# Análise de Variância com dois ou mais factores - planeamento factorial

Em muitas experiências interessa estudar o efeito de mais do que um factor sobre uma variável de interesse. Quando uma experiência envolve dois ou mais factores diz-se que temos uma ANOVA múltipla. Uma ANOVA em que todas as combinações de todos os níveis de todos os factores são consideradas diz-se ANOVA factorial. Na maioria das situações, quando estamos interessados em estudar a influência de dois ou mais factores numa variável, utilizamos uma ANOVA factorial.

Exemplo: Pretende-se estudar a concentração de cálcio no sangue de uma população de aves parte da qual foi sujeita a um tratamento hormonal. Os investigadores pretendem averiguar se existem diferenças na concentração média de cálcio dependendo do tratamento hormonal e também dependendo do sexo das aves. Os factores deste estudo são o tratamento hormonal (presente ou ausente) e o sexo (feminino e masculino).

# Análise de Variância múltipla - planeamento hierárquico

Em geral o número de níveis de cada factor bem como o seu valor não depende dos restantes factores. Quando o número de níveis ou o seu valor varia consoante os níveis considerados nos restantes factores diz-se que temos uma ANOVA hierárquica. Nestes casos deixamos de ter uma ANOVA factorial. Enquanto numa ANOVA factorial os factores são cruzados (dando origem a todas as possíveis combinações dos seus níveis), numa ANOVA hierárquica os factores são encaixados uns nos outros (dando origem a uma estrutura tipo árvore).

Exemplo: Pretende-se fazer um estudo sobre os níveis de uma dada substância no sangue (usada como anti-epiléptico) e para tal várias amostras de sangue foram enviadas para 4 laboratórios. Cada laboratório utiliza diferentes técnicas para fazer a análise e o número de técnicas disponíveis também varia de laboratório para laboratório. Neste caso temos dois factores: o laboratório (com 4 níveis) e a técnica de análise (com um número de níveis que depende do laboratório). Este último factor encontra-se encaixado no primeiro.

# ANOVA múltipla - factores fixos, aleatórios e mistos

Vimos que numa ANOVA simples o factor em causa podia ter os efeitos fixos ou os efeitos aleatórios. O mesmo se vai passar com os modelos de ANOVA com dois ou mais factores.

Quando um modelo tem todos os factores com efeitos fixos diz-se que temos uma ANOVA de efeitos fixos ou um Modelo I de ANOVA.

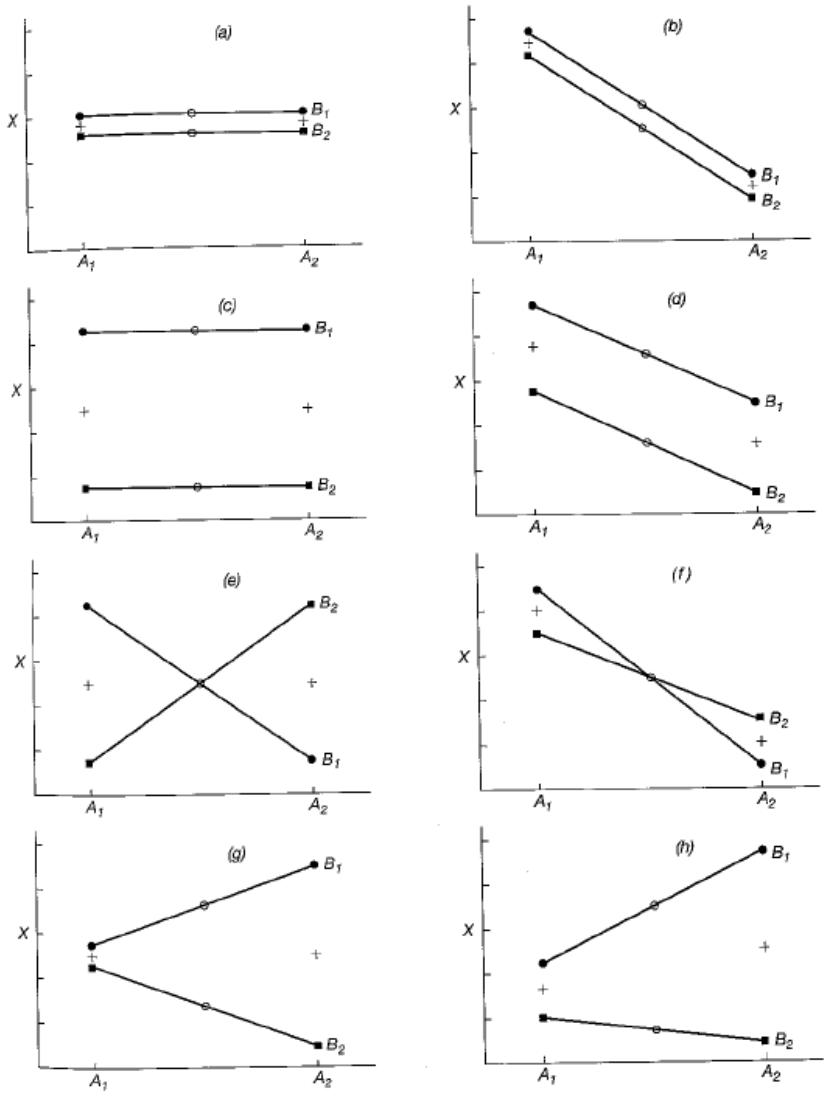
Quando um modelo tem todos os factores com efeitos aleatórios diz-se que temos uma ANOVA de efeitos aleatórios ou um Modelo II de ANOVA.

Quando um modelo tem alguns factores com efeitos fixos e outros com efeitos aleatórios diz-se que temos uma ANOVA de efeitos mistos ou um Modelo III de ANOVA.

# Análise de Variância múltipla - interacção entre factores

Quando temos dois ou mais factores há que ter em conta que estes podem interagir entre si, i.e., a variação na variável resposta produzida por uma alteração do nível de um dos factores pode variar consoante os níveis dos restantes factores. Assim há que prestar atenção às possíveis interacções entre os vários factores, dois a dois, três a três, etc.. Quanto mais factores existirem no estudo mais complexo se torna o modelo, porque o número de interacções possíveis aumenta muito rapidamente. Quando não existe interacção entre os factores o valor esperado de cada combinação de níveis dos factores é a soma dos valores esperados de cada nível separadamente e o modelo diz-se aditivo.

Seguidamente apresenta-se um conjunto de gráficos que pretende ilustrar diferentes comportamentos de ANOVA's com 2 factores ($A$ e $B$), tendo cada um deles apenas 2 níveis ($A_1$ e $A_2$, $B_1$ e $B_2$). Quando as linhas são paralelas temos modelos sem interacção entre os factores (modelo aditivo). Este tipo de gráficos permite ao investigador ter uma ideia se a interacção está presente ou não.


(a)
X
$B_1$
$B_2$
$A_1$
$A_2$
(b)
X
$B_1$
$B_2$
$A_1$
$A_2$
(c)
X
$B_1$
$B_2$
$A_1$
$A_2$
(d)
X
$B_1$
$B_2$
$A_1$
$A_2$
(e)
X
$B_2$
$B_1$
$A_1$
$A_2$
(f)
X
$B_2$
$B_1$
$A_1$
$A_2$
(g)
X
$B_1$
$B_2$
$A_1$
$A_2$
(h)
X
$B_1$
$B_2$
$A_1$
$A_2$

# Análise de Variância múltipla no SPSS

A ANOVA com dois ou mais factores pode ser realizada no SPSS no menu `Analyze / General Linear Model / Univariate`. (Atenção que ANOVA múltipla significa que temos apenas uma variável dependente e múltiplos factores a influenciar essa variável. Daí o menu ser identificado por `Univariate`. ANOVA multivariada (`Multivariate`) refere-se a experiências em que temos várias variáveis de resposta que interessa analisar em simultâneo.)

Na janela principal selecciona-se a variável em estudo (dependent variable) e seleccionam-se os factores (fixos ou aleatórios) para as respectivas janelas.

Por defeito o SPSS assume o modelo factorial completo (com todas as interacções entre os factores). Se quisermos especificar um modelo que não seja este podemos fazê-lo através do botão `Model`

# Análise de Variância dupla

Uma ANOVA com dois factores diz-se ANOVA dupla. Em seguida iremos considerar o modelo geral de uma ANOVA factorial dupla (planeamento completamente aleatorizado).

Iremos designar os factores por $A$ e $B$ sendo que $A$ tem $a$ níveis e $B$ tem $b$ níveis. Existem portanto $ab$ combinações possíveis dos níveis dos factores. Tal como foi feito para a ANOVA simples iremos considerar o planeamento equilibrado, ou seja, para cada combinação de níveis dos factores existem $n$ observações (réplicas) independentes. No total são necessárias $N = abn$ observações.

As observações da variável de interesse $Y$ são indexadas por 3 índices, $Y_{ijk}$, $i$ representa o nível do factor $A$, $j$ representa o nível do factor $B$, e $k$ representa a posição dentro do grupo $ij$.

# Modelo de ANOVA factorial dupla

$$Y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \begin{cases} i = 1, 2, \ldots, a, \\ j = 1, 2, \ldots, b, \\ k = 1, 2, \ldots, n, \end{cases}$$

onde

- $\mu$ representa a média global,
- $\tau_i$ representa o efeito do nível $i$ do factor $A$,
- $\beta_j$ representa o efeito do nível $j$ do factor $B$,
- $\gamma_{ij}$ representa o efeito da interacção dos factores $A$ e $B$,
- $\epsilon_{ijk}$ representa um erro aleatório de cada observação sendo estes erros independentes entre si e todos com distribuição Normal de média 0 e variância $\sigma^2$.

# ANOVA factorial dupla - pressupostos e hipóteses a testar

Pressupostos exigidos:

1. Temos $ab$ grupos de observações independentes ($ab$ amostras aleatórias) sendo os grupos independentes entre si.
2. Cada grupo de observações deve provir de uma distribuição Normal.
3. A variância de todas as populações deve ser a mesma.

Hipóteses a testar (se os efeitos forem fixos)

1. $H_0: \ \tau_1 = \tau_2 = \ldots = \tau_a = 0 \quad vs \quad H_1: \ \tau_i \neq 0$ pelo menos para um $i$ (efeito principal do factor $A$)

2. $H_0: \ \beta_1 = \beta_2 = \ldots = \beta_b = 0 \quad vs \quad H_1: \ \beta_j \neq 0$ pelo menos para um $j$ (efeito principal do factor $B$)

3. $H_0: \ \gamma_{11} = \gamma_{12} = \ldots = \gamma_{ab} = 0 \ vs \ H_1: \ \gamma_{ij} \neq 0$ pelo menos para um par $i, j$ (interacção entre os factores $A$ e $B$)

# Partição da soma de quadrados

Seja

$$y_{i\cdot\cdot} = \sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk} \qquad \bar{y}_{i\cdot\cdot} = \frac{y_{i\cdot\cdot}}{bn}$$

$$y_{\cdot j\cdot} = \sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk} \qquad \bar{y}_{\cdot j\cdot} = \frac{y_{\cdot j\cdot}}{an}$$

$$y_{ij\cdot} = \sum_{k=1}^{n} y_{ijk} \qquad \bar{y}_{ij\cdot} = \frac{y_{ij\cdot}}{n}$$

$$y_{\cdots} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk} \qquad \bar{y}_{\cdots} = \frac{y_{\cdots}}{abn}$$

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{...})^2}_{SS_{total}} = \underbrace{bn\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})^2}_{SS_A}$$

$$+\underbrace{an\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})^2}_{SS_B}$$

$$+\underbrace{n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2}_{SS_{AB}}$$

$$+\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{ij.})^2}_{SS_E}$$

$SS_{total}$ tem $N - 1 = abn - 1$ graus de liberdade.

$SS_A$ tem $a - 1$ graus de liberdade.

$SS_B$ tem $b - 1$ graus de liberdade.

$SS_{AB}$ tem $(a - 1)(b - 1)$ graus de liberdade.

$SS_E$ tem $ab(n - 1)$ graus de liberdade.

# Tabela de ANOVA

| Fonte de Variação | Soma de quadrados | g.l. | Média de quadrados | $F_{obs}$ | $p$ |
|---|---|---|---|---|---|
| factor A | $SS_A$ | $a-1$ | $MS_A = \frac{SS_A}{a-1}$ | $\frac{MS_A}{MS_E}$ | $(\cdot)$ |
| factor B | $SS_B$ | $b-1$ | $MS_B = \frac{SS_B}{b-1}$ | $\frac{MS_B}{MS_E}$ | $(\cdot)$ |
| Interacção | $SS_{AB}$ | $(a-1)(b-1)$ | $MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$ | $\frac{MS_{AB}}{MS_E}$ | $(\cdot)$ |
| Erros | $SS_E$ | $ab(n-1)$ | $MS_E$ | | |
| Total | $SS_{total}$ | $abn-1$ | | | |

Através desta tabela podemos testar as hipóteses referidas anteriormente através dos *p-values* da última coluna. Neste caso:

a estatística de teste para as hipóteses 1 (efeito principal do factor $A$) é $F = \frac{MS_A}{MS_E} \sim F_{a-1,ab(n-1)}$, sob $H_0$;

a estatística de teste para as hipóteses 2 (efeito principal do factor $B$) é $F = \frac{MS_B}{MS_E} \sim F_{b-1,ab(n-1)}$, sob $H_0$;

a estatística de teste para as hipóteses 3 (interacção) é $F = \frac{MS_{AB}}{MS_E} \sim F_{(a-1)(b-1),ab(n-1)}$, sob $H_0$;

# Verificação dos pressupostos da ANOVA

Deve-se sempre verificar os pressupostos de realização da ANOVA.

Para averiguar se podemos considerar que a variância de todos os grupo é constante podemos utilizar um teste de homogeneidade de variâncias, como por exemplo o teste de Levene disponível no SPSS (menu `Analyze / General Linear Model / Univariate`, botão `Options` opção `Homogeneity tests`).

Para averiguar se os erros se podem considerar como sendo provenientes de uma população Normal faz-se uma análise de resíduos. Se conhecêssemos os erros que afectam as observações poderíamos construir QQ-plots e fazer testes de ajustamento à Normal. Mas não conhecemos os erros pois estes são dados por

$$\epsilon_{ijk} = Y_{ijk} - (\mu + \tau_i + \beta_j + \gamma_{ij})$$

e os parâmetros $\mu$, $\tau_i$, $\beta_j$ e $\gamma_{ij}$ são desconhecidos.

Ora, $\mu + \tau_i + \beta_j + \gamma_{ij}$ representa o valor médio da combinação dos níveis $A_i$ e $B_j$, que podemos representar por $\mu_{ij}$. Este valor médio pode ser estimado pela média das observações deste grupo, $\bar{Y}_{ij\cdot\cdot}$. Assim, os erros podem ser estimados por

$$\hat{\epsilon}_{ijk} = Y_{ijk} - \bar{Y}_{ij\cdot\cdot}$$

Estas diferenças chamam-se resíduos e costumam-se representar por $e_{ijk}$.

Uma análise de resíduos consiste em estudar o conjunto de todos os resíduos $e_{ijk}$, $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, n$, no sentido de averiguar se podemos considerar que essa amostra é aleatória e proveniente de uma população Normal. Para averiguar a Normalidade, constroem-se QQ-plots e fazem-se testes de ajustamento.

No SPSS, podemos guardar os resíduos, para seguidamente os analisar, através do botão `Save` do menu da ANOVA, `Analyze / General Linear Model / Univariate`.

# Análise de Variância com blocos aleatorizados

Em certas experiências podem existir factores (externos) que introduzem variabilidade nos dados e que interessa controlar. Por exemplo, se estivermos interessados em comparar 3 variedades de trigo através do peso médio dos grãos, pode ter influência o tipo de solo em que as plantas vão crescer. Em vez de seleccionarmos ao acaso um certo número de campos para semear as várias sementes, podemos seleccionar um conjunto de campos (possivelmente com características de solo diferentes) e dividir cada campo em três parte de modo a semear as três variedades de trigo em cada campo. Neste tipo de planeamento designa-se cada campo por bloco. Os blocos constituem o factor externo cuja variabilidade induzida vai ser possível controlar do ponto de vista estatístico.

Assim, num planeamento com blocos aleatorizados temos um factor de interesse que possui $g$ níveis (tratamentos) e temos $b$ blocos prefazendo um total de $N = gb$ observações. Os tratamentos são distribuídos aleatoriamente pelos $g$ elementos de cada bloco.

# Modelo de ANOVA com blocos aleatorizados

O modelo para este planeamento é uma extensão do modelo de ANOVA simples e é também um caso particular do modelo de ANOVA factorial dupla. Neste caso não existe interacção entre os factores e apenas dispomos de uma observação por célula.

As observações são designadas por $Y_{ij}$ onde $i = 1, \ldots, g$ identifica o grupo e $j = 1, \ldots, b$ identifica o bloco.

$$Y_{ij} = \mu_i + \beta_j + \epsilon_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij},$$

onde

- $\mu_i$ representa a média de cada grupo,
- $\mu$ representa a média de todos os grupos,
- $\tau_i$ representa o efeito do tratamento $i$
- $\beta_j$ representa o efeito do bloco $j$ e
- $\epsilon_{ij}$ representa um erro aleatório de cada observação sendo estes erros aleatórios e independentes entre si.

# ANOVA com blocos aleatorizados - pressupostos

Pressupostos exigidos:

1. O modelo descrito anteriormente é válido.

2. Os erros são aleatórios e independentes entre si, com distribuição Normal, $\epsilon_{ij} \frown N(0, \sigma)$.

3. O factor em estudo e o factor bloco não têm interacção (resulta do pressuposto 1.).

## Hipóteses a testar

No caso de o factor em estudo ser de feitos fixos temos

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_g = \mu \quad vs \quad H_1: \ \mu_i \neq \mu \ \text{ pelo menos para um } i$$

ou equivalentemente

$$H_0: \ \tau_1 = \tau_2 = \ldots = \tau_g = 0 \quad vs \quad H_1: \ \tau_i \neq 0 \ \text{pelo menos para um } i$$

No caso de o factor em estudo ser de efeitos aleatórios temos

$$H_0: \ \sigma_\tau^2 = 0 \quad vs \quad H_1: \sigma_\tau^2 > 0,$$

onde $\sigma_\tau^2$ representa a variância associada ao factor de interesse.

## Hipóteses a testar

Também podemos testar se os blocos produzem diferenças na variável resposta (ou seja, se vale a pena considerar os blocos como um factor)

O factor associado aos blocos tanto pode ser considerado fixo como aleatório (a situação mais habitual é ser aleatório). A tabela de ANOVA é igual em todos os casos.

No caso de o factor bloco ser de feitos fixos temos

$$H_0: \ \beta_1 = \beta_2 = \ldots = \beta_g = 0 \quad vs \quad H_1: \ \beta_i \neq 0 \text{ pelo menos para um } i$$

No caso de o factor bloco ser de feitos aleatórios temos

$$H_0: \ \sigma^2_\beta = 0 \quad vs \quad H_1: \sigma^2_\beta > 0,$$

onde $\sigma^2_\beta$ representa a variância associada ao factor bloco.

# Partição da soma de quadrados

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{..})^2}_{SS_{total}} = \underbrace{b\sum_{i=1}^{g}(\bar{y}_{i.}-\bar{y}_{..})^2}_{SS_{Trat}} + \underbrace{g\sum_{j=1}^{b}(\bar{y}_{.j}-\bar{y}_{..})^2}_{SS_B}$$

$$+\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})^2}_{SS_E}$$

$SS_{total}$ tem $N-1=gb-1$ graus de liberdade.

$SS_{Trat}$ tem $g-1$ graus de liberdade.

$SS_B$ tem $b-1$ graus de liberdade.

$SS_E$ tem $(g-1)(b-1)$ graus de liberdade.

# Tabela de ANOVA

| Fonte de Variação | Soma de quadrados | g.l. | Média de quadrados | $F_{obs}$ | $p$ |
|---|---|---|---|---|---|
| Tratamentos | $SS_{Trat}$ | $g-1$ | $MS_{Trat}$ | $\frac{MS_{Trat}}{MS_E}$ | $(\cdot)$ |
| Blocos | $SS_B$ | $b-1$ | $MS_B$ | $\frac{MS_B}{MS_E}$ | $(\cdot)$ |
| Erros | $SS_E$ | $(g-1)(b-1)$ | $MS_E$ | | |
| Total | $SS_{total}$ | $gb-1$ | | | |

Através desta tabela podemos testar as hipóteses referidas anteriormente através do *p-value* associado aos tratamentos. Neste caso a estatística de teste é $F = \frac{MS_{Trat}}{MS_E} \sim F_{g-1,(g-1)(b-1)}$, sob $H_0$.

Também podemos testar se os blocos influenciam os resultados (em média) através do *p-value* associado aos blocos. Neste caso a estatística de teste é $F = \frac{MS_B}{MS_E} \sim F_{b-1,(g-1)(b-1)}$, sob $H_0$.

# ANOVA múltipla - Comparações múltiplas

Tal como foi descrito para a ANOVA simples, quando se rejeita a hipótese nula de igualdade das médias (em pelo menos um dos factores) pode-se proceder a uma análise de comparações múltiplas para averiguar quais os pares de níveis que apresentam diferenças significativas entre si (dois a dois). Os métodos de Bonferroni, Tuckey, Dunnett (entre outros) podem ser generalizados para planeamentos com dois ou mais factores.

No SPSS estes procedimentos encontram-se disponíveis no botão `Post Hoc` do menu da ANOVA.

# Análise de Variância com medições repetidas

Um planeamento em que cada unidade experimental é medida duas ou mais vezes (em geral sequencialmente), diz-se que contém medições repetidas. Trata-se de uma generalização do conceito de amostras emparelhadas. Em inglês este tipo de planeamento é designado por *repeated measures experimental design* ou *within-subjects* ou *treatment-by-treatment design*.

Por exemplo: um investigador está interessado em comparar 3 drogas para reduzir a tensão arterial em doentes hiper-tensos. Um planeamento completamente aleatorizado consiste em alocar cada uma das drogas (aleatoriamente) a 15 doentes (3 grupos de 5 doentes cada) e após um período de tratamento efectuar uma ANOVA simples. Outra possibilidade consiste em seleccionar apenas 5 doentes e administrar as três drogas a cada doente, de forma sequencial no tempo. Assim, cada doente dá origem a três medições e no total continuamos a ter 15 observações. Neste caso passamos a ter uma experiência com medições repetidas , ou seja, um planeamento do tipo dos blocos aleatórios em que cada doente funciona como um bloco.

As vantagens destes planeamentos são geralmente as seguintes:

1. A experiência torna-se mais económica pois exige um menor número de unidades experimentais.
2. A variabilidade entre unidades experimentais é reduzida relativamente a um planeamento completamente aleatorizado.
3. A potência é superior à do planeamento completamente aleatorizado. (tal como acontecia ao compararmos duas média com amostras emparelhadas ou com amostras independentes).

As desvantagens são geralmente as seguintes:

1. A experiência pode tornar-se muito demorada.
2. Podem existir efeitos de alguns dos tratamentos que preduram no tempo (*carryover*) e que afectam os resultados dos outros tratamentos.

No SPSS existe um menu específico para ANOVA com observações repetidas, `Analyze / General Linear Model / Repeated Measures`.