toma o valor 1 se a observação corresponde ao nível i do factor, e o valor 0 caso contrário ($i = 1, \dots, k$).

$\boldsymbol{\varepsilon}$ o vector dos n erros aleatórios.

As variáveis indicatrizes

Por exemplo, se se fizerem $n = 9$ observações, com $n_1 = 3$ observações no primeiro nível do factor, $n_2 = 4$ no segundo nível e $n_3 = 2$ observações no terceiro nível, os vectores \mathcal{I}_2 e \mathcal{I}_3 serão:

$$\mathcal{I}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} , \quad \mathcal{I}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

A relação de base em notação vectorial

Em notação vectorial, a equação de base que descreve as n observações de Y pode escrever-se como no Modelo Linear:

$$\mathbf{Y} = \mu \cdot \mathbf{1}_n + \alpha_1 \cdot \mathcal{I}_1 + \alpha_2 \cdot \mathcal{I}_2 + \alpha_3 \cdot \mathcal{I}_3 + \boldsymbol{\varepsilon} .$$

No exemplo com as $n_1 = 3$, $n_2 = 4$ e $n_3 = 2$ observações:

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

$$\Leftrightarrow \mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

O problema do excesso de parâmetros

Existe um problema "técnico": as colunas da matriz X são linearmente dependentes, pelo que a matriz $\mathbf{X}'\mathbf{X}$ não é invertível.

Existe um excesso de parâmetros no modelo. Soluções possíveis:

1. retirar o parâmetro μ do modelo.
 - ▶ corresponde a retirar a coluna de uns da matriz X;
 - ▶ cada α_i equivale a μ_i , a média do nível;
 - ▶ não se pode generalizar a situações mais complexas;
 - ▶ mais difícil de encaixar na teoria já dada.
2. tomar $\alpha_1 = 0$: será a solução utilizada.
 - ▶ corresponde a excluir a 1a. variável indicatriz do modelo (e de X);
 - ▶ permite aproveitar a teoria do modelo RLM e é generalizável.
3. impor restrições aos parâmetros: e.g., $\sum_{i=1}^k \alpha_i = 0$.
 - ▶ Foi a solução clássica, ainda hoje frequente em livros de ANOVA;
 - ▶ mais difícil de encaixar na teoria já dada.

Cada solução tem implicações na forma de interpretar os parâmetros.

A relação de base para o nosso exemplo (cont.)

Admitindo $\alpha_1 = 0$, re-escrevemos o modelo como:

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \mu_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

Agora μ_1 é o valor médio das observações do nível $i = 1$:

$$E[Y_{1j}] = \mu_1, \quad \forall j = 1, \dots, n_1$$

$$E[Y_{2j}] = \mu_1 + \alpha_2, \quad \forall j = 1, \dots, n_2$$

$$E[Y_{3j}] = \mu_1 + \alpha_3, \quad \forall j = 1, \dots, n_3$$

Cada α_i ($i > 1$) representa um **acréscimo** à média do primeiro nível.

A matriz X numa ANOVA a um factor

Agora, a matriz X tem nas colunas os vectores $\mathbf{1}_n, \mathbf{1}_2, \mathbf{1}_3, \dots, \mathbf{1}_k$.

Na ANOVA a um factor, a matriz do modelo X indica quais as observações correspondentes a cada nível do factor.

Esta natureza especial da matriz X na ANOVA faz com que **resultados gerais válidos** para qualquer Modelo Linear tenham expressões específicas no contexto da ANOVA.

Exploraremos essas expressões específicas.

Os estimadores dos parâmetros

Como a equação do modelo ANOVA é um caso particular da equação do Modelo Linear, a fórmula dos estimadores de mínimos quadrados dos parâmetros é igualmente

$$\hat{\beta} = (X^t X)^{-1} X^t Y.$$

Devido à natureza das colunas da matriz X , tem-se:

$$X^t X = \begin{bmatrix} n & n_2 & n_3 & n_4 & \dots & n_k \\ n_2 & n_2 & 0 & 0 & \dots & 0 \\ n_3 & 0 & n_3 & 0 & \dots & 0 \\ n_4 & 0 & 0 & n_4 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ n_k & 0 & 0 & 0 & \dots & n_k \end{bmatrix}$$

Os estimadores dos parâmetros (cont.)

Tem-se também:

$$(X^t X)^{-1} = \frac{1}{n_1} \begin{bmatrix} 1 & -1 & -1 & -1 & \dots & -1 \\ -1 & \frac{n_1+n_2}{n_2} & 1 & 1 & \dots & 1 \\ -1 & 1 & \frac{n_1+n_3}{n_3} & 1 & \dots & 1 \\ -1 & 1 & 1 & \frac{n_1+n_4}{n_4} & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 1 & 1 & 1 & \dots & \frac{n_1+n_k}{n_k} \end{bmatrix}$$

$$X^t Y = \begin{bmatrix} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij} \\ \sum_{j=1}^{n_2} Y_{2j} \\ \sum_{j=1}^{n_3} Y_{3j} \\ \vdots \\ \sum_{j=1}^{n_k} Y_{kj} \end{bmatrix}$$

Os estimadores dos parâmetros (cont.)

Logo,

$$\hat{\mu}_1 = \bar{Y}_1.$$

$$\hat{\alpha}_2 = \bar{Y}_2 - \bar{Y}_1.$$

$$\hat{\alpha}_3 = \bar{Y}_3 - \bar{Y}_1.$$

$$\vdots \quad \quad \quad \vdots$$

$$\hat{\alpha}_k = \bar{Y}_k - \bar{Y}_1.$$

onde $\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ é a média das n_i observações de Y no nível i .

Ou seja, os parâmetros são estimados pelas quantidades amostrais correspondentes.

Os estimadores das médias de nível

Dados os estimadores referidos no acetato anterior, e uma vez que as médias de cada nível (além do primeiro) são dadas por $\mu_i = \mu_1 + \alpha_i$, temos que os estimadores de cada média de nível são

$$\hat{\mu}_1 = \bar{Y}_1.$$

$$\hat{\mu}_2 = \bar{Y}_2.$$

$$\hat{\mu}_3 = \bar{Y}_3.$$

$$\vdots \quad \quad \quad \vdots$$

$$\hat{\mu}_k = \bar{Y}_k.$$

sendo \bar{Y}_i a média das n_i observações de Y no nível i do factor.

Os valores ajustados \hat{Y}_{ij}

Do que foi visto, decorre que qualquer observação tem valor ajustado:

$$\hat{Y}_{ij} = \hat{\mu}_i = \hat{\mu}_1 + \hat{\alpha}_i = \bar{Y}_{i.}.$$

Ou seja, os valores ajustados \hat{Y}_{ij} são iguais para todas as observações num mesmo nível i do factor, e são dadas pela média amostral das observações nesse nível.

Tal como na Regressão, os valores ajustados de Y resultam de projectar ortogonalmente os valores observados da variável resposta Y sobre o subespaço de \mathbb{R}^n gerado pelas colunas da matriz \mathbf{X} .

Numa ANOVA a um factor, o subespaço $\mathcal{C}(\mathbf{X})$ tem natureza especial.

O subespaço $\mathcal{C}(\mathbf{X})$ numa ANOVA a um factor

Qualquer vector no subespaço $\mathcal{C}(\mathbf{X})$ tem de ter valores iguais para todas as observações dum mesmo nível do factor:

$$a_1 \cdot \mathbf{1}_n + a_2 \cdot \mathcal{J}_2 + a_3 \cdot \mathcal{J}_3 + \dots + a_k \cdot \mathcal{J}_k = \begin{bmatrix} a_1 \\ \dots \\ a_1 \\ \hline a_1 + a_2 \\ \dots \\ a_1 + a_2 \\ \hline a_1 + a_3 \\ \dots \\ a_1 + a_3 \\ \hline (\dots) \\ \hline a_1 + a_k \\ \dots \\ a_1 + a_k \end{bmatrix}$$

O vector $\hat{\mathbf{Y}}$ pertence a $\mathcal{C}(\mathbf{X})$, logo tem essa característica, como se viu.

O modelo ANOVA a 1 factor para efeitos inferenciais

Para se poder fazer inferência no modelo ANOVA a um factor, admite-se não apenas que cada observação individual Y_{ij} é da forma

$$Y_{ij} = \mu_1 + \alpha_i + \varepsilon_{ij}, \quad \forall i = 1, \dots, k, \quad \forall j = 1, \dots, n_i,$$

com $E[\varepsilon_{ij}] = 0$ e $\alpha_1 = 0$.

Admite-se ainda que os erros aleatórios ε_{ij} têm as mesmas propriedades que no modelo de regressão linear: **Normais, de variância constante e independentes.**

O modelo ANOVA a um factor

Modelo ANOVA a um factor, com k níveis

Existem n observações, Y_{ij} , n_i das quais associadas ao nível i ($i = 1, \dots, k$) do factor. Tem-se:

- 1 $Y_{ij} = \mu_1 + \alpha_i + \varepsilon_{ij}, \quad \forall i = 1, \dots, k, \quad \forall j = 1, \dots, n_i \quad (\alpha_1 = 0).$
- 2 $\varepsilon_{ij} \cap \mathcal{N}(0, \sigma^2)$
- 3 $\{\varepsilon_{ij}\}_{i=1}^n$ v.a.s independentes.

O modelo tem k parâmetros desconhecidos: a média de Y no primeiro nível do factor, μ_1 , e os acréscimos α_i ($i > 1$) que geram as médias de cada um dos $k - 1$ restantes níveis do factor. Ou seja,

$$\boldsymbol{\beta} = (\mu_1, \alpha_2, \alpha_3, \dots, \alpha_k)^t.$$

O modelo ANOVA a um factor - notação vectorial

De forma equivalente, em notação vectorial,

Modelo ANOVA a um factor - notação vectorial

O vector \mathbf{Y} das n observações verifica:

- 1 $\mathbf{Y} = \mu_1 \cdot \mathbf{1}_n + \alpha_2 \cdot \mathcal{J}_2 + \alpha_3 \cdot \mathcal{J}_3 + \dots + \alpha_k \cdot \mathcal{J}_k + \boldsymbol{\varepsilon}$, sendo $\mathbf{1}_n$ o vector de n uns e $\mathcal{J}_2, \mathcal{J}_3, \dots, \mathcal{J}_k$ as variáveis indicatrizes dos níveis indicados.
- 2 $\boldsymbol{\varepsilon} \cap \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, sendo \mathbf{I}_n a matriz identidade $n \times n$.

Trata-se de um modelo análogo a um modelo de Regressão Linear Múltipla, diferindo apenas na natureza das variáveis predictoras, que são aqui variáveis indicatrizes dos níveis 2 a k do factor.

Versão vectorial/matricial do modelo a um factor

Uma terceira forma equivalente de escrever o Modelo:

Modelo ANOVA a um factor - notação vectorial/matricial

O vector \mathbf{Y} das n observações verifica:

- 1 $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,
onde $\mathbf{X} = [\mathbf{1}_n | \mathcal{J}_2 | \mathcal{J}_3 | \dots | \mathcal{J}_k]$ e $\boldsymbol{\beta} = (\mu_1, \alpha_2, \alpha_3, \dots, \alpha_k)^t$,
sendo $\mathbf{1}_n$ o vector de n uns e $\mathcal{J}_2, \mathcal{J}_3, \dots, \mathcal{J}_k$ as variáveis indicatrizes dos níveis indicados.
- 2 $\boldsymbol{\varepsilon} \cap \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, sendo \mathbf{I}_n a matriz identidade $n \times n$.

O teste aos efeitos do factor

A hipótese de que nenhum dos níveis do factor afecte a média da variável resposta corresponde à hipótese

$$\alpha_2 = \alpha_3 = \dots = \alpha_k = 0 \\ \Leftrightarrow \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

Dado o paralelismo com os modelos de Regressão Linear, esta hipótese corresponde a dizer que todos os coeficientes das "variáveis predictoras" (na ANOVA, as variáveis indicatrizes \mathcal{I}_i) são nulos. Logo, **é possível testar esta hipótese, através dum teste F de ajustamento global do modelo** (ver acetato 208).

Tratando-se dum caso particular do modelo linear, **neste contexto há fórmulas específicas.**

Os graus de liberdade

Numa ANOVA a um factor, o **número de parâmetros do modelo** é $p + 1 = k$. Logo, os graus de liberdade associados a cada Soma de Quadrados são:

| SQxx | g.l. |
|------|---------|
| SQF | $k - 1$ |
| SQRE | $n - k$ |

No contexto da ANOVA a um factor, utiliza-se **SQF** em vez de **SQR**, para indicar a Soma de Quadrados relacionada com o **Factor** (embora a sua definição seja idêntica).

Os **Quadrados Médios** continuam a ser os quocientes das Somas de Quadrados a dividir pelos respectivos graus de liberdade.

O Teste F aos efeitos do factor numa ANOVA

Sendo válido o Modelo de ANOVA a um factor, tem-se então:

Teste F aos efeitos do factor

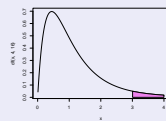
Hipóteses: $H_0 : \alpha_i = 0 \quad \forall i=2, \dots, k$ vs. $H_1 : \exists i=2, \dots, k \text{ t.q. } \alpha_i \neq 0$.
[FACTOR NÃO AFECTA] vs. [FACTOR AFECTA Y]

Estatística do Teste: $F = \frac{QMF}{QMRE} \cap F_{(k-1, n-k)}$ se H_0 .

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rej. H_0 se $F_{calc} > f_{\alpha(k-1, n-k)}$



Também as Somas de Quadrados e Quadrados Médios têm fórmulas específicas a este contexto.

Os resíduos e SQRE

Viu-se antes (acetato 272) que $\hat{Y}_{ij} = \hat{\mu}_i = \bar{Y}_i$, pelo que o resíduo da observação Y_{ij} é dado por:

$$E_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_i,$$

Logo, a **Soma de Quadrados dos Resíduos** é dada por:

$$SQRE = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 = \sum_{i=1}^k (n_i - 1) \cdot S_i^2,$$

onde $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$ é a **variância amostral** das n_i observações no i -ésimo nível do factor.

SQRE mede **variabilidade no seio dos k níveis.**

A Soma de Quadrados associada ao Factor

A Soma de Quadrados associada à Regressão toma, neste contexto, a designação **Soma de Quadrados associada ao Factor** e será representada por **SQF**. É dada por:

$$SQF = \sum_{i=1}^k \sum_{j=1}^{n_i} (\hat{Y}_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{Y}_i - \bar{Y}_{..})^2$$

$$\Leftrightarrow SQF = \sum_{i=1}^k n_i \cdot (\bar{Y}_i - \bar{Y}_{..})^2$$

sendo $\bar{Y}_{..} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$ a média da totalidade das n observações.

SQF mede **variabilidade entre as médias amostrais de cada nível.**

A relação entre Somas de Quadrados

A relação fundamental entre as três Somas de Quadrados ganha, neste contexto, um **significado particular**:

$$SQT = SQF + SQRE \\ \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y}_{..})^2 + \sum_{i=1}^k (n_i - 1) \cdot S_i^2.$$

onde:

SQT – numerador da **variância amostral** S_Y^2 da totalidade das n observações de Y ;

SQF – medida da **variabilidade das k médias de nível** (**variabilidade inter-níveis**);

SQRE – soma ponderada das **variâncias amostrais de Y em cada um dos k níveis** (**variabilidade intra-níveis**).

O quadro-resumo da ANOVA a 1 Factor

Pode-se coleccionar esta informação numa [tabela-resumo da ANOVA](#).

| Fonte | g.l. | SQ | QM | f_{calc} |
|----------|---------|--|---------------------------|--------------------|
| Factor | $k - 1$ | $SQF = \sum_{i=1}^k n_i \cdot (\bar{y}_i - \bar{y}..)^2$ | $QMF = \frac{SQF}{k-1}$ | $\frac{QMF}{QMRE}$ |
| Resíduos | $n - k$ | $SQRE = \sum_{i=1}^k (n_i - 1) s_i^2$ | $QMRE = \frac{SQRE}{n-k}$ | |
| Total | $n - 1$ | $SQT = (n - 1) s_y^2$ | - | - |

Factores no R

O R tem uma estrutura de dados específica para variáveis qualitativas (categóricas), designada **factor**.

Um **factor**, é criado pelo comando **factor**, aplicado a um vector de tipo character contendo os nomes dos vários níveis:

```
> factor(c("Adubo 1", "Adubo 1", ... , "Adubo 5"))
```

NOTA: Explore o comando **rep** para instruções curtas que criam repetições de valores.

E.g., no objecto **iris**, a coluna **Species** é um factor. Vejamos como a função **summary** lida com factores:

```
> summary(iris)
Sepal.Length Sepal.Width Petal.Length Petal.Width Species
Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100 setosa :50
1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300 versicolor:50
Median :5.800 Median :3.000 Median :4.350 Median :1.300 virginica :50
Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199
3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800
Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500
```

ANOVAs a um Factor no R

Para efectuar uma ANOVA a um Factor no R, convém **organizar os dados numa data.frame com duas colunas**:

- 1 uma para os valores (numéricos) da **variável resposta**;
- 2 outra para o **factor** (com a indicação dos seus níveis).

As fórmulas usadas no R para indicar uma ANOVA a um factor são semelhantes às da regressão linear, indicando o factor preditor.

Por exemplo, para efectuar uma ANOVA de comprimentos das pétalas sobre espécies, nos dados dos $n = 150$ lírios, a fórmula é:

$$\text{Petal.Length} \sim \text{Species}$$

uma vez que a **data frame** **iris** contém uma coluna de nome **Species** que foi definida como **factor**.

ANOVAs a um factor no R (cont.)

Embora seja possível usar o comando **lm** para efectuar uma ANOVA (a ANOVA é caso particular do Modelo Linear), existe outro **comando que organiza a informação da forma mais tradicional numa ANOVA**: **aov**.

E.g., a ANOVA de comprimento de pétalas sobre espécies para os lírios invoca-se da seguinte forma:

```
> aov(Petal.Length ~ Species, data=iris)
```

É produzido o seguinte resultado (diferente do do comando **lm**):

```
Call: aov(formula = Petal.Length ~ Species, data=iris)
Terms:
          Species Residuals
Sum of Squares 437.1028 27.2226
Deg. of Freedom      2      147

Residual standard error: 0.4303345
```

ANOVAs a um factor no R (cont.)

A função **summary** também pode ser aplicada ao resultado de uma ANOVA, produzindo o **quadro-resumo da ANOVA**:

```
> iris.aov <- aov(Petal.Length ~ Species, data=iris)
> summary(iris.aov)

          Df Sum Sq Mean Sq F value    Pr(>F)
Species    2  437.10   218.55  1180.2 < 2.2e-16 ***
Residuals 147   27.22    0.19
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Neste caso, rejeita-se claramente a hipótese de que os acréscimos de nível, α_i , sejam todos nulos, pelo que se rejeita a hipótese de comprimentos médios de pétalas iguais em todas as espécies.

O **factor** afecta a **variável resposta**.

Os parâmetros estimados, no R

Para obter as estimativas dos parâmetros $\mu_1, \alpha_2, \alpha_3, \dots, \alpha_k$, pode aplicar-se a função **coef** ao resultado da ANOVA.

No exemplo dos lírios, temos:

```
> coef(iris.aov)
(Intercept) Speciesversicolor Speciesvirginica
          1.462                2.798                4.090
```

Estes são os **valores estimados** dos parâmetros

- $\hat{\mu}_1$: **média amostral** de comprimentos de pétalas **setosa**;
- $\hat{\alpha}_2$: **acréscimo** que, somado à média amostral da 1a. espécie, nos dá a média amostral dos comprimentos de pétalas **versicolor**;
- $\hat{\alpha}_3$: **acréscimo** que, somado à média amostral da 1a. espécie, nos dá a média amostral dos comprimentos de pétalas **virginica**.

Parâmetros estimados no R (cont.)

Para melhor interpretar os resultados, vejamos as **médias por nível do factor** da variável resposta, através da função `model.tables`, com o argumento `type="means"`:

```
> model.tables(iris.aov , type="mean")
Tables of means
Grand mean
3.758

Species
Species
  setosa versicolor  virginica
  1.462      4.260      5.552
```

O R ordena os níveis de um factor por ordem alfabética.

ANOVAs como modelo Linear no R

Também é possível estudar uma ANOVA através do comando `lm`, nomeadamente para fazer inferência sobre os parâmetros do modelo:

```
> summary(lm(Petal.Length ~ Species , data=iris))
Call: lm(formula = Petal.Length ~ Species, data=iris)
Residuals:
    Min       1Q   Median       3Q      Max
-1.260 -0.258  0.038  0.240  1.348

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.46200    0.06086   24.02  <2e-16 ***
Speciesversicolor 2.79800    0.08607   32.51  <2e-16 ***
Speciesvirginica  4.09000    0.08607   47.52  <2e-16 ***
---
Residual standard error: 0.4303 on 147 degrees of freedom
Multiple R-squared:  0.9414, Adjusted R-squared:  0.9406
F-statistic: 1180 on 2 and 147 DF,  p-value: < 2.2e-16
```

A exploração ulterior de H_1

A Hipótese Nula, no teste F numa ANOVA a 1 Factor, afirma que todos os níveis do factor têm efeito nulo, isto é, que a média da variável resposta Y é igual nos k níveis do Factor:

$$\alpha_2 = \alpha_3 = \dots = \alpha_k = 0 \\ \Leftrightarrow \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

A Hipótese Alternativa diz que **pelo menos um** dos níveis do factor tem uma média de Y diferente do primeiro nível:

$$\exists i \text{ tal que } \alpha_i \neq 0 \quad (i > 1) \\ \Leftrightarrow \exists i \text{ tal que } \mu_1 \neq \mu_i \quad (i > 1)$$

Ou seja, nem todas as médias de nível de Y são iguais

A exploração ulterior de H_1 (cont.)

Caso se opte pela Hipótese Alternativa, fica em aberto (excepto quando $k = 2$) a questão de **saber quais os níveis do factor cujas médias diferem entre si**.

Mesmo com $k = 3$, a rejeição de H_0 pode dever-se a:

$$\begin{aligned} \mu_1 = \mu_2 \neq \mu_3 \quad \text{i.e., } \alpha_2 = 0; \alpha_3 \neq 0 \\ \mu_1 = \mu_3 \neq \mu_2 \quad \text{i.e., } \alpha_3 = 0; \alpha_2 \neq 0 \\ \mu_1 \neq \mu_2 = \mu_3 \quad \text{i.e., } \alpha_2 = \alpha_3 \neq 0; \\ \mu_i \text{ todos diferentes} \quad \text{i.e., } \alpha_2 \neq \alpha_3 \text{ e } \alpha_2, \alpha_3 \neq 0. \end{aligned}$$

Como optar entre estas diferentes alternativas?

A exploração ulterior de H_1 (cont.)

Uma hipótese consiste em efectuar testes aos α_i s, com base na teoria já estudada anteriormente.

Mas quanto maior for k , mais sub-hipóteses alternativas existem, mais testes haverá para fazer.

Não se trata apenas de uma questão de serem necessários muitos testes. A multiplicação do número de testes faz perder o controlo do nível de significância α global para o conjunto de todos os testes.

As comparações múltiplas

É possível construir testes de hipóteses relativos a todas as diferenças $\mu_i - \mu_j$, definidas pelas médias populacionais de Y nos níveis i, j de um factor ($i, j = 1, \dots, k$, com $i \neq j$), **controlando o nível de significância global α do conjunto dos testes**. Tais testes chamam-se **testes de comparações múltiplas** de médias.

O **nível de significância α** nos testes de comparação múltipla é a **probabilidade de rejeitar qualquer das hipóteses $\mu_i = \mu_j$, caso ela seja verdade**, ou seja, é um nível de significância **global**.

Alternativamente, podem-se construir **intervalos de confiança** para cada diferença $\mu_i - \mu_j$, com um nível $(1 - \alpha) \times 100\%$ de confiança de que os verdadeiros valores de $\mu_i - \mu_j$ pertencem a todos os intervalos.

Distribuição de Tukey para Amplitudes Studentizadas

O mais usado teste de comparações múltiplas é o **teste de Tukey**, que se baseia no seguinte resultado.

Teorema (Distribuição de Tukey)

Sejam $\{\mathbf{W}_i\}_{i=1}^k$ variáveis aleatórias independentes, com distribuição Normal, de iguais parâmetros: $\mathbf{W}_i \cap \mathcal{N}(\mu_W, \sigma_W^2), \forall i = 1, \dots, k$.

- Seja S_W^2 um estimador da variância comum σ_W^2 , tal que $\frac{v S_W^2}{\sigma_W^2} \cap \chi_v^2$.
- Seja $R_W = \max_i \mathbf{W}_i - \min_i \mathbf{W}_i$ a **amplitude amostral**.
- Sejam S_W e R_W independentes.

Então, a **amplitude Studentizada**, $\frac{R_W}{S_W}$, tem a **distribuição de Tukey**, que depende de dois parâmetros: k e v .

A utilidade da distribuição de Tukey

Numa ANOVA a um factor, admitimos que

$$Y_{ij} = \underbrace{\mu_1 + \alpha_j}_{=\mu_i} + \varepsilon_{ij}, \quad (\alpha_1 = 0),$$

pelo que (com os pressupostos relativos aos erros aleatórios do modelo ANOVA)

$$Y_{ij} \cap \mathcal{N}(\mu_i, \sigma^2).$$

Logo, a **média amostral de cada nível**, $\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$, tem distribuição

$$\bar{Y}_i \cap \mathcal{N}\left(\mu_i, \frac{\sigma^2}{n_i}\right) \Leftrightarrow \bar{Y}_i - \mu_i \cap \mathcal{N}\left(0, \frac{\sigma^2}{n_i}\right)$$

A utilidade da distribuição de Tukey (cont.)

Caso o **delineamento seja equilibrado**, isto é,

$$n_1 = n_2 = \dots = n_k (= n_c),$$

as k diferenças $\bar{Y}_i - \mu_i$ terão a mesma distribuição $\mathcal{N}(0, \sigma^2/n_c)$, e serão as variáveis \mathbf{W}_i do Teorema no acetato (297).

Um **estimador da variância comum** σ^2/n_c é dado por $QMRE/n_c$, e:

$$(n-k) \cdot \frac{QMRE/n_c}{\sigma^2/n_c} = \frac{SQRE}{\sigma^2} \cap \chi_{n-k}^2,$$

(acetatos 188 e 189, pois no modelo ANOVA há k parâmetros). Os valores ajustados \bar{Y}_i e os resíduos que definem $SQRE$ são independentes, logo, a **amplitude amostral**

$$R = \max_i (\bar{Y}_i - \mu_i) - \min_j (\bar{Y}_j - \mu_j)$$

é independente do estimador da variância comum, $QMRE/n_c$.

Aplica-se o Teorema do acetato (297).

A utilidade da distribuição de Tukey (cont.)

Assim,

$$\frac{R}{S} = \frac{\max_i (\bar{Y}_i - \mu_i) - \min_j (\bar{Y}_j - \mu_j)}{\sqrt{\frac{QMRE}{n_c}}}$$

tem a **distribuição de Tukey**, com parâmetros k e $n-k$.

O quociente $\frac{R}{S}$ não pode ser negativo, por definição.

Este resultado pode ser usado para construir testes de hipóteses ou intervalos de confiança para o conjunto de todas as diferenças de médias de nível de Y , $\mu_i - \mu_j$.

Intervalos de Confiança para $\mu_i - \mu_j$

Seja $q_{\alpha(k, n-k)}$ o valor que numa distribuição de Tukey com parâmetros k e $n-k$, deixa à direita uma região de probabilidade α . Então, por definição:

$$P\left[\frac{R}{S} < q_{\alpha(k, n-k)}\right] = 1 - \alpha$$

Logo, um intervalo de confiança a $(1 - \alpha) \times 100\%$ para a amplitude R é dado por:

$$R < q_{\alpha(k, n-k)} \cdot \sqrt{\frac{QMRE}{n_c}}$$

Os valores da função distribuição cumulativa e os quantis $q_{\alpha(k, n-k)}$ duma distribuição de Tukey são calculados no \mathbb{R} , através das funções **ptukey** e **qtukey**, respectivamente.

Intervalos de Confiança para $\mu_i - \mu_j$ (cont.)

Mas $R = \max_i (\bar{Y}_i - \mu_i) - \min_j (\bar{Y}_j - \mu_j)$ é a **maior de todas as diferenças** do tipo $|(\bar{Y}_i - \mu_i) - (\bar{Y}_j - \mu_j)|$, para qualquer $i, j = 1, \dots, k$.

Logo, para todos os pares de níveis i e j , tem-se, com grau de confiança global $(1 - \alpha) \times 100\%$,

$$\begin{aligned} |(\bar{Y}_i - \bar{Y}_j) - (\mu_i - \mu_j)| &\leq R < q_{\alpha(k, n-k)} \cdot \sqrt{\frac{QMRE}{n_c}} \\ \Leftrightarrow (\bar{Y}_i - \bar{Y}_j) - q_{\alpha(k, n-k)} \cdot \sqrt{\frac{QMRE}{n_c}} &< (\mu_i - \mu_j) < \\ &(\bar{Y}_i - \bar{Y}_j) + q_{\alpha(k, n-k)} \cdot \sqrt{\frac{QMRE}{n_c}} \end{aligned}$$

Testes de Hipóteses para $\mu_i - \mu_j = 0, \forall i, j$

Alternativamente, a partir do resultado do acetato (300) é possível testar a Hipótese Nula de que **todas** as diferenças de pares de médias de nível, $\mu_i - \mu_j$, sejam nulas, em cujo caso

$$|\bar{Y}_i - \bar{Y}_j| < q_{\alpha(k, n-k)} \cdot \sqrt{\frac{QMRE}{n_c}},$$


com probabilidade $(1 - \alpha)$. Qualquer diferença de médias amostrais de nível, $\bar{Y}_i - \bar{Y}_j$, que exceda o limiar

$$q_{\alpha(k, n-k)} \cdot \sqrt{\frac{QMRE}{n_c}}$$

indica que, para esse par de níveis i, j , se deve considerar $\mu_i \neq \mu_j$.

O nível (global) de significância de todas estas comparações é α , ou seja, a probabilidade de se concluir que $\mu_i \neq \mu_j$ (para algum par i, j), se em todos os casos $\mu_i = \mu_j$, é α .

Comparações Múltiplas de Médias no

As comparações múltiplas de médias de nível, com base no resultado de Tukey, podem ser facilmente efectuadas no .

Para se obter o termo de comparação nos testes de hipóteses a que $\mu_i - \mu_j = 0$, o quantil de ordem $1 - \alpha$ na distribuição de Tukey é obtido a partir do comando

```
> qtkey(1-alpha, k, n-k)
```

(com os valores numéricos de α , k e $n - k$).

O valor de \sqrt{QMRE} é dado pelo comando `aov`, sob a designação "Residual standard error".

Comparações Múltiplas de Médias no (cont.)

Os intervalos de Confiança a $(1 - \alpha) \times 100\%$ para as diferenças de médias são obtidos através do comando `TukeyHSD`. Por exemplo, para os dados dos lírios,

```
> TukeyHSD(aov(Sepal.Width ~ Species, data=iris))
Tukey multiple comparisons of means
 95% family-wise confidence level

$Species
      diff      lwr      upr    p adj
versicolor-setosa -0.658 -0.81885528 -0.4971447 0.0000000
virginica-setosa   -0.454 -0.61485528 -0.2931447 0.0000000
virginica-versicolor 0.204  0.04314472  0.3648553 0.0087802
```

Neste exemplo, nenhum dos intervalos inclui o valor zero, pelo que consideramos que $\mu_i \neq \mu_j$, para qualquer $i \neq j$, ou seja, todas as médias de espécie são diferentes.

Comparações Múltiplas de Médias no (cont.)


O valor de prova indicado (`p adj`) deve ser interpretado como o valor de α para o qual cada diferença de médias, $\bar{Y}_i - \bar{Y}_j$, seria, pela primeira vez, considerado não significativo.

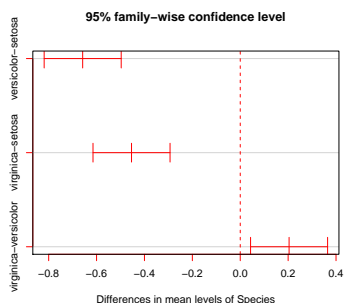
```
> TukeyHSD(aov(Sepal.Width ~ Species, data=iris))
Tukey multiple comparisons of means
 95% family-wise confidence level

$Species
      diff      lwr      upr    p adj
versicolor-setosa -0.658 -0.81885528 -0.4971447 0.0000000
virginica-setosa   -0.454 -0.61485528 -0.2931447 0.0000000
virginica-versicolor 0.204  0.04314472  0.3648553 0.0087802
```

Assim, para $\alpha = 0.00878$, a diferença de médias amostrais para as espécies *virginica* e *versicolor* já seria considerada não significativa. Ou seja, o intervalo a $(1 - \alpha) \times 100\% = 99.122\%$ de confiança para essa diferença de médias já conteria o valor zero.


Representação gráfica das comparações múltiplas

O  disponibiliza ainda um auxiliar gráfico para visualizar as comparações das médias de nível, através da função `plot`, aplicada ao resultado da função `TukeyHSD`.



Delineamentos não equilibrados

Quando o delineamento da ANOVA a um Factor não é equilibrado (isto é, existe diferente número de observações nos vários níveis do factor), os resultados agora enunciados não são, em rigor, válidos.

Mas, para delineamentos em que o desequilíbrio no número de observações não seja muito acentuado, é possível ajustar os valores da distribuição de Tukey. A função `TukeyHSD` do  incorpora essas correções.

Análise de Resíduos na ANOVA a 1 Factor

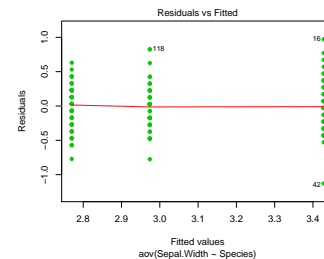
A validade dos pressupostos do modelo estuda-se de forma idêntica ao que foi visto na Regressão Linear. Mas há algumas particularidades.

Numa ANOVA a um factor, os resíduos aparecem empilhados em k colunas nos gráficos de \hat{y}_{ij} vs. e_{ij} , porque qualquer valor ajustado \hat{y}_{ij} é igual para observações num mesmo nível do factor.

Este padrão **não** indicia qualquer violação aos pressupostos do modelo.

Análise de Resíduos na ANOVA a 1 Factor (cont.)

Padrão de resíduos numa ANOVA a 1 Factor (o exemplo considerado é $\text{Sepal.Width} \sim \text{Species}$, nos lírios)



Inspeccionando a homogeneidade de variâncias

Outra particularidade da ANOVA, resultante do facto de haver n_i repetições em cada um dos k níveis do factor: **é possível testar formalmente se as variâncias dos erros aleatórios diferem entre os níveis do factor.**

O **Teste de Bartlett** testa as hipóteses

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

vs.

$$H_1 : \exists i, i' \text{ t.q. } \sigma_i^2 \neq \sigma_{i'}^2,$$

sendo σ_i^2 a variância comum dos erros aleatórios ε_{ij} do nível i .

Médias aritméticas e médias geométricas

Relação geral entre a média aritmética e a média geométrica (mesmo que ponderadas) de quaisquer k números **positivos**.

Sejam $\tau_1, \tau_2, \dots, \tau_k$ números positivos, e p_1, p_2, \dots, p_k números entre 0 e 1, de soma 1.

A **média aritmética** (ponderada com pesos p_i) dos τ_i s é

$$MA = \sum_{i=1}^k p_i \tau_i.$$

A **média geométrica** (ponderada com pesos p_i) dos τ_i s é

$$MG = \prod_{i=1}^k \tau_i^{p_i}.$$

Quando $p_i = \frac{1}{k}, \forall i$, temos as médias aritmética e geométrica simples.

A desigualdade entre média aritmética e geométrica

Quaisquer que sejam os valores (positivos) dos τ_i e das ponderações p_i , tem-se a seguinte desigualdade entre a média aritmética e geométrica dos k valores de τ :

$$MG \leq MA \quad (4)$$

A igualdade em (4) verifica-se se e só se os k valores de τ são iguais:

$$\tau_1 = \tau_2 = \dots = \tau_k.$$

Quanto maior fôr a dispersão dos τ , maior será a diferença entre média geométrica e média aritmética.

O nosso contexto

Admita-se que os erros aleatórios, e portanto as observações Y_{ij} , do nível i do factor têm **variância comum** $V[\varepsilon_{ij}] = V[Y_{ij}] = \sigma_i^2$, podendo, no entanto os σ_i^2 diferir entre níveis.

A ideia subjacente à estatística do teste de Bartlett é a de comparar uma média aritmética (MA) e geométrica (MG) das variâncias amostrais

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2,$$

que estimam as variâncias **populacionais** σ_i^2 .

Se fôr verdadeira a Hipótese Nula do teste de Bartlett (σ_i^2 todos iguais), é natural que as variâncias amostrais S_i^2 sejam aproximadamente iguais e $\frac{MA}{MG}$ seja próximo de 1. Quanto maior esta razão das médias, mais duvidosa se torna H_0 .

Estimando as variâncias de nível

No cálculo das médias aritmética e geométrica das variâncias amostrais de nível S_i^2 , vamos utilizar ponderações apropriadas ao contexto.

Se usarmos como ponderações

$$p_i = \frac{n_i - 1}{\sum_i (n_i - 1)} = \frac{n_i - 1}{n - k},$$

a média aritmética ponderada dos estimadores S_i^2 é o Quadrado Médio Residual da ANOVA (ver o Acetato 282):

$$MA = \sum_{i=1}^k \frac{n_i - 1}{n - k} \cdot S_i^2 = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{n - k} = QMRE.$$

A ideia subjacente ao teste de Bartlett

A média geométrica dos k estimadores de variâncias de nível é:

$$MG = \prod_{i=1}^k (S_i^2)^{\frac{n_i - 1}{n - k}}.$$

Sabemos que $MA/MG \geq 1$. Quanto maior for este quociente, maior será a variabilidade dos S_i^2 , e portanto mais duvidosa será a Hipótese Nula da igualdade dos σ_i^2 .

Logo, o quociente $\frac{MA}{MG}$ é um candidato a estatística do teste à igualdade de variâncias, com Região Crítica unilateral direita. Mas é necessário conhecer a distribuição de probabilidades duma estatística do Teste, sob H_0 .

O teste de Bartlett

Bartlett demonstrou que, sob H_0 , uma transformação monótona crescente do quociente MA/MG tem distribuição assintoticamente χ^2 , caso as variáveis subjacentes às variâncias tenham distribuição Normal. Concretamente, demonstrou que

$$K^2 = \frac{n - k}{C} \cdot \ln \left[\frac{MA}{MG} \right] = \frac{n - k}{C} \cdot (\ln MA - \ln MG),$$

tem, assintoticamente distribuição χ_{k-1}^2 , sendo

$$C = 1 + \frac{1}{3(k-1)} \left[\sum_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{n - k} \right].$$

O Teste de Bartlett

Teste de Bartlett à homogeneidade de variâncias

Hipóteses: $H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$ vs. $H_1 : \exists i, i' \text{ t.q. } \sigma_i^2 \neq \sigma_{i'}^2$
[Variâncias homogêneas] [Var. heterogêneas]

Estatística do Teste:

$$K^2 = \frac{(n - k) \ln QMRE - \sum_{i=1}^k (n_i - 1) \ln S_i^2}{C} \sim \chi_{k-1}^2$$

$$\text{onde } C = 1 + \frac{1}{3(k-1)} \left[\sum_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{n - k} \right].$$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se $K_{\text{calc}}^2 > \chi_{\alpha(k-1)}^2$

O Teste de Bartlett no \mathbb{R}

No \mathbb{R} , o teste de Bartlett é invocado pelo comando `bartlett.test`, tendo por argumento uma fórmula (análoga à usada no comando `aov` para indicar a variável resposta e o factor). E.g.,

```
> bartlett.test(Sepal.Width ~ Species, data=iris)
```

```
Bartlett test of homogeneity of variances
```

```
data: Sepal.Width by Species
```

```
Bartlett's K-squared = 2.0911, df = 2, p-value = 0.3515
```

Neste caso, o teste de Bartlett indica a não rejeição de H_0 , ou seja, é admissível a hipótese de igualdade nas variâncias em cada nível do factor.

Precauções

Duas precauções na utilização do teste de Bartlett:

- O teste de Bartlett é fortemente sensível à Normalidade das observações subjacentes.
- A distribuição χ^2 é apenas assintótica. Uma regra comum é considerar que o teste apenas deve ser usado caso $n_i \geq 5$, $\forall i = 1, \dots, k$.

Violações aos pressupostos da ANOVA

Violações aos pressupostos do modelo não têm sempre igual gravidade. Alguns comentários gerais:

- O teste F da ANOVA e as comparações múltiplas de Tukey são relativamente robustos a desvios à hipótese de normalidade.
- As violações ao pressuposto de variâncias homogêneas são em geral pouco graves no caso de delineamentos equilibrados, mas podem ser graves em delineamentos não equilibrados.
- A falta de independência entre erros aleatórios é a violação mais grave dos pressupostos e deve ser evitada, o que é em geral possível com um delineamento experimental adequado.

Uma advertência

Na formulação clássica do modelo ANOVA a um Factor, e a partir da equação-base

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

em vez de impor a condição $\alpha_1 = 0$, impõe-se a condição $\sum_i \alpha_i = 0$.

Esta condição alternativa:

- muda a forma de interpretar os parâmetros (μ é agora uma espécie de média geral das observações e α_i o desvio médio das observações do nível i em relação a essa média geral);
- Muda os estimadores dos parâmetros.
- Não muda o resultado do teste F à existência de efeitos do factor, nem a qualidade global do ajustamento.
- A nossa formulação, além de generalizável a modelos com mais Factores, permite aproveitar directamente os resultados da Regressão Linear Múltipla.

Delineamentos e Unidades experimentais

No delineamento das experiências para posterior análise através duma ANOVA (ou regressão linear), é frequente que as n observações da variável resposta correspondam a n diferentes unidades experimentais (indivíduos, parcelas de terreno, locais, etc.).

É conveniente que as unidades experimentais nas quais se recolhem os dados sejam tão homogêneas quanto possível, excepto em aspectos associados aos factores incorporados no modelo.

Unidades experimentais (cont.)

Qualquer variabilidade não controlada nas unidades experimentais (isto é, que não se pode atribuir aos preditores) é considerada no modelo como variação aleatória, pelo que irá contribuir para aumentar o valor de $SQRE$ e de $QMRE$.

Aumentar $QMRE$ significa, no teste aos efeitos do factor, diminuir o valor calculado da estatística F , afastando-a da região crítica. Assim,

numa ANOVA

heterogeneidade não controlada nas unidades experimentais contribui para esconder a presença de eventuais efeitos do factor.

numa Regressão Linear

heterogeneidade não controlada nas unidades experimentais contribui para piorar a qualidade de ajustamento do modelo, diminuindo o seu Coeficiente de Determinação.

Controlar a heterogeneidade

Na prática, é frequentemente impossível tornar as unidades experimentais totalmente homogêneas.

A natural variabilidade de plantas, animais, terrenos, localidades geográficas, células, etc. significa que em muitas situações existirá variabilidade não controlável entre unidades experimentais.

Alguma protecção contra efeitos não controlados resulta dos princípios de:

- repetição;
- casualização.

Deve-se associar níveis do factor às unidades experimentais de forma aleatória (casualizada).

Criar factores para controlar variabilidade

Mesmo que seja possível encontrar unidades experimentais homogêneas, isso pode ter um efeito indesejável: restringir a validade dos resultados ao tipo de unidades experimentais com as características utilizadas na experiência.

Caso se saiba que existe um factor de variabilidade importante nas unidades experimentais, a melhor forma de controlar os seus efeitos consiste em contemplar a existência desse factor de variabilidade no delineamento e no modelo, de forma a filtrar os seus efeitos.

Um exemplo

Pretende-se analisar o rendimento de 5 diferentes variedades de trigo. Os rendimentos são também afectados pelos tipo de solos usados.

Nem sempre é possível ter terrenos homogéneos numa experiência. Mesmo que seja possível, **pode não ser desejável**, por se limitar a validade dos resultados a um único tipo de solos.

Admita-se que existem terrenos com quatro diferentes tipos de solos. Cada terreno pode ser dividido em cinco parcelas viáveis para o trigo. Em vez de repartir aleatoriamente as 5 variedades pelas 20 parcelas, é preferível forçar cada tipo de terreno a conter uma parcela com cada variedade. Apenas dentro dos terrenos haverá casualização.

Num delineamento experimental deste tipo, os terrenos designam-se **blocos casualizados**.

Um exemplo (cont.)

A situação descrita no acetato anterior é a seguinte:

Bloco 1 (Solo 1)

| | | | | |
|-------|-------|-------|-------|-------|
| Var.1 | Var.3 | Var.4 | Var.5 | Var.2 |
|-------|-------|-------|-------|-------|

Bloco 2 (Solo 2)

| | | | | |
|-------|-------|-------|-------|-------|
| Var.4 | Var.3 | Var.5 | Var.1 | Var.2 |
|-------|-------|-------|-------|-------|

Bloco 3 (Solo 3)

| | | | | |
|-------|-------|-------|-------|-------|
| Var.2 | Var.4 | Var.1 | Var.3 | Var.5 |
|-------|-------|-------|-------|-------|

Bloco 4 (Solo 4)

| | | | | |
|-------|-------|-------|-------|-------|
| Var.5 | Var.2 | Var.4 | Var.1 | Var.3 |
|-------|-------|-------|-------|-------|

Houve uma **restrição à casualização total**: dentro de cada bloco há casualização, mas obriga-se cada bloco a ter uma parcela associada a cada nível do factor **variedade**.

Delineamentos factoriais a dois factores

O delineamento agora exemplificado é um caso particular de um **delineamento factorial a dois factores**, sendo um dos factores a variedade de trigo e a outra o tipo de solos.

A existência de mais do que um factor pode resultar de:

- a tentativa de controlar a variabilidade experimental;
- pretender-se realmente estudar eventuais efeitos de mais do que um factor sobre a variável resposta.

Historicamente, a primeira situação ficou associada à designação **blocos**, e na segunda fala-se apenas em **factores**. Mas são **situações análogas**.

Um **delineamento factorial** é um delineamento em que há observações para todas as possíveis combinações de níveis de cada factor.

Modelo ANOVA a 2 Factores (sem interacção)

A um delineamento com 2 factores pode ser associado um modelo ANOVA que prevê a existência de **dois diferentes tipos de efeitos**: os efeitos associados aos níveis de cada um dos factores.

Admita-se a existência de:

- Uma **variável resposta** Y , da qual se efectuem n observações.
- Um **Factor A**, com a níveis.
- Um **Factor B**, com b níveis.

Modelo ANOVA a 2 Factores (sem interacção)

Notação: Cada observação da variável resposta será agora identificada com **três índices**, Y_{ijk} , onde:

- i indica o nível i do Factor A.
- j indica o nível j do Factor B.
- k indica a repetição k no nível i do factor A e nível j do Factor B.

Cada **situação experimental** é dada pelo **cruzamento dum nível dum Factor com um nível do outro Factor**, cruzamento chamado **célula**.

Modelo ANOVA a 2 Factores (sem interacção)

O número de observações na célula (i, j) é representado por n_{ij} .

Tem-se

$$\sum_{i=1}^a \sum_{j=1}^b n_{ij} = n.$$

Se o número de observações for igual em todas as células,

$$n_{ij} = n_c, \quad \forall i, j,$$

estamos perante um **delineamento equilibrado**.

A matriz do delineamento na ANOVA a 2 Factores (sem interacção)

$$X = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 & 1 & \dots & 0 \\ 1 & 0 & \dots & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 & 0 & \dots & 1 \\ 1 & 0 & \dots & 0 & 0 & \dots & 1 \\ \hline 1 & 1 & \dots & 0 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 0 & 0 & \dots & 1 \\ 1 & 1 & \dots & 0 & 0 & \dots & 1 \\ \hline \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 1 & 0 & \dots & 1 \\ 1 & 0 & \dots & 1 & 0 & \dots & 1 \end{bmatrix}$$

\uparrow \uparrow \uparrow \uparrow \uparrow \uparrow \uparrow
 1_n A_2 A_a B_2 B_b

A natureza do parâmetro

Uma observação de Y efectuada na célula $(1, 1)$, correspondente ao cruzamento do primeiro nível de cada factor será da forma:

$$Y_{11k} = \mu + \varepsilon_{11k} \implies E[Y_{11k}] = \mu$$

O parâmetro μ corresponde ao valor esperado da variável resposta Y na célula cujas indicatrizes foram excluídas da matriz do delineamento. Será doravante chamado μ_{11} .

A natureza dos parâmetros α_i

Uma observação de Y efectuada na célula $(i, 1)$, com $i > 1$, correspondente ao cruzamento dum nível do factor A diferente do primeiro, com o primeiro nível do Factor B será da forma:

$$Y_{i1k} = \mu_{11} + \alpha_i + \varepsilon_{i1k} \implies \mu_{i1} = E[Y_{i1k}] = \mu_{11} + \alpha_i$$

O parâmetro $\alpha_i = \mu_{i1} - \mu_{11}$ corresponde ao acréscimo no valor esperado da variável resposta Y associado a observações do nível $i > 1$ do Factor A (relativamente às observações do primeiro nível do Factor A). Designa-se o efeito do nível i do factor A.

A natureza dos parâmetros β_j

Uma observação de Y efectuada na célula $(1, j)$, com $j > 1$, correspondente ao cruzamento do primeiro nível do factor A com um nível do Factor B diferente do primeiro será da forma:

$$Y_{1jk} = \mu_{11} + \beta_j + \varepsilon_{1jk} \implies \mu_{1j} = E[Y_{1jk}] = \mu_{11} + \beta_j$$

O parâmetro $\beta_j = \mu_{1j} - \mu_{11}$ corresponde ao acréscimo no valor esperado da variável resposta Y associado a observações do nível j do Factor B (relativamente às observações do primeiro nível do Factor B). Designa-se o efeito do nível j do factor B.

Observações de Y no caso geral

Estas interpretações dos parâmetros α_i e β_j confirmam-se para observações de Y efectuadas numa célula genérica (i, j) , com $i, j > 1$, correspondente ao cruzamento de níveis diferentes do primeiro, quer no Factor A, quer no Factor B. Essas observações serão da forma:

$$Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk} \implies E[Y_{ijk}] = \mu_{11} + \alpha_i + \beta_j.$$

Os valores esperados de Y são, neste caso, acrescidos em relação ao valor esperado numa observação na célula $(1, 1)$, quer pela parcela α_i , quer pela parcela β_j .

O modelo ANOVA a dois factores, sem interacção

Juntando os pressupostos necessários à inferência,

Modelo ANOVA a dois factores, sem interacção

Existem n observações, Y_{ijk} , n_{ij} das quais associadas à célula (i, j) ($i = 1, \dots, a; j = 1, \dots, b$). Tem-se:

- 1 $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk}, \forall i=1, \dots, a; j=1, \dots, b; k=1, \dots, n_{ij} \quad (\alpha_1 = 0; \beta_1 = 0).$
- 2 $\varepsilon_{ijk} \cap \mathcal{N}(0, \sigma^2), \forall i, j, k$
- 3 $\{\varepsilon_{ijk}\}_{i,j,k}$ v.a.s independentes.

O modelo tem $a + b - 1$ parâmetros desconhecidos:

- o parâmetro μ_{11} ;
- os $a - 1$ acréscimos α_i ($i > 1$); e
- os $b - 1$ acréscimos β_j ($j > 1$).

Testando a existência de efeitos

Um teste global de ajustamento do modelo não distinguiria entre os efeitos do Factor A e os efeitos do Factor B.

Mais útil será **testar a existência dos efeitos de cada factor separadamente**. Seria útil dispôr de testes para as hipóteses:

- $H_0 : \alpha_i = 0, \quad \forall i = 2, \dots, a$; e
- $H_0 : \beta_j = 0, \quad \forall j = 2, \dots, b$.

Teste aos efeitos do Factor B

O modelo do Acetato ANOVA a 2 Factores, sem interacção (Acetato 344) tem equação de base, em notação vectorial,

$$Y = \mu 1_n + \alpha_2 \mathcal{I}_{A_2} + \dots + \alpha_a \mathcal{I}_{A_a} + \beta_2 \mathcal{I}_{B_2} + \dots + \beta_b \mathcal{I}_{B_b} + \epsilon$$

O facto de ser um Modelo Linear permite aplicar a teoria já conhecida para este tipo de modelos, para testar as hipóteses

$$H_0 : \beta_j = 0, \quad \forall j = 2, \dots, b \quad \text{vs.} \quad H_1 : \exists j \text{ tal que } \beta_j \neq 0.$$

Trata-se dum teste F parcial comparando o modelo

$$\text{(Modelo } M_{A+B}) \quad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \epsilon_{ijk},$$

com o submodelo de equação de base

$$\text{(Modelo } M_A) \quad Y_{ijk} = \mu_{11} + \alpha_i + \epsilon_{ijk},$$

que é um modelo ANOVA a 1 Factor (factor A).

A construção do teste aos efeitos do Factor B

Seja o delineamento equilibrado, ou não, podemos:

- construir as matrizes X do delineamento para os dois modelos (M_{A+B} e M_A).
- Obter as respectivas estimativas de parâmetros, $\hat{\beta} = (X'X)^{-1} X'Y$, para a matriz X correspondente a cada modelo.
- Obter as respectivas Somas de Quadrados Residuais.
- Efectuar o teste F parcial indicado, com a estatística de Teste apropriada:

$$\text{(Efeitos Factor B)} \quad F = \frac{\overbrace{SQRE_A - SQRE_{A+B}}^{=SQB}}{\frac{b-1}{SQRE_{A+B}}} = \frac{SQRE_A - SQRE_{A+B}}{\frac{SQRE_{A+B}}{n-(a+b-1)}}$$

Testando os efeitos principais de cada Factor

Consideremos também um teste aos efeitos do Factor A. Definindo:

- SQA como a Soma de Quadrados do Factor no Modelo M_A ; e
- SQB como no acetato anterior,

temos:

$$\begin{aligned} SQB &= SQRE_A - SQRE_{A+B} \\ SQA &= SQF_A = SQT - SQRE_A \end{aligned}$$

Somando estas SQs a $SQRE_{A+B}$, obtém-se:

$$SQRE_{A+B} + \underbrace{SQA + SQB}_{=SQF_{A+B}} = SQT$$

que é uma decomposição de SQT .

Usamos as Somas de Quadrados de cada factor para definir os numeradores das estatísticas dos dois testes, e o Quadrado Médio Residual do modelo $A+B$ para definir o denominador das duas estatísticas.

O Teste F aos efeitos do factor A

Sendo válido o Modelo de ANOVA a dois factores, sem interacção, e definindo $QMA = \frac{SQA}{a-1}$, temos:

Teste F aos efeitos do factor A

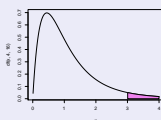
Hipóteses: $H_0 : \alpha_i = 0 \quad \forall i=2, \dots, a$ vs. $H_1 : \exists i=2, \dots, a \text{ t.q. } \alpha_i \neq 0$.
[A NÃO AFECTA Y] vs. [A AFECTA Y]

Estatística do Teste: $F = \frac{QMA}{QMRE} \cap F_{(a-1, n-(a+b-1))}$ se H_0 .

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se
 $F_{calc} > f_{\alpha(a-1, n-(a+b-1))}$



O Teste F aos efeitos do factor B

Sendo válido o Modelo de ANOVA a dois factores, sem interacção definindo $QMB = \frac{SQB}{b-1}$, temos:

Teste F aos efeitos do factor B

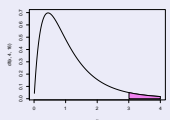
Hipóteses: $H_0 : \beta_j = 0 \quad \forall j=2, \dots, b$ vs. $H_1 : \exists j=2, \dots, b \text{ t.q. } \beta_j \neq 0$.
[B NÃO AFECTA Y] vs. [B AFECTA Y]

Estatística do Teste: $F = \frac{QMB}{QMRE} \cap F_{(b-1, n-(a+b-1))}$ se H_0 .

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se
 $F_{calc} > f_{\alpha(b-1, n-(a+b-1))}$



Fórmulas para delineamentos equilibrados

Sejam:

$\bar{Y}_{i..}$ a média amostral das $b n_c$ observações do nível i do

$$\text{Factor A, } \bar{Y}_{i..} = \frac{1}{b n_c} \sum_{j=1}^b \sum_{k=1}^{n_c} Y_{ijk}$$

$\bar{Y}_{.j.}$ a média amostral das $a n_c$ observações do nível j do

$$\text{Factor B, } \bar{Y}_{.j.} = \frac{1}{a n_c} \sum_{i=1}^a \sum_{k=1}^{n_c} Y_{ijk}$$

$\bar{Y}_{...}$ a média amostral da totalidade das $n = a b n_c$

$$\text{observações, } \bar{Y}_{...} = \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_c} Y_{ijk}$$

Se o delineamento é equilibrado, ou seja, $n_{ij} = n_c, \forall i, j$, tem-se:

- $\hat{\mu}_{11} = \bar{Y}_{1..} + \bar{Y}_{.1.} - \bar{Y}_{...}$
- $\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}$
- $\hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...}$

Fórmulas para delineamentos equilibrados (cont.)

Tendo em conta estas fórmulas e a equação base do Modelo, tem-se que os valores ajustados de cada observação dependem apenas das médias dos respectivos níveis em cada factor e da média geral de todas as observações:

$$\hat{Y}_{ijk} = \hat{\mu}_{11} + \hat{\alpha}_i + \hat{\beta}_j = \bar{Y}_{i..} + \bar{Y}_{.j.} - \bar{Y}_{...}, \quad \forall i, j, k$$

Consideremos agora as fórmulas das três Somas de Quadrados no Modelo M_{A+B} .

As Somas de Quadrados (delineamento equilibrado)

É preciso somar variando os 3 índices:

$$SQT = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_c} (Y_{ijk} - \bar{Y}_{...})^2$$

$$SQF = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_c} (\hat{Y}_{ijk} - \bar{Y}_{...})^2$$

$$SQRE = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_c} (Y_{ijk} - \hat{Y}_{ijk})^2$$

A Soma de Quadrados dos Factores

No Modelo M_{A+B} , a Soma de Quadrados associada aos Factores (SQF_{A+B}) tem, para delineamentos equilibrados, a seguinte decomposição:

$$\begin{aligned} SQF_{A+B} &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_c} [(\bar{Y}_{i..} + \bar{Y}_{.j.} - \bar{Y}_{...}) - \bar{Y}_{...}]^2 \\ &= \sum_{i=1}^a \sum_{j=1}^b n_c \cdot [(\bar{Y}_{i..} - \bar{Y}_{...}) + (\bar{Y}_{.j.} - \bar{Y}_{...})]^2 \\ &= \underbrace{b n_c \cdot \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2}_{= SQA} + \underbrace{a n_c \cdot \sum_{j=1}^b (\bar{Y}_{.j.} - \bar{Y}_{...})^2}_{= SQB} \end{aligned}$$

SQA e SQB em delineamentos equilibrados

A Soma de Quadrados associada ao factor A obtida no acetato 354 e usado no teste aos efeitos do Factor A é a Soma de Quadrados do Factor (SQF_A) do Modelo M_A , apenas com o Factor A.

Nesse modelo, os valores ajustados são $\hat{Y}_{ijk} = \bar{Y}_{i..}$ (acetato 273), logo:

$$SQF_A = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_c} (\hat{Y}_{ijk} - \bar{Y}_{...})^2 = b n_c \cdot \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2 = SQA$$

Da mesma forma, num delineamento equilibrado, SQB é a Soma de Quadrados do Factor (SQF_B) do Modelo M_B , apenas com o Factor B:

Nesse modelo, os valores ajustados são $\hat{Y}_{ijk} = \bar{Y}_{.j.}$ (acetato 273), logo:

$$SQF_B = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_c} (\hat{Y}_{ijk} - \bar{Y}_{...})^2 = a n_c \cdot \sum_{j=1}^b (\bar{Y}_{.j.} - \bar{Y}_{...})^2 = SQB$$

O quadro-resumo da ANOVA a 2 Factores (sem interacção; delineamento equilibrado)

| Fonte | g.l. | SQ | QM | f_{calc} |
|----------|-------------------|---|---------------------------------------|--------------------|
| Factor A | $a - 1$ | $SQA = b n_c \cdot \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2$ | $QMA = \frac{SQA}{a-1}$ | $\frac{QMA}{QMRE}$ |
| Factor B | $b - 1$ | $SQB = a n_c \cdot \sum_{j=1}^b (\bar{Y}_{.j.} - \bar{Y}_{...})^2$ | $QMB = \frac{SQB}{b-1}$ | $\frac{QMB}{QMRE}$ |
| Resíduos | $n - (a + b - 1)$ | $SQRE = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_c} (Y_{ijk} - \hat{Y}_{ijk})^2$ | $QMRE = \frac{SQRE}{n - (a + b - 1)}$ | |
| Total | $n - 1$ | $SQT = (n - 1) s_y^2$ | - | - |

ANOVA a dois Factores, sem interacção no R

Para efectuar uma ANOVA a dois Factores (sem interacção) no R, convém **organizar os dados numa data.frame com três colunas:**

- 1 uma para os valores (numéricos) da variável resposta;
- 2 outra para o factor A (com a indicação dos seus níveis);
- 3 outra para o factor B (com a indicação dos seus níveis).

As fórmulas utilizadas no R para indicar uma ANOVA a dois Factores, sem interacção, são semelhantes às usadas na Regressão Linear com dois preditores, devendo o nome dos dois factores ser separado pelo símbolo +:

$$y \sim fA + fB$$

Um exemplo

O rendimento de cinco variedades de aveia (*manchuria*, *svansota*, *velvet*, *trebi* e *peatland*) foi registado em seis diferentes localidades¹. Em cada localidade foi semeada uma e uma só parcela com cada variedade (havendo casualização em cada localidade).

```
> summary(aov(Y1 ~ Var + Loc, data=immer))
              Df Sum Sq Mean Sq F value    Pr(>F)
Var             4  2756.6    689.2  4.2309  0.01214 *
Loc             5 17829.8   3566.0 21.8923 1.751e-07 ***
Residuals      20  3257.7    162.9
```

Há alguma indicação de efeitos significativos entre variedades, e muita entre localidades. E num modelo sem efeito de localidades (blocos)?

```
> summary(aov(Y1 ~ Var, data=immer))
              Df Sum Sq Mean Sq F value Pr(>F)
Var             4  2756.6    689.2  0.817  0.5264
Residuals      25 21087.6    843.5
```

¹Dados em Immer, Hayes e LeRoy Powers, Statistical adaptation of barley varietal adaptation, Journal of the American Society for Agronomy, 26, 403-419, 1934.

Delineamentos não equilibrados

Se um delineamento **não** é equilibrado, as fórmulas do acetato 351, e as que delas decorrem, não se aplicam.

É possível manter uma decomposição do tipo

$$SQT = SQA + SQB + SQRE$$

e justificar testes análogos aos considerados nos acetatos (349) e (350), mas de duas formas alternativas e diferentes:

- Tomar

$$SQA = SQF_A \quad \text{e} \quad SQB = SQRE_A - SQRE_{A+B} \quad (\neq SQF_B)$$

- Tomar

$$SQB = SQF_B \quad \text{e} \quad SQA = SQRE_B - SQRE_{A+B} \quad (\neq SQF_A)$$

Modelos com interacção

Um modelo ANOVA a 2 Factores, sem interacção, foi considerado para um **delineamento factorial**, isto é, em que se cruzam todos os níveis de um e outro factor.

Um **modelo sem efeitos de interacção** é utilizado sobretudo quando existe **uma única observação em cada célula**, i.e., $n_{ij} = 1, \forall i, j$.

Na presença de repetições nas células, a forma mais natural de modelar um delineamento com dois factores é a de prever a existência de **um terceiro tipo de efeitos**: os **efeitos de interacção**.

A ideia é incorporar na equação base do modelo para Y_{ijk} uma parcela $(\alpha\beta)_{ij}$ que permita que em cada célula haja um **efeito específico da combinação dos níveis i do Factor A e j do Factor B**:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

Os valores esperados de Y_{ijk}

Vamos admitir as seguintes **restrições aos parâmetros**:

$$\alpha_1 = 0 \quad ; \quad \beta_1 = 0 \quad ; \quad (\alpha\beta)_{1j} = 0, \forall j \quad ; \quad (\alpha\beta)_{i1} = 0, \forall i.$$

Tem-se:

- Para a primeira célula ($i = j = 1$): $\mu_{11} = E[Y_{11k}] = \mu$.
- Nas restantes células $(1, j)$ do primeiro nível do Factor A: $\mu_{1j} = E[Y_{1jk}] = \mu_{11} + \beta_j$.
- Nas restantes células $(i, 1)$ do primeiro nível do Factor B: $\mu_{i1} = E[Y_{i1k}] = \mu_{11} + \alpha_i$.
- Nas células genéricas (i, j) , com $i > 1$ e $j > 1$, $\mu_{ij} = E[Y_{ijk}] = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$.

Os efeitos α_i e β_j designam-se **efeitos principais** de cada Factor.

Variáveis indicatrizes de célula

A versão vectorial do modelo com interacção associa os novos efeitos $(\alpha\beta)_{ij}$ a **variáveis indicatrizes de cada célula**, excluindo, mais uma vez, as células associadas ao primeiro nível de qualquer factor.

A equação-base do modelo ANOVA a 2 Factores, com interacção, é:

$$Y = \mu \mathbf{1}_n + \alpha_2 \mathcal{I}_{A_2} + \dots + \alpha_a \mathcal{I}_{A_a} + \beta_2 \mathcal{I}_{B_2} + \dots + \beta_b \mathcal{I}_{B_b} + (\alpha\beta)_{22} \mathcal{I}_{A_2:B_2} + (\alpha\beta)_{23} \mathcal{I}_{A_2:B_3} + \dots + (\alpha\beta)_{ab} \mathcal{I}_{A_a:B_b} + \epsilon$$

onde $\mathcal{I}_{A_i:B_j}$ representa a **variável indicatriz da célula** correspondente ao nível i do Factor A e nível j do factor B.

Existem neste modelo **ab** parâmetros.

Cada indicatriz de célula é da forma $\mathcal{I}_{A_i:B_j} = \mathcal{I}_{A_i} * \mathcal{I}_{B_j}$, com o **operador *** a indicar uma **multiplicação, elemento a elemento**, entre dois vectores.

Modelo ANOVA a 2 factores, com interacção (cont.)

O ajustamento deste modelo faz-se de forma análoga ao ajustamento de modelos anteriores.

A matriz X do delineamento é agora constituída por ab colunas:

- uma coluna de uns, $\mathbf{1}_n$, associada ao parâmetro μ_{11} .
- $a-1$ colunas de indicatrizes de nível do factor A, \mathcal{I}_{A_i} , ($i > 1$), associadas aos parâmetros α_i .
- $b-1$ colunas de indicatrizes de nível do factor B, \mathcal{I}_{B_j} , ($j > 1$), associadas aos parâmetros β_j .
- $(a-1)(b-1)$ colunas de indicatrizes de célula, $\mathcal{I}_{A_i B_j}$, ($i, j > 1$), associadas aos efeitos de interacção $(\alpha\beta)_{ij}$.

Como em modelos anteriores, $\hat{Y} = HY$, sendo H a matriz que projecta ortogonalmente sobre o espaço $\mathcal{C}(X)$ gerado pelas colunas desta

matriz X . E também, $SQRE_{A \times B} = \|Y - \hat{Y}\|^2 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (Y_{ijk} - \hat{Y}_{ijk})^2$.

Os três testes ANOVA

Neste delineamento, desejamos fazer um teste à existência de cada um dos três tipos de efeitos:

- $H_0 : (\alpha\beta)_{ij} = 0, \forall i = 2, \dots, a, \forall j = 2, \dots, b$;
- $H_0 : \alpha_i = 0, \forall i = 2, \dots, a$; e
- $H_0 : \beta_j = 0, \forall j = 2, \dots, b$.

As estatísticas de teste para cada um destes testes obtêm-se a partir da decomposição da Soma de Quadrados Total em parcelas convenientes.

O modelo ANOVA a dois factores, com interacção

Juntando os pressupostos necessários à inferência,

Modelo ANOVA a dois factores, com interacção (Modelo $M_{A \times B}$)

Existem n observações, Y_{ijk} , n_{ij} das quais associadas à célula (i, j) ($i = 1, \dots, a; j = 1, \dots, b$). Tem-se:

- 1 $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \forall i=1, \dots, a; j=1, \dots, b; k=1, \dots, n_{ij}$
($\alpha_1=0; \beta_1=0; (\alpha\beta)_{1j}=0, \forall j; (\alpha\beta)_{i1}=0, \forall i$).
- 2 $\varepsilon_{ijk} \cap \mathcal{N}(0, \sigma^2)$
- 3 $\{\varepsilon_{ijk}\}_{i,j,k}$ v.a.s independentes.

O modelo tem ab parâmetros desconhecidos: μ_{11} ; os $a-1$ acréscimos α_i ($i > 1$); os $b-1$ acréscimos β_j e os $(a-1)(b-1)$ efeitos de interacção $(\alpha\beta)_{ij}$, para $i > 1, j > 1$.

Testando efeitos de interacção

Para testar a existência de efeitos de interacção,

$$H_0 : (\alpha\beta)_{ij} = 0, \forall i = 2, \dots, a, \forall j = 2, \dots, b,$$

pode efectuar-se um teste F parcial comparando o modelo

$$\text{(Modelo } M_{A \times B}) \quad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

com o submodelo

$$\text{(Modelo } M_{A+B}) \quad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk},$$

Designa-se Soma de Quadrados associada à interacção à diferença

$$SQAB = SQRE_{A+B} - SQRE_{A \times B}$$

Testando os efeitos principais de cada Factor

Para testar os efeitos principais do Factor B,

$H_0 : \beta_j = 0, \forall j = 2, \dots, b$, pode partir-se dos modelos

$$\text{(Modelo } M_{A+B}) \quad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk}$$

$$\text{(Modelo } M_A) \quad Y_{ijk} = \mu_{11} + \alpha_i + \varepsilon_{ijk},$$

e tomar

$$SQB = SQRE_A - SQRE_{A+B}$$

$$SQA = SQF_A$$

A decomposição de SQT

Definimos :

$$SQAB = SQRE_{A+B} - SQRE_{A \times B}$$

$$SQB = SQRE_A - SQRE_{A+B}$$

$$SQA = SQF_A$$

Somando estas Somas de Quadrados a $SQRE_{A \times B}$, obtêm-se:

$$SQRE_{A \times B} + \underbrace{SQAB + SQA + SQB}_{= SQF_{A \times B}} = SQT$$

Esta decomposição de SQT gera as quantidades nas quais se baseiam as estatísticas dos três testes associados ao Modelo $M_{A \times B}$.

O quadro-resumo

Com base na decomposição do acetato 368 podemos construir o quadro resumo da ANOVA a 2 Factores, com interacção.

| Fonte | g.l. | SQ | QM | f_{calc} |
|------------|--------------|--------------------|----------------------------------|---------------------|
| Factor A | $a - 1$ | SQA | $QMA = \frac{SQA}{a-1}$ | $\frac{QMA}{QMRE}$ |
| Factor B | $b - 1$ | SQB | $QMB = \frac{SQB}{b-1}$ | $\frac{QMB}{QMRE}$ |
| Interacção | $(a-1)(b-1)$ | SQAB | $QMAB = \frac{SQAB}{(a-1)(b-1)}$ | $\frac{QMAB}{QMRE}$ |
| Resíduos | $n - ab$ | SQRE | $QMRE = \frac{SQRE}{n-ab}$ | |
| Total | $n - 1$ | $SQT = (n-1)s_y^2$ | - | - |

O Teste F aos efeitos de interacção

Sendo válido o Modelo ANOVA a dois factores, com interacção:

Teste F aos efeitos de interacção

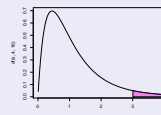
Hipóteses: $H_0 : (\alpha\beta)_{ij} = 0 \quad \forall i, j$ vs. $H_1 : \exists i, j \text{ t.q. } (\alpha\beta)_{ij} \neq 0$.
[NÃO HÁ INTERACÇÃO] vs. [HÁ INTERACÇÃO]

Estatística do Teste: $F = \frac{QMAB}{QMRE} \cap F_{((a-1)(b-1), n-ab)}$ se H_0 .

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se
 $F_{calc} > f_{\alpha((a-1)(b-1), n-ab)}$



O Teste F aos efeitos principais do factor A

Sendo válido o Modelo ANOVA a dois factores, com interacção (delineamento equilibrado) tem-se então:

Teste F aos efeitos principais do factor A

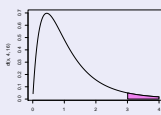
Hipóteses: $H_0 : \alpha_i = 0 \quad \forall i=2, \dots, a$ vs. $H_1 : \exists i=2, \dots, a \text{ t.q. } \alpha_i \neq 0$.
[≠ EFEITOS DE A] vs. [∃ EFEITOS DE A]

Estatística do Teste: $F = \frac{QMA}{QMRE} \cap F_{(a-1, n-ab)}$ se H_0 .

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se
 $F_{calc} > f_{\alpha(a-1, n-ab)}$



O Teste F aos efeitos principais do factor B

Sendo válido o Modelo ANOVA a dois factores, com interacção (delineamento equilibrado) tem-se então:

Teste F aos efeitos principais do factor B

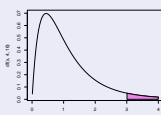
Hipóteses: $H_0 : \beta_j = 0 \quad \forall j=2, \dots, b$ vs. $H_1 : \exists j=2, \dots, b \text{ t.q. } \beta_j \neq 0$.
[≠ EFEITOS DE B] vs. [∃ EFEITOS DE B]

Estatística do Teste: $F = \frac{QMB}{QMRE} \cap F_{(b-1, n-ab)}$ se H_0 .

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se
 $F_{calc} > f_{\alpha(b-1, n-ab)}$



ANOVA a dois Factores, com interacção no R

Para efectuar uma ANOVA a dois Factor, com interacção, no R, convém organizar os dados numa data.frame com três colunas:

- 1 uma para os valores (numéricos) da variável resposta;
- 2 outra para o factor A (com a indicação dos seus níveis);
- 3 outra para o factor B (com a indicação dos seus níveis).

As fórmulas utilizadas no R para indicar uma ANOVA a dois Factores, com interacção, recorrem ao símbolo *:

$y \sim fA * fB$

Estimação da interacção necessita de repetições

Para se poder estudar efeitos de interacção, é necessário que haja repetições nas células.

Os graus de liberdade do SQRE são $n - ab$. Se houver uma única observação em cada célula, tem-se $n = ab$, ou seja, tantos parâmetros quantas as observações existentes.

Num delineamento com uma única observação por célula é obrigatório optar por um modelo sem interacção. Havendo repetições, é mais natural considerar um modelo com interacção.

Valores ajustados de Y

Sejam

- \bar{Y}_{ij} : a média amostral das n_{ij} observações da célula (i, j) ,
- $\bar{Y}_{i.}$: a média amostral das $\sum_j n_{ij}$ observações do nível i do Factor A,
- $\bar{Y}_{.j}$: a média amostral das $\sum_i n_{ij}$ observações do nível j do Factor B,
- $\bar{Y}_{...}$: a média amostral da totalidade das $n = \sum_i \sum_j n_{ij}$ observações.

Os valores ajustados \hat{Y}_{ijk} são iguais para todas as observações numa mesma célula, e são dados pela média amostral da célula:

$$\hat{Y}_{ijk} = \bar{Y}_{ij..}$$

Comparações múltiplas de médias de células

O número potencialmente grande de comparações possíveis entre médias de célula aconselha a utilização de métodos de comparação múltipla, que permitam controlar globalmente o nível de significância do conjunto de testes de hipóteses (ou grau de confiança do conjunto de intervalos de confiança).

O mais utilizado dos métodos de comparação múltipla está associado ao nome de Tukey. Foi já introduzido no estudo de delineamentos a 1 Factor. Adapta-se facilmente à comparação múltipla de médias de células.

O Teste de Tukey

Teste de Tukey para médias de células

Admite-se que o delineamento é equilibrado, com n_c repetições em todas as ab células.

Rejeita-se a igualdade das médias das células (i, j) e (i', j') , a favor da hipótese $\mu_{ij} \neq \mu_{i'j'}$, se

$$|\bar{Y}_{ij.} - \bar{Y}_{i'j'.}| > q_{\alpha(ab, n-ab)} \cdot \sqrt{\frac{QMRE}{n_c}},$$

sendo $q_{\alpha(ab, n-ab)}$ o valor que deixa à direita uma região de probabilidade α numa distribuição de Tukey com parâmetros $k = ab$ (o número total de médias de célula) e $v = n - ab$ (os graus de liberdade associados ao QMRE).

Intervalos de Confiança para $\mu_{ij} - \mu_{i'j'}$

Com grau de confiança global $(1 - \alpha) \times 100\%$, todas as diferenças de médias de pares de células, $\mu_{ij} - \mu_{i'j'}$, estão em intervalos da forma:

$$\left[(\bar{Y}_{ij.} - \bar{Y}_{i'j'.}) - q_{\alpha(ab, n-ab)} \cdot \sqrt{\frac{QMRE}{n_c}}, (\bar{Y}_{ij.} - \bar{Y}_{i'j'.}) + q_{\alpha(ab, n-ab)} \cdot \sqrt{\frac{QMRE}{n_c}} \right]$$

Conclui-se que $\mu_{ij} \neq \mu_{i'j'}$ se o intervalo correspondente a este par de células não contém o valor zero.

Tukey no R

A obtenção dos Intervalos de Confiança de Tukey no R, para a diferença da média de células, no caso de um delineamento a dois Factores, é análogo ao caso de um único factor:

> TukeyHSD(aov(y ~ fA * fB))

O R produz também intervalos de confiança para as médias de nível de cada Factor isoladamente.

É possível representar graficamente estes Intervalos de Confiança encaixando o comando anterior na função plot.

Estimadores de parâmetros

Os estimadores dos parâmetros num modelo ANOVA a 2 Factores, com interacção, são:

- $\hat{\mu}_{11} = \bar{Y}_{11.}$
- $\hat{\alpha}_i = \bar{Y}_{i1.} - \bar{Y}_{11.} \quad (i > 1)$
- $\hat{\beta}_j = \bar{Y}_{1j.} - \bar{Y}_{11.} \quad (j > 1)$
- $(\hat{\alpha}\hat{\beta})_{ij} = (\bar{Y}_{ij.} + \bar{Y}_{11.}) - (\bar{Y}_{i1.} + \bar{Y}_{1j.}) \quad (i, j > 1).$

Intervalos de confiança ou testes de hipóteses para qualquer dos parâmetros individuais, ou combinações lineares desses parâmetros, podem ser efectuados utilizando a teoria geral do Modelo Linear.

Soma de Quadrados Residual

Tendo em conta que os valores ajustados correspondem às medias amostrais da célula onde se efectuaram as observações, $\hat{Y}_{ijk} = \bar{Y}_{ij.}$, verifica-se que:

$$SQRE = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (Y_{ijk} - \hat{Y}_{ijk})^2 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij.})^2$$

$$\Leftrightarrow SQRE = \sum_{i=1}^a \sum_{j=1}^b (n_{ij} - 1) S_{ij}^2,$$

sendo S_{ij}^2 a variância amostral das observações da célula (i, j) .

Outras SQs para delineamentos equilibrados

Para delineamentos equilibrados (com n_c observações por célula) é possível dar igualmente fórmulas simples para as Somas de Quadrados associadas aos efeitos principais de cada factor.

Estas fórmulas correspondem às Somas de Quadrados associadas a cada factor caso se ajustasse (aos mesmos dados) um modelo ANOVA apenas com esse factor:

$$SQA = bn_c \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2$$

$$SQB = an_c \sum_{j=1}^b (\bar{Y}_{.j.} - \bar{Y}_{...})^2$$

Análise dos Resíduos

A validade dos pressupostos do Modelo relativos aos erros aleatórios pode ser estudada de forma análoga ao que foi visto para um delineamento a 1 Factor.

Os resíduos relativos a uma mesma célula aparecem em ab colunas verticais num gráfico de E_{ijk} vs. \hat{Y}_{ijk} .

A hipótese de heterogeneidade de variâncias entre diferentes células pode ser testada recorrendo ao Teste de Bartlett, caso a dimensão da amostra seja grande (e.g., $n_{ij} \geq 5$ em todas as células).

O Teste de Bartlett para delineamentos a dois factores

Teste de Bartlett à homogeneidade de variâncias

Hipóteses: $H_0: \sigma_{11}^2 = \sigma_{12}^2 = \dots = \sigma_{ab}^2$ vs. $H_1: \exists i, j, j' : \sigma_{ij}^2 \neq \sigma_{ij'}^2$
 [Variâncias homogéneas] [Var. heterogéneas]

Estatística do Teste:

$$K^2 = \frac{(n-ab) \ln QMRE - \sum_{i=1}^a \sum_{j=1}^b (n_{ij} - 1) \ln S_{ij}^2}{C} \sim \chi_{ab-1}^2$$

$$\text{onde } C = 1 + \frac{1}{3(ab-1)} \left[\sum_{i=1}^a \sum_{j=1}^b \frac{1}{n_{ij}-1} - \frac{1}{n-ab} \right].$$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se $K_{calc}^2 > \chi_{\alpha, (ab-1)}^2$

O Teste de Bartlett no \mathbb{R} , para 2 Factores

No \mathbb{R} , o comando `bartlett.test` apenas aceita a indicação de um factor. Mas a extensão do teste de Bartlett às variâncias de células é imediata se as ab células forem identificadas como ab níveis de 1 Factor.

Um comando que permite criar um vector que distinga entre células definidas por factores fA e fB para posterior utilização num teste de Bartlett é:

```
> celulas <-paste( fA , fB , sep="0")
> bartlett.test( y ~ celulas)
```

Uma advertência

Na formulação clássica do modelo ANOVA a dois Factores, com interacção, e a partir da equação-base $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, em vez de impor as condições $\alpha_1 = \beta_1 = (\alpha\beta)_{11} = (\alpha\beta)_{1j} = 0$ ($\forall i, j$), admite-se a existência de acréscimos de todos os tipos para qualquer valor de i e j e impõe-se as condições:

- $\sum_i \alpha_i = 0$;
- $\sum_j \beta_j = 0$;
- $\sum_j (\alpha\beta)_{ij} = 0, \quad \forall j$;
- $\sum_i (\alpha\beta)_{ij} = 0, \quad \forall i$.

Esta condição alternativa:

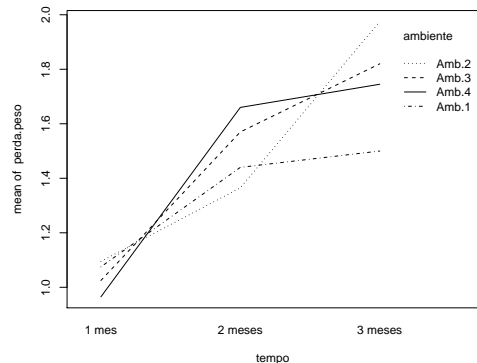
- muda a forma de interpretar os parâmetros;
- Muda os estimadores dos parâmetros.
- Não muda o resultado dos testes F à existência de efeitos.

Visualização gráfica de efeitos de interação

A existência de **efeitos de interação** transparece em gráficos onde:

- O eixo horizontal é associado aos níveis de um factor (e.g., fA);
- o eixo vertical é associado a valores (médios) da variável resposta (Y);
- para cada nível do segundo factor (e.g., fB), indica-se um ponto para cada nível do primeiro factor e respectiva média de célula da variável resposta;
- unem-se os pontos correspondentes a um mesmo nível do segundo factor.

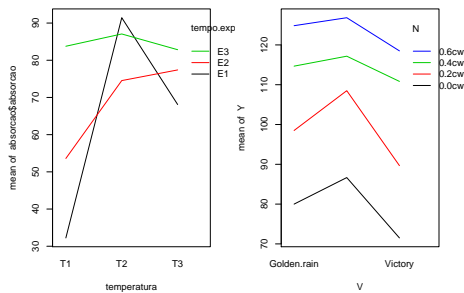
Exemplo (Dados do Exercício 8 ANOVA)



Como ler os gráficos de interação

A inexistência de interação produz linhas "paralelas" (ver exemplo da direita).

Havendo interação, as linhas estarão longe de qualquer paralelismo (ver exemplo da esquerda).



Delineamentos hierarquizados

Delineamentos que, superficialmente, podem confundir-se com os delineamentos factoriais são delineamentos onde surgem dois (ou mais) factores, mas em que **os níveis de um dos factores variam consoante os níveis do outro factor**.

Por exemplo, considere uma variável resposta *rendimento* de trigo, que se pretende modelar com os factores *variedade* e *adubação*. Suponha que

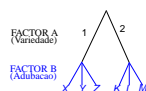
- na variedade 1 as adubações mais frequentes são X, Y e Z;
- na variedade 2 as adubações mais utilizadas são K, L e M.

Um delineamento factorial obriga a ter $ab = 2 \times 6 = 12$ células, sabendo-se de antemão que não interessam as células que combinam a variedade 1 com as adubações K,L,M e as células que combinam a variedade 2 com as adubações X,Y,Z.

Delineamentos hierarquizados (cont.)

Uma alternativa é considerar uma **hierarquia** dos factores: só identificamos os níveis do factor adubo após ter identificado o nível do factor variedade com que se trabalha. O número total de células ficou reduzido a $3+3=6$.

| | K | L | M | X | Y | Z |
|-------------|---|---|---|---|---|---|
| Variedade 1 | - | - | - | x | x | x |
| Variedade 2 | x | x | x | - | - | - |



Um tal delineamento diz-se **hierarquizado** (*nested*, em inglês).

O modelo a 2 Factores, hierarquizados

Cada observação é representada por uma v.a com **três índices**, Y_{ijk} :

- i nível do factor dominante ($i = 1, \dots, a$);
- j nível do factor subordinado ($j = 1, \dots, b_i$);
- k repetição para a célula (i, j) , com $k = 1, \dots, n_{ij}$.

Nota: b_i pode ser diferente para cada nível i do factor dominante.

A equação base do modelo inclui **efeitos de nível do Factor A** e **efeitos de nível do factor B (subordinado)**:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk},$$

com $\alpha_1 = 0$ e $\beta_{1(i)} = 0, \forall i$.

Não faz sentido falar em efeitos do nível j do Factor B, sem especificar qual o nível do Factor A a que nos referimos. Não faz sentido falar em efeitos de interação.

Variáveis indicatrizes e número de parâmetros

Como em modelos anteriores, a cada parâmetro associa-se uma variável indicatriz das observações correspondentes. Assim:

- um parâmetro μ_{11} , associado à **coluna de uns**, 1_n .
- $(a - 1)$ parâmetros α_i , associados às indicatrizes \mathcal{I}_{A_i} de cada nível $i > 1$ do Factor A.
- $\sum_{i=1}^a (b_i - 1)$ parâmetros $\beta_{j(i)}$, associados às indicatrizes $\mathcal{I}_{B_{j(i)}}$ de cada nível $j > 1$ do Factor B, para $i = 1, \dots, a$.

O no. de parâmetros é igual ao no. de situações experimentais:

$$1 + (a - 1) + \sum_{i=1}^a (b_i - 1) = \sum_{i=1}^a b_i$$

Se houver sempre $b = b_j$ níveis do Factor B, em cada nível i do Factor A, haverá ab parâmetros no modelo.

Os valores esperados de Y_{ijk}

Tem-se:

- Para a primeira célula ($i = j = 1$): $E[Y_{ijk}] = \mu = \mu_{11}$.
- Nas restantes células ($i = 1; j > 1$) do primeiro nível do Factor A: $\mu_{1j} = E[Y_{ijk}] = \mu_{11} + \beta_{j(1)}$.
- Nos restantes primeiros níveis do factor B ($i > 1; j = 1$): $\mu_{i1} = E[Y_{ijk}] = \mu_{11} + \alpha_i$.
- Nas células genéricas (i, j), com $i > 1$ e $j > 1$, $\mu_{ij} = E[Y_{ijk}] = \mu_{11} + \alpha_i + \beta_{j(i)}$.

Os efeitos α_i e $\beta_{j(i)}$ designam-se **efeitos dos níveis de cada Factor**.

O modelo ANOVA a dois factores, hierarquizados

Juntando os pressupostos necessários à inferência,

Modelo ANOVA a dois factores, hierarquizados (Modelo $M_{A/B}$)

Seja A o Factor dominante e B o Factor subordinado.

Existem n observações, Y_{ijk} , n_{ij} das quais associadas à célula (i, j) ($i = 1, \dots, a$; $j = 1, \dots, b_i$). Tem-se:

- 1 $Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$, $\forall i=1, \dots, a$; $j=1, \dots, b_i$; $k=1, \dots, n_{ij}$
($\alpha_1 = 0$; $\beta_{1(i)} = 0$, $\forall i$).
- 2 $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$, $\forall i, j, k$
- 3 $\{\varepsilon_{ijk}\}_{i,j,k}$ v.a.s independentes.

Os dois testes ANOVA

Neste delineamento, desejamos fazer um teste à existência de cada um de dois tipos de efeitos:

- $H_0 : \alpha_i = 0$, $\forall i = 2, \dots, a$; e
- $H_0 : \beta_{j(i)} = 0$, $\forall i = 1, \dots, a$ e $j = 2, \dots, b_i$.

As estatísticas de teste para cada um destes testes obtêm-se a partir da decomposição da Soma de Quadrados Total em parcelas convenientes.

Como em delineamentos anteriores, as Somas de Quadrados associadas a cada tipo de efeito resultam de tomar as diferenças das Somas de Quadrados Residuais de modelos onde se vão sucessivamente omitindo os efeitos correspondentes.

A decomposição de SQT

Para efectuar a decomposição da Soma de Quadrados Total, consideremos os modelos

$$\begin{aligned} \text{(Modelo } M_{A/B}) \quad Y_{ijk} &= \mu_{11} + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}, \\ \text{(Modelo } M_A) \quad Y_{ijk} &= \mu_{11} + \alpha_i + \varepsilon_{ijk}, \end{aligned}$$

Designa-se **Soma de Quadrados associada aos efeitos de B** a

$$SQB(A) = SQRE_A - SQRE_{A/B}$$

e **Soma de Quadrados associada aos efeitos de A** à diferença

$$SQA = SQF_A = SQT - SQRE_A$$

Juntamente com $SQRE_{A/B}$, tem-se:

$$SQT = SQA + SQB(A) + SQRE_{A/B}$$

Graus de liberdade

Os **graus de liberdade** associados a cada tipo de efeito são dados por:

- $g.l.(SQA) = a - 1$, o número de parâmetros associados aos efeitos de nível de A.
- $g.l.[SQB(A)] = \sum_{i=1}^a (b_i - 1)$, o número de parâmetros associados aos efeitos de nível de B.
- $g.l.(SQRE) = n - \sum_{i=1}^a b_i$, o número de observações menos o número total de parâmetros do modelo.

Quadro-resumo da ANOVA a 2 Factores hierarquizados

| Fonte | g.l. | SQ | QM | f_{calc} |
|-------------|--------------------------|--------------------|--|-----------------------|
| Factor A | $a - 1$ | SQA | $QMA = \frac{SQA}{a-1}$ | $\frac{QMA}{QMRE}$ |
| Factor B(A) | $\sum_{i=1}^a (b_i - 1)$ | SQB(A) | $QMB(A) = \frac{SQB(A)}{\sum_{i=1}^a (b_i - 1)}$ | $\frac{QMB(A)}{QMRE}$ |
| Resíduos | $n - \sum_{i=1}^a b_i$ | SQRE | $QMRE = \frac{SQRE}{n - \sum_{i=1}^a b_i}$ | |
| Total | $n - 1$ | $SQT = (n-1)S_y^2$ | - | - |

O Teste F aos efeitos do factor A (dominante)

Sendo válido o Modelo de ANOVA a dois factores hierarquizados, tem-se:

Teste F aos efeitos do factor A (dominante)

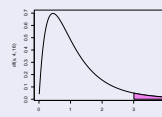
Hipóteses: $H_0 : \alpha_i = 0 \quad \forall i=2, \dots, a$ vs. $H_1 : \exists i=2, \dots, a$ t.q. $\alpha_i \neq 0$.
[FACTOR A NÃO AFECTA] vs. [FACTOR A AFECTA Y]

Estatística do Teste: $F = \frac{QMA}{QMRE} \cap F_{(a-1, n-\sum_i b_i)}$ se H_0 .

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se
 $F_{calc} > f_{\alpha(a-1, n-\sum_i b_i)}$



O Teste F aos efeitos do factor B (subordinado)

Sendo válido o Modelo de ANOVA a dois factores hierarquizado,

Teste F aos efeitos do factor B (subordinado)

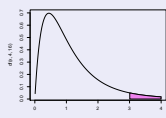
Hipóteses: $H_0 : \beta_{j(i)} = 0 \quad \forall j=2, \dots, b_i, i=1, \dots, a$ vs. $H_1 : \exists i, j$ t.q. $\beta_{j(i)} \neq 0$.
[FACTOR B NÃO AFECTA] vs. [FACTOR B AFECTA Y]

Estatística do Teste: $F = \frac{QMB(A)}{QMRE} \cap F_{(\sum_i (b_i - 1), n - \sum_i b_i)}$ se H_0 .

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se
 $F_{calc} > f_{\alpha(\sum_i (b_i - 1), n - \sum_i b_i)}$



ANOVA a dois Factores hierarquizados no R

Para efectuar uma ANOVA a dois Factor hierarquizados no R, convém **organizar os dados numa data.frame com três colunas:**

- 1 uma para os valores (numéricos) da variável resposta;
- 2 outra para o factor A (com a indicação dos seus níveis);
- 3 outra para o factor B (com a indicação dos seus níveis).

As fórmulas utilizadas no R para indicar uma ANOVA a dois Factores, sem interação, são semelhantes às usadas na Regressão Linear com dois preditores, devendo o nome dos dois factores ser separado pelo símbolo /. Se o factor fA é dominante:

$$y \sim fA / fB$$

Um exemplo

Um estudo sobre rendimentos (Y), de várias variedades de aveia (factor V), tendo sido usadas várias adubações azotadas (factor N), mas nem sempre iguais para cada variedade.

```
> summary(aov(Y ~ V/N, data=oats))
          Df Sum Sq Mean Sq F value    Pr(>F)
V           2  1786.4    893.2  1.7949 0.1749504
V:N         9 20342.2   2260.2  4.5421 0.0001397 ***
Residuals  60 29857.3    497.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Neste caso, apenas o factor subordinado parece ter efeitos sobre a variável resposta.

Comparações múltiplas de médias

Caso se conclua pela existência de efeitos do factor subordinado, é natural querer comparar médias da variável resposta nas $\sum_{j=1}^a b_j$ diferentes situações experimentais.

Os testes/intervalos de confiança de Tukey podem ser utilizados, caso o delineamento seja equilibrado, isto é, se houver o mesmo número de observações em cada situação experimental.

Neste caso, os parâmetros da distribuição de Tukey serão

- o número de situações experimentais, $k = \sum_{i=1}^a b_i$; e
- os graus de liberdade associados ao QMRE, $v = n - \sum_{i=1}^a b_i$.

Análise de resíduos

Também no que respeita à **análise de resíduos** para validar os pressupostos do modelo, a situação é **análoga** à de casos anteriores.

Pode efectuar-se um **teste de Bartlett** para testar a hipótese que as variâncias populacionais são iguais em cada uma das $k = \sum_{i=1}^a b_i$ diferentes situações experimentais. A estatística de teste e os graus de liberdade da respectiva distribuição assintótica são iguais aos casos anteriores (ver acetato 318), com este valor de k .

Comentários finais ANOVA

1. Outros tipos de delineamentos experimentais

Existem numerosos outros tipos de delineamentos experimentais mais complexos.

Alguns delineamentos visam reduzir o número de situações experimentais que seria necessário estudar (objectivo que também pode motivar um **delineamento hierarquizado**). Entre estes, refiram-se:

- Os **quadrados latinos**; ou
- os **delineamentos em blocos incompletos**.

Outros delineamentos visam ultrapassar dificuldades práticas na execução de uma experiência, como é o caso dos delineamentos em **parcelas divididas (split plots)**.

2. ANOVAs como comparação de k amostras

Alguns testes ANOVA generalizam os testes t de comparação de médias de duas amostras, estudados na disciplina de Estatística, para o caso de **haver mais do que duas amostras**.

Na disciplina de Estatística estudaram-se testes para comparar:

- As médias de 2 populações, com **amostras independentes**; e
- As médias de 2 populações, com **amostras emparelhadas**.

Em ambos os casos efectuava-se um **teste t** .

2. ANOVAs como comparação de k amostras (cont.)

- A estatística F do teste aos efeitos do factor, num **modelo ANOVA a 1 Factor com $k = 2$ níveis**, é o **quadrado da estatística t** à diferença de médias, no caso de **amostras independentes**.
- A estatística F do teste aos efeitos do Factor, num **modelo ANOVA a 1 Factor com blocos casualizados** (i.e., sem interacção e uma única observação por célula), **quando $a = 2$** , é o **quadrado da estatística t** à diferença de médias, no caso de **amostras emparelhadas**.

3. Delineamentos factoriais a vários factores

Um delineamento com observações para todas as combinações de níveis de cada factor (células) designa-se um delineamento **factorial**. Delineamentos factoriais podem ter qualquer número de factores.

Num delineamento **factorial a três factores – A, B e C** – cada observação da variável resposta indexa-se com **quatro índices**: Y_{ijkl} indica a observação l no nível i do Factor A, nível j do Factor B e nível k do Factor C. A equação de base para Y_{ijkl} prevê a existência de **sete tipos de efeitos**:

- três **efeitos principais de cada factor**, α_i , β_j e γ_k .
- três **efeitos de interacção dupla** associados a cada combinação de níveis de dois Factores diferentes: $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$ e $(\beta\gamma)_{jk}$.
- um **efeito de tripla interacção** para as **células** onde se cruzam níveis dos três factores: $(\alpha\beta\gamma)_{ijk}$

3. O modelo a três factores

A equação de base do modelo é agora da forma:

$$Y_{ijkl} = \mu_{111} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl},$$

excluindo-se efeitos sempre que um dos índices fôr 1.

O modelo tem abc parâmetros.

A Soma de Quadrados Total vai ser agora decomposta em **oito parcelas**: SQA, SQB, SQC, SQAB, SQAC, SQBC, SQABC e SQRE. As sete SQs associadas a efeitos são **definidas pela diferença das Somas de Quadrados Residuais de modelos onde se vão sucessivamente omitindo os efeitos correspondentes**.

3. O modelo a três factores (cont.)

Os **graus de liberdade** associados a cada tipo de efeito generalizam conceitos anteriores:

- Para as **SQs de efeitos principais de factor**, são os números de níveis, menos um: $a - 1$, $b - 1$ e $c - 1$.
- para as **interacções duplas**, são o produto dos graus de liberdade de cada factor: $(a - 1)(b - 1)$, $(a - 1)(c - 1)$ e $(b - 1)(c - 1)$.
- para as **interacções triplas**, são o produto dos graus de liberdade dos três efeitos principais: $(a - 1)(b - 1)(c - 1)$.
- para o residual, o número de observações menos o número de parâmetros, $n - abc$.

3. O modelo a três factores (cont.)

Haverá **sete testes**: um para cada tipo de efeitos.

As estatísticas desses sete testes são todas do tipo $\frac{QMx}{QMRE}$, onde x designa o tipo de efeitos em questão.

As estatísticas desses testes terão, sob H_0 , distribuição F com graus de liberdade dados pelos g.l. do numerador e do denominador, respectivamente.

3. Um exemplo

No **R**, ANOVAs factoriais a 3 Factores fazem-se de forma análoga às de dois factores:

```
> summary(aov(yield ~ N*P*K, data=npk))
              Df Sum Sq Mean Sq F value Pr(>F)
N              1  189.28   189.28   6.1608 0.02454 *
P              1    8.40     8.40   0.2735 0.60819
K              1   95.20    95.20   3.0986 0.09746 .
N:P            1   21.28    21.28   0.6927 0.41750
N:K            1   33.14    33.14   1.0785 0.31448
P:K            1    0.48     0.48   0.0157 0.90192
N:P:K         1   37.00    37.00   1.2043 0.28870
Residuals    16  491.58    30.72
```

4. Comparações múltiplas alternativas na ANOVA

A comparação de múltiplas médias, que abordámos pela teoria de Tukey, tem alternativas.

A alternativa mais conceituada baseia-se na teoria de **Scheffé**. Produz intervalos de confiança maiores (ao mesmo nível $(1 - \alpha) \times 100\%$ de confiança) do que os intervalos de Tukey.

Quer Tukey, quer Scheffé, podem ser generalizados para obter testes/intervalos de confiança sobre **combinações lineares genéricas das médias** de nível ou de células. Neste caso, a teoria de Scheffé tem melhor desempenho.

5. Métodos não paramétricos de tipo ANOVA

Uma forma alternativa de estudar problemas análogos aos objectivos de ANOVAs resulta da utilização de **métodos não paramétricos**.

Métodos não paramétricos são métodos em que não se exigem hipóteses tão fortes como os métodos clássicos, (e.g., a hipótese de normalidade). A sua maior generalidade tem como contrapartida uma menor capacidade de rejeitar as hipóteses nulas caso elas sejam falsas (i.e., têm menor **potência**), quando os pressupostos adicionais dos métodos clássicos são válidos.

Embora nem sempre, com grande frequência os métodos não paramétricos substituem os valores observados da variável resposta pelas **ordens (ranks)** dessas observações. As estatísticas de teste são então funções dessas ordens.

5. Métodos não paramétricos de tipo ANOVA (cont.)

O **teste de Kruskal-Wallis** é uma alternativa não paramétrica à ANOVA a 1 Factor, em que:

- Cada observação é substituída pela sua ordem;
- A estatística de teste compara as ordens médias em cada nível do factor com a ordem média global.
- A hipótese nula é que nos vários níveis do factor as observações seguem a mesma distribuição.
- A hipótese alternativa é que a distribuição dos vários níveis difere apenas nas suas localizações (medianas).

5. Métodos não paramétricos de tipo ANOVA (cont.)

O teste de Friedman é uma alternativa não paramétrica à ANOVA a 1 Factor, com blocos casualizados, ou seja, a dois Factores, sem interação, nem repetições nas células, em que:

- Cada observação é substituída pela sua ordem no seio do seu bloco;
- A estatística de teste compara as ordens médias em cada nível do factor com a ordem média global.
- A hipótese nula é que nos vários níveis do factor as observações seguem a mesma distribuição, excepto devido a translações associadas a cada bloco.
- A hipótese alternativa é que a distribuição dos vários níveis difere também devido a translações associadas aos níveis do factor.

5. Pontes entre ANOVAs e métodos não paramétricos

Em ambos os casos, as estatísticas de teste podem ser escritas como funções das Somas de Quadrados usuais, aplicadas às ordens, em vez de aos valores observados de Y .

Os métodos não paramétricos são uma alternativa viável quando haja violação grave dos pressupostos dos modelos ANOVA clássicos.

6. Efeitos aleatórios em modelos tipo ANOVA

Nos modelos ANOVA estudados nesta disciplina, admitiu-se sempre que as parcelas de efeitos nas equações dos modelos eram constantes. Este tipo de modelos dizem-se de efeitos fixos.

Uma outra grande classe de modelos alternativos designam-se modelos de efeitos aleatórios.

6. Modelos ANOVA com efeitos aleatórios (cont.)

Se um factor tem um grande número (ou mesmo uma infinidade) de possíveis níveis, não sendo possível estudar todos, pode ter de se estudar apenas uma amostra aleatória de níveis do factor, na tentativa de extrair conclusões para o factor na sua totalidade.

Esta situação surge com frequência quando os níveis de um factor são plantas, animais, pessoas, terrenos, ou outras entidades para as quais se admite variabilidade, mas em que não é possível estudar a totalidade dos possíveis casos (níveis do factor).

Efeitos de blocos, ou de factores hierarquizados (subordinados) são, com muita frequência, mais correctamente descritos por efeitos aleatórios.

6. Modelos ANOVA com efeitos aleatórios (cont.)

Nesse caso, os níveis seleccionados aleatoriamente para o estudo terão efeitos que são melhor descritos por variáveis aleatórias, e não constantes.

Por exemplo, a equação base de um modelo a um factor com efeitos aleatórios, com k níveis do factor, será

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

sendo α_i uma variável aleatória que indica o efeito do nível que vier a ser aleatoriamente seleccionado como nível i do factor.

Podem ser considerados modelos com mais do que um factor em que todos, ou alguns, factores têm efeitos aleatórios. Um modelo com alguns efeitos fixos e outros aleatórios diz-se um modelo misto.

6. Modelos ANOVA com efeitos aleatórios (cont.)

Numa situação em que apenas se observa uma amostra aleatória de níveis de um (ou mais) factores,

- Considerar efeitos fixos corresponde a condicionar a análise aos níveis do factor que foram escolhidos. Neste caso, as conclusões apenas se referem, rigorosamente falando, aos níveis do factor observados (plantas, animais, terrenos, etc.).
- Caso se pretenda tirar conclusões válidas para a totalidade dos níveis do factor, deverão ser usados modelos com efeitos aleatórios.